dossiês

12
265
5987
569
48
4
641
986

Um mundo para conhecer os números

Escola Secundária de Tomaz Pelayo

Instituto Nacional de Estatística
Statistics Portugal

Direcção Regional de Educação do Norte
DREN

10º aniversário
ALEA

Um mundo para conhecer os números

Ficha Técnica

Título

Um mundo para conhecer os números

Editores

Instituto Nacional de Estatística, I.P.
Av. António José de Almeida
1000-043 Lisboa
Portugal

Escola Secundária Tomaz Pelayo
Rua Prof. Doutor Fernando Pires de Lima
4780-430 Santo Tirso
Portugal

Direcção Regional de Educação do Norte
Rua António Carneiro, 98
4349-003 Porto
Portugal

Design, Composição e Impressão

Instituto Nacional de Estatística, I.P.

Tiragem

300 exemplares

ISBN

978-98925-0043-0

Depósito Legal

300079/09

Periodicidade

Irregular

Índice

Este trabalho contém uma compilação de cinco dossiês publicados no ALEA. A lista completa de dossiês disponíveis, até agora, na área Dossiês & Recursos do ALEA, é a seguinte:

I - População e Demografia: População e Demografia - Quantos somos e como somos?

II - Ambiente e Recursos

III - A Inflação e o Índice de Preços no Consumidor

IV - Estatística com Excel

V - Censos 2001 - «Tu Também Contas!»

VI - Notas sobre a História da Estatística

VII – Probabilidades com Excel

VIII – Números do Cinema

IX – Representações Gráficas

X – EuropALEA - Uma visão Estatística da Europa : dos 6 aos 28

XI – Inquérito Estatístico

XII- Software Estatístico

XIII – Estatística Descritiva com Excel – Complementos

XIV – Estatística com R - Uma iniciação para o Ensino Básico e Secundário

Estes recursos, assim como outros conteúdos (cursos, informação estatística, jogos, anedotas, etc.), podem ser consultados na página do ALEA em **www.alea.pt**. Visite o ALEA. Um mundo para conhecer os números.

Prefácio

O ALEA faz 10 anos. E 10 anos notáveis.

Estão de parabéns os seus responsáveis e colaboradores. Estão de parabéns a Escola Secundária Tomaz Pelayo, o Instituto Nacional de Estatística e a Direcção Regional de Educação do Norte, instituições que são o sustentáculo deste projecto. Está também de parabéns a supervisora científica, Profª Doutora Maria Eugénia Graça Martins. Mas estão principalmente de parabéns todos os seus utilizadores, quer sejam alunos ou professores dos Ensinos Básico e Secundário, aos quais o projecto principalmente se dirige, quer sejam cidadãos interessados em melhorar a sua literacia estatística.

O ALEA assume-se efectivamente como um projecto ao serviço da literacia estatística, indispensável nos nossos dias ao exercício pleno da cidadania. De facto, não basta ao cidadão dispor de informação, não lhe basta dispor também de informação estatística, é necessário ainda que ele saiba compreender e interpretar essa informação e a saiba utilizar na tomada de decisões úteis, quer na sua vida pessoal quer na sua intervenção na sociedade. A literacia estatística é assim um instrumento poderoso ao serviço da qualidade da democracia.

O ALEA é um exemplo vivo do que podem fazer a vontade e a determinação de alguns quando postas ao serviço da comunidade. Quem visita a página *web* **www.alea.pt** do ALEA fica encantado com o que lá vê e seguro de que este projecto

é um instrumento muito útil para a melhoria da qualidade do ensino da Estatística em Portugal (e noutros países, quer de língua portuguesa, quer outros, já que a página tem uma versão em língua inglesa). A qualidade do projecto é notável e isso mesmo foi reconhecido a nível internacional, tendo o ALEA sido galardoado em 2007 com o Prémio "Best Cooperative Project Award" que, pela primeira vez, o International Statistical Literacy Program (ISLP) atribuiu. Note-se que o ISLP é um projecto da International Association for Statistical Education, a secção de educação estatística do centenário International Statistical Institute. Curiosamente, a 56ª Sessão do International Statistical Institute (reunião científica internacional que decorre de dois em dois anos) teve lugar em 2007 em Lisboa. Portugal está assim a afirmar-se internacionalmente na área da Estatística, quer nos aspectos científicos, quer também, através do ALEA, nos aspectos educativos. E obviamente, o progresso científico na área da Estatística só é sustentável se estiver assente numa educação estatística de qualidade. Daí a Sociedade Portuguesa de Estatística (SPE), a que tenho a honra de presidir, se preocupar não apenas com o desenvolvimento científico, mas também com o progresso educativo, com iniciativas várias, de que destacamos os Prémios Estatístico Júnior. É, pois, com grande satisfação que registamos o valioso trabalho desenvolvido pelo ALEA.

Mas não contente com tão valiosos contributos, o ALEA oferece-nos agora esta publicação comemorativa do seu 10º aniversário. Ela contém 5 dossiers dos muitos mais produzidos pelo ALEA.

Os temas versados são: O Inquérito Estatístico (com importantes considerações metodológicas e práticas de como organizar e interpretar inquéritos estatísticos), Estatística com R (que nos ensina a utilizar este software livre para os cálculos e gráficos estatísticos), Notas sobre a História da Estatística (desde os primórdios à actualidade, não esquecendo a Estatística em Portugal), Representações Gráficas (atraentes e facilmente compreensíveis e bem sabemos que uma boa imagem vale mais do que 1000 palavras) e Estatística Descritiva com EXCEL (que põe ao alcance de todos os cálculos básicos e os gráficos estatísticos). Foram temas muito bem escolhidos e muito bem desenvolvidos, agora postos em forma de livro, já que não só de *internet* vive o homem e é muito mais agradável ler e estudar num livro do que num ecrã. Parabéns por mais esta utilíssima iniciativa, que, além do tudo o mais, tem um grafismo muito atraente.

E é aqui uma boa oportunidade para registar o importante apoio que o INE tem prestado a muitas iniciativas que visam o desenvolvimento científico e educacional da Estatística em Portugal e de que a SPE tem frequentemente beneficiado. Mais uma vez o País conta com o apoio do INE, agora nesta publicação. O seu lançamento vai decorrer na cerimónia de abertura do XVII Congresso Anual da SPE. Que excelente ocasião para sentar à mesma mesa três aliados ao serviço da Estatística em Portugal: a SPE, o INE e o ALEA.

Para o leitor apenas um voto que certamente se cumprirá, o de que desfrute este livro com prazer e proveito.

Carlos Braumann

(Presidente da Sociedade Portuguesa de Estatística)

Notas sobre a História da Estatística

Maria João Ferreira
Isabel Tavares

com a colaboração da Prof.ª Doutora Maria Antónia Amaral Turkman

Notas sobre a História da Estatística

Maria João Ferreira
Isabel Tavares

Sumário:

1. Introdução

Este dossiê inclui os factos considerados mais importantes da história da Estatística e das Estatísticas, desde as Antigas Civilizações até aos nossos dias. Alguns cientistas também são aqui mencionados, não todos, mas sim aqueles que deram um maior contributo para o desenvolvimento da Estatística. No último capítulo é apresentado um pouco da evolução da Estatística no nosso país. No final, a rubrica "Ver Também" contém ligações para outros estudos de interesse relacionados com as temáticas em causa (publicações e páginas na internet).

2. As Civilizações Antigas

2.1 Introdução

Desde o começo da civilização que a Estatística tem estado sempre presente: nos primórdios mais oculta e na actualidade mais visível.

Contar, enumerar e recensear sempre foi uma preocupação permanente em todas as culturas. Em civilizações como a antiga Grécia, Roma, Egipto, Israel, Índia, Japão, China, etc, o Estado tinha necessidade de conhecer a sua população, tanto a nível económico como a nível social. Os Imperadores da altura ordenavam os recenseamentos da população com vista à cobrança de impostos e ao recrutamento militar, pois as guerras eram constantes e havia necessidade de conseguir jovens rapazes para serem treinados fisicamente para a guerra.

Nas civilizações antigas quem não respondesse aos Censos era punido com a morte.

Estes recenseamentos não podem ser comparados com os da actualidade, pois não assentavam em princípios estatísticos credíveis ou não eram feitos exaustivamente. Pode dizer-se contudo que o princípio da Estatística começou com estas sociedades, não como hoje é conhecida entre nós mas de uma maneira mais simples e rudimentar.

2.2 As Civilizações Antigas e os Censos

2.2.1 A Grécia Antiga (2100 a.C. a 146 a.C.)

A Grécia Antiga abrangia um vasto território. Era formada por um conjunto de cidades-estado, politicamente autónomas, possuindo em comum os costumes e a língua. No século V a.C. entre estas cidades sobressaía Atenas. A sua cultura era a mais brilhante de todas as cidades gregas, em particular nas artes, no teatro, na história e na filosofia. Também possuía o governo mais democrático de todas as cidades gregas. Além de Atenas destacavam-se as cidades de Esparta e Corinto.

Como se refere em Bedarida et al, 1987, Atenas era a cidade grega que melhor conhecia a sua população. Aristóteles dá-nos a conhecer que em cada nascimento se oferecia à sacerdotisa de Atenas uma medida de frumento (uma espécie de trigo candial), e em cada falecimento uma medida de cevada. Além disso, todos os jovens quando atingiam a idade de 18 anos eram inscritos na qualidade de cidadãos e eram colocados na lista de homens em estado de apresentar armas. Até esta idade, somente estudavam aritmética, literatura, música, escrita e educação física. As jovens não recebiam qualquer educação formal, mas aprendiam os ofícios domésticos e os trabalhos manuais com as mães. É através destas descrições feitas por historiadores que conseguimos aperceber-nos dos primeiros recenseamentos efectuados nas antigas civilizações. Também é sabido que os estrangeiros eram recenseados, através do seu tributo particular que era cobrado por cabeça.

É curioso constatar que no quadro descritivo de Atenas, já Aristóteles descrevia não só a situação de uma cidade ou de um país por si só, do ponto de vista do governo, da justiça, das ciências e das artes, dos museus e dos costumes, mas também por comparação com outros Estados. Deste modo, podemos observar nesta parte da obra de Aristóteles, o princípio da Estatística Descritiva.

Estatística Descritiva:

Estudo descritivo de dados de uma amostra (ou de uma população) em que se resume toda a informação recolhida em gráficos e tabelas, calculando algumas das suas características, por exemplo a moda, a média, frequências, etc.

2.2.2. A Antiga Civilização Egípcia (5000 a 30 a.C.)

A cultura egípcia é uma das mais antigas e mais duradouras, com uma duração de quase cinco milénios. Beneficiou de uma abundância de boas terras, de recursos minerais próximos e de uma boa posição estratégica.

Localização

O EGIPTO antigo ocupava quase a mesma área que o Egipto actual ocupa hoje. A sua civilização, muito perto do Rio Nilo, era cercada quase completamente pelo deserto.

2.2.2.1 Os recenseamentos e a estatística de "massa"

Se o cálculo remonta às mais antigas comunidades humanas, a estatística de "massa" teve início com os grandes Impérios da Antiguidade, preocupados em administrar os seus bens, os seus homens, as suas armas e as suas imensas obras públicas. Esta enumeração presume uma organização complexa e uma forte estrutura administrativa. Mas os recenseamentos já eram praticados por uma das mais antigas civilizações conhecidas: o Egipto, provocado em parte por falta de mão de obra ligada à construção das pirâmides. Um registo de Pierre de Palerme datado de 2900 A.C. fez, de facto, alusão ao recenseamento de pessoas. No período de 2700 a 2500 A.C., já existiam recenseamentos bianuais, depois anuais, sobre os diferentes bens que tinham como destino a fiscalização. Por volta de 1900 A. C., são estabelecidas as listas dos familiares dos soldados; estas informações destinavam-se ao uso fiscal e militar. Em meados de 1200 A.C. apareceram as listas das casas, dos chefes de família e seus parentes , com a indicação do nome do pai e da mãe de cada ocupante. No tempo de Amasis II (Século VI a.C.) todos os indivíduos tinham de declarar todos os anos ao governo da sua província (incorrendo na pena de morte, caso não o fizessem) a sua profissão e suas fontes de rendimento.

Recenseamento:

Estudo de um universo de pessoas, instituições ou objectos físicos com o objectivo de obter conhecimentos quantitativos acerca das características importantes dessa população.

Os antigos egípcios acreditavam que poderiam comunicar com os deuses através do rei. O rei tinha poder absoluto, dirigia o governo, o comércio e a política externa, aplicava as leis e conduzia o exército.

Todos os trabalhadores pagavam impostos, calculados a partir de uma percentagem de sua produção. Além disso, cada casa tinha que disponibilizar um trabalhador por várias semanas em cada ano para a realização de obras públicas. As pirâmides provavelmente foram construídas por trabalhadores que contribuíam com os seus serviços anuais. De facto, o rigor da sua construção e as suas dimensões implicavam uma organização de trabalho humano nunca antes demonstrada em nenhuma outra civilização.

Ora, a administração deste Estado, constituída essencialmente pelos numerosos «escribas», só era possível graças a um grande número de funcionários muito eficazes e . Estes usavam caracteres hieroglíficos que apareceram na Fenícia no ano 3000 a.C., escritos a partir de imagens e que vigoraram até ao fim do Império Egípcio.

2.2.3. Israel e os Números (1700 a.C. a 70 d.C)

As pessoas confundem os termos "Hebreu", "Judeu" e "Israelita". Os Hebreus são os primeiros judeus, os primeiros habitantes da Terra de Israel, aqueles que usaram pela primeira vez a língua hebraica. O termo tem um sentido mais étnico e tribal do que religioso. Quanto a israelitas e judeus, fazia-se uma distinção no período entre os séculos X e VIII a.C., quando dez tribos se estabeleceram no norte da Terra Santa (Reino de Israel) e duas no sul (Reino de Judá). Hoje, porém, os dois termos são sinónimos.

Em "Pour une Histoire de la Statistique" (Bedarida et al, 1987), refere-se que a atitude dos Hebreus relativamente aos censos contribuiu, largamente, para modelar a opinião ocidental durante quase 2000 anos.

O legado cultural hebreu foi importante para a formação de vários traços da cultura ocidental, pois a produção cultural hebraica está ligada com a sua vida religiosa.

Dos hebreus guardamos também sua cultura e a crença em um Deus único, Criador de todo o Universo e de todas as coisas. Boa parte da Bíblia foi escrita por eles.

Deste modo, a história do povo hebreu não pode ser dissociada da história da sua religião, pois o que sabemos sobre o povo Hebreu deve-se sobretudo às informações da Bíblia, principalmente do Antigo testamento. Assim sendo, a referida obra chama a atenção para um facto curioso a observar, que é a atitude ambígua, hesitante e contraditória que reporta a Bíblia. Na maior parte das vezes , os recenseamentos eram tidos como sacrílegos porque se declaravam contra o segredo da vida e da criação, do qual Deus era o único detentor. É claro que aqui como noutros lugares, a população receava ver-se recenseada para fins fiscais e militares, e parecia-lhes, por outro lado, que fazer inventários da sua riqueza, tanto de homens como de bens, podia trazer desgraça.

Por todas estas razões, os recenseamentos não parecem ser admissíveis senão quando ordenados pelo próprio Deus. E além disso, são por vezes atribuídos a Satanás, o que parece ser o único meio para explicar os males que lhes aconteciam, como se as vidas recenseadas não pudessem ser resgatadas e para sempre ficassem condenadas.

Segundo os hebreus antigos, os recenseamentos não parecem ser admissíveis senão quando ordenados pelo próprio Deus. E além disso, são por vezes atribuídos a Satanás, o que parece ser o único meio para explicar os males que lhes aconteciam, como se as vidas recenseadas não pudessem ser resgatadas e para sempre ficassem condenadas.

O recenseamento ordenado por Deus em Sinai é relatado em duas passagens, no livro de Moisés ao qual foi dado o nome de «Números». Iavé impôs a Moisés no deserto de Sinai: «fazei o recenseamento geral de toda a comunidade dos filhos de Israel, clã por clã, família por família» (Números, 1, 2). No livro do Êxodo (30, 12-15), está escrito que, quando Moisés fez o recenseamento daqueles que deviam ser numerados, «cada um deveria pagar a Iavé para redenção da sua vida, para que esse recenseamento não lhe trouxesse calamidades». Iavé exigia homenagens e oferendas exclusivas em sua honra, e, em troca, seria o Todo-Poderoso protector do povo hebreu.

2.2.4. A máquina de recensear Chinesa

Localização

A CHINA localiza-se no extremo sul do continente asiático. O País é cortado por grandes rios: rio Amarelo e Azul, que com outros rios, Branco e Vermelho, formam longos vales que fertilizam os campos do coração da China.

No Extremo Oriente também se desenvolveram civilizações antigas perfeitamente acostumadas com a prática dos recenseamentos.

Os registos históricos mais antigos dizem-nos que o primeiro recenseamento foi realizado no ano 2238 a.C., pelo primeiro imperador da China, Yu ou Yao. O regime chinês desejava conhecer com exactidão o número de habitantes, a fim de poder repartir o território, de distribuir as terras, estabelecer os rolos de pergaminho de impostos e de proceder ao recrutamento militar.

Foram vários os recenseamentos efectuados na China:

- Os recenseamentos ligados a um sistema de recrutamento (época da dinastia dos Han, 200 a.C. – 200 d. C.). O Estado, como meio de centralização, procura avaliar o número de soldados disponíveis para as guerras e para o trabalho público.
- Os recenseamentos ligados ao sistema de distribuição das terras (do terceiro reino à quinta dinastia: 221-959 d.C.). Para encorajar a produção agrícola e restringir os grandes domínios o soberano redistribui, com efeito, as terras em troca de serviços e de pagamentos em prazos fixos e surge a necessidade de conhecer a dimensão e a composição das famílias.
- De 960 a 1368 d.C. os recenseamentos têm como objectivo principal a fiscalização. A noção de família ainda prevalece.
- Na época de Ming (1368-1844 d.C.), funciona o que M. Cartier chamou de uma «admirável máquina» de recenseamento. Até ao fim da dinastia, procede-se à redacção dos «registos de cartas» da população. Estes registos continham o nome, a profissão, o sexo e a idade.

• A partir do 1644 d.C.(Ching) houve um período de registos para a policia, destinados a vigiar a deslocação dos habitantes e a despistar os indivíduos pouco recomendáveis. Em 1741 são modificados os métodos de estimação. Em 1975, vigorou o sistema pao-chia, que exigia a aposição em todas as casas de um cartaz indicando o número de ocupantes, o sexo, a idade, a profissão e o montante dos seus tributos. Este sistema permitiu obter séries demográficas desde 1750 a 1850.

Em suma, durante um longo período, o imenso império Chinês esforçou-se por se recensear apesar das dificuldades com uma "paciência" comparável ao rigor científico dos Estados modernos.

2.2.5. O Japão até a Tokugawa

Localização

O JAPAO localiza-se no extremo leste da Asia sendo formado por quatro ilhas principais e 3 mil ilhas mais pequenas. O país está exposto a terramotos e erupções vulcânicas. E a segunda potência económica mundial.

O Japão parece ter conhecido os recenseamentos numa época bem remota da história, mas os resultados desses recenseamentos não foram divulgados. O primeiro surgiu no ano de 86 a.C., no tempo do imperador Soujin. As actividades da população, nesse tempo, eram registadas de modo a permitir examinar a sua evolução. A meio do século VII a.C a reforma de Taika que visava submeter toda a população a um tributo coincide com a redistribuição das terras, o que necessitava do estabelecimento de um cadastro e de registos de direitos civis revistos todos os 6 anos. As famílias eram recenseadas pela casa da câmara e arquivadas em função dos seus recursos, com distinção do sexo e do grupo etário. Este recenseamento não tinha somente como objectivo a tributação de impostos, mas também facilitar o recrutamento militar e o trabalho forçado.

Segundo este livro, no tempo dos Tokugawa (séculos XVII-XIX), no fim do século XVII (1665), efectuaram-se recenseamentos locais. Em 1721, procedeu-se ao primeiro recenseamento geral, operação que deveria ser repetida de 6 em 6 anos. Neste recenseamento eram excluídas certas categorias da população, assim como os nobres, os habitantes mais pobres ou as crianças com menos de 15 anos. Como é evidente, este registo comportava um certo número de lacunas. Daí se compreende a grande ansiedade que os Japoneses tinham pelo desenvolvimento da demografia.

2.2.6. Um tratado de recenseamento na Índia Antiga

Localização

A INDIA é um país situado a sul da Ásia, com forma de losango. É limitado a Norte pela China, Nepal e Butão; a Este por Myanmar; a Noroeste pelo Paquistão; e a Sudeste, Sul e Sudoeste pelo oceano Índico.

Um outro exemplo, muito conhecido, de interesse demonstrado pelos impérios asiáticos na enumeração da sua população é o tratado redigido pelo hindu Kautilya, ministro do rei Candragupta (313-289 a.C.), fundador da dinastia e do primeiro império indiano os Maurya (313-226 a.C.), no século IV antes da nossa era. Este tratado era extremamente original e avançado para a época. Sendo de ciência política é também um tratado de economia: o seu nome correcto era Arthasástra, ou seja tratado ou ciência (sástra) do progresso (artha).

Nesta obra, que descreve o estado centralizador e expansionista que era o império Maurya, Kautilya, (mais tarde Machiaval), reflecte sobre a arte de governar e indica ao soberano como aumentar incessantemente o seu reino. Exactamente como Kautilya, o Estado deverá dirigir e controlar tudo. Mestre absoluto da economia, ele governa com o auxílio de um aparelho administrativo muito extenso, desempenhado pelo exército e pela polícia secreta. Para se realizar um "rol planificador", o Estado, segundo Kautilya, terá de recorrer aos recenseamentos, à estatística e ao cadastro. "Tudo o que for feito terá que ser conhecido: do efectivo da população até o número de elefantes, passando pelas matérias-primas, os produtos fabricados, os preços e os salários".

Arthasástra: O Tratado do Progresso

Em Arthasástra, Kautilya descreve com muita precisão as tarefas dos revisores nos diferentes escalões territoriais. Em cada estado o revisor deve dividir o país em quatro províncias, recensear e transferir para a escrita o número de aldeias e ordená-las conforme a sua riqueza (ricas, médias e pobres), de modo a melhor contabilizar o trabalho e os produtos que, em grande parte, eram entregues sob a forma de impostos. Por outro lado, com esta orientação pretendia-se, também, fazer um melhor recrutamento de soldados.

O revisor provincial assegurava a escrituração dos registos, nomeadamente das casas e das pessoas que não pagavam os impostos. Por outro lado, estavam também registados o nome das pessoas pertencentes a cada uma das quatro classes (varsa), o número de feitores, de pastores, de comerciantes, de artesãos, de trabalhadores livres ou escravos, o número de animais, e ainda a quantidade de dinheiro, de trabalho, de direitos e coimas. O revisor registava igualmente, em cada família, o número de mulheres e de homens, de crianças, de pessoas idosas, e os seus ofícios, os seus modos de vida, o montante dos seus recursos e das suas despesas.

Por sua vez, o governador geral do país mantinha o registo do número de habitantes, o sexo, a casta, o nome de família e o ofício, e também o domicílio, os recursos e as despesas.

Assim informado e apetrechado, o Estado, segundo Kautilya, poderia, mais eficazmente executar o seu rol de previsões e de racionalizações.

2.2.7. O recenseamento em Roma (750 a.C. a 476 d.C.)

Localização

A ITÁLIA estende-se no centro do mar Mediterrâneo, tendo a Sul e a Oeste duas grandes ilhas: Sicília e Sardehna. Cerca de 80% do território é montanhoso ou colinoso, sendo a maior extensão de terra plana a da planície Padana, atravessada pelo Rio Pó.

A cidade de Roma foi fortemente influenciada, em matéria de recenseamentos, no que respeita a conceitos e práticas, pelo pensamento Oriental. No fim do século VI antes de Cristo, os recenseamentos eram feitos de 5 em 5 anos, até ao ano 68 a.C. e, depois de uma interrupção de uma vintena de anos, foram retomados por Augusto sob uma forma decenal.

Segundo a tradição, o primeiro recenseamento autorizava a repartição entre as tarefas civis e as militares não por cabeça, mas segundo a fortuna.

Os cidadãos romanos eram obrigados a declarar as suas fortunas, o seu nome, o dos seus pais, a idade, o nome da sua esposa assim como o dos seus filhos, a tribo onde residiam e o número de escravos. Caso não fornecessem algumas destas informações poderiam ficar sem os seus bens ou sem os direitos de cidadão.

Os censos permitiam não só classificar os cidadãos segundo os seus rendimentos, mas também cobrar impostos sobre os seus rendimentos e determinar a condição social que lhes permitisse ter funções a nível político e militar na cidade.

2.2.8. As Estatísticas na Era de Cristo...

A data do nascimento de Cristo é hoje bastante controversa, pois o governador romano da Síria que incluía a Judeia e a Galileia, por ordem do Senado, teve de fazer um recenseamento para o qual utilizou uma técnica, talvez a mais absurda de todas (Collected Works: obras de J.Tiago de Oliveira, Volume II,1995). A Bíblia conta que São José e a virgem Maria saíram de Nazareth, na Galileia, para Belém, na Judeia, para responder ao Censo ordenado por César Augusto (as pessoas tinham que ser entrevistadas no local de sua origem). Foi enquanto estavam na cidade que Jesus nasceu.

Em Portugal está escrito em Diário da República e portanto constitui lei, que os jogos de azar são, pura e simplesmente, jogos de Acaso. O que não significa, portanto, jogos de má sorte.

3. ...Até à Idade Moderna

3.1. As Estatísticas e os Jogos de Azar

Os jogos sempre tiveram grande interesse e foram largamente praticados em todas as Civilizações. Eram de tal maneira importantes que, no Olimpo grego, havia uma Deusa "encarregada" das artes do Acaso, que era a Deusa Thykhe, parente da Deusa da fortuna do Panteão romano, de todos conhecida pela chamada roda da fortuna, que era o seu símbolo (Oliveira, 1995). O termo Acaso, ou mais propriamente o termo Azar não significa aqui má sorte ou má fortuna; a palavra azar vem do árabe e significa exactamente Acaso.

O termo "azar", usado na expressão "jogos de azar" não significa má sorte ou má fortuna mas simplesmente Acaso.

3.2. O início das Probabilidades

Como refere J. Tiago de Oliveira, em Jerusalém ainda existe um traçado no chão da prisão em que esteve Cristo, formando um quadrado dividido em nove partes iguais, relativo ao velho jogo do galo. Do mesmo modo os jogos estiveram sempre presentes em quase todas as civilizações, como o mostram vários documentos do tipo arqueológicos ou históricos. Curiosamente, os jogos nunca foram objecto de estudo até à Idade Média.

A abordagem matemática do acaso, do azar e do risco só se iniciou há pouco mais de 500 anos. A disciplina que assim foi constituída, a Teoria das Probabilidades, nasceu das tentativas de quantificação dos riscos dos seguros e de avaliar as possibilidades de se ganhar em jogos de azar.

Com o término da Idade Média, o crescimento dos centros urbanos levou ao aparecimento do seguro de vida. Foi em torno desses assuntos que surgiram os primeiros estudos matemáticos sobre seguros. Mas, só passados quase 250 anos, com Daniel Bernoulli, é que a matemática dos seguros atingiu um estado suficientemente maduro. Ele retomou um problema clássico

de, a partir de um número determinado de recém nascidos, calcular o número esperado de sobreviventes após n anos. Bernoulli deu também os primeiros passos em direcção a novos tipos de seguros calculando a mortalidade causada pela varíola em pessoas com uma dada idade.

Técnicas Combinatórias:

Técnicas de contagem que nos permitem saber quantos são os resultados possíveis de uma experiência. Não interessa saber quais são os resultados (enumeração directa), mas sim qual o número de resultados.

Girolano Cardano (1501/1576) foi um matemático notável, vigarista notável, médico notável, probabilista notável, algebrista notável e escreveu um pequeno manual de jogos de azar "Liber de Ludo Aleae", que é, talvez o primeiro sobre probabilidades, que analisa jogos e possibilidades. Cardano foi o primeiro a introduzir técnicas combinatórias para calcular a quantidade de possibilidades favoráveis num evento aleatório. Limitou-se a resolver alguns problemas concretos, isto é, problemas com dados estritamente numéricos, mas nunca chegou a produzir nenhum teorema. Podemos considerar Pascal (1623/1662) e Fermat (1601/1665) como sendo os fundadores do Cálculo das Probabilidades.

O primeiro grande problema das Probabilidades, que foi proposto pelo Cavaleiro de Méré a Pascal, surgiu na corte dos reis de França onde a nobreza se divertia, entre outras actividades, a jogar. Tratava-se da procura da compreensão de um determinado jogo com três dados de que Méré não conseguia entender os resultados empiricamente observados. Pascal e Fermat, separadamente, encontraram a solução do problema, mas a solução de Pascal era muito específica enquanto que a de Fermat constituiu talvez o primeiro método geral das probabilidades. Naquele problema surgiam duas situações que se punham com a mesma probabilidade mas que diferiam na verificação empírica da análise de frequência. Começa aqui a surgir a ideia da Lei dos Grandes Números e a identificação "automática" entre probabilidade e frequência num elevado número de provas.

Blaise Pascal nasceu em 1623 em Clermont. Filósofo, matemático, físico, teólogo e escritor deu uma grande contribuição para o desenvolvimento do estudo das probabilidades, descobrindo novas propriedades do triângulo aritmético, conhecido entre nós como o Triângulo de Pascal.

Pierre de Fermat, nasceu em 1601 em Beaumont. Conhecido como o "Príncipe dos Amadores em Matemática", estudou matemática por vocação, tendo sido, como advogado, conselheiro do Parlamento de Toulouse desde 1631. É considerado o criador da teoria dos números e precursor da geometria analítica, cálculo das probabilidades e cálculo diferencial. O seu contributo para o cálculo das probabilidades derivou da correspondência estabelecida com o seu colega Pascal para tentarem resolver os problemas expostos pelo Cavaleiro de Méré.

Inicia-se então um período, que termina no princípio do século xx, em que a Estatística é marginalizada e em que o que se desenvolve é o Cálculo das Probabilidades.

A LEI DOS GRANDES NUMEROS, em linguagem simplista diz-nos que a frequência de um acontecimento, numa longa série de experiências, se aproxima, cada vez mais, da probabilidade desse acontecimento, probabilidade que assim surge como uma frequência –limite. Ou seja, a Lei dos Grandes Números exprime-se pela ideia de que se a probabilidade de uma face de um dado é 1/6, em 100 experiências sucessivas independentes cerca de 100/6 vezes essa face aparecerá, em 1000 experiências sucessivas independentes cerca de 1000/6 vezes essa face aparecerá, etc.

3.2.1 A curiosidade do "passe-dix"

"A incerteza tem sido, desde há longos tempos, uma preocupação do homem. E foi a arte lúdica dos jogos que, através das probabilidades, construiu os instrumentos e as regras que permitem à Estatística medir a intensidade de incerteza (ou de realização) dos fenómenos." (Oliveira, 1995)

O "Passe - Dix"

Na corte de França era comum o jogo do "passe–dix" em que o jogador atira 3 dados simultaneamente e ganha se a soma dos pontos passa de 10, perdendo se a soma for 9 ou inferior. Um inteligente e culto jogador inveterado, o Cavaleiro de Méré, ao tempo de Luís XIV, tinha observado que saía mais vezes a soma 11 do que a soma 12, facto que lhe parecia estranho pois as formas que lhe levavam às somas 11 e 12 são as seguintes:

Quadro 1

Soma 11	Soma 12
6+4+1	6+5+1
6+3+2	6+4+2
5+5+1	6+3+3
5+4+2	5+5+2
5+3+3	5+4+3

e portanto em número igual (6) o que devia dar frequência igual ou muito aproximada. Todavia é fácil ver que enquanto a forma (6,4,1) se pode dar de 6 modos (pense-se, por exemplo, que os dados são de cores diferentes e que 6,4,1 pode sair com 6 no dado branco, 4 no azul, 1 no verde ou com 6 no azul, 4 no verde e 1 no branco, etc., ao todo de 6 maneiras), já o mesmo não sucede para a forma (4,4,3) que só pode acontecer dos 3 modos em que o "3" sai com um dos três dados e os "4" nos outros dois. Feitas agora as contas com cuidado (o número total de modos está entre parêntesis, a seguir a cada forma) vê-se que 12 só pode acontecer de 25 modos enquanto que 11 pode ser observado de 27 maneiras diferentes. Méré tinha, pois, verificado correctamente que no jogo de "passe–dix" a soma 11 era mais frequente (provável) do que a soma 12, em contradição com o que à primeira vista parecia dever acontecer.

Quadro 2

Soma 11	Soma 12
6+4+1(6)	6+5+1(6)
6+3+2(6)	6+4+2(6)
5+5+1(3)	6+3+3(3)
5+4+2(6)	5+5+2(3)
5+3+3(3)	5+4+3(6)
4+4+3(3)	4+4+4(1)
(27)	(25)

3.3 O desenvolvimento da Estatística

É a partir do século XVIII que a Estatística começa a caminhar para a ciência que conhecemos hoje em dia.

Nessa altura apareceram duas Escolas, uma na Alemanha e outra em Inglaterra. A Escola Descritiva Alemã, assim como ficou conhecida, afastou-se das ideias que fundamentaram a Estatística Moderna. O representante mais conhecido da Escola Alemã foi Gottfried Achenwall (1719-1772), o qual é considerado por alguns autores o "pai" da palavra Estatística. Mas, na opinião de Sir Maurice Kendall (Pearson e Kendall, 1820), esta palavra já tinha sido utilizada em Itália, num trabalho do historiador Girolamo Ghilini, em 1589 que se refere a um registo da "civile, politica, statistica e militare scienza". Segundo Kendall, a palavra utilizada na Escola Alemã denotava apenas o método utilizado nos estudos dedicados à descrição dos estados políticos e, se alguma informação numérica aparecia nesses registos era somente por acaso ou conveniência. A Escola Inglesa, "Escola de Aritméticos Políticos", preocupava-se com o estudo numérico dos fenómenos sociais e políticos.

A Escola de Aritméticos Políticos preocupava-se com o estudo numérico dos fenómenos sociais e políticos, enquanto que a Escola Alemã somente fazia a descrição dos estados.

Da Escola Inglesa surgiram dois Estatísticos importantes para o desenvolvimento da Estatística Moderna, sendo eles, John Graunt (1620-1674) e William Petty (1623-1687).


Natural and Political
OBSERVATIONS
Bills of Mortality,
By JOHN GRAUNT,
LONDON.

O trabalho desenvolvido por John Graunt (Seneta e Heyde, 2001) constituiu a base da Estatística Moderna. Graunt estudou a mortalidade da cidade de Londres e as incidências das causas naturais, sociais e políticas nesse fenómeno. Através das Tábuas de Mortalidade realizadas na altura da peste na cidade de Londres, Graunt fez uma análise exaustiva do número de pessoas que morriam de várias doenças e estimou o número de nascimentos de homens e mulheres. Foi a primeira pessoa a fazer observações entre sexos e mostrou que nasciam mais homens que mulheres e que por cada 100 pessoas nascidas, 36 morriam aos 6 anos e 7 sobreviviam até aos 70 anos.

John Graunt nasceu em 1620 em Londres. Homem bem conceituado e muito estudioso, ocupou cargos muito importantes na cidade de Londres. Herdou a loja do seu pai e conseguiu por o negócio em grande evolução. Foi Capitão da banda militar e, nos últimos anos, Major. Um dos fundadores da Royal Society, viveu numa época marcada pelo nascimento da ciência moderna. Em 1662, Graunt publicou a sua grande obra *Natural and Political Observations on the London Bills of Mortality* o qual foi o seu primeiro tratamento estatístico de dados demográficos e a tentativa de aplicar a teoria a problemas reais.

Graunt publicou a sua obra Natural and Political Observation Made Upon The Bills of Mortality em 1662, a qual deu um grande impulso à análise quantitativa dos fenómenos sociais e ao desenvolvimento das Estatísticas Demográficas. O trabalho realizado por John Graunt chamou a atenção de Carlos III (Rei de Inglaterra), que propôs a Graunt ser sócio fundador da Royal Society.

William Petty trabalhou em conjunto com John Graunt durante três anos e, também ele pode ser considerado como um impulsionador da Estatística Moderna.

John Graunt nasceu em 1620 em Londres. Homem bem conceituado e muito estudioso, ocupou cargos muito importantes na cidade de Londres. Herdou a loja do seu pai e conseguiu por o negócio em grande evolução. Foi Capitão da banda militar e nos últimos anos Major. Um dos fundadores da Royal Society, viveu numa época marcada pelo nascimento da ciência moderna. Em 1662, Graunt publicou a sua grande obra Natural and Political Observations on the London Bills of Mortality o qual foi o seu primeiro tratamento estatístico de dados demográficos e a tentativa de aplicar a teoria a problemas reais.

Antes de aparecer a Empresa Geral de Registos em Inglaterra, Petty já tinha proposto uma empresa de Estatística Central. Esta empresa não tinha só como objectivo o registo dos baptismos, casamentos e mortes, mas também as características das casas, o tamanho das famílias, o sexo, a idade, a forma de ocupação e nível de estudos de cada membro da família. Propôs a elaboração de Tábuas de Sobrevivência baseadas em taxas de mortalidade por grupos etários. A ligação das probabilidades com os conhecimentos estatísticos veio dar uma nova dimensão à Estatística. Considera-se uma nova fase, em que se começa a fazer Inferência Estatística. Neste período alguns estudiosos evidenciam-se. É o caso de Christian Huygens (1629-1695) que introduz a noção de valor médio ou esperança matemática, em 1654.

Outro dos estudiosos foi Abraham De Moivre (1667-1754) que abriu caminho ao desenvolvimento da geometria analítica e da teoria das probabilidades; publicou em 1718 o célebre Doctrine of Chances sobre a teoria do acaso, onde expôs a definição de independência estatística junto com muitos problemas relacionados com dados e outros jogos, por exemplo a probabilidade de tirar bolas de cores diferentes de uma urna. É atribuído a De Moivre o princípio segundo o qual a probabilidade de um acontecimento composto é o produto das probabilidades das componentes, embora essa ideia já tivesse aparecido em trabalhos anteriores. Também ele se interessou pelas estatísticas demográficas e fundou a teoria das pensões.

Inferência Estatística

Fase fundamental da análise estatística, durante a qual, conhecidas certas propriedades (obtidas a partir de uma análise descritiva da amostra), expressas por meio de proposições, se imaginam proposições mais gerais, que exprimam a existência de leis (na população).

Mas as três grandes figuras da Teoria das Probabilidades foram, na verdade, Jacob Bernoulli, Thomas Bayes e Pierre Simon Laplace.

Jacob Bernoulli (1654-1705) em 1713, de quem é editada "posmortem", a "Ars Conjectandi", mostra, ao mesmo tempo que Leibniz, uma consciência do que vai ser ou deve ser a ciência Estatística. Uma das grandes contribuições para a Estatística, foi a distribuição de Bernoulli, que consiste em dizer que cada tentativa tem duas possibilidades de ocorrência chamadas: sucesso e insucesso (ex.: no lançamento de uma moeda ou sai cara ou coroa). Esta distribuição foi a base da distribuição binomial.

Todos estes contributos foram extremamente importantes para a Estatística porque começaram a levantar os grandes problemas da Teoria das Probabilidades. Problemas que só foram resolvidos de maneira completa, metódica e sistemática em 1933 por Kolmogorov.

Prova de Bernoulli:

1. Considera-se à partida um número fixo, n, de observações, a que é usual chamar provas;
2. As observações são independentes umas das outras;
3. Em cada observação pode-se obter um de dois resultados possíveis a que chamamos sucesso ou insucesso;
4. A probabilidade de sucesso, p, é constante de observação para observação.

Posteriormente surge Bayes (1701-1761) que, segundo Tiago de Oliveira, foi o primeiro a lançar claramente o problema fundamental da Estatística: de que maneira, a partir das observações, é possível saber alguma coisa relativamente a um certo universo. Em 1762 Bayes demonstrou o método que ficou conhecido pela Regra de Bayes, a qual consiste na partição do espaço amostral em diversos subconjuntos cujas probabilidades são conhecidas e é representada pela seguinte fórmula:

$$P(A_i / B) = \frac{P(B / A_i)P(A_i)}{\sum P(B / A_j)P(A_j)}$$

As ideias de Thomas Bayes não foram muito bem aceites pelos cientistas daquela época pois as equações resultantes da Estatística Bayesiana eram por vezes bastante difíceis de resolver. Já no século XX, a partir da década de 90, com o crescente desenvolvimento dos computadores, essas ideias foram recuperadas e são frequentemente aplicadas em estudos estatísticos.

Entretanto, surge uma outra figura de grande relevo, Pierre Simon de Laplace (1749-1827), que publicou em 1812 o tratado "Teoria Analítica das Probabilidades" (Théorie Analytique des Probabilités), constituindo um grande marco da Teoria das Probabilidades. Neste tratado Laplace definiu probabilidade como o número de vezes em que um dado acontecimento pode ocorrer, dividido pelo número total dos casos que podem acontecer, considerando-se que estes têm possibilidades iguais de acontecer.

Pierre Simon de Laplace, nasceu em 1749 na Normandia (França). Astrónomo e matemático francês, estudou em Beumont-en-Auge, onde começou a despertar o seu interesse pela matemática. O seu grande contributo para o desenvolvimento da Estatística deve-se à publicação do tratado "Teoria Analítica das Probabilidades" onde descreveu um cálculo útil para assegurar um "grau de credibilidade racional" a proposições sobre acontecimentos aleatórios.

"…É notável que tal ciência, que começou nos estudos sobre jogos de azar, tenha alcançado os mais altos níveis do conhecimento humano."

Laplace

Tiago de Oliveira (1995), refere que a Estatística está por vezes reduzida, como sucede nos países menos desenvolvidos, a uma contabilidade dos factos, a uma listagem de acontecimentos, como por exemplo, sobre o número de indivíduos que morreram com a doença A ou B, sem a análise das causas desses factos.

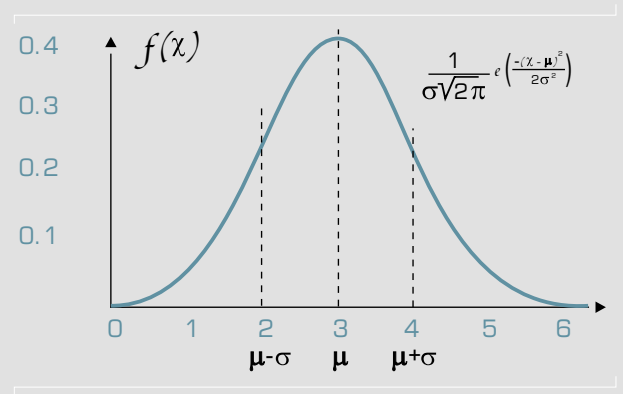
O primeiro a abordar o problema com bastante clareza e a defender a criação de um serviço autónomo de Estatística, foi o belga Adolph Quételet (1796-1874) que em 1846 propôs a organização de censos e preparou a organização do grande serviço belga de Estatística. Quételet generalizou o uso da distribuição normal além da sua aplicação para a análise de erros e, em particular, a aplicação da distribuição normal para o estudo das características humanas, tais como altura e peso. Quételet melhorou os métodos para a recolha de dados e trabalhou na análise estatística de dados que envolvem crime, mortalidade, geofísica e astronomia, organizou a primeira conferência de estatística em 1853 e escreveu "Sur l'homme et le développement de ses facultés, essai d'une physique sociale", publicado nesse ano.

"...todas as ciências de observação, no princípio, passaram pelas mesmas fases; foram artes, porque elas se limitavam a agrupar duma maneira mais ou menos feliz colecções de factos pertencendo a uma mesma ordem de coisas; e foi pela comparação e estudo destes factos que foram elevadas de seguida, à posição em que as vemos hoje. Porquê mostrar-se mais exigente para com a Estatística?"

Adolph Quételet

Outro matemático que deu um grande contributo para o desenvolvimento da Estatística foi o "Príncipe dos Matemáticos", Carl Friedrich Gauss (1777-1855). Forneceu o ponto de partida para algumas das principais áreas de pesquisa da matemática moderna; formulou a chamada lei de Gauss, que trata da distribuição de certos valores ao longo de uma curva em forma de sino (contribuição extremamente valiosa no campo da Estatística).

Exemplo de uma Curva de Gauss



A distribuição normal é uma aproximação à distribuição de valores de uma característica. A forma exacta da distribuição depende da média e do desvio padrão da distribuição.

Duas figuras igualmente importantes para o desenvolvimento da estatística foram: Siméon Denis Poisson (1781-1840), que em 1810 descobriu a forma limitada da distribuição binomial que posteriormente recebeu o seu nome; e Marquês de Condorcet (1743-1794), que é o primeiro a fazer a aplicação destas « artes mágicas do Acaso» aos problemas de carácter social e a analisar metodicamente o problema das votações.

Estes dois homens foram os primeiros a preocuparem-se com as aplicações sociais da estatística.

A partir da segunda década do século XIX, dá-se uma explosão no desenvolvimento da estatística moderna, tendo como principal responsável, Ronald A. Fisher, conhecido entre nós como o pai da estatística moderna. Quanto a este célebre matemático, vamos conhecê-lo no capítulo seguinte.

4. A Estatística nos dias de hoje

4.1. Introdução

É na segunda metade do século XIX, que se dá a viragem da Estatística Descritiva ou Gráfica para o estudo metodológico, a qual se iniciou a partir do Primeiro Congresso de Estatística que teve lugar em Bruxelas, em 1853 (Oliveira, 1995). Até aqui, a Estatística era vista somente como uma mera compilação de dados, a sua disposição em tabelas, uns tantos cálculos de médias e outras estatísticas simples...e pouco mais. A decisão Estatística era, tantas vezes, feita de um modo intuitivo, vendo se o valor calculado a partir da amostra estava próximo ou distante daquele que teoricamente se esperava. É nesta altura que surgem novos nomes importantes para o desenvolvimento da Estatística, sendo eles Galton, Karl Pearson, "Student", Lexis e Von Bortkiewicz. Estes matemáticos, "abrem" caminho para Fisher, Neyman e Wald, lançarem os fundamentos da Estatística Moderna, a procura dos métodos óptimos da inferência, o estudo do comportamento indutivo, rigorizando a comparação indutiva e vaga.

4.2. A Estatística no Estudo da Hereditariedade Humana

Na área da hereditariedade pode afirmar-se que os "pais" da Inferência Estatística, foram J. Neyman e Karl Pearson. Embora os estudos estivessem associados a questões relacionadas com a Biologia e a Genética, os métodos que criaram, tais como a "hipótese nula" e "nível de significância", fazem hoje parte da rotina diária de todo o estatístico e cientista que precisa da Estatística.

Francis Galton

Francis Galton, um dos grandes fundadores da ciência moderna e da ciência humana, em particular no século XIX, foi o fundador da antropologia, do estudo da natureza humana e de suas origens, autor de muito do estudo da meteorologia (descobriu e introduziu o termo anticiclone) e instituiu o começo do estudo da genética.

Fundador do termo Eugenia e activamente envolvido na sua prática, a qual propunha o melhoramento genético da espécie humana, Francis Galton, acreditava que as características físicas e mentais dos seres humanos seriam devidas à hereditariedade. Idealizou instrumentos para medir a capacidade sensitiva, a memória e a imaginação. Publicou, em 1865, um livro "Hereditary Talent and Genius" onde defende a ideia de que a inteligência é predominantemente herdada e não fruto de acção ambiental.

A ambição principal de Galton era provar como é que o carácter e os talentos foram transmitidos pela reprodução através de sucessivas gerações. Instalou o seu laboratório em Londres, onde os visitantes podiam fazer-se examinar desfilando perante os seus instrumentos. A altura, o peso, a envergadura do palmo, a capacidade respiratória, a força, etc., eram medidos no laboratório

Francis Galton nasceu a 16 de Fevereiro de 1822 perto de Birminghan, Inglaterra. Afirma-se que, antes de completar 3 anos, foi capaz de ler um livro simples, e desde muito jovem deu provas de engenho para a mecânica e para as matemáticas. Fundador da escola biométrica, interessou-se pelos métodos estatísticos e pela sua aplicação a todas as espécies de domínios. Os trabalhos de Galton são baseados na medição quantitativa feita a partir da lei normal de Gauss. A sua contribuição essencial na Estatística é o conceito de correlação e a sua medição pelo coeficiente de correlação.

de Galton. Com os dados recolhidos elaborou gráficos, curvas de probabilidade, valores médios, entre outros cálculos. Galton criou um esquema explicativo que mais tarde viria a dar lugar à medida da correlação entre duas variáveis. Seria Pearson a formular, mais tarde, o coeficiente de correlação. Por volta de 1870, Galton teve a ideia de modificar um dispositivo que tinha criado e usado em lições para ilustrar as bases da lei do erro. A este dispositivo chamou-o de quincunx. (ver caixa explicativa)

EUGENIA:

Termo definido por Francis Galton como sendo o estudo dos agentes sob o controlo social que podem melhorar ou empobrecer as qualidades raciais das futuras gerações seja física ou mentalmente.

Galton modificou o quincunx para demonstrar que as distribuições normais eram habitualmente uma mistura de distribuições normais. Por outras palavras, com a força da experimentação e o dispositivo que ele inventou, chamado quincunx, concluiu que possuía uma clara prova experimental de que as causas significativas dos fenómenos poderiam, de facto, ser isoladas em conformidade com a lei do erro.

Numa primeira fase Galton inspirou-se no mundo natural, inicialmente reflectindo em pomares de fruta, e como é que factores específicos, tais como o aspecto, podem afectar o tamanho da fruta.

Galton Quincunx

Este aparelho consiste num conjunto de bolas de chumbo que descem por um rampa com grande inclinação. Estas, durante o seu percurso, colidem com pregos colocados ao longo da rampa.

Não é difícil imaginar condições nas quais as bolas têm igual probabilidade de ressaltar à esquerda ou à direita do prego. Se por baixo de cada prego estão colocados dois pregos numa linha horizontal e o declive da rampa estiver correctamente ajustado, a bola baterá num ou noutro depois de ressaltar do primeiro prego. Novamente a bola deve ter igual probabilidade de queda à esquerda ou à direita desses pregos.

As probabilidades de queda à esquerda de ambos ou entre eles ou à direita de ambos, deveriam estar na proporção 1:2:1. O processo pode ser continuado e está claro que as probabilidades de um bola passar entre os pregos diferentes de uma fila são proporcionais aos números no Triângulo de Pascal:

1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
... ...

A distribuição de probabilidades ao longo da n-ésima fila é assim proporcional aos coeficientes de $(1+t)^n$. Uma tal distribuição é chamada distribuição binomial.

Uma rampa deste tipo é chamada Galton Quincunx, depois do nome do seu inventor, Galton; Quincunx é o nome latino para a face 5 de um dado, ou qualquer padrão semelhante.

Na base da rampa foram feitas partições para as bolas e foi colocado um vidro para que as bolas não passem de uma para outra. Na parte superior da rampa foi construído um reservatório para colocar as bolas, que se encontra fechado por uma pequena porta que pode ser removida. Quando a porta é removida as bolas descem pela rampa abaixo e são desviadas pelos pregos que se encontram distribuídos de forma conveniente. Se o ângulo for ajustado adequadamente, o número de bolas nos compartimentos pode aproximar-se muito da distribuição binomial.

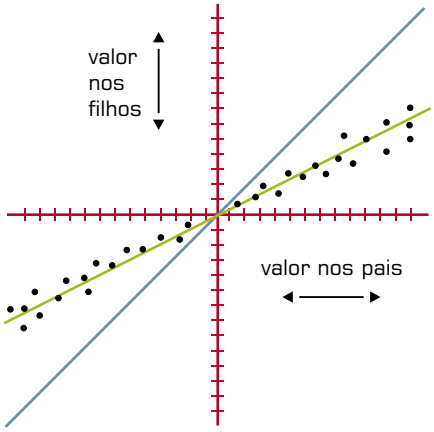
Para um grande número de bolas e de filas de pregos esta distribuição aproxima-se da curva erro padrão $y=Ke^{-\frac{\chi^2}{2s^2}}$, com k e s constantes.

A curva formada pelas colunas de bolas nos compartimentos deveria dar uma ideia grosseira da sua forma.

4.2.1 Lei da regressão para a mediocridade

O investigador britânico, Francis Galton, a partir de um estudo com pares pais-filhos, propôs a "lei da regressão para a mediocridade".

Lei da regressão para a mediocridade de Galton:



No gráfico acima está representada a relação de uma variável métrica entre pais e filhos (por exemplo, altura). A linha azul representa o esperado se os filhos tiverem exactamente o valor da média dos pais. Note-se que pais que apresentam valores maiores da característica têm descendência com um valor médio da característica menor que a média observada medida entre os pais. Por outro lado, os pais que têm o valor menor da característica têm os filhos com valores maiores que o resultante da média entre os pais. Por isso a lei foi chamada de "regressão para a média". Como curiosidade, o método estatístico de ajuste de linhas pelo método dos mínimos quadrados é até hoje chamado de "regressão linear" devido a Pearson, um dos seguidores de Galton. O índice r, que mostra quão bem os pontos experimentais se ajustam a uma recta, é o coeficiente de regressão linear de Pearson.

Os resultados e suas interpretações aparentemente antagónicas originaram uma disputa de natureza científica que durou as primeiras décadas do século XX. Essa disputa teve importância na discussão a respeito do processo de evolução biológica, pois Charles Darwin, um dos criadores da teoria da evolução por selecção natural junto com Alfred Russell Wallace, também inglês, acreditava que a evolução por selecção natural era um processo que ocorria sobre a variação genética de natureza contínua, sendo portanto um processo gradual.

4.3 De Karl Pearson a Ronald A. Fisher

É a meados do século XIX que se dá o aparecimento da Estatística Moderna. Pode-se dizer que esta nova etapa da Estatística nasceu nos laboratórios de pesquisas biométricas.

Comecemos por falar de Karl Pearson; Matemático britânico, foi o fundador da "Biometrika" (revista sobre Biometria muito conhecida a nível internacional) e seguidor de Francis Galton. É conhecido entre nós como o "criador da Estatística Aplicada". Formou-se na Universidade de Cambridge e inicialmente dedicou-se ao estudo da hereditariedade aplicando métodos estatísticos e desenvolvendo a teoria de Galton. O trabalho de Karl Pearson é constituído por uma enorme quantidade de trabalhos publicados principalmente na revista "Biometrika", a qual foi fundada em conjunto com Walter Weldon e Francis Galton.

Desenvolveu a teoria da regressão e da correlação aplicada aos problemas da hereditariedade, criou o teste do "qui quadrado" e foi um dos defensores do reconhecimento da Estatística como uma disciplina autónoma e introduzida no ensino secundário. (Galeria dos Matemáticos 1991).

Karl Pearson nasceu em Londres a 27 de Março de 1857 é considerado o "criador da Estatística Aplicada". Seguidor de Francis Galton no seu trabalho de hereditariedade. Apesar de todo o seu trabalho ser ligado à biologia, o seu grande contributo para a Estatística deve-se a descobertas feitas para explicar os problemas biológicos relacionados com a evolução e com a hereditariedade.

Criou o "método dos momentos" e o sistema de "curvas de frequência", que ainda hoje são usados para a descrição matemática dos fenómenos naturais. A distribuição de Pearson, mais conhecida entre nós como a distribuição do "qui quadrado" ( ), constitui a base da Estatística das pequenas amostras de populações normais, servindo para medir a confiança de resultados estatísticos, testar hipóteses, etc.

Outro matemático importante para a evolução da estatística moderna é o inglês William Sealey Gosset, mais conhecido como Student. Student trabalhou como químico na Cervejaria Guiness, onde começou a fazer várias experiências relacionadas com o controlo de qualidade da cerveja. Student no início das suas experiências aplicou a distribuição Normal, começando a sentir dificuldades na utilização da "Lei do Erro" em amostras pequenas. Para resolver esse problema entrou em contacto com o grande estatístico da altura, Karl Pearson, o qual já tinha desenvolvido as ideias que o levaram à distribuição do mas, tal como todos os estatísticos da altura, estava mais interessado em grandes amostras. Contudo, Student desenvolveu o teste t de Student e os resultados forma publicados na revista "Biometrika".

William Sealey Gosset nasceu a 13 de junho de 1876 em Canterbury Inglaterra. Estudou química e matemática e contribuiu para a Estatística com a descoberta da distribuição *t student*. Devido à fábrica onde trabalhava não deixar publicar o seu nome verdadeiro, pois tinha medo de que as fábricas concorrentes soubessem das descobertas feitas sobre a qualidade do produto, Gosset é conhecido entre nós como Student, pseudónimo modesto utilizado por este grande estatístico.

Utilizou o pseudónimo de Student, devido à Cervejaria Guiness não desejar que os seus concorrentes soubessem dos métodos estatísticos utilizados para melhorar a qualidade da sua cerveja. Apesar da grande importância desta descoberta, o seu trabalho foi ignorado e só redescoberto por Fisher. A distribuição t é uma distribuição de probabilidade teórica e semelhante à curva normal reduzida, diferenciando-se desta com a introdução de um parâmetro chamado grau de liberdade. Estes graus de liberdade podem ser qualquer número real maior que zero.

Falemos agora do grande Estatístico Ronald A. Fisher, um dos fundadores da Estatística Moderna.

Interessou-se pela teoria da evolução e selecção, sobretudo em genética, tal como Francis Galton e foi com este tema que se interessou pela Estatística e que desenvolveu grande parte dos seus trabalhos. Mantendo correspondência com o seu grande amigo Student, Fisher acabou por fazer a distinção entre a média amostral e a média da população. Interessou-se pelas amostras relativamente pequenas e não pelas infinitivamente grandes. Era uma pessoa que não gostava de cometer erros e sofria bastante quando os tinha de admitir. Por isso, pensou em várias teorias que mais tarde ele e outros tentaram desenvolver. Foi rejeitado para o serviço nacional na 1ª Grande Guerra devido à fraca visão que possuía e então começou a leccionar numa escola secundária como forma de serviço comunitário.

... apesar de haver sempre incerteza na estatística isto não implica que haja falta de precisão. - a incerteza pode ser alvo de precisão quantitativa. Fisher fez muito para dar forma e realidade a esta ideia.

G.A. Barnard

Professor Universitário em ESSEX

Nessa altura, o seu trabalho na área de Estatística chamou a atenção de Karl Pearson, famoso estatístico da altura. Pearson, criticou o trabalho de Fisher, talvez por inveja, ferindo o seu orgulho, o que acabou por gerar um grande conflito entre estes dois estatísticos pois ambos começaram a reparar nos erros que cada um cometia.2 Em 1919 teve duas propostas de emprego: ou iria trabalhar para Inglaterra com Pearson ou para a Estação Agrícola Experimental de Rothamsted. Como não tinha grande amizade por Pearson, optou pela segunda proposta, a qual também o entusiasmou bastante, pois na Estação Agrícola existiam observações adquiridas há mais de cem anos. Procedeu à análise desses dados e introduziu um novo conjunto de métodos, como por exemplo o da máxima verosimilhança, (procedendo ao estudo de todas as suas propriedades), a análise de variância, os testes de hipóteses, e o planeamento de experiências.

Ronald Aylmer Fisher, nasceu a 17 de Fevereiro de 1890 em East Finchley Londres e é considerado um dos pais e fundadores da Estatística Moderna. Licenciou-se em astronomia na Universidade de Cambridge, tendo-se interessado desde muito novo pela matemática. O seu contributo para a evolução da Estatística é baseado, na maior parte, em experiências realizadas na Estação Agrícola Experimental de Rothamsted. Aí desenvolveu alguns métodos estatísticos tal como o método da máxima verosimilhança, a análise de variância, os testes de hipótese, e o planeamento de experiências.

Estas ideias deram aos investigadores muitos instrumentos para lidar com variáveis, amostras pequenas e estimativas mais precisas.

Fisher recebeu três medalhas da Royal Statistical Society: a Medalha Real (1938), a Medalha de Darwin (1948) e a Medalha de Copley (1955), tendo sido nomeado Cavaleiro pela Rainha Isabel em 1952.

Nunca deixou de parte os seus estudos realizados em genética, tendo mesmo previsto dois novos anticorpos ao avaliar os tipos de sangue. Toda esta estatística é estudada hoje em quase todos os cursos universitários e faz parte do nosso dia-a-dia.

4.4 Andrei Nicolaevitch Kolmogorov

Nasceu no dia 25 de Abril de 1903 em Tambov, Rússia e desde muito cedo, Kolmogorov interessou-se pela matemática. Com cinco ou seis anos, descobriu que a sucessão de somas de números ímpares é igual à sucessão de quadrados de números inteiros.

$$1=1^2$$
$$1+3=2^2$$
$$1+3+5=3^2$$
$$1+3+5+7=4^2$$
$$\dots$$
$$1+3+\dots+(2n-1)=n^2$$

Na escola, Kolgomorov era uma criança que inventava vários problemas de matemática, sendo muitos deles publicados no jornal da escola.

Tal como foi referido no capítulo 3, Kolmogorov lançou as bases axiomáticas das probabilidades e desenvolveu toda uma teoria que constituiu um enorme avanço na área, estabelecendo um marco histórico. Essencialmente, os axiomas de Kolmogorov estabelecem que:

Os Axiomas das Probabilidades

- Associados aos possíveis resultados de uma experiência aleatória, existe sempre um espaço amostral e uma álgebra de acontecimentos;
- Para todos os acontecimentos da álgebra, existe um número não-negativo (maior ou igual a zero), chamado probabilidade, que se atribui a tal acontecimento;
- A probabilidade do espaço amostral é igual a 1;
- Para quaisquer dois acontecimentos disjuntos (que não compartilham nenhum resultado) a probabilidade da reunião é igual à soma das suas probabilidades;
- O Axioma anterior é verdadeiro para infinitas uniões, desde que todos os pares de acontecimentos sejam disjuntos.

A aplicação da lógica matemática aos princípios acima leva às seguintes propriedades fundamentais da probabilidade:

Propriedades Fundamentais das Probabilidades:

- A probabilidade de qualquer acontecimento é sempre um número maior ou igual a zero e menor ou igual a um;
- A probabilidade de um acontecimento impossível é zero;
- Se a ocorrência de um acontecimento implica a ocorrência de um outro, então a probabilidade

do primeiro é menor do que a probabilidade do segundo;

• A probabilidade da união de dois acontecimentos é igual à probabilidade do primeiro mais a probabilidade do segundo menos a probabilidade da ocorrência simultânea dos dois.

4.5 O Século XX

4.5.1 Berço das Aplicações da Estatística

A Estatística encontra aplicações em quase todos os campos da actividade humana. No sector agrícola Fisher deu um grande contributo devido ao emprego na Estação Agrária Experimental de Rothamsteed. Os métodos de análise estatística permitiram a melhoria da produtividade, o aumento da eficácia, o estudo cuidado e metódico das condições de produção, etc. "As aplicações industriais surgem por volta da década de 30: as cartas de controle, o controle dos lotes (estes tão ligados ao desenvolvimento dos testes de hipóteses) são talvez os primeiros contributos da Estatística ao aperfeiçoamento tecnológico da sociedade industrial; no domínio das aplicações médicas, o estudo da eficácia dos fármacos, da qualidade dos tratamentos, a detecção de causas possíveis de doença, são algumas das aplicações da estatística" (Oliveira,1995). O Estado tem necessidade de conhecer a população; para isso recorre à Estatística, nomeadamente aos recenseamentos, para tomar decisões a nível governamental, por exemplo, para saber quantos indivíduos dos 15 aos 18 anos existem numa certa localidade: a partir daí vai saber se há necessidade de construir uma escola secundária nessa localidade ou não. Os serviços de Meteorologia, tão importantes para a navegação aérea e marítima, são essencialmente estatísticos. A Informática também encontra aplicações estatísticas , por exemplo, na Inteligência Artificial, na avaliação de desempenho de redes de computadores, etc. A Medicina recorre à Estatística para prever determinadas doenças e quais os efeitos que determinado medicamento pode ter em certos doentes. Na Engenharia, a Estatística é aplicada mais a nível do controlo de qualidade, por exemplo, na obtenção da percentagem de peças defeituosas que uma máquina pode produzir.

4.5.2 Análise Exploratória da Dados

As técnicas clássicas de estatística foram concebidas para serem as melhores possíveis, assumindo um conjunto de pressupostos rígidos. Experiência e investigação posterior levaram-nos a reconhecer que as técnicas clássicas se comportam deficientemente quando a situação real se afasta do ideal descrito por esse conjunto de pressupostos. Desenvolvimentos recentes, tais como métodos robustos e de análise exploratória de dados, contribuem para aumentar a eficácia da análise estatística.

O principal objectivo de uma análise exploratória é extrair informações dos dados, estabelecendo relações entre objectos e variáveis. A análise exploratória não estabelece modelos à priori, mas permite que, a partir das relações observadas nos dados, sejam levantadas hipóteses e propostos modelos.

Existem duas fases na prática de análise de dados: exploratória e confirmatória. A análise exploratória de dados realça a procura flexível de pistas e da evidência, enquanto a análise confirmatória de dados realça a avaliação da evidência disponível.3

4.6 - Tendências para o Futuro

Actualmente as informações estatísticas são obtidas, classificadas e armazenadas em meio magnético e disponibilizadas em diversos sistemas de informações abrangentes que fornecem aos pesquisadores/cidadãos e às organizações da sociedade informações estatísticas inteligentes e necessárias ao desenvolvimento de suas actividades. A expansão no processo de obtenção, armazenamento e disseminação de informações estatísticas, extensivamente facilitadas pelo uso dos recursos computacionais, tem sido acompanhada pelo rápido desenvolvimento de

novas técnicas e metodologias estatísticas de análise estatística de dados.

Uma nova área em que a informática deu um forte impulso foi a da "Engenharia de dados".

Com a descoberta do cálculo computacional, desenvolveram-se famílias de algoritmos para tratamento de dados, que se podem agrupar na área do *Data Mining*.

5. A Estatística em Portugal

5.1 Portugal e a Estatística: os números e a história

"A aplicação da Estatística em Portugal começou, tal como nos outros países da Europa, com a necessidade de o Estado conhecer melhor as características da sua população. A partir do século XVI, factores como a afirmação do Estado Absolutista, o desenvolvimento da administração, de um mercado cada vez mais amplo e dinâmico, implicaram o recurso ao quantitativo como elemento que começou a ser decisivo na administração.

Tratava-se de contar a riqueza em tempos mercantilistas, fosse em homens, fosse em géneros, estimando a grandeza das potencialidades militares, avaliando os recursos tributários, esboçando orçamentos estatais" (Sousa, 1995).

Território

Portugal está situado a sudoeste da Península Ibérica. Este país de configuração rectangular, é limitado a oriente e ao norte pela Espanha. A fronteira terrestre de Portugal segue ocasionalmente o curso dos rios, mas na sua maior extensão não existem barreiras naturais. Esta fronteira, que remonta ao ano 1297, é a mais antiga da Europa.

Segundo a obra "História da Estatística em Portugal" (Fernando Sousa, 1995), o registo de acontecimentos, especialmente a contagem de forças militares, a enumeração de bens, rendimentos e despesas, constituem os objectos de notação que mais se destacam na Idade Média portuguesa, marcada pela grande escassez de dados de natureza quantitativa estatística.

O rei tinha necessidade de conhecer o seu exército e a sua população a defender, e por isso logo havia necessidade de quantificar a sociedade. Os primeiros registos encontrados são relativos aos besteiros (soldados cuja arma principal era uma Besta), os quais eram objecto de listagens de controlo e mais tarde estabeleceu-se uma relação quantitativa entre o número de besteiros de cada concelho ("conto") e a respectiva população. Com base no papel da Igreja, também na Idade Média, produziram-se numerosos documentos (censuais e tombos de propriedades) relativamente ao conhecimento da realidade económico-social de áreas por si controladas. A crise instalada nos séculos XIV e XV, exigiu dos senhores eclesiásticos e laicos um melhor aproveitamento dos seus patrimónios fundiários, levando-os à elaboração de inventários sistemáticos de bens e rendimentos, aos tombos, que permitiam não só conhecer e dominar melhor a situação económica de cada senhorio, mas também prever os rendimentos de cada ano.

Eram feitas Inquirições, isto é, inquéritos feitos pelos monarcas portugueses, nos quais eram investigados os estados dos direitos reais e a legitimidade das possessões dos nobres. Destas inquirições também se podia tirar conclusões acerca da organização profissional e económica, bem como detectar alguns níveis de estratificação social. Com base no resultados destas Inquirições, D. Dinis mandou fazer um cadastro geral, ou seja, um registo escrito, para evitar que os ambiciosos se apoderassem de terrenos e direitos que não lhes pertenciam. Naturalmente surgiram protestos, reclamações, algumas tentativas de revolta, mas a vontade e as ordens do rei prevaleceram.

Com a aproximação do Estado Liberal e a afirmação do conceito de Nação como base da administração, a cobertura estatística generalizada para o país começa a ser reclamada, pois o governo não se pode exercer eficazmente sobre o incerto, o desconhecido. Surgem planos para o cadastro do Reino, levantam-se numeramentos de carácter sistemático, inicia-se a primeira grande série estatística sobre o comércio externo – Balança Geral do Comércio do Reino de Portugal, 1776-1831, que podemos adoptar como o símbolo do início de um novo período.

Numeramentos:

Contagem do número de fogos (casas) feita com o objectivo de recolher dados para lançar impostos ou recrutar militares.

Multiplicam-se os quadros estatísticos em diversas áreas da realidade social, apontam-se números globais, mas a informação é ainda, em grande parte, dispersa, recolhida em segunda mão, produzida por terceiros e nem sempre de acordo com os requisitos de qualidade e exigência que a estatística requer – por exemplo, os dados da população são solicitados aos párocos – no comércio externo (1842), nas contribuições municipais (1845), no movimento da alfândegas de Lisboa e Porto (1856-1857), na área demográfica, com a realização do primeiro censo digno desse nome (1864), noutras áreas, com a publicação do Annuario Estatistico (1875), a que se seguirão séries autónomas para outros sectores, (contribuições, movimento bancário, transportes, etc.).

A entrada na era estatística faz-se, portanto, gradualmente, ao longo do século XIX, com a criação de organismos que se fazem representar nos respectivos Congressos Internacionais.

Mas só no século XX surge uma eficaz utilização dos dados recolhidos, com o desenvolvimento da estatística como ramo aplicado da matemática, ligando ao cálculo das probabilidades, que vai permitir o fornecimento regular de indicadores de síntese, a perspectiva sequencial das tendências de desenvolvimento, a possibilidade de prospectiva. Situação apenas possível com a criação do Instituto Nacional de Estatística (INE), em 1935." (Fernando Sousa,1995)

Trabalhos estatísticos importantes e conhecidos depois da fundação da nacionalidade portuguesa e antes da criação do INE

- Rol de Besteiros do Conto, de D. Afonso III (1260-1279);
- Rol de Besteiros do Conto, de D. João I (1421-1422);
- Numeramento ou Cadastro Geral do Reino, de D. João III (1527);
- Resenha de Gente de Guerra, de D. Filipe III (1639);
- Lista dos Fogos e Almas que há nas Terras de Portugal, de D. João V (1732), também conhecida por Censo do Marquês de Abrantes;
- Numeramento de Pina Manique, de D. Maria I (1798);
- Recenseamento Geral do Reino, de D. João VI, também conhecido por Censo do Conde de Linhares (1801);
- Recenseamentos Gerais de 1835 e 1851.

5.2 Os Recenseamentos em Portugal

Os primeiros censos portugueses foram realizados de 31 de Dezembro de 1863 para 1 de Janeiro de 1864, tendo por base as orientações do Congresso Internacional de Estatística realizado em Bruxelas, em 1853. Antes desta data, tal como foi referido anteriormente, já se realizavam em Portugal recenseamentos, mas por não serem exaustivos e/ou não se apoiarem em princípios estatísticos credíveis, não podem ser considerados equivalentes aos iniciados em 1864.

A palavra Censo deriva da palavra Censere que em latim significa Taxar.

Nestes censos foi optado o método de recolha directa sendo todas as pessoas recenseadas no mesmo dia e nos lugares onde passaram a noite. Os recenseamentos a partir daqui deveriam ser realizados de 10 em 10 anos, mas o recenseamento seguinte foi em 1878 ao qual se seguiria o Censo de 1890. A partir de então, os recenseamentos populacionais têm vindo a realizar-se, com algumas excepções, regularmente com intervalos de 10 anos.

Desde 1940 (inclusive), os recenseamentos passaram a ser realizados pelo Instituto Nacional de Estatística e a partir de 1970 realizou-se em simultâneo o I Recenseamento Geral da Habitação.

Até aos dias de hoje, já foram realizados catorze recenseamentos da população e quatro da habitação.

Apresentam-se de seguida todos os recenseamentos efectuados em Portugal, e os seus antecedentes históricos resumidos:

1864 - **1 de Janeiro (I Recenseamento Geral da População):**

Realizou-se o I Recenseamento Geral da População, tendo por base as orientações do Congresso Internacional de Estatística, que teve lugar em Bruxelas, em 1853.

1878 - **1 de Janeiro (II Recenseamento Geral da População):**

Efectuou-se o II Recenseamento Geral da População; embora mais completo que o anterior, quanto às variáveis observadas e aos apuramentos efectuados, ainda tem um conteúdo bastante reduzido.

1890 - **1 de Dezembro (III Recenseamento Geral da População):**

Realizou-se já com novas orientações metodológicas, de acordo com o Congresso Internacional de Estatística de S. Petersburgo, realizado em 1872; a caracterização da população e das famílias foi bastante mais completa.

1900 - **1 de Dezembro (IV Recenseamento Geral da População):**

A metodologia da recolha de dados, do seu tratamento e apresentação foi semelhante à do censo anterior, tendo-se, no entanto, registado algumas inovações.

1911 - **1 de Dezembro (V Recenseamento Geral da População)**

Manteve-se a metodologia e as variáveis observadas.

1920 - **1 de Dezembro (VI Recenseamento Geral da População):**

Manteve-se a metodologia e as variáveis observadas.

1930 - **1 de Dezembro (VII Recenseamento Geral da População):**

Não houve grandes alterações nas características observadas, continuando mal coberta a parte referente às características económicas.

1940 - **12 de Dezembro (VIII Recenseamento Geral da População):**

Este foi o primeiro censo efectuado pelo Instituto Nacional de Estatística e é aceite como um marco na história dos recenseamentos portugueses. Adoptou-se uma nova metodologia de execução. As características económicas são definidas com maior rigor e consideradas como um elemento importante de observação.

1950 - **15 de Dezembro (IX Recenseamento Geral da População):**

Seguiu a metodologia do censo anterior mas com algumas inovações como, por exemplo, a melhoria da técnica das perguntas fechadas.

1960 - **15 de Dezembro (X Recenseamento Geral da População):**

Publicaram-se pela primeira vez dados retrospectivos. Os recenseamentos de 1950 e 1960 seguem, de perto, o conteúdo do de 1940.

1970 - **15 de Dezembro (XI Recenseamento Geral da População) (I Recenseamento Geral da Habitação):**

Realizou-se o I Recenseamento Geral da Habitação, juntamente com o da População; contudo, o programa audacioso que procurava dar resposta às inúmeras solicitações governamentais não teve sucesso no plano executivo, em especial na totalidade dos resultados a divulgar.

1981 - **16 de Março (XII Recenseamento Geral da População) (II Recenseamento Geral da Habitação):**

Realizaram-se os recenseamentos da População e Habitação que seguiram, de perto, as recomendações internacionais (CEE/ ONU) e fazem, em quase todas as áreas, uma aplicação rigorosa dos conceitos e uma grande desagregação geográfica dos respectivos dados.

1991 - **15 de Abril (XIII Recenseamento Geral da População) (III Recenseamento Geral da Habitação):**

> Seguiu-se a metodologia do censo anterior, desenvolvendo-se no entanto algumas das vertentes de preparação da operação e do tratamento dos dados já iniciados em 1981. Construiu-se uma Base Geográfica de Referenciação Espacial, constituída por um conjunto de suportes cartográficos contendo a informação que permite a divisão das freguesias em secções e subsecções estatísticas.

2001 - **12 de Março (XIV Recenseamento Geral da População) (IV Recenseamento Geral da Habitação):**

> A grande diferença prende-se essencialmente com a inovação das tecnologias utilizadas (digitalização cartográfica, utilização de sistemas de informação geográfica, leitura óptica dos questionários, codificação assistida por computador e o reforço da correcção automática das respostas incoerentes). Também é introduzida uma nova questão no questionário individual que diz respeito à deficiência.

5.3 O Ensino da Estatística em Portugal

Não só em Portugal, mas em muitos outros países a Estatística é um ramo da Matemática Aplicada. O seu estudo e desenvolvimento como ciência tem vindo a crescer com o progresso social e hoje a Estatística está presente em quase todas as áreas do saber.

Como refere João Branco (JME-190), no final do século XIX assistiu-se a uma generalizada emergência e reconhecimento de problemas de natureza estatística nos vários ramos científicos, na indústria e em actividades governamentais o que fez crescer o interesse pela actividade estatística. A rapidez com que estes desenvolvimentos ocorreram gerou uma crise de falta de pessoal técnico com conhecimentos de estatística que foi intensamente procurado pelas instituições que desejavam usufruir da nova metodologia para fazer avançar as suas actividades. É neste contexto que surgiu a necessidade de ensinar estatística a um número de pessoas cada vez maior. Inicialmente a prioridade foi dada ao ensino avançado com vista a aperfeiçoar os conhecimentos daqueles com interesse na profissão de estatístico ou dos que se encontravam a apoiar actividades de investigação nos vários ramos da ciência. Só depois se passou a pensar no ensino da Estatística elementar destinado a fornecer conhecimentos básicos a estudantes das ciências naturais e sociais e ainda a estudantes interessados em seguir uma actividade de estatístico profissional. Apesar de elementares estes conhecimentos começaram a ser introduzidos nos cursos de pós-graduação ou nos últimos anos da graduação. Porém depressa se concluiu que estes cursos de estatística elementar deviam ser introduzidos mais cedo, numa fase mais inicial do ensino universitário.

5.3.1. Estatística no Secundário

Segundo João Branco (JME-190), o ensino da Estatística no Secundário, surgiu como uma necessidade de proporcionar à população em geral um sistema coerente de ideias estatísticas e de capacidades para usar essas ideias, com naturalidade, numa sociedade cada vez mais baseada em dados e informação numérica. Uma reunião de grande importância para o desenvolvimento do ensino desta disciplina, teve lugar em Royaumont, em 1959 sob os auspícios dos directores da Organização Europeia da Cooperação Económica (OECE), organização a que sucedeu a Organização para a Cooperação e Desenvolvimento Económico (OCDE), em 1961.

A este acontecimento compareceram matemáticos de todo o mundo com o fim de estudar uma reforma profunda do ensino da Matemática ao nível do ensino pré-universitário, tendo-se concluído que se deveria introduzir no plano de estudos secundários o ensino do Cálculo das Probabilidades e da Estatística.

O movimento que começa a registar-se em alguns países com o objectivo de modificar os programas e métodos de ensino da Matemática nas escolas secundárias chega também a Portugal, sobretudo através de publicações e reuniões promovidas pela OCDE. E é José Sebastião e Silva, um dos mais importantes matemáticos portugueses de todos os tempos, que fica com a responsabilidade do projecto de modernização do ensino da Matemática no 3º ciclo.

A modificação dos programas com vista a adaptá-los às exigências da revolução científica e tecnológica que caracteriza a época levam à introdução, pela primeira vez, nos liceus portugueses, de vários temas entre os quais elementos de Cálculo das Probabilidades e de Estatística. Em 1963/64 são criadas as três primeiras turmas para funcionarem a título experimental. Foi esta experiência, repetida sucessivamente ao longo de vários anos e estendida a várias dezenas de turmas espalhadas pelos liceus do País, que preparou o terreno para a introdução definitiva destas matérias no currículo do ensino secundário.

É curioso saber que o movimento que leva à introdução da Estatística no secundário ocorre sensivelmente no mesmo período em que se dão passos definitivos para implantar o ensino da Estatística nas licenciaturas de matemática nas universidades. O movimento para o ensino da Estatística na universidade foi particularmente activo na Faculdade de Ciências de Lisboa tendo levado à criação da Primeira licenciatura em Probabilidades e Estatística em 1982. Neste movimento destaca-se José Tiago de Oliveira, grande cientista que se apaixona pela Estatística e seus problemas a todos os níveis incluindo também o ensino no secundário. (JME-190)


PRIMEIROS ELEMENTOS

ESTATISTICA,

Segundo Adrião Ferreira da Cunha (2001), verificou-se em 1841 o início do ensino da Estatística em Portugal na Faculdade de Direito da Universidade de Coimbra. Foi introdutor deste ensino o Professor Adrião Sampaio com a sua obra Primeiros Elementos da Ciência Estatística que utilizou para auxílio das suas aulas suplementares ao Curso de Economia Política de que era encarregado na referida Faculdade.

5.4 O INE e o Sistema Estatístico Nacional

Em Portugal o INE é o organismo operacional responsável pela recolha, apuramento e difusão das estatísticas oficiais nacionais. No entanto, existem organismos que gerem a atribuição de competências em todo o processo estatístico.

Composição do Sistema Estatístico Nacional

Nos termos da Lei nº 22/2008, de 13 de Maio, o SEN é constituído pelo Conselho Superior de Estatística (CSE), pelo Instituto Nacional de Estatística (INE), pelo Banco de Portugal e pelos Serviços Regionais de Estatística das Regiões Autónomas dos Açores e da Madeira.

- O CSE é o órgão do estado que superiormente orienta e coordena o Sistema Estatístico Nacional.
- O INE é o orgão central de produção e difusão de estatísticas oficiais que assegura a supervisão tecnico-cientifica do SEN.
- O Banco de Portugal no âmbito das suas atribuições de recolha e elaboração de estatísticas monetárias, financeiras, cambiais e da balança de pagamentos.
- Os Serviços Regionais de Estatística dos Açores e da Madeira, que funcionam em relação às estatísticas oficiais de âmbito nacional, como delegações do INE, IP.
- As entidades produtoras de estatísticas oficiais por delegação do INE, IP.

O Instituto Nacional de Estatística (INE) foi criado em 1935 numa tentativa de dar resposta à procura cada vez maior da informação estatística. Tem como objectivo o exercício de funções tais como efectuar inquéritos, recenseamentos e outras operações estatísticas; criar gerir e centralizar ficheiros de unidades estatísticas; aceder aos dados individuais (excepto dados de pessoas singulares) disponíveis nas entidades encarregadas da gestão de serviços públicos; realizar estudos de estatística pura e aplicada e proceder à análise económico-social de dados estatísticos disponíveis; promover a formação de quadros do SEN e cooperar com organizações estatísticas estrangeiras.

Em 1989 o INE passou a ser um instituto público, ao qual foi concedida personalidade jurídica, autonomia administrativa, financeira e património próprio.

Nos dias de hoje, o INE tem dezenas de publicações oficiais, não só em estudos demográficos mas em diversos campos de aplicação, tal como indústria, comércio, educação, etc.

Ver também...

Publicações

• BÉDARIDA et al (1987), Pour Une Histoire De La Statistique, Economica.

• CUNHA, Adrião Simões Ferreira (2001), Nótulas Históricas em Torno do Sistema Estatístico Nacional, Lisboa, Instituto Nacional de Estatística.

• DAVID, F.N.(1998), Games, Gods and Gambling, A History of Probability and Statistical Ideas, Dover Publications, Inc. Mineola, New York.

•Galeria dos Matemáticos do Jornal de Matemática Elementar (2º Volume), (1994), Lisboa.

•Galeria de Matemáticos do Jornal de Matemática Elementar, (1991), Lisboa.

•HEYDE, C.C., SENETA, E. (2001), Statisticians of the Centuries, Springer, New York.

•HOAGLIN, David C., MOSTELLER, Frederick, TUKEY, John W. (1983), Novas Tecnologias/ Estatística: Análise Exploratória de Dados. Técnicas Robustas, Edições Salamandra.

•INE, Programa Global dos Censos 2001, Instituto Nacional de Estatística, Lisboa, disponível em: http://www.ine.pt/censos2001/Organizacao/programaglobal.asp

•Jornal de Matemática Elementar nº190, Lisboa.

•LOUREIRO, Cristina; OLIVEIRA, Fernanda; BRUNHEIRA, Lina, Ensino e Aprendizagem da Estatística, Sociedade Portuguesa de Estatística, Associação de Professores de Matemática, Departamento de Educação e de Estatística e Investigação Operacional da Faculdade de Ciências da Universidade de Lisboa, 2000.

•OLIVEIRA, J. Tiago (1981), O Ensino Iniciário da Estatística. Actas do II Colóquio de Estatística e Investigação Operacional, Fundão.

•OLIVEIRA, J. Tiago (1995), Collected works (Volume II), Évora.

•PERSON, E. S., KENDALL, Sir Maurice (1820), Studies in the History of Statistics and Probability, volume I, Charles Griffin & Co Ltd, London.

• SOUSA, Fernando (1995), História da Estatística em Portugal, Instituto Nacional de Estatística, Lisboa.

• STIGLER, Stephen M. (1986), The History of Statistics, The Measurement of Uncertainty before 1900, Belknap Harvard.

Web sites:

http://www.sobiografias.hpg.com.br (algumas biografias de personalidades históricas);

http://users.hotlink.com.br/marielli/ (neste site de matemática, encontra várias biografias de matemáticos famosos, bem como histórias sobre os números, aritmética, etc);

http://www.educ.fc.ul.pt/semtem/semtem99/sem21/framegeral.htm (história do triângulo de Pascal);

http://www.mala.bc.ca/~johnstoi/darwin/sect4.htm

http://www.mat.uc.pt/~bebiano/Atractor/esta.htm (alguns modelos matemáticos, entre os quais o Quincunx);

http://www.geocities.com/g10ap/matematicos/os_grandes_genios.htm (biografias de grandes génios matemáticos);

http://www.ib.usp.br/evolucao/QTL/historiaqtl.html (inclui a explicação da lei da regressão para a mediocridade de Galton)

O Inquérito Estatístico

Maria João Ferreira
Pedro Campos

O Inquérito Estatístico

Uma introdução à elaboração de questionários, amostragem, organização e apresentação dos resultados

Maria João Ferreira
Pedro Campos

Sumário:

1. Introdução

Neste Dossiê, que teve a colaboração e supervisão da Profª Doutora Maria Eugénia Graça Martins, Professora da Faculdade de Ciências da Universidade de Lisboa e consultora científica do ALEA, poderá encontrar uma pequena introdução às fases de um inquérito por questionário, as regras de construção de um questionário, noções sobre como seleccionar os elementos da amostra e ainda a preparação do relatório para apresentação final dos resultados. No final, a rubrica Ver Também contém ligações para outros estudos de interesse relacionados com as temáticas em causa (publicações e páginas na internet).

2. Porque Fazemos Inquéritos Estatísticos?

O Inquérito é um dos instrumentos mais utilizados no domínio da investigação aplicada, nomeadamente na área social. Desde os estudos de mercado às pesquisas puramente teóricas, passando pelas sondagens de opinião, poucos são os estudos que não se apoiam, parcial ou totalmente, em informações recolhidas com base em inquéritos.

Sondagem:

Estudo científico de uma parte de uma população com o objectivo de estudar atitudes, hábitos e preferências da população relativamente a acontecimentos, circunstâncias e assuntos de interesse comum.

2.1. O que é um Inquérito Estatístico?

É a necessidade de conhecer uma população no que se refere a uma ou várias características, que nos leva a recorrer à realização de inquéritos.

A alternativa da observação directa, mesmo que viável, em certos casos, levaria demasiado tempo, ou seria impossível quando os fenómenos em estudo se reportam ao passado (Ghiglione e Matalon, 1992).

Um inquérito pode ser considerado como uma interrogação particular acerca de uma situação englobando indivíduos, com o objectivo de generalizar.

O recurso ao inquérito é necessário de cada vez que temos necessidade de informação sobre uma grande variedade de comportamentos de um mesmo indivíduo, ou quanto pretendemos conhecer o mesmo tipo de variável para muitos indivíduos.

População:

Colecção de unidades individuais, que podem ser pessoas, empresas ou resultados experimentais, com uma ou mais caracteristicas comuns, que se pretendem estudar.

Inquérito:

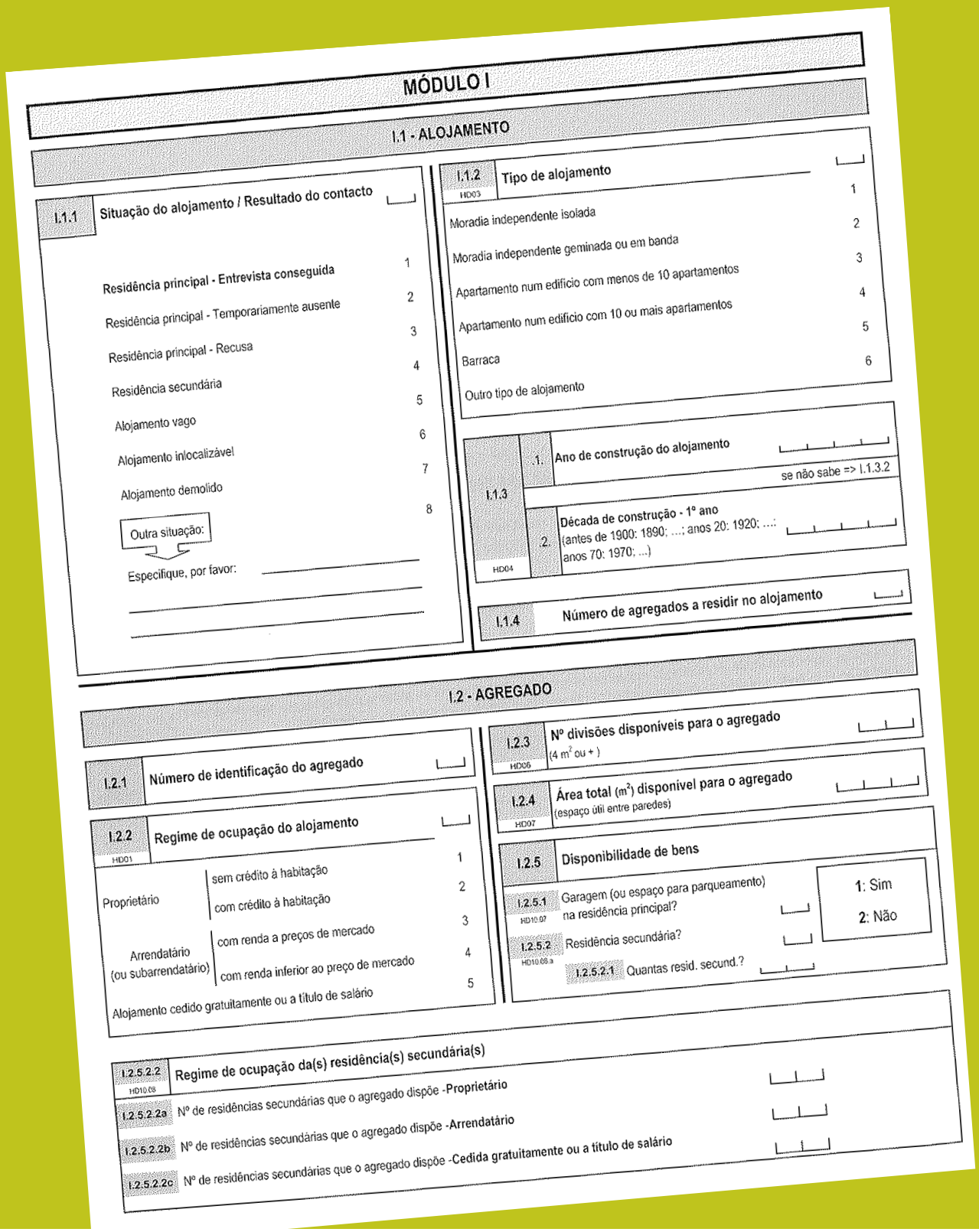
Um inquérito pode ser considerado como uma interrogação particular acerca de uma situação englobando indivíduos, com o objectivo de generalizar.

Exemplo de um dos Inquéritos realizado pelo INE:

O Inquérito aos Orçamentos Familiares, actualmente denominado IDF, realizado pelo INE, tem como objectivo conhecer a origem e o valor dos rendimentos dos agregados e a forma como se transformam em despesas de consumo. É através deste inquérito que se pode actualizar o Índice de Preços no Consumidor, desenvolver e construir um sistema de Indicadores de Pobreza, a análise da concentração da despesa e do rendimento dos agregados familiares, bem como a realização de outros estudos sócio-económicos.

A figura 1 contém uma das partes do questionário que tinha de ser preenchida todos os dias por uma pessoa do agregado familiar, de preferência a pessoa que efectuava as compras. Neste caso, o método de recolha de informação (ou dados) utilizado neste inquérito, conciliou a recolha através do auto-preenchimento (preenchimento feito pelo próprio inquirido) com a recolha por entrevista. Mais à frente abordamos todas estas técnicas de recolha de informação.

Fig. 1 - Questionário utilizado no Inquérito às Despesas das Famílias (Fonte: INE)



MÓDULO I

I.1 - ALOJAMENTO

I.1.1 Situação do alojamento / Resultado do contacto

1 Residência principal - Entrevista conseguida
2 Residência principal - Temporariamente ausente
3 Residência principal - Recusa
4 Residência secundária
5 Alojamento vago
6 Alojamento inlocalizável
7 Alojamento demolido
8 Outra situação:
Especifique, por favor:

I.1.2 Tipo de alojamento (HD03)

1 Moradia independente isolada
2 Moradia independente geminada ou em banda
3 Apartamento num edifício com menos de 10 apartamentos
4 Apartamento num edifício com 10 ou mais apartamentos
5 Barraca
6 Outro tipo de alojamento

I.1.3 (HD04)
.1 Ano de construção do alojamento
se não sabe => I.1.3.2
.2 Década de construção - 1º ano (antes de 1900: 1890; ...; anos 20: 1920; ...; anos 70: 1970; ...)

I.1.4 Número de agregados a residir no alojamento

I.2 - AGREGADO

I.2.1 Número de identificação do agregado

I.2.2 Regime de ocupação do alojamento (HD01)

1 Proprietário: sem crédito à habitação
2 Proprietário: com crédito à habitação
3 Arrendatário (ou subarrendatário): com renda a preços de mercado
4 Arrendatário (ou subarrendatário): com renda inferior ao preço de mercado
5 Alojamento cedido gratuitamente ou a título de salário

I.2.3 Nº divisões disponíveis para o agregado (4 m² ou +) (HD06)

I.2.4 Área total (m²) disponível para o agregado (espaço útil entre paredes) (HD07)

I.2.5 Disponibilidade de bens

I.2.5.1 Garagem (ou espaço para parqueamento) na residência principal? (HD10.07)
I.2.5.2 Residência secundária? (HD10.08.a)
I.2.5.2.1 Quantas resid. secund.?

1: Sim
2: Não

I.2.5.2.2 Regime de ocupação da(s) residência(s) secundária(s) (HD10.08)
I.2.5.2.2a Nº de residências secundárias que o agregado dispõe - Proprietário
I.2.5.2.2b Nº de residências secundárias que o agregado dispõe - Arrendatário
I.2.5.2.2c Nº de residências secundárias que o agregado dispõe - Cedida gratuitamente ou a título de salário

Na figura 2, podemos observar um esquema das técnicas de recolha de informação.

Uma das técnicas mais utilizada na realização de Inquéritos é o Questionário. Trata-se de uma técnica não documental, de observação indirecta, que pode ser feita através de uma entrevista. O inquérito muitas vezes é visto como um processo completo (desde a recolha, à análise, utilizando várias técnicas). O questionário é o instrumento de notação.

3. O Questionário e as Fases de um Inquérito

3.1 Inquérito e Questionário

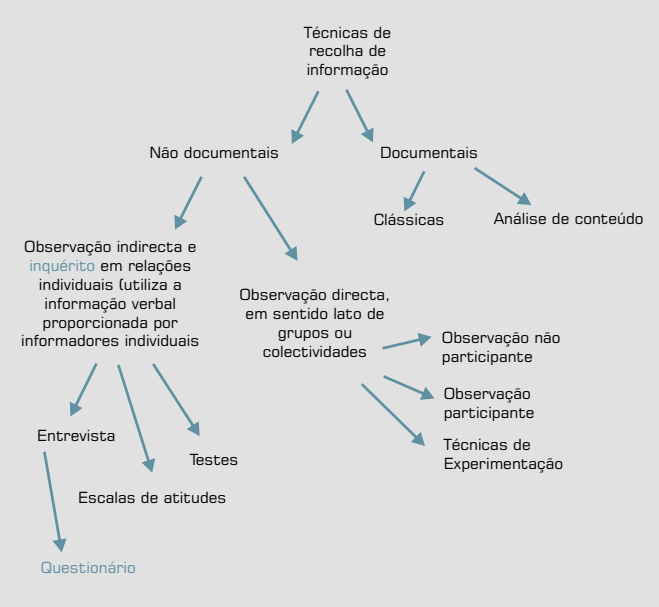
Neste ponto, faremos uma aproximação às noções de Inquérito e Questionário, enquadrando os vários métodos de recolha de informação.

Pode-se considerar que existem dois tipos de técnicas de recolha de informação: as documentais e não documentais. Nas técnicas documentais o objectivo é a recolha de informação a partir de suportes bibliográficos já existentes. É o caso da pesquisa bibliográfica e da análise de textos. Nas técnicas não documentais o investigador realiza observação directa (como por exemplo, a medição da altura do salto de um atleta ou o número de flexões por minuto) ou indirecta - podendo ser feita, neste caso, através da administração de um questionário.

Questionário:

É um dos suportes de registo de informação nos Inquéritos, feito ou não através de uma entrevista

Fig. 2
(adapt. Lima, 1981)



Tal como foi referido anteriormente, recorremos ao inquérito para compreender fenómenos como as atitudes, as opiniões ou as preferências, que só são acessíveis de uma forma prática pela linguagem, e que só raramente se exprimem de forma espontânea. É através do inquérito, e por vezes através da observação, que podemos obter informações do que se passa num determinado momento. Colocando um maior número de questões podem-se fazer análises mais profundas, descrever de forma mais perspicaz as opiniões e os comportamentos que procuramos estudar, verificar hipóteses mais complexas, etc.

Independentemente de todas as vantagens que um questionário tem, existem sempre algumas desvantagens, das quais se destacam o facto de o questionário depender totalmente da linguagem - tudo o que dispomos é do que o inquirido pôde ou quis dizer.

Portanto, as perguntas de um questionário devem recorrer a palavras simples e a uma linguagem acessível, clara e precisa (eliminando a possibilidade de interpretações subjectivas por parte do inquirido). As questões devem ser curtas e directas (evitando as negações e sobretudo as duplas-negações).

No próximo capítulo exploraremos melhor as regras para a construção de questionários.

3.2. Etapas do desenvolvimento de um inquérito

As etapas de desenvolvimento de um inquérito não se descrevem segundo uma ordem linear constante. Segundo Giglione e Matalon (1992), antes de realizarmos um inquérito devemos saber quem queremos inquirir e o que devemos perguntar. Podemos dizer que ao elaborarmos um inquérito, devemos ter em consideração algumas preocupações: ao planear o inquérito já deve estar definida a população que se pretende inquirir e o que se quer saber acerca dela, quais os objectivos do inquérito e como vai ser aplicado; depois, deve-se preparar o instrumento de notação (questionário), para o qual é necessário ter-se em atenção o tipo de perguntas, a ordem pela qual ocorrem, a linguagem aplicada e a apresentação final; por último, surge o trabalho no terreno (recolha de dados), onde se recolhe toda a informação necessária para concretizar o objectivo do inquérito. A recolha dos dados pode ser feita de várias formas, que veremos mais adiante.

4. Como perguntar? - Regras gerais para a construção de um questionário

O questionário é um dos instrumentos de notação mais utilizado para obter informação acerca de uma dada população. A construção do questionário e a formulação das questões constituem uma fase fundamental do desenvolvimento de um inquérito. Para construir um questionário é necessário saber com exactidão o que procuramos, garantir que as questões tenham a mesma interpretação em todos os inquiridos e que todos os aspectos das questões tenham sido bem abordados, etc. Estas condições resultam da realização das entrevistas e do teste às primeiras versões do questionário (pré-teste).

Pré-teste:

Consiste em testar o questionário junto de uma parte da amostra, antes deste ser utilizado em definitivo.

4.1 Os diferentes tipos de questões

As primeiras questões de um questionário são muito importantes. São elas que indicam às pessoas inquiridas o estilo geral do questionário, o género de resposta que delas se espera e o tema que vai ser abordado. É também a partir delas que se estabelece a relação entrevistador-entrevistado, pois determinam a forma de reacção do entrevistado, nomeadamente se este sente que a sua vida privada está a ser incomodada. Normalmente é preferível começar por questões que despertem interesse no entrevistado e não o assustem.

As questões de um questionário podem ser fechadas, abertas e semi-abertas.

Questões fechadas:

São questões onde existe uma lista pré-estabelecida de respostas, a qual é apresentada ao inquirido, para ele indicar a que melhor corresponde à resposta que deseja dar.

4.1.1 Questões fechadas

Diz-se que uma questão é fechada se as modalidades de resposta são impostas (Grangé e Lebart, 1994).

Por exemplo:

Qual é a sua situação de estado civil ?

[1] Solteiro
[2] Casado ou a viver maritalmente
[3] Divorciado ou separado
[4] Viúvo

Este tipo de questões autoriza uma pré-codificação, ou seja, uma tradução imediata da resposta sob a forma de um código alfanumérico. Estas questões limitam as pessoas inquiridas a responder somente àquilo que lhes é apresentado como modalidades de resposta.

Podemos distinguir vários tipos de questões fechadas:

- Questões de resposta única (o inquirido escolhe apenas uma modalidade de resposta).
- Questões de resposta múltipla (o inquirido escolhe de várias modalidades de respostas em número limitado ou não), por exemplo:

Quais são, na sua opinião, os pontos fortes do produto X? (indique no máximo 3 escolhas)

[1] apresentação geral
[2] forma
[3] comodidade de emprego
[4] variedade de utilizações
[5] eficácia
[6] robustez
[7] preço
[8] duração da garantia
[9] serviço pós-venda

- Classificação (o inquirido ordena as várias modalidades de respostas por ordem de importância), por exemplo:

Para o produto Y, classifique as seguintes características, partindo daquilo que considera como os seus pontos mais fortes até aos pontos mais fracos, utilizando a numeração de 1 a 9, sendo o 1 o ponto mais forte e o 9 o mais fraco.

[ ] apresentação geral
[ ] forma
[ ] comodidade de emprego
[ ] variedade de utilizações
[ ] eficácia
[ ] robustez
[ ] preço
[ ] duração da garantia
[ ] serviço pós-venda

As questões em escala também são um tipo de questões fechadas. Este tipo de questões permite atenuar as respostas quando estamos na presença de questões do tipo concordo/não concordo. Para uma situação deste tipo, poderíamos estabelecer uma escala completa de respostas do tipo:

Concordo plenamente / concordo um pouco / indiferente / não concordo muito / em desacordo total

Um questionário composto, na sua maioria, por questões fechadas, não deve ultrapassar os 45 minutos quando a sua aplicação é feita em boas condições, ou seja, em casa do inquirido ou num lugar tranquilo (Ghiglione e Matalon, 1992). Ultrapassando esse limite, o interesse perde-se, o que se nota através de sinais como a rapidez das respostas indicando pouca reflexão sobre as mesmas.

Do ponto de vista da análise de resultados, as questões fechadas são, em principio, as mais cómodas. Quando se trata de um inquérito de aplicação e exploração rápida, como uma sondagem de opinião, esforçamo-nos por aplicar apenas este tipo de questões.

4.1.2 Questões abertas

Para estas questões não existe qualquer tipo de restrição à resposta, devendo esta ser transcrita literalmente, através do modo mais fiável.

O espaço reservado para esta restrição deverá ser medido previamente para facilitar a exploração das respostas (Grangé, 1994).

Questões abertas:

São questões às quais o inquirido responde como quer, utilizando o seu próprio vocabulário.

Exemplo de uma questão aberta:

Qual o tipo de detergente que usa para a máquina da louça?

_____

Há várias razões para se formularem questões abertas. Muitas vezes não se tem tempo para elaborar uma lista de respostas-tipo a apresentar às pessoas e, por essa razão, deixa-se um espaço aberto para registar a resposta do inquirido. Por outro lado, podemos ter que recorrer a questões abertas quando os pré-testes (ver 4.5) do questionário forem insuficientes, ou ainda quando as respostas a esses pré-testes pareçam demasiado complexas para poderem ser resumidas numa lista de tamanho aceitável (Ghiglione e Matalon, 1992). Por último, há uma razão forte para nos levar a preferir deixar uma questão aberta: é que um questionário totalmente fechado torna-se rapidamente fastidioso. Apoiando-se nas listas de respostas que lhes apresentamos, as pessoas podem reflectir cada vez menos e tomar cada vez menos cuidado com o que dizem. Outro motivo para se escolher a forma aberta é que esta permite várias codificações. Depois de analisarmos todas as respostas, estas vão ser codificadas mediante a construção de um livro de códigos (também designado por tabela de classificação).

4.1.3 Questões semi-abertas

Num questionário podem ocorrer simultaneamente modalidades de resposta fechada e aberta na mesma questão:

Qual é o nome da companhia de seguros do seu veículo?

[1] companhia A
[2] companhia B
[...] ...
[10] outra: _____

Esta forma mista tende a resolver os problemas de pertinência e de exaustividade das questões fechadas, reduzindo fortemente os custos de codificação pós-inquérito de uma resposta "literal".

4.2 Ordem das questões

Na elaboração de um questionário deve ter-se em consideração um princípio, meio e fim. Não existe uma regra para a ordem das perguntas, mas sim alguns conselhos que podem ser seguidos. No princípio deve existir uma pequena introdução sobre a entidade que promove o estudo, qual o objectivo do questionário e as vantagens que esse estudo pode trazer para a sociedade.

As primeiras questões devem ser simples pois elas vão determinar a condução do questionário.

As primeiras questões devem ser simples pois vão determinar a condução do questionário. Se as primeiras questões forem complicadas, o inquirido pode perder o interesse de responder, o que dificulta o trabalho do entrevistador. Com o decorrer do questionário as perguntas devem ser mais específicas, por exemplo, abordar temas embaraçantes ou íntimos, por exemplo "Lava os dentes todos os dias?", temas que podem levar a um esforço mental, como por exemplo, pedir para ordenar por ordem de preferência os produtos que gosta mais, etc. Os dados pessoais podem tanto vir no princípio como no fim, dependendo do critério do investigador. Todas as questões devem ser claras, nunca devem sugerir nenhuma resposta particular e não devem exprimir nenhuma expectativa (Ghiglione e Matalon, 1992).

Um questionário deve parecer uma troca de palavras tão natural quanto possível. Se possível deve elaborar-se como um guião.

É certo que não é possível perguntarmos tudo num questionário, pois os vários temas de estudo podem originar muitas questões; logo deve-se ter sensibilidade suficiente para escolher as questões mais importantes para o estudo.

Questões de filtro:

Servem para filtrar as pessoas para as quais certas questões não fazem qualquer sentido ou não são aplicáveis.

As questões devem ser curtas e sequenciais, sem repetições nem descontextualizadas. Por exemplo, antes de perguntarmos a uma pessoa se gostou do filme X deve-se perguntar se alguma vez viu o filme X, pois assim poderemos ter uma questão de filtro que vai avaliar a informação que o entrevistado tem acerca do filme. Caso a sua informação seja nula isto é, que nunca tenha visto o filme X, as questões seguintes que poderiam ser acerca do filme já não fazem sentido para este entrevistado, logo esta questão tem de ser um filtro, passando assim a outra questão acerca de outro tema.

Exemplo de questão de filtro:

1. Alguma vez viu o filme X?

☐ Sim ☐ Não (passe à questão 2)

1.1 Gostou do filme? → **Questão de filtro**

☐ Sim ☐ Não

1.1 Era capaz de voltar a ver o filme X?

☐ Sim ☐ Não

2. Alguma vez viu a série Y?

4.3. Outras sugestões na elaboração de questões

Um questionário não deve conter só perguntas abertas ou ou só perguntas fechadas. Deve-se alternar as questões para não tornar o questionário maçador. Como já foi referido, um questionário só com questões fechadas pode por vezes levar a que o entrevistado fique um pouco "irritado", pois tem a sensação que as respostas lhe estão a ser impostas. Não se devem utilizar questões duplas, isto é, não devemos introduzir mais do que uma ideia em cada pergunta. Antes de elaborarmos algumas questões que podem provocar o embaraço do entrevistado, tais como por exemplo, questões sobre religião ou consumo de determinados produtos, devemos fazer uma pequena introdução ao inquirido, pois muitas pessoas podem ter receio de fornecer respostas erradas ou confessar a sua ignorância. Por isso, uma regra consiste em abordar essas questões da seguinte forma:

"...no seu caso pessoal poderia dizer-me..."; "Gostaria de saber a sua opinião...".

4.4 Os diferentes tipos de escalas

Se um questionário contiver perguntas fechadas, é necessário escolher sempre um conjunto de alternativas para cada questão (conforme Hill e Hill, 2000). Por exemplo, na questão Sexo, as alternativas são homem e mulher. Convém codificar as respostas (associar números a cada resposta) para que estas possam ser analisadas posteriormente por meio de técnicas estatísticas. Os dois tipos de escala frequentemente usados em questionários são as escalas nominais e as escalas ordinais. Mas há, no entanto, outros tipos de escalas igualmente utilizadas: as escalas de intervalo e de rácio.

4.4.1 Escala nominal

Este tipo de escala é utilizado em questões como a deste exemplo:

Na empresa onde trabalha qual é o cargo que ocupa?

Gerente	Técnico	Administrador	Operário
1	2	3	4

A estas questões é possível atribuir um número a cada categoria para codificar a resposta. Estes números só servem para identificar as categorias. Aliás, as diferentes modalidades ou categorias poderiam ser codificadas por outros símbolos, não necessariamente numéricos – por exemplo as categorias da variável sexo, masculino e feminino, poderiam ser representadas por M e F, respectivamente. Numa escala nominal não faz sentido calcular a média das variáveis, mas sim calcular as frequências das suas modalidades . Para se saber mais sobre o cálculo de frequências numa escala nominal, consulte o curso de Noções de Estatística existente na página do ALEA (página 2 do capítulo III, Dados, Tabelas e Gráficos - 1. Tipos de Dados, em: www.alea.ine.pt/html/nocoes/html/cap3_1_1.html).

4.4.2 Escala ordinal

Este tipo de escala é utilizado em questões como a que se segue:

Indique o seu grau de concordância ou discordância das seguintes afirmações relativas ao produto X

	Discordo totalmente	Discordo	Não concordo nem discordo	Concordo	Concordo totalmente
O produto X tem uma embalagem atractiva.	1	2	3	4	5
O produto X tem um preço muito caro.	1	2	3	4	5

Para as variáveis ordinais, do mesmo modo que para as nominais, também se utilizam as categorias mas, no entanto, existe uma relação de ordem entre elas. Se um júri ordenar 5 candidatos de 1 – mais fraco, a 5 – mais forte, podemos dizer que o candidato que ficou em 4º lugar é melhor do que o que ficou em 3º lugar. No entanto, não poderemos dizer que o candidato classificado com o número 4, é duas vezes melhor que o classificado com o número 2, isto é, não é possível medir a magnitude das diferenças entre as categorias (Hill e Hill, 2000). Do mesmo modo que para as variáveis nominais, continua a não ter sentido o cálculo da média mas, já que existe uma ordenação, pode-se calcular a mediana.

4.5 O Pré-teste

No início do capítulo falamos sobre o pré-teste. Mas afinal para que serve o pré-teste?

Quando uma primeira versão do questionário fica redigida, ou seja, quando a formulação de todas as questões e a sua ordem são provisoriamente fixadas, é necessário garantir que o questionário seja de facto aplicável e que responda efectivamente aos problemas colocados pelo investigador (Ghiglione e Matalon, 1992). Então, o questionário deve ser aplicado a um pequeno grupo de pessoas, com o objectivo de saber se elas entenderam o significado do questionário e das perguntas. Esta situação permite-nos saber como as questões e respostas são compreendidas, permite-nos evitar erros de vocabulário e de formulação e salientar recusas, incompreensões e equívocos (Ghiglione e Matalon, 1992). Com a elaboração do pré-teste podemos avaliar a taxa de recusas, conhecer a forma como as pessoas reagem ao questionário e se a ordem das questões não coloca nenhum problema. Podemos também constatar se há questões às quais quase todas as pessoas respondem da mesma forma, o que as torna muito pouco úteis para análises mais finas, realizadas através do cruzamento com outras questões. Neste caso é necessário rectificar a forma como as questões estão colocadas. Pode também recorrer-se a técnicas como a análise factorial, para identificar questões redundantes. Depois da análise do pré-teste, caso existam muitas alterações, é necessário voltar a testar o questionário quantas vezes for preciso.

5. Como Seleccionar os Elementos para a Amostra

De cada vez que se faz uma sondagem, é necessário seleccionar uma amostra da população que se pretende estudar, à qual se aplica depois um inquérito, para eventualmente se extrapolarem os resultados para toda a população (Vicente, Reis e Ferrão, 1996).

A necessidade de conhecer uma população no que respeita a uma ou várias características, impulsiona um processo de recolha e análise de informação. A dificuldade e mesmo nalguns casos, a impossibilidade de estudar a totalidade da população ditou a importância do estudo do recurso a amostras. É impossível assegurar a qualidade de uma sondagem, se não houver um conhecimento dos problemas e do impacto que eles podem ter nos resultados do estudo.

Amostra:

É uma parte da população que é observada com o objectivo de obter informação para estudar a característica pretendida.

5.1 Sondagens *versus* Recenseamentos

Quando precisamos de fazer um estudo sobre uma população, nem sempre é possível fazer um recenseamento, isto é, inquirir todos os elementos e, mesmo que fosse possível, este processo demoraria muito tempo, o que tornaria o estudo muito caro e possivelmente já sem nenhum sentido, pois tornar-se-ia desactualizado. As sondagens são mais baratas, menos demoradas, sendo muito mais fácil aceder a todos os elementos de uma amostra do que aos de uma população inteira.

Recenseamento:

Estudo de um universo de pessoas, instituições ou objectos físicos com o propósito de adquirir conhecimentos, observando todos os seus elementos e fazer juízos acerca de características importantes desse universo.

É certo que os recenseamentos são importantes pois são úteis na actualização de bases de dados para a realização de sondagens. Em Portugal, os Censos ou recenseamentos são realizados de dez em dez anos o que faz com que consigamos ter uma actualização exaustiva, tanto do parque habitacional como das características da população residente. Com o decorrer do tempo, essa base de dados vai ficando desactualizada, pois num curto espaço de tempo existem mudanças, tanto a nível habitacional como populacional. Por isso, conforme vão decorrendo os inquéritos por amostragem, a actualização da base de dados vai sendo feita.

5.2 Fases de realização de uma sondagem

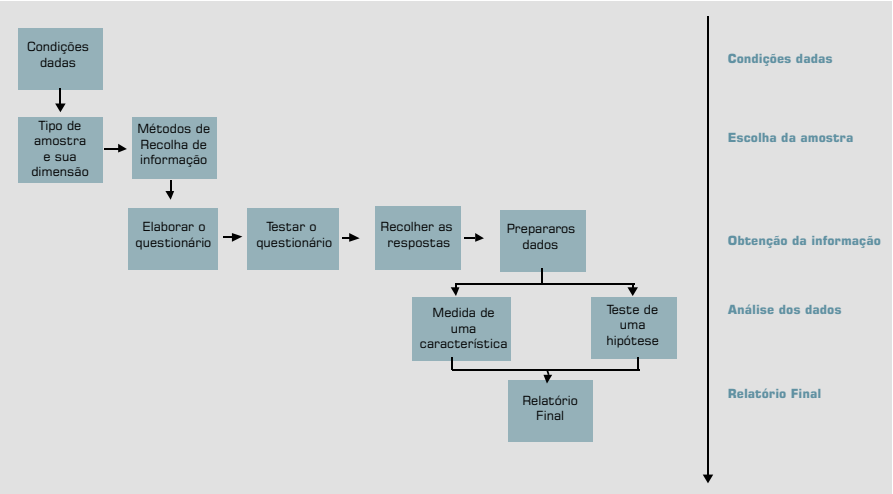
Como é habitual numa sondagem, o inquérito é aplicado a uma amostra retirada de uma população (Vicente, Reis e Ferrão, 1996). Conceber e levar à prática um estudo por sondagem é um processo complexo envolvendo diversas fases interdependentes.

A vantagem deste esquema é a ilustração clara da fase de **amostragem** (nível "escolha da **amostra**"), dentro do processo de **sondagem**.

É sobre os **métodos de amostragem** que a seguir nos vamos debruçar.

Fig. 4 - O recenseamento é uma fotografia instantânea da população num determinado momento.

Fig. 5 - (adapt. Vicente, Reis e Ferrão, 1996)



5.3 Etapas do Plano Amostral

Segundo Vicente Reis e Ferrão (1996), "o plano amostral é o momento da sondagem onde se seleccionam os elementos a partir dos quais se vão recolher os dados necessários".

Assim sendo, os passos requeridos para recolher a amostra podem ser descritos através da seguinte sequência:

- Definir a População Alvo
- Identificar a base de Sondagem
- Escolher uma técnica amostral
- Determinar a dimensão da amostra
- Seleccionar os elementos da amostra
- Recolher a informação necessária dos elementos da amostra

5.3.1 Definir a População Alvo

A definição da população alvo é uma das fases mais importantes na realização de uma sondagem. É sobre essa população que o nosso estudo vai incidir. A maior parte dos autores (Cochran (1963), Stuart (1984) e Barnett (1991)) definem como população alvo a totalidade dos elementos sobre os quais incide a nossa análise e dos quais se pretende obter informação. Para definir correctamente a população alvo, primeiro temos de ter a certeza qual é o objectivo do nosso inquérito, e depois, podemo-nos perguntar: sobre quem incide o inquérito? quem são os elementos de referência acerca dos quais se pretende obter a informação? Por exemplo, vamos supor que o objectivo do nosso inquérito era caracterizar o emprego e o desemprego em Portugal. Este estudo tem de ser feito junto das famílias mas, como através dos alojamentos é mais fácil detectar as famílias, devido à existência das moradas de residência, então a nossa população alvo é a dos alojamentos.

População Alvo:

Totalidade dos elementos sobre os quais incide a nossa análise e dos quais se pretende obter informação.

Alojamento:

Local distinto e independente construído, reconstruído, ampliado ou transformado para habitação humana e que, no período de observação, não está a ser utilizado, totalmente, para outro fim.

Base de Sondagem:

Diz respeito a listas, mapas ou qualquer outro registo da população de onde será retirada a amostra.

5.3.2 Identificação da Base de Sondagem

A base de sondagem é uma listagem dos elementos da qual se vai seleccionar a amostra (Vicente et al, 1996). Para utilizar a base de sondagem como a fonte para a recolha da amostra é necessário que se possam identificar as unidades amostrais, sendo estas, elementos ou grupo de elementos da população.

Pela dificuldade em construir essas listagens, é em muitos casos impossível fazer coincidir a população alvo com a população a inquirir. Trata-se dos casos em que a população é muito grande, tornando-se incomportável fazer selecções sucessivas de amostras. Nestes casos opta-se por considerar uma grande amostra, bem representativa da população, a que se chama base de sondagem. A partir desta população, que reúne características da população-alvo inicial, extraem-se, de seguida, várias amostras. No Instituto Nacional de Estatística, por exemplo, a Amostra-Mãe (utilizada em diversos inquéritos como por exemplo o Inquérito ao Emprego) é uma grande amostra extraída da população-alvo, a partir da qual se extraem outras amostras, relacionadas com os inquéritos às famílias. Posteriormente, quando esta base de sondagem começa a ficar saturada, pois certos indivíduos já foram inquiridos um determinado número de vezes, a base de sondagem é actualizada, através da substituição por novos indivíduos. Gomes (1998) explica claramente esta estratégia que consiste em actualizar uma parte "representativa" da população alvo, que assume o papel de base de sondagem. Tal como se referiu, em Portugal o INE actualiza a Amostra-Mãe de 5 em 5 anos e a partir de 1998 recorre-se a uma actualização parcial em cada ano.

5.3.3 Escolha de uma técnica amostral

Depois de definida a população-alvo, o problema que se levanta é o da selecção dos elementos da amostra. Nesta fase da sondagem importa distinguir os métodos probabilísticos ou aleatórios (em que aos elementos da população está associada uma probabilidade de inclusão na amostra) dos não probabilísticos (onde essa probabilidade não é determinada).

Os métodos probabilísticos estão associados à selecção de amostras aleatórias. No momento da selecção de uma amostra aleatória tem de se considerar toda a população, (ou, quando tal se justifica, uma base de sondagem).

Uma amostra é considerada não aleatória quando determinados elementos da população não têm possibilidade de serem escolhidos. Por exemplo, nas entrevistas de rua, apesar das pessoas serem escolhidas aleatoriamente, a amostra que se obtém é uma amostra não aleatória, visto que nem todos os indivíduos da população têm a mesma possibilidade de passar no local no momento em que se estão a realizar as entrevistas.

Amostragem Aleatória:

Procedimento de selecção dos elementos ou grupo de elementos de um modo tal que dá a cada elemento da população uma probabilidade de inclusão na amostra calculável e diferente de zero, ou seja, cada elemento da populaçao tem uma probabilidade conhecida de ser escolhido.

Amostragem não Aleatória:

Procedimento de selecção de elementos da população que permite a escolha dos indivíduos a incluir na amostra segundo determinado critério mais ou menos subjectivo. Nesta forma de amostragem não se conhece a probabilidade de determinado elemento ser seleccionado.

Importa salientar que só com a utilização de **amostras aleatórias** é possível conhecer o grau de confiança (grau de certeza que se tem a respeito da precisão da estimativa) dos resultados, mas em contrapartida são as **amostras não aleatórias** que possibilitam a conclusão mais rápida do estudo e com menor custo (Vicente, Reis e Ferrão, 1996). Quer se escolha uma **amostra aleatória ou não**, o importante é obter estimativas próximas dos parâmetros a estimar e isto só se consegue se tivermos uma **amostra** o mais representativa possível do universo.

Depois de feita uma pequena introdução acerca dos tipos de **amostras** veremos a seguir, com mais pormenor, as várias técnicas amostrais. Os principais tipos de **Amostragem Aleatória** são: simples, sistemática, estratificada, por *Clusters*, multi-etapas e multi-fases.

Fig. 6 - Entrevista de rua realizada porta a porta

A - Métodos Probabilísticos

5.3.3.1 Amostragem Aleatória Simples

O tipo de **amostragem probabilística** mais conhecido é o da **amostragem aleatória simples**. Segundo Stuart (1984), uma **amostra aleatória simples** (a.a.s.) de dimensão *n* é uma **amostra** seleccionada por um processo que confere a cada conjunto possível de n elementos a mesma probabilidade de ser seleccionado.

Pode-se mostrar que neste **plano de amostragem**, todos os elementos da população têm a mesma probabilidade de serem escolhidos para fazer parte da amostra.

Plano de Amostragem:

Metodologia adoptada para obter a amostra da população.

A obtenção de uma **amostra aleatória simples** pode ser feita mediante os seguintes passos (Vicente, Reis e Ferrão, 1996):

Passos para obtenção de uma amostra aleatória simples:

1. Numerar consecutivamente os elementos da população de 1 a N;
2. Escolher n elementos mediante o uso de um procedimento aleatório como seja o método da lotaria ou utilizando tabelas de números aleatórios, que podem ser geradas por computador. Os números têm que ser diferentes e não superiores a N;
3. Uma vez escolhidos os números, os elementos da população que lhes correspondem constituirão a amostra.

A escolha das **a.a.s.** nem sempre é a melhor opção. Devido a todos os indivíduos da **população** terem a mesma possibilidade de pertencerem à **amostra**, pode resultar em **amostras** muito dispersas geograficamente e, se forem exigidas **entrevistas pessoais**, a amostra obtida torna-se dispendiosa e morosa. Estas **amostras** podem ser uma óptima escolha se a **população** for reduzida; existirem listas com os elementos da **população**, sendo portanto possível a definição da **base de sondagem** e se a dispersão geográfica dos elementos não for um problema.

Exemplo de utilização da amostragem aleatória simples:

Considere-se uma população constituída por 20 nomes, de onde se pretende seleccionar aleatoriamente 10 nomes. O investigador associa cada nome da lista inicial a um número de 1 a 20, por exemplo, por ordem alfabética, sendo os números representados por dois dígitos - como por exemplo o 1, que será escrito 01. Depois, com o auxílio de uma tabela de números aleatórios (que se encontra praticamente em todos os livros de Estatística), o investigador vai seleccionando números de dois dígitos, até completar a dimensão da amostra necessária. Repare-se que haverá necessidade de seleccionar mais de 10 números, pois alguns não terão contrapartida na população considerada – por exemplo, se seleccionar o 56, terá de o deitar fora e seleccionar um outro número. Um outro processo consiste em gerar aleatoriamente, pelo computador (folha de cálculo, etc.) 10 números aleatórios entre 1 e 20.

Numa **população** com *N* elementos, o número total de amostras possíveis de n elementos, retirados sem reposição é dado por: $C_n^N = \frac{N!}{n!(N-n)!}$, pelo que a probabilidade de cada uma ser seleccionada é $(\frac{N!}{n!(N-n)!})^{-1}$

(ver "combinatória" no curso de Noções Probabilidades do ALEA em: www.alea.pt/html/probabil/html/cal_combinatorio/html/calcomb.html)

5.3.3.2 Amostragem Sistemática

Dada uma população de dimensão *N*, ordenada por algum critério, uma **amostra sistemática**, de dimensão n, é obtida seleccionando aleatoriamente um elemento de entre os primeiros K da **base de sondagem**, onde K é a parte inteira do quociente N/n, e adicionando todos os K-ésimos elementos seguintes (Vicente, Reis e Ferrão, 1996).

Passos para obtenção de uma amostra sistemática de dimensão *n*:

1. Calcular o intervalo *k* da amostra (obtido pelo quociente *N/n*, em que *k* representa a parte inteira desse quociente).
2. Escolher aleatoriamente um número *j* entre 1 e *k*.

Partindo desse número, adicionar sucessivamente o valor *k*, ficando assim seleccionados os elementos *j*, *j*+*k*, *j*+2*k*, *j*+3*k*, …, *j*+(*n*-1)*k*, perfazendo um total de *n* observações seleccionadas para a amostra.

A selecção de um elemento, na **amostra sistemática**, depende do que foi anteriormente seleccionado. De facto só o primeiro elemento é que é seleccionado aleatoriamente, sendo os restantes dependentes dessa primeira escolha. Neste tipo de amostra a probabilidade de selecção não é igual para todos os elementos.

Exemplo de utilização da amostragem sistemática (população conhecida)

- retirado de Vicente, Reis e Ferrão, (1996)

Consideremos uma população com 5135 indivíduos e pretende-se uma amostra aleatória sistemática de dimensão 100. Então o intervalo da amostra será 5135/100 ou seja 51,35, originando k=51; seguidamente, escolhe-se aleatoriamente um número entre 1 e 51 (por exemplo o 2) e por fim, todos os 51-ésimos da lista. Neste caso a amostra seria composta pelos elementos 2, 53, 104, 155, ... ,5051.

Por vezes a **amostragem sistemática** (a.s.) é preferível à **amostragem aleatória simples** (a.a.s.), por ser mais fácil de realizar devido ao facto de precisar de menos tempo do que o método de a.a.s. que utiliza o método da lotaria. Por outro lado, tem como desvantagens a dificuldade de atribuir números ao acaso, quando a população é desconhecida. Nestes casos, o valor j é escolhido ao acaso, mas os restantes elementos ($j+k$, $j+2k$, etc) são escolhidos por aplicação de um intervalo fixo, e portanto, não são escolhidos aleatoriamente (Hill, Hill, 2000).

Exemplo de utilização da amostragem sistemática (população desconhecida):

Suponhamos que queremos extrair uma amostra de 20 pessoas compradoras de um determinado estabelecimento comercial.

Como não sabemos qual a dimensão da nossa população, não podemos aplicar a a.a.s., logo vamos ter de aplicar a amostragem sistemática. Como fazemos para obter a nossa amostra?

Podemos optar pelo critério de escolher um comprador de 5 em 5 pelo que, o 5º, 10º, 15º, 20º, etc. são os elementos pertencentes à nossa amostra.

Outra desvantagem é que se deve ter em conta os padrões de repetição que podem enviesar a amostra. Imaginemos, por exemplo, que existe a necessidade de controlar a pontualidade e a assiduidade de um determinado funcionário. A população em estudo é composta pelos registos diários de entrada e saída do livro de ponto. Suponhamos que este funcionário está autorizado a chegar mais tarde às quartas-feiras por imperativos familiares. Se optarmos pela amostragem sistemática para a recolha da amostra e se k=7, sendo o primeiro dia uma quarta-feira, teremos de seleccionar apenas as quartas-feiras, o que enviesará a amostra. Este tipo de problemas surge sempre que a população está associada a padrões de repetição, como acontece neste caso com os dias da semana.

5.3.3.3 Amostragem Aleatória Estratificada

Enquanto as duas formas de **amostragem** anteriores consideram a **população** como um todo, existem situações em que conseguem identificar-se subdomínios ou subgrupos, que resultam da divisão da **população** em grupos ou **estratos** (Vicente, Reis e Ferrão, 1996). É o caso da **amostragem estratificada**. Nesta, cada **estrato** é tomado como uma **população** separada e a selecção dos elementos dentro de cada um dos estratos é feita à parte.

A **amostragem estratificada** tem, assim, por princípio, dividir a **população** em subconjuntos chamados estratos, de forma a realizar uma sondagem em cada um deles.

Estrato:

Subgrupo de elementos da população, que se pretende que sejam o mais homogéneos possível entre si no que respeita à característica em estudo.

Passos para obtenção de uma amostra estratificada:

1. Definir os estratos. Os estratos têm de ser bastante diferentes uns dos outros, mas os elementos dentro de cada estrato têm de ter características comuns (ex. sexo, grupo etário).
2. Seleccionar os elementos dentro de cada estrato, independentes uns dos outros.
3. Conjugar os elementos seleccionados em cada estrato, que na sua totalidade constituem a amostra.

Este tipo de **amostragem** é muito usado, visto que a maioria das **populações** podem ser divididas em **estratos** (por exemplo, homens/mulheres, alunos do ensino superior/não superior, etc) e conduz-nos a análises de subgrupos com variabilidades inferiores do que na a.a.s. Este tipo de amostragem tem como desvantagem ser muito caro e moroso quando existem muitos estratos.

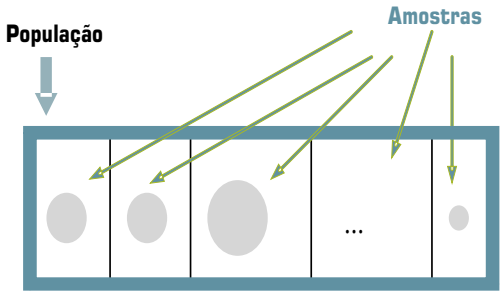
A **população** com N unidades é, assim, dividida em subpopulações ou estratos com $N1$, $N2$, …, Nk elementos, onde $N1+N2+\ldots+Nk = N$. Os **estratos** assim formados são mutuamente exclusivos e exaustivos.

Como já foi referido, a lógica que assiste à **estratificação** de uma **população** é a da identificação de grupos que variam muito entre si, ou seja, no que diz respeito ao **parâmetro** em estudo, mas muito pouco dentro de si, ou seja, cada grupo é homogéneo e com pouca variabilidade (Vicente, Reis e Ferrão, 1996). Cada **estrato** é tomado como uma **população** separada, de onde se retira uma **amostra**, que fornece uma estimativa. As estimativas obtidas a partir dos k estratos servem de base à construção de estimativas do parâmetro populacional em estudo.

Parâmetro:

Indicador quantitativo referente a um atributo ou característica da população (ex. média de idade das mulheres, total de pequenas empresas, etc.).

Fig. 7 - Esquema da amostragem aleatória estratificada



Exemplo de utilização da amostragem aleatória estratificada:

Suponhamos que se pretendia estudar o volume das vendas de prestação de serviços das empresas de construção civil. Podemos à partida considerar a População das empresas dividas em 3 estratos quanto ao número de trabalhadores que emprega: pequenas – 10 ou menos trabalhadores, médias – entre 11 e 40 e grandes – mais de 41 trabalhadores. Uma vez identificados os estratos, procede-se numa segunda etapa à recolha de uma a.a.s. dentro de cada estrato. Admitindo que a população em estudo é constituída por 500 empresas, das quais 55% são pequenas, 35% são médias e 10% são grandes e que a dimensão da amostra pretendida é de 85, seleccionaríamos, amostras de dimensão 47, 30 e 8, respectivamente do conjunto das pequenas, das médias e das grandes empresas. Esta selecção teve em conta a manutenção da igualdade da proporção do tamanho da amostra em cada estrato. Existem outros métodos de estratificação que podem ser consultados em Cochran

5.3.3.4 Amostragem Aleatória por Cachos

Um **cacho**, **grupo** ou **"cluster"**, é uma entidade que ocorre naturalmente associada a uma realidade. Uma escola, por exemplo (composta por várias salas, alunos e professores) pode ser considerada um **"cluster"** ou **cacho**. Podem ser considerados **"clusters"** universidades, hospitais, cidades, países, etc, onde existam réplicas da população a estudar. Estes grupos são seleccionados aleatoriamente e todos os **elementos** desse grupo são incluídos na **amostra**.

Cacho ou Cluster:

Grupo de unidades elementares da população, idealmente com a mesma variabilidade da população.

A preferência por este tipo de **amostragem** em muitos casos deve-se muitas vezes ao facto de esta ter um custo reduzido relativamente a outros tipos de **amostragem**.

Passos para obtenção de uma amostra por cachos:

1. Especificar os cachos, isto é, geralmente os elementos dos cachos estão fisicamente muito próximos e por isso apresentam características muito similares. Assim, pode não ter interesse definirmos cachos muito grandes.
2. Seleccionar uma amostra de cachos aleatoriamente e incluir na amostra todos os elementos que pertencem aos cachos seleccionados.

Como nem sempre é fácil obtermos **bases de sondagens**, a utilização da **amostragem por cachos** torna-se mais económica e é muito utilizada quando queremos fazer uma **sondagem** que cobre uma grande área geográfica.
Para exemplificarmos melhor este tipo de **amostragem**, consideremos um cacho de uvas. Se nós retirarmos uma uva do cacho, ficamos a saber se o resto das uvas desse mesmo cacho é de boa qualidade ou não, não precisando de comer o cacho todo, pelo que a selecção de todos os elementos do cacho para pertencerem à amostra resulta numa certa redundância.

Verifica-se que o princípio que torna eficiente a **amostra estratificada** torna ineficiente a **amostra por cachos** (Vicente, Reis e Ferrão, 1996). Quanto mais semelhantes forem os elementos dentro de um **cacho**, melhores serão os resultados se esse cacho for usado como um estrato na amostra estratificada e piores se forem usados como unidades amostrais na **amostragem por cachos**.

Exemplo: diferenças entre a amostragem estratificada e amostragem por cachos

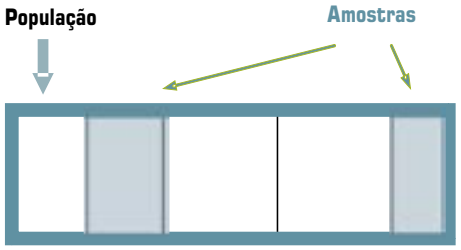
Caso 1: amostragem estratificada

Os empregados da firma XYZ são agrupados de acordo com os departamentos onde trabalham (vendas, marketing, investigação e produção). Seleccionam-se, em seguida, 10 empregados, aleatoriamente, de cada grupo.

Caso 2: amostragem por cachos

Cinco hotéis da cadeia Lírios (que é composta por 10 hotéis) foram seleccionados aleatoriamente. Todos os empregados desses 5 hotéis foram considerados na amostra.

Fig. 8 - Esquema da amostragem aleatória por cachos. Podemos imaginar que os cahos (aqui representados pelas células) são os hotéis do exemplo acima referido . Neste caso, apenas dois dos hotéis foram seleccionados, num total dos 5 que existem na população.



5.3.3.5 Amostragem Aleatória Multi-Etapas

A amostragem multi-etapas pode ser considerada como uma extensão da amostragem por cachos em que só alguns dos cachos são seleccionados e dos grupos ou cachos só se retiram alguns através de amostragem aleatória simples.

Exemplos de cachos numa amostragem multi-etapas (Vicente, Reis e Ferrão, 1996):

Clusters ou unidade amostral primária	Unidade amostral secundária	Unidade amostral terciária	Unidade amostral quaternária
Freguesia	Quarteirão	Prédio	Habitação
Página	Linha de Texto		
País	Centro urbano	Estab. comercial	

A **amostra** do Inquérito ao Emprego realizado pelo INE, por exemplo, é recolhida com base num processo de **amostragem multi-etápica**. De acordo com a sua metodologia (INE, 1998) a **população** é repartida num certo número de **unidades primárias** (freguesias). Cada unidade primária é, por sua vez repartida por secções estatísticas (áreas geográficas contíguas e uma única freguesia com cerca de 300 **alojamentos**). Cada secção estatística constitui uma **unidade secundária**. Em cada secção são listadas todas as unidades de **alojamento** que a constituem.

Unidade Amostral:

Elemento ou grupo de elementos da população. Uma amostra é constituída por unidades amostrais baseada em métodos probabilísticos.

B - Métodos Não Probabilísticos

Depois de termos abordado algumas técnicas de **amostragem aleatória**, vamos ver alguns tipos de **amostragem não aleatória**. Segundo Bacelar (1999), ao contrário das **técnicas aleatórias**, estas técnicas não têm "garantia estatística" de que a **amostra** seleccionada seja representativa. Não existe, nestes casos, uma teoria estatística de suporte à obtenção de **amostras** representativas, mas pode existir uma probabilidade significativamente elevada de que a **amostra** obtida seja representativa, embora esta probabilidade não seja muitas vezes de determinar. Estas técnicas de **amostragem não aleatória** são muito utilizadas e muito úteis quando não é possível usar **amostras aleatórias**, no âmbito de estudos preliminares ou exploratórios.

5.3.3.6 Amostragem por Conveniência

Uma **amostra por conveniência** consiste num grupo de indivíduos que se encontram disponíveis no momento da investigação. Estas **amostras** não são representativas da **população** (Vicente et al, 1996). Apesar da sua fragilidade científica, este tipo de **amostragem** pode ser usada com êxito em situações nas quais captar ideias gerais e identificar aspectos críticos pode ser mais importante do que a objectividade científica, como é o caso da realização de **pré-testes** de um **questionário**. Devido ao caracter "oportunista" da amostra, os seus elementos podem não ser representativos da população.

Exemplo de utilização da amostragem por conveniência:

Consideremos um estudo sobre a associação entre o rendimento das famílias e o acesso a serviços de saúde mental (psicanálise, psicologia médica, etc.). Para um estudo deste tipo, um investigador colocou 5 entrevistadores, em frente a 5 supermercados e 5 igrejas de um bairro degradado nos subúrbios de Nova Iorque.

5.3.3.7 Amostragem "Bola de Neve"

Este tipo de **amostragem** recai nos indivíduos que foram previamente identificados como pertencentes à **amostra**. É uma técnica utilizada nos casos em que não existe informação disponível sobre a **população**, ou torna-se impossível disponibilizá-la. Este tipo de amostragem é utilizado quando se pretende analisar populações pequenas ou com características muito específicas.

Para construir uma **amostra** baseada nesta técnica, o **entrevistador** pede ajuda ao inquirido, após ser entrevistado, para que este forneça nomes de outros indivíduos que possam ser igualmente inquiridos (Vicente et al, 1996). Um inconveniente deste processo é que as pessoas que são entrevistadas, têm tendência a indicarem amigos o que leva por vezes a termos uma **amostra** de pessoas que pensam e agem de forma idêntica.

Exemplo de utilização da amostragem "Bola de Neve":

Vamos supor que queremos uma amostra de toxicodependentes que residem no Porto. Como não temos nenhuma listagem, o que fazemos é tentar encontrar uma pessoa com essa característica e, depois de a entrevistarmos, pedirmos para nos indicar o nome de outras pessoas toxicodependentes residentes no Porto e garantir que não referimos qual a fonte dessa informação

5.3.3.8 Amostragem por quotas

Este é o **método não aleatório** de amostragem mais utilizado. É muito semelhante à **amostragem aleatória estratificada**, mas a selecção dos elementos da amostra não é **aleatória**. A existência deste método de **amostragem** justifica-se fundamentalmente pela inexistência de listagens da **população** (Vicente et al, 1996). A **amostragem por quotas** conduz a uma **amostra** onde a proporção de elementos que possuem uma determinada característica é aproximadamente igual à proporção de indivíduos na população que possuem essa mesma característica. Por exemplo, se a **população** tem tantos homens como mulheres, o mesmo vai acontecer na **amostra**.

Passos para obtenção de uma amostra por quotas:

1. Definir as quotas, isto é, dividir a população em categorias. A escolha das variáveis é feita na maioria dos casos com base no recenseamento da população, quando se trata de variáveis sócio-demográficas.
2. Seleccionar os elementos, cabendo ao entrevistador tomar a decisão de quem é escolhido. A única obrigatoriedade é que respeite as quotas estabelecidas no plano de amostragem. Muitas vezes definem-se planos para seleccionar os elementos, tais como circuitos urbanos ou fórmulas para encontrar o andar e o alojamento a inquirir num prédio.

A qualidade de uma **amostra por quotas** depende da forma como os **entrevistadores** procuram os indivíduos e entram em contacto com eles (Ghiglione e Matalon, 1992). Para assegurar uma melhor representatividade, os **entrevistadores** devem ser enviados para zonas tiradas à sorte. Aí, eles poderão, ou abordar quem passa, ou utilizar o método porta-a-porta, ou eventualmente, combinar os dois. A reprodução das distribuições da **população** deve ser considerada como uma condição necessária, mas não suficiente, da qualidade de uma **amostra**.

Neste método o tempo de realização do trabalho de campo é inferior ao dos **métodos aleatórios**, pois não há necessidade de contactar mais do que uma vez o entrevistado (Vicente et al, 1996). Se no primeiro contacto o indivíduo não se encontra é automaticamente substituído por outro. Esta pode ser uma vantagem expressiva se existir uma grande urgência na obtenção da informação.

Exemplo de utilização da amostragem por quotas:

Suponhamos que queremos fazer uma pesquisa sobre "quem pratica exercício físico". É certo que temos de ter em conta a idade, o sexo, tempo livre, etc. O primeiro passo que tem de ser dado é saber a proporção existente na população dessas características. Vamos supor que existem na população 40% de homens e 60% mulheres. Então, o entrevistador terá de inquirir 40% de homens e 60% de mulheres, o que será a sua "quota".

De seguida, apresentamos um quadro comparativo de alguns métodos probabilísticos e não probabilísticos, mais utilizados.

Fig. 9 - Métodos de amostragem probabilísticos e não probabilísticos mais utilizados – quadro resumo

Método/descrição	Vantagens	Desvantagens
Métodos Probabilísticos		
Amostragem Simples (Qualquer conjunto de n elementos tem a mesma probabilidade de ser seleccionado, de onde resulta que os elementos têm igual probabilidade de serem seleccionados)	Utilização fácil.	Os membros de alguns grupos de interesse menos representativos podem não ocorrer nas proporções desejadas.
Amostragem Estratificada (a população estudada é agrupada de acordo com características de interesse ou estratos)	Conduz a análises por subgrupos com variâncias inferiores do que na a mostragem simples.	Caro e moroso quando existem muitos estratos
Amostragem Sistemática (todo o x-ésimo elemento da população é seleccionado até perfazer o tamanho da amostra, de acordo com um passo fixo. Esse passo é determinado dividindo o tamanho da população pelo tamanho da amostra desejado).	Conveniente quando existe uma listagem de nomes como suporte da amostra.	Dever-se-á ter em conta os padrões de repetição que podem enviesar a amostra.
Amostragem por Cachos e Multi-etápica (Dos grupos formados naturalmente e que fazem parte da amostra serão inquiridos todos os seus elementos).	Utilização conveniente quando existem unidades estatísticas que correspondem aos grupos desejados (escolas, hospitais, etc.)	
Métodos não Probabilísticos		
Amostragem por Conveniência (utilização de indivíduos que se encontram disponíveis).	Método prático pois a investigação recai em unidades já disponíveis (estudantes nas escolas, doentes na sala de espera, etc.).	Devido ao caracter "oportunista" da amostra, os seus elementos podem não ser representativos da população.
Amostragem "Bola de neve" (Elementos previamente identificados identificam outros membros da população)	Útil quando não existem referências sobre a população ou essas referências são muito difíceis de obter.	A amostra pode resultar bastante enviesada.
Amostragem por quotas (A população é dividida em grupos, com base em características que só são identificáveis através da entrevista).	Torna-se prático quando existe informação fiável sobre as proporções dos atributos que interessam na população.	Neste processo o entrevistador pode conferir involutariamente enviesamentos na selecção dos inquiridos.

5.3.4 Como determinar a dimensão da amostra

A questão da dimensão a considerar para **amostra** é sempre uma decisão importante no processo de **sondagem**. Há dois aspectos muito importantes a ter em conta nesta fase: a **precisão** requerida para os resultados (pois existe sempre um erro que se pretende que seja o mais reduzido possível) e as **limitações de tempo e de custo** envolvidas na sondagem.

Também temos de ter em conta que quanto maior for a **amostra**, maior é a precisão, mas também maior é o custo. Por isso, devemos conjugar bem as duas situações.

A dimensão da **amostra** necessária para obter uma determinada precisão nos resultados só pode ser calculada matematicamente se as **amostras** forem escolhidas por um processo **aleatório**. Caso contrário, segundo Weiers (1998) temos três opções: adoptar a dimensão já utilizada, com sucesso, em estudos anteriores das mesmas características, ter em conta o orçamento disponível para o estudo e os custos envolvidos e por fim supormos que a **amostra** é **aleatória** e ver qual a dimensão que seria necessária, sendo o valor encontrado meramente indicativo. Uma **amostra** deve ser representativa da **população**, isto é, tem de apresentar os aspectos típicos, pois a **amostra** é um modelo em miniatura da **população**. Deve-se ter presente que a dimensão da amostra a recolher não é directamente proporcional ao tamanho da população e que essa dimensão depende fundamentalmente da variabilidade existente na população. Por exemplo, se relativamente à população constituída pelos alunos do 10º ano de uma escola secundária, estivermos interessados em estudar a sua idade média, a dimensão da amostra a recolher não necessita de ser muito grande já que a variável idade apresenta valores semelhantes, numa classe etária restrita. No entanto se a característica a estudar for o tempo médio que os alunos levam a chegar de casa à escola, já amostra terá de ter uma dimensão maior, uma vez que a variabilidade da população é muito maior. Cada aluno pode apresentar um valor diferente para esse tempo. Num caso extremo, se numa população a variável a estudar tiver o mesmo valor para todos os elementos, então bastaria recolher uma amostra de dimensão 1 para se ter informação completa sobre a população; se, no entanto, a variável assumir valores diferentes para todos os elementos, para se ter o mesmo tipo de informação seria necessário investigar todos os elementos (Graça Martins, 2001).

Exemplo: Determinação do tamanho da amostra num problema de estimação de uma proporção *p*

Pretende-se determinar a verdadeira proporção p de indivíduos com rendimento inferior a 1000 contos por ano numa região portuguesa. O intervalo de confiança para uma proporção tem a seguinte forma (admitindo uma dimensão da amostra maior que 100):

sendo:

c = parâmetro determinado pelo nível de confiança desejado

n = tamanho da amostra

f = frequência relativa do atributo na amostra (proporção)

$$\left[f - c\sqrt{\frac{f(1-f)}{n}}; f + c\sqrt{\frac{f(1-f)}{n}} \right]$$

Assim, a dimensão da amostra é determinada fixando a amplitude (A) e o nível de confiança desejados.

$$n = \frac{4c^2 f(1-f)}{A^2}$$

1. Consideremos uma população de dimensão **N** e seja **p** a proporção (desconhecida) de elementos da população que verificam determinada característica. Para estimar esta proporção **p**, recolhe-se uma amostra de dimensão **n** e calcula-se a proporção **p'** de elementos nessa amostra, que verificam a característica em estudo. Então o estimador **p'** é um bom estimador de **p**, com algumas propriedades muito interessantes, entre as quais sobressai o facto de ter uma variância (medida da variabilidade entre **p** e **p'**) igual a

$$\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)$$

Repare-se que se a dimensão n da amostra, for muito pequena quando comparada com a dimensão da população, N-n é aproximadamente igual a N-1, e fica unicamente o 1º factor da expressão que mede a variabilidade. É por esta razão que se diz que "quando a dimensão da população é muito grande quando comparada com a dimensão da amostra, pode-se considerar infinita".

2. Intervalo de confiança para a proporção p

Independentemente de como se chega lá, a forma do intervalo de confiança para p, com uma confiança de 100(1-α)% (α é um valor normalmente considerado da ordem de 0,05, e daí ser costume apresentar o intervalo de 95% de confiança!) é

$$\left(p' + z_\alpha\sqrt{\frac{p'(1-p')}{n}},\ p' + z_\alpha\sqrt{\frac{p'(1-p')}{n}}\right)$$

Amplitude do intervalo = $2\, z_\alpha\sqrt{\frac{p'(1-p')}{n}}$

À quantidade $z_\alpha\sqrt{\frac{p'(1-p')}{n}}$ chama-se a *margem de erro* ou precisão da sondagem.

3. Qual a dimensão da amostra que é necessário recolher para obter um intervalo com uma determinada precisão d e com um nível de confiança 100(1-α)%?

Teremos de resolver a seguinte equação em ordem a n: $z_\alpha\sqrt{\frac{p'(1-p')}{n}} < d$

$$n > \left(\frac{z\alpha}{d}\right)^2 p'(1-p')$$

Como o p´só é conhecido depois de recolhermos a amostra, temos de nos precaver para o valor máximo de p'(1-p') que acontece quando p'=1/2, de onde vem

$$n > \left(\frac{z\alpha}{2d}\right)^2$$

Apresenta-se a seguir a tabela dos valores de Zα, para alguns valores de α:

Confiança 100(1-α)%	z_α
90%	1.645
95%	1.960
98%	2.326
99%	2.576

Exemplo: Pretende-se saber se a população em geral tem confiança nos professores. Pretende-se obter uma estimativa com uma confiança de 95% e uma margem de erro no máximo de 0.05. Qual a dimensão da amostra que se deve recolher?

$$n > \left(\frac{1.96}{2\times 0.05}\right)^2$$

n=385

Se para a mesma confiança pretendermos uma margem de erro de 0.02, virá que a dimensão da amostra é muito maior, pois terá de ser igual a 2401!

5.3.5 Seleccionar os elementos da amostra

Tal como vimos nos itens anteriores, existem várias formas de seleccionarmos os **elementos** de uma **amostra**. Nas **amostras aleatórias** o esquema de selecção designa objectivamente qual o **elemento** a ser escolhido. Nestes casos, devido à existência de listagens prévias que contêm as referências sobre os elementos incluídos na amostra, é possível identificar cada um dos inquiridos e estabelecer contactos (pessoais, via telefone, ou por correio) de modo a desencadear o processo de recolha de dados. No caso do Inquérito ao Emprego do INE, por exemplo, os seleccionados são contactados por correio, seguindo-se um conjunto de várias visitas pessoais dos entrevistadores. Se a amostra for não aleatória, o entrevistador tem de seleccionar os elementos a incluir e, para tal, devido à inexistência de uma base de sondagem, é necessário recorrer ao julgamento humano (Vicente, Reis e Ferrão, 1996). No caso da amostragem por quotas, por exemplo, existem guiões ou planos que constituem um bom auxílio, pois ajudam o entrevistador a introduzir alguma aleatoriedade no processo de selecção dos entrevistados. Estes guiões ou planos contêm fórmulas para seleccionar as ruas dentro de uma freguesia, ou para seleccionar alojamentos dentro de um edifício.

6. Recolha da informação necessária dos elementos da amostra

Uma vez seleccionados os elementos da amostra há que os contactar no sentido de obter os dados necessários para a concretização do objectivo do estudo. Num estudo por sondagem existem essencialmente três métodos de recolha de informação: a entrevista pessoal, entrevista telefónica e o questionário por correio. Cada um destes métodos tem as suas vantagens e desvantagens, as quais passam a ser mencionadas.

6.1 Entrevista Pessoal

A entrevista pessoal pode ser considerada como uma conversa entre duas pessoas, face a face, iniciada e conduzida pelo entrevistador com o propósito particular de obter informação relevante, no sentido de concretizar os objectivos do estudo (Mayer, 1974). Este tipo de recolha de informação, foi durante muito tempo o mais utilizado, sendo hoje em dia, bastante importante na realização de alguns inquéritos realizados pelo INE. Este método de recolha de informação pode ser um bocado dispendioso, visto haver necessidade de formação prévia do entrevistador e este ter de se deslocar ao local do inquirido para obter a entrevista. Por vezes estas deslocações têm de ser feitas várias vezes, porque os entrevistados

não se encontram em casa, ou porque naquele momento não estão disponíveis para responder ao questionário. Por vezes pode também ocorrer uma recusa, o que torna este método mais dispendioso do que os outros dois métodos seguintes. Segundo Aaker e Day (1990) só 30% a 40% do tempo do entrevistador é gasto com a entrevista propriamente dita, pois o restante tempo é ocupado em deslocações, localização dos inquiridos, etc. É certo, que este método tem vantagens em relação ao questionário por correio, pois a entrevista pode ser conseguida em poucos minutos enquanto que o questionário por correio pode demorar semanas. A taxa de respostas é mais elevada na entrevista pessoal, devido ao facto de haver maior incentivo para a resposta por parte do entrevistador para com o entrevistado.

Entrevistador:

Pessoa responsável pela recolha de informação que vai de encontro aos objectivos particulares de cada estudo, realizando as entrevistas de acordo com as regras estabelecidas.

Entrevista Pessoal:

Pode ser considerada como uma conversa entre duas pessoas, face a face, iniciada e dirigida pelo entrevistador com o propósito particular de obter informação relevante, no sentido de concretizar os objectivos do estudo.

6.2 A Entrevista Telefónica

A **entrevista telefónica** é uma alternativa à **entrevista pessoal**. A recolha desta informação é feita pelo telefone, tal como o nome diz, onde o **entrevistador** realiza o **questionário** ao entrevistado. Este método torna-se por vezes mais barato do que o anterior. Por exemplo, se tivermos em conta que não é necessário fazer várias deslocações aos **alojamentos** para conseguirmos as entrevistas sendo o tempo que se gasta a fazer uma **entrevista por telefone** menor do que no caso da entrevista pessoal, este método é muito mais vantajoso. Mas, nem tudo são vantagens, pois se o **questionário** for muito longo, pode fatigar-se mais depressa e a interacção com o entrevistador é menor.

6.3 O Questionário por Correio ou de auto-preenchimento

A característica deste método é que aquele que vai responder ao **questionário**, após ter lido as questões e explicações que as acompanham, deverá por si só redigir as suas respostas sem poder recorrer a um **entrevistador**. Este método é aconselhável no caso de populações geograficamente dispersas. Os custos de recolha de informação são reduzidos. Os **questionários** são pré-testados várias vezes para se ter a certeza que as questões são entendidas e que todas as pessoas as entendem da mesma maneira. Apesar dos custos serem reduzidos, a questão do tempo nem sempre é muito favorável, portanto quando se tem de obter respostas rápidas este método não é aconselhável. Além do mais, deve-se ter em conta a taxa de não respostas que neste tipo de recolha de informação pode ser sempre mais elevado face aos anteriores.

Hoje em dia, com o desenvolvimento dos Call Centers (locais onde se realizam e recebem chamadas telefónicas), muitos inquéritos passavam a fazer-se no modo CATI-COMPUTER ASSISTED TELEPHONE INTERVIEW-. Têm profilerado, também, os inquéritos via web, realizados no modo CAWI- COMPUTER ASSISTED WEB INTERVIEW.

7. Organização e apresentação dos dados

Depois da definição do problema a estudar, da planificação do inquérito e da recolha dos dados temos o problema da organização os dados. A organização dos dados consiste em "resumir" os resultados obtidos de uma forma simples e clara para melhor serem interpretados. A apresentação dos dados pode ser feita de várias maneiras. Por exemplo, numa abordagem inicial, os dados podem ser apresentados em tabelas de frequências, diagramas de barras, diagramas circulares, histogramas, etc. Para obter mais informações sobre a organização dos dados ao nível da estatística descritiva introdutória, pode consultar os Dossiês sobre Estatísticas com Excel e Gráficos, disponíveis na página do ALEA (www.alea.pt/html/statofic/html/dossier/html/dossier.html) e neste livro.

Pode também consultar os resultados dos Mini-Censos realizados a várias escolas do nosso país, disponíveis na página:www.alea.pt/html/statofic/html/dossier/doc/Dossier5_2.PDF, onde encontrará um exemplo de formas de organização dos dados. Pode ainda consultar a Galeria Virtual (www.alea.pt/html/galvirt/html/galeriavirt.html) que contém exemplos de gráficos e quadros que sintetizam a informação principal dos inquéritos realizados.

Numa última fase, é necessário ter em atenção a apresentação do relatório final. Segundo Hill e Hill (2000) existem vários tipos de relatórios: por exemplo o académico e o interno. Ambos têm estruturas semelhantes e contêm os itens que a seguir se apresentam.

7.1 Algumas recomendações

Qualquer relatório deve conter um título que identifique qual o conteúdo apresentado no relatório. O índice deve conter todos os capítulos existentes no relatório. Devem ser enumerados e conter o número da página onde começam.

Embora o resumo seja a primeira parte do relatório, é normal não o escrever até que todas as outras componentes estejam escritas, revistas, "polidas" e existam nas suas versões finais. (Hill e Hill, 2000). O resumo deve conter a informação sobre qual a razão que levou a fazer a investigação, como foi feita, quais os resultados mais importantes e as conclusões tiradas acerca da sua investigação e como podem ajudar a resolver o problema. A introdução tem como objectivo explicar qual a natureza da investigação e as razões que a justificaram e deve apresentar uma breve panorâmica sobre os restantes capítulos do relatório.

7.2 Os resultados

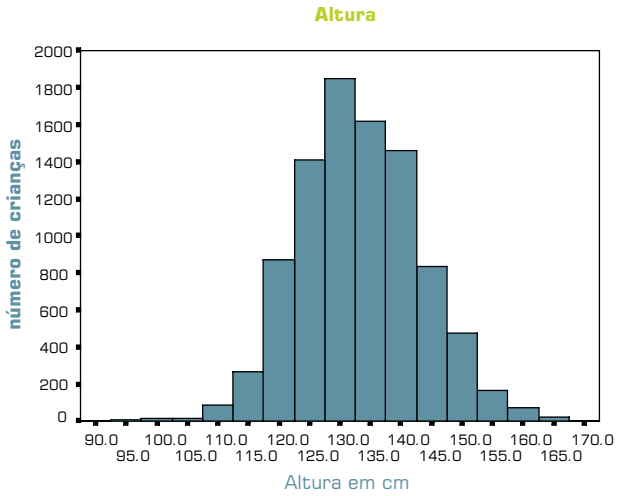
Existem várias maneiras de apresentar resultados numéricos. Devemos sempre apresentar uma análise exploratória inicial dos dados, com particular incidência num resumo das principais variáveis analisadas.

Por exemplo, nos "Mini-Censos", uma das variáveis analisadas foi a altura dos indivíduos[1] . No relatório que apresenta os resultados deste trabalho, um dos quadros contém uma síntese descritiva desta variável:

ALTURA

N	9171
Minimo	92
Máximo	170
Média	133.21
Desvio padrão	9.917

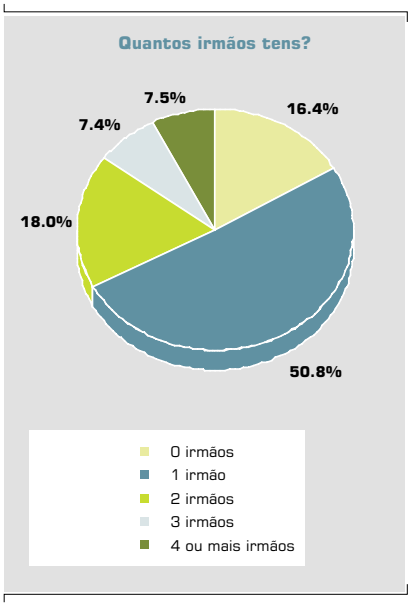
Para os mesmos dados optou-se por fazer igualmente uma representação gráfica, sob a forma de um histograma (ver regras de construção de histogramas nas Noções de Estatística do ALEA).



Para a variável "Número de irmaõs", apresentou-se a tabela de frequências e o gráfico circular correspondente.

Neste quadro podemos ver o número de irmãos que cada criança tem. Podemos observar que cerca de metade das crianças que responderam a esta questão têm mais um irmão e que 16% são filhos únicos. 18% das crianças têm 2 irmãos e as restantes têm 3 ou mais.

		Frequências Absolutas	Frequências Relativas (%)	Frequências Relativas Acumuladas (%)
número de irmãos	0	1403	16.4	16.4
	1	4356	50.8	67.1
	2	1540	18.0	85.1
	3	636	7.4	92.5
	4 ou mais	643	7.5	100.0
	Total	8578	100.0	
Não responderam		593		
Total		9171		



[1] Uma das principais iniciativas realizadas pelo ALEA em 2001 foi a do "Mini-Censos" destinado às escola básicas. Remeteram-se os inquéritos a escolas do 1º ciclo e toda a informação recolhida foi organizada e tratada por uma equipa conjunta envolvendo também técnicos do INE e da Sociedade Portuguesa de Estatística. os "Mini-Censos" tiveram com principal propósito dar a conhecer aos alunos o que são, para que servem e como se fazem os Censos. O relatório com os resultados deste encontra-se disponível em: www.alea.pt/Html/statofic/html/censos2001/html/censos2001.html

Segundo Hill e Hill (2000) quando apresentamos os resultados, devemos ter em atenção qual o nosso público alvo, para assim escolhermos o método mais adequado de apresentação. Quando o público alvo está habituado a ler e interpretar quadros, devemos utilizá-los mas de uma forma a facilitar a sua interpretação. Por outro lado, quando o público alvo não está habituado a ler e interpretar quadros, devemos utilizar gráficos para apresentar a informação mais importante. Ambas as escolhas de apresentação dos resultados devem ser acompanhadas por uma explicação em forma de texto para melhor compreensão do leitor. Os quadros e gráficos apresentados devem ser todos numerados e conterem um título.

Para esta fase do trabalho recomendamos uma consulta aos dossiês didácticos "Estatística com Excel" e "Representações gráficas".

8. Ver também...

Publicações

• ALEA, "Estatística com Excel", Dossiê Didáctico nº IV, disponível em: http://alea.ine.pt/html/statofic/html/dossier/html/meio_dossier4.html

• ALEA, "Representações Gráficas - notas sobre a criação e apresentação de alguns tipos de gráficos", Dossiê Didáctico nº IX, disponível em: http://alea.ine.pt/html/statofic/html/dossier/html/meio_dossier9.html

• BACELAR, S. (1999), Relatório de Aula Teórico-Prática sobre Amostragem nas Ciências Sociais, PAPCC, FEP, Porto, Universidade do Porto;

• CAMPOS, P. (1997), Relatório de aula teórica -prática sobre Teoria da Amostragem, PAPCC, FEP, Universidade do Porto.

• CAMPOS, P. (2000), Módulo 2 - da Concepção ao Tratamento Estatístico de Questionários - Apontamentos do curso de Análise Estatística de Dados com SPSS. Escola Superior de Biotecnologia da Universidade Católica, Porto.

• GHIGLIONE, R. e MATALON, B. (1992), O Inquérito, Teoria e Prática, Oeiras, Celta Editora;

• GOMES, P. (1998), Tópicos de Sondagens, (Curso apresentado no âmbito do VI Congresso da Sociedade Portuguesa de Estatística - Tomar, 9 a 12 de Junho de 1998);

• GRANGÉ, D., LEBART, L. (1994), Traitements Statistiques des Ênquetes, Paris, Edições Dunod;

• HILL, M. M., Hill, A. (2000), Investigação por Questionário, Lisboa, Edições Sílabo;

• INE (1998), Inquérito ao Emprego - Série - 1998; também disponível na Internet na publicação referente 1º Trimestre de 1998 das Estatísticas do Emprego.

• LIMA, M. P. (1981), O Inquérito Sociológico - Problemas de Metodologia, 2ª Ed., Editorial Presença;

• MARTINS, E. G., (2001), Noções Básicas sobre Amostragem - Introdução à Inferência Estatística, Departamento de Estatística e Investigação Operacional, Faculdade de Ciências da Universidade de Lisboa;

• STUART, A., (1984), The Ideas of Sampling, Monograph no. 4, Charles Griffin and Company Ltd, London;

• VICENTE, P., REIS, E. e FERRÃO, F. (1996), Sondagens - A amostragem como factor decisivo da qualidade, Lisboa, Edições Sílabo;

• WEIERS, R.M. (1998), Marketing Research, 2nd Ed., Prentice-Hall, London.

Web sites:

http://www.socio-estatistica.com.br/

http://www.fecap.br/portal/index.asp

Nestes dois sites pode encontrar algumas sugestões sobre a construção de questionários e algumas referências bibliográficas.

Estatística Descritiva com EXCEL

Luísa Canto E Castro Loura
Maria Eugénia Graça Martins

Estatística Descritiva com EXCEL

Complementos

Luísa Canto E Castro Loura
Maria Eugénia Graça Martins

Sumário

Nota Introdutória

Este dossiê é constituído por 5 capítulos, cada um autónomo dos restantes. Assim, um leitor interessado em saber como construir uma tabela de frequências ou um histograma vai directamente para o Capítulo 2, "Representação e redução de dados. Tabelas e gráficos", sem necessitar de passar pelo Capítulo 1. Do mesmo modo, se estiver interessado em utilizar o Excel, por exemplo, no cálculo da média ou da mediana, vai directamente para o Capítulo 3. Assim, é fundamental a consulta do índice, para situar os seus interesses imediatos.

Este dossiê começou por ter como objectivo servir de apoio à interpretação do programa do módulo B2 dos cursos profissionais. Nestes cursos era pressuposto que os alunos tivessem um módulo de Estatística onde aprenderiam os principais conceitos e técnicas subjacentes ao tratamento e redução de colecções de dados.

Entretanto fizemos uma revisão do texto inicial e pensamos que a sua utilização poderá ser útil a todos os que pretenderem implementar as referidas técnicas. O software escolhido é o Excel (versão em Inglês) que, embora não seja um software estatístico, inclui funções para cálculo das principais estatísticas descritivas, permite realizar as principais representações gráficas e, mediante recurso a outras funções predefinidas, permite ainda efectuar procedimentos não imediatos como seleccionar aleatoriamente uma amostra, construir histogramas com classes de diferente amplitude, organizar os dados em tabelas de contingência ou, até mesmo, proceder à simulação de pequenas experiências aleatórias.

A abordagem foi feita de uma forma que se pretendeu simples, pois o nosso objectivo foi fazer uma introdução à utilização do Excel. Para a resolução de alguns dos exemplos tratados, haverá outros tipos de abordagem, ainda utilizando o Excel e incentivamos fortemente os leitores a enveredarem e ensaiarem outras alternativas, que possam eventualmente ser utilizadas.

Não é demais repetir a ideia de que a Estatística é uma ciência e também é uma arte. Assim, cada utilizador da Estatística pode dar um pouco de si ao fazer um tratamento de dados, mesmo que esse tratamento seja só exploratório ou descritivo.

1. Noções básicas sobre amostragem

1.1 - Introdução[1]

Não é uma tarefa simples definir o que é a Estatística. Por vezes define-se como sendo um conjunto de técnicas de tratamento de dados, mas é muito mais do que isso! A Estatística é uma "arte" e uma ciência que permite tirar conclusões e de uma maneira geral fazer inferências a partir de conjuntos de dados.

Até 1900, a Estatística resumia-se ao que hoje em dia se chama Estatística Descritiva ou Análise de Dados. Apesar de tudo, deu contribuições muito positivas em várias áreas científicas.

A necessidade de uma maior formalização nos métodos utilizados, fez com que, nos anos seguintes, a Estatística se desenvolvesse numa outra direcção, nomeadamente no que diz respeito ao desenvolvimento de métodos e técnicas de Inferência Estatística. Assim, por volta de 1960 os textos de Estatística debruçam-se especialmente sobre métodos de estimação e de testes de hipóteses, assumindo determinadas famílias de modelos, descurando os aspectos práticos da análise dos dados.

Porém, na última década, em grande parte devido às facilidades computacionais postas à sua disposição, os Estatísticos têm-se vindo a preocupar cada vez mais, com a necessidade de desenvolver métodos de análise e exploração dos dados, que dêem uma maior importância aos dados e que se traduz na seguinte frase "Devemos deixar os dados falar por si".

Do que dissemos anteriormente, podemo-nos aperceber que a Estatística é uma ciência que trata de dados e que num procedimento estatístico estão envolvidas duas fases importantes, nomeadamente a fase que diz respeito à organização de dados – Análise de Dados, e a fase em que se procura retirar conclusões a partir dos dados, dando ainda informação de qual a confiança que devemos atribuir a essas conclusões – Inferência Estatística. Existe, no entanto, uma fase pioneira, que diz respeito à Produção ou Aquisição de Dados. Para realçar a importância desta fase consideremos, por analogia, o que se passa quando se pretende realizar um determinado cozinhado. Começa-se por seleccionar os ingredientes, que serão depois manipulados de acordo com determinada receita. O resultado do cozinhado pode ser desastroso, embora de aspecto agradável. Efectivamente se os ingredientes não estiverem em condições, resulta um prato de aspecto semelhante ao que se obteria com ingredientes bons, mas de sabor intragável. O mesmo se passa com o procedimento estatístico. Se os dados não forem bons, embora se aplique a técnica correcta, o resultado pode ser desastroso, na medida em que se pode ser levado e retirar conclusões erradas.

Hoje em dia com a utilização cada vez maior de dados nas mais variadas profissões e nas mais diversas situações do dia a dia, torna-se necessário acompanhar este processo de uma cultura estatística que cada vez mais abarque um maior número de pessoas, para que mais facilmente se consiga compreender o mundo que nos rodeia.

[1] Este capítulo segue de perto o texto Introdução à Probabilidade e à Estatística – Com complementos de Excel, de Maria Eugénia Graça Martins, edição da Sociedade Portuguesa de Estatística, 2005.

Sendo a Estatística a ciência que trata dos dados, gostaríamos desde já de chamar a atenção para que fazer estatística é muito mais do que fazer cálculos e manipular fórmulas. Também não é matemática, embora utilize a matemática. Efectivamente, ao fazer estatística trabalhamos com dados, que são mais do que números! Como diz David Moore (1997) " Data are numbers, but they are not "just numbers". Data are numbers with a context. The number 10.5, for example, carries no information by itself. But if we hear that a friend's new baby weighed 10.5 pounds at birth, we congratulate her on the healthy size of the child. The context engages our background knowledge and allows us to make judgements. We know that a baby weighing 10.5 pounds is quite large, and that it isn't possible for a human baby to weigh 10.5 ounces or 10.5 kilograms. The context makes the number informative".

Da experiência que temos no dia a dia com os dados já concluímos, com certeza, que estes apresentam variabilidade. Por exemplo é comum que um pacote de açúcar que na embalagem tenha escrito um quilograma, não pese exactamente um quilograma. Por outro lado ao pesar duas vezes o mesmo pacote possivelmente não obteremos o mesmo valor. Assim, ao dizermos que o peso do pacote é um determinado valor, não podemos ter a certeza que esse valor seja correcto. Esta variabilidade está presente em todas as situações do mundo que nos rodeia, pelo que as conclusões que tiramos a partir dos dados que se nos apresentam, têm inerente um certo grau de incerteza.

A Estatística trata e estuda esta variabilidade apresentada pelos dados. Permite-nos a partir dos dados retirar conclusões, mas também exprimir o grau de confiança que devemos ter nessas conclusões. É precisamente nesta particularidade que se manifesta toda a potencialidade da Estatística.

Podemos então, e tal como refere David Moore em Perspectives on Contemporary Statistics, considerar três grandes áreas nesta ciência dos dados:

1. Aquisição de dados
2. Análise dos dados
3. Inferência a partir dos dados

Neste capítulo vamos abordar o primeiro tema considerado, ou seja o que diz respeito à Aquisição de Dados, numa perspectiva de que pretendemos obter dados, a partir dos quais seria possível responder a determinadas questões, isto é, posteriormente retirar conclusões para as Populações a partir das quais esses dados são adquiridos – contexto em que tem sentido fazer inferência estatística. Vamos assim, preocupar-nos em obter amostras representativas de Populações que se pretendem estudar.

1.2 – Aquisição de dados: sondagens e experimentações. População e amostra. Parâmetro e Estatística.

O mundo que nos rodeia será mais facilmente compreendido se puder ser quantificado. Em todas as áreas do conhecimento é necessário saber "o que medir" e "como medir". Na Estatística ensina-se a recolher dados válidos, assim como a interpretá-los.

Perante um conjunto de dados podem-se distinguir duas situações:

• Aquela em que o estatístico é confrontado com conjuntos de dados sem ter qualquer ideia preconcebida sobre o que é que vai encontrar e então procede a uma análise exploratória de dados, quase sempre utilizando processos

gráficos, análise esta que revelará aspectos do comportamento dos dados. Neste caso não se fala em amostras, mas sim conjuntos de dados (Murteira, 1993) e de uma maneira geral a análise exploratória é suficiente para os fins que se têm em vista;

• Uma outra em que procede à análise de dados com propósitos bem definidos no sentido de responder a questões específicas. Neste caso os dados têm que ser produzidos ou adquiridos por meio de técnicas adequadas de forma a que resultem dados válidos (amostras representativas). Estas técnicas, em que é fundamental a intervenção do acaso, revolucionaram e fizeram progredir a maior parte dos campos da ciência aplicada. Pode-se dizer que hoje em dia não existe área do conhecimento para cujo progresso não tenha contribuído a Estatística.

Abordaremos de seguida algumas das técnicas de aquisição de dados, que se enquadram nesta última situação, em que se distinguem as **Sondagens e Experimentações (aleatoriezadas).**

Gostaríamos desde já de realçar que o objectivo deste texto é o de explorar, de uma forma simples, algumas das técnicas de amostragem, com vista à realização de sondagens, situações que se encontram de um modo geral nas Ciências Sociais, ao contrário das Ciências experimentais, tais como Física ou Química, em que a recolha de dados se faz fundamentalmente recorrendo a experiências. Por exemplo, a população constituída pelos eleitores, a população constituída pela contas sedeadas num banco, etc., que só contêm um número finito de elementos, ao contrário da População conceptual de respostas geradas por um processo químico.

Não é demais realçar a importância desta fase, a que chamamos de Produção ou Aquisição de Dados. Como é referido em Tannenbaum (1998), página 426: "Behind every statistical statement there is a story, and like a story it has a beginning, a middle, an end, and a moral. In this first statistics chapter we begin with the beginning, which in statistics typically means the process of gathering or collecting data. Data are the raw material of which statistical information is made, and in order to get good statistical information one needs good data".

1.2.1 – Sondagens. População e amostra. Parâmetro e Estatística.

Estas noções, que já foram dadas num módulo anterior, são aqui de novo apresentadas, unicamente com o objectivo de enquadrar o estudo seguinte, ou seja, o de introduzir algumas noções de Amostragem.

O objectivo de uma sondagem é o de recolher informação acerca de uma população, seleccionando e observando um conjunto de elementos dessa população.

Sondagem

Estudo estatístico de uma população, feiro através de uma amostra, destinado a estudar uma ou mais características tais como elas de apresenta nessa população.

Por exemplo, numa fábrica de parafusos o departamento de controlo de qualidade pretende saber qual a percentagem de parafusos defeituosos. Tempo, custos e outros inconvenientes impedem a inspecção de todos os parafusos. Assim, a informação pretendida será obtida à custa de uma parte do conjunto – amostra, mas com o objectivo de tirar conclusões para o conjunto todo – população. Se se observarem todos os elementos da população tem-se um recenseamento. Por vezes confunde-se sondagem com amostragem. No entanto a amostragem diz respeito ao procedimento da recolha da amostra qualquer que seja o estudo estatístico que se pretenda fazer, pelo que a amostragem é uma das fases das sondagens, já que estas devem incluir ainda o estudo dos dados recolhidos, assim como a elaboração do relatório final.

População, unidade, amostra

População é o conjunto de objectos, indivíduos ou resultados experimentais acerca do qual se pretende estudar alguma característica comum. As populações podem ser finitas ou infinitas, existentes ou conceptuais. Aos elementos da população chamamos unidades estatísticas.

Amostra

É uma parte da população que é observada com o objectivo de obter informação para estudar a característica pretendida.

Geralmente, há algumas quantidades numéricas acerca da população que se pretendem conhecer. A essas quantidades chamamos parâmetros.

Por exemplo, ao estudar a população constituída por todos os potenciais eleitores para as legislativas, dois parâmetros que podem ter interesse são:

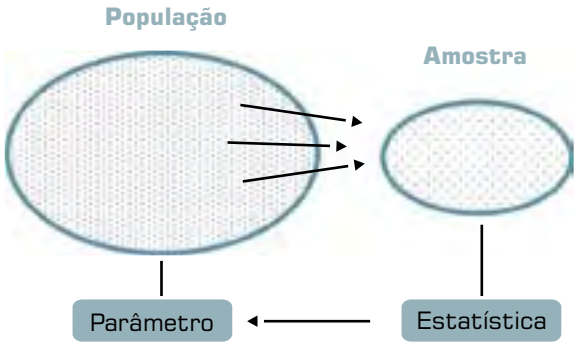
- idade média dos potenciais eleitores que estão decididos a votar;
- percentagem de eleitores que estão decididos a votar.

Para conhecer aqueles parâmetros, teria de se perguntar a cada eleitor a sua idade, assim como a sua intenção no que diz respeito a votar ou não. Esta tarefa seria impraticável, nomeadamente por questões de tempo e de dinheiro.

Os parâmetros são estimados por estatísticas, que são números calculados a partir dos dados que constituem a amostra. No caso do exemplo anterior, se se tivesse recolhido uma amostra de dimensão 1000, à característica populacional "percentagem de eleitores que estão decididos a votar" corresponde a característica amostral "percentagem dos 1000 eleitores, que interrogados disseram estar decididos a votar". Estas quantidades são conceptualmente distintas, pois enquanto a característica populacional (parâmetro) pode ser considerada um valor exacto, embora desconhecido, a característica amostral (estatística) é conhecida, embora difira de amostra para amostra, mas que todavia pode ser considerada uma estimativa útil da característica populacional respectiva.

Parâmetro

É uma característica numérica da população, enquanto que a estatística é uma característica numérica da amostra.



No entanto, para se poder utilizar as estatísticas, para estimar parâmetros é necessário que as amostras sejam representativas das populações de onde foram retiradas.

Observação – Anteriormente dissemos que uma estatística é um número calculado a partir dos dados da amostra, que se utiliza para estimar um parâmetro. Como, de um modo geral, podemos recolher muitas amostras diferentes, embora da mesma dimensão, teremos muitas estatísticas diferentes, como estimativas do parâmetro em estudo. Tantas as amostras diferentes (2 amostras da mesma dimensão serão diferentes se diferirem pelo menos num dos elementos) que se puderem obter da população, tantas as estimativas eventualmente diferentes que se podem calcular para o parâmetro. Então podemos considerar que todas estas estimativas são os valores observados de uma função dos elementos da amostra, a que se dá o nome de estimador. A esta função também se dá o nome de estatística,

utilizando-se assim, indevidamente, o mesmo termo para a variável e o valor observado da variável.

É oportuno chamar a atenção para o seguinte: por vezes a População que se estuda, ou seja a População inquirida, não é a objecto do estudo – População alvo ou População objectivo. Por exemplo, se se pretende estudar a População constituída pelos indivíduos adultos de nacionalidade portuguesa - População alvo, a População inquirida pode, no entanto, ser constituída pelos indivíduos adultos de nacionalidade portuguesa e residentes no território português, à data do inquérito.

1.2.1.1 – Amostra enviesada. Amostra aleatória e amostra não aleatória.

Uma amostra que não seja representativa da População diz-se enviesada e a sua utilização pode dar origem a interpretações erradas, como se sugere nos seguintes exemplos:

- utilizar uma amostra constituída por 10 benfiquistas, para prever o vencedor do próximo Benfica - Sporting!
- utilizar uma amostra constituída por leitores de determinada revista especializada, para tirar conclusões sobre a opinião da população em geral.

Um processo de amostragem diz-se enviesado quando tende sistematicamente a seleccionar elementos de alguns segmentos da População, e a não seleccionar sistematicamente elementos de outros segmentos da População.

Surge assim, a necessidade de fazer um planeamento da amostragem, onde se decide quais e como devem ser seleccionados os elementos da População, com o fim de serem observados, relativamente à característica de interesse. De um modo geral, o trabalho do Estatístico deve começar antes de os dados serem recolhidos. Deve planear o modo de os recolher, de forma a que, posteriormente, se possa extrair o máximo de informação relevante para o problema em estudo, ou seja para a população de onde os dados foram recolhidos e de modo a que os resultados obtidos possam ser considerados válidos. Vem a propósito referir a seguinte frase de Fisher: "Ao pedir a um Estatístico que diagnostique dados já recolhidos, muitas vezes só se obtém uma autópsia".

O planeamento de um estudo estatístico, que começa com a forma de seleccionar a amostra, deve ser feito de forma a evitar amostras enviesadas. Alguns processos que provocam quase sempre amostras enviesadas são, por exemplo, a amostragem por conveniência e a obtenção de uma amostra por resposta voluntária. Este último processo é usado, com muita frequência, pelas estações de televisão ou jornais, com resultados por vezes contraditórios com os que se obtêm quando se utiliza um processo correcto (aleatório) de seleccionar a amostra.

A utilização de uma amostragem por conveniência também se realiza frequentemente, quando se selecciona a amostra a partir de uma listagem dos elementos de determinado clube ou grupo, como por exemplo a Ordem dos Engenheiros. A seguir apresentamos exemplos de más amostras ou amostras enviesadas e resultado da sua aplicação:

Amostra 1 – A SIC pretende saber qual a percentagem de pessoas que é a favor da despenalização do aborto. Para isso indicou dois números de telefone, um dos quais para as respostas SIM e o outro para a resposta NÃO. Resultado – A utilização da percentagem de respostas positivas como indicação da percentagem da população portuguesa que é a favor da despenalização do aborto é enganadora. Efectivamente só uma pequena percentagem da população responde a estas questões e de um modo geral tendem a ser pessoas com a mesma opinião.

Amostra 2 – Uma estação de televisão preparou um debate sobre o aumento de criminalidade, onde enfatizou o facto de ter aumentado o número de crimes violentos. Ao

mesmo tempo, e inserida no mesmo programa, decorria uma sondagem de opinião sobre se as pessoas eram a favor da implementação da pena de morte. Esta recolha de opiniões era feita no molde descrito no exemplo anterior, isto é, por resposta voluntária. Resultado – A utilização da percentagem de SIM´s, que naturalmente se espera elevada, dá uma indicação errada sobre a opinião da população em geral. As pessoas influenciadas pelo debate e pelo medo da criminalidade serão levadas a telefonar dando indicação de estarem a favor da pena de morte.

Amostra 3 – Recolha de opiniões de alguns leitores de determinada revista técnica, para representar as opiniões dos portugueses em geral.

Resultado – Diferentes tipos de pessoas lêem diferentes tipos de revistas, pelo que a amostra não é representativa da população. Basta pensar que, de um modo geral, a população feminina ainda não adere às revistas técnicas como a população masculina. A amostra daria unicamente indicações sobre a população constituída pelos leitores da tal revista.

Amostra 4 – Utilização de alguns alunos de uma turma, para tirar conclusões sobre o aproveitamento de todos os alunos da escola. Resultado – Poderíamos concluir que o aproveitamento dos alunos é pior ou melhor do que na realidade é. As turmas de uma escola não são todas homogéneas, pelo que a amostra não é representativa dos alunos da escola. Poderia servir para tirar conclusões sobre a população constituída pelos alunos da turma.

Amostra 5 – Utilização dos jogadores de uma equipa de basquete de uma determinada escola para estudar as alturas dos alunos dessa escola. Resultado – O estudo concluiria que os estudantes são mais altos do que na realidade são.

Os exemplos que apresentámos anteriormente são exemplos de amostras enviesadas porque tiveram a intervenção do factor humano. Com o objectivo de minimizar o enviesamento, no planeamento da escolha da amostra deve ter-se presente o princípio da aleatoriedade de forma a obter uma amostra aleatória.

Amostra aleatória e amostra não aleatória – Dada uma população, uma amostra aleatória é uma amostra tal que qualquer elemento da população tem alguma probabilidade de ser seleccionado para a amostra. Numa amostra não aleatória, alguns elementos da população podem não poder ser seleccionados para a amostra.

Amostra aleatória e amostra não aleatória

Dada uma população, uma amostra aleatória é uma amostra tal que qualquer elemento da população tem alguma probabilidade de ser seleccionado para a amostra. Numa amostra não aleatória, alguns elementos da população podem não poder ser seleccionados para a amostra.

Quando se pretende recolher uma amostra de dimensão n, de uma População de dimensão N, podemos recorrer a vários processos de amostragem. Como normalmente o objectivo é, a partir das propriedades estudadas na amostra, inferir propriedades para a População, gostaríamos de obter processos de amostragem que dêem origem a "bons" estimadores. Embora a classificação de um estimador como "bom" ou não, saia fora do âmbito deste trabalho, podemos adiantar que essa análise só pode ser efectuada se conseguirmos estabelecer um plano de amostragem que atribua a cada amostra seleccionada uma determinada probabilidade, e esta atribuição só pode ser feita com planos de amostragem aleatórios. Assim, é importante termos sempre presente o princípio da aleatoriedade, quando vamos proceder a um estudo em que procuramos alargar para a População as propriedades estudadas na amostra. Numa secção posterior apresentaremos técnicas para obter amostras aleatórias.

doença, pelo que se seleccionaram 20 doentes padecendo dessa doença; h) Pretendia-se averiguar o nº de carros vendidos num dia por um stand de automóveis, pelo que se investigou junto de por cada um dos 5 empregados desse stand, quantos carros tinha vendido; i) Pretendia-se averiguar o número de leitores dos jornais diários, pelo que se investigou junto de 6 jornais diários, o número de leitores. j) Pretendia-se averiguar a percentagem de raparigas que frequentam o tronco comum de Matemática Aplicada da FCUL, pelo que se seleccionaram 50 alunos do dito curso.

Exercícios

População e Amostra

Identifique, no que se segue, População e Amostra:

a) Numa determinada empresa, pretende-se saber qual o salário médio dos seus empregados, pelo que se recolheu informação sobre os salários mensais, auferidos pelos empregados dessa empresa;

b) Prendia-se saber a nota média obtida na prova global de Matemática no ano lectivo 2000¬2001, dos alunos do 10º ano da Escola Secundária Prof. Herculano de Carvalho, pelo que se recolheu informação sobre as notas obtidas nessa disciplina por todos os alunos da Escola; c) Pretendia-se averiguar a idade média dos alunos do 10º ano da Escola Secundária Prof. Herculano de Carvalho, pelo que se recolheu informação sobre a idade de 45 alunos do 10º ano dessa Escola; d) Pretendia-se averiguar a quantidade de vinho produzida no Alentejo, no ano de 1999, pelo que se recolheu informação sobre as quantidades de vinho produzidas por 10 agricultores da região do Alentejo; e) Pretendia-se estudar o salário médio auferido pelos trabalhadores da indústria têxtil, pelo que se recolheu informação sobre os salários mensais auferidos por 250 desses trabalhadores; f) Pretendia-se averiguar a quantidade mensal de batata consumida nos lares portugueses, pelo que se recolheu informação sobre as quantidades de batata consumidas mensalmente em 100 lares portugueses; g) Pretendia-se estudar a eficácia de um medicamento novo para curar determinada

Parâmetro e Estatística

1. Diga se são verdadeiras ou falsas as seguintes afirmações:

a) Uma estatística é um número que se calcula a partir da amostra; b) Os parâmetros utilizam-se para estimar estatísticas; c) A média populacional é um parâmetro; d) Um parâmetro é uma característica numérica da variável que se está a estudar na População.

2. Identifique cada uma das quantidades seguintes, a negrito, como parâmetro ou estatística:

a) Nas últimas eleições para a Associação de Estudantes da Escola, 67% dos estudantes que votaram, fizeram-no na lista vencedora; b) Para obter uma estimativa do número de irmãos dos alunos que frequentam o 4.º ano de uma escola básica, perguntou-se a 30 alunos, escolhidos ao acaso, quantos irmãos tinham. Verificou¬se que em média, tinham 1.5 irmãos. c) Dos 230 deputados que compõem a VIII legislatura, 21.3% são mulheres. d) Perguntou-se a 80 deputados qual o partido que representavam, tendo-se concluído que 49% representavam o PS. e) Perguntou-se a 10 deputados qual a sua idade, tendo-se concluído que a idade média era de 45 anos.

Amostras enviesadas e amostras aleatórias

1. (Adaptado de Rossman, 2001) Considere a População constituída pelos deputados da VIII legislatura, que se encontra em anexo. Seleccione 5 deputados de que já tenha ouvido falar.

a) Estes deputados constituem uma amostra ou uma população? b) Quantos deputados, nos 5 seleccionados, pertencem ao círculo eleitoral da sua residência? c) Suponha que está interessada em estudar o n.º médio de anos de serviço dos deputados que constituem a VIII legislatura. Considera o conjunto de deputados seleccionados representativos da população? Porquê? d) Se calculasse a média dos anos de serviço dos deputados seleccionados esperava obter um valor superior ou inferior ao da média populacional? e) Se na sua aula ou outros colegas seleccionassem conjuntos de 5 deputados, pelo mesmo processo, isto é, deputados que lhe sejam familiares, espera que a média dos anos de serviço, tenha a mesma tendência, de sistematicamente exibir um enviesamento em determinado sentido? Explique. f) Se tivesse seleccionado pelo mesmo processo 10 deputados, obteria uma amostra mais representativa do que a constituída pelos 5 deputados? Explique.

1.2.2 - Experimentações

Enquanto que o objectivo de uma sondagem é o de recolher informação acerca de uma população seleccionando e observando uma amostra da população tal qual ela se apresenta, pelo contrário, uma experimentação impõe um tratamento às unidades experimentais com o fim de observar a resposta. O princípio base de uma experimentação é o método da comparação, em que se comparam os resultados obtidos na variável resposta de um grupo de tratamento com um grupo de controlo.

Exemplo 1.2.2.1 (Moore, 1997) – Será que a aspirina reduz o perigo de um ataque cardíaco? O estudo conhecido por Physicians' Health Study, foi uma experimentação médica levada a cabo com o objectivo de responder a esta questão específica.

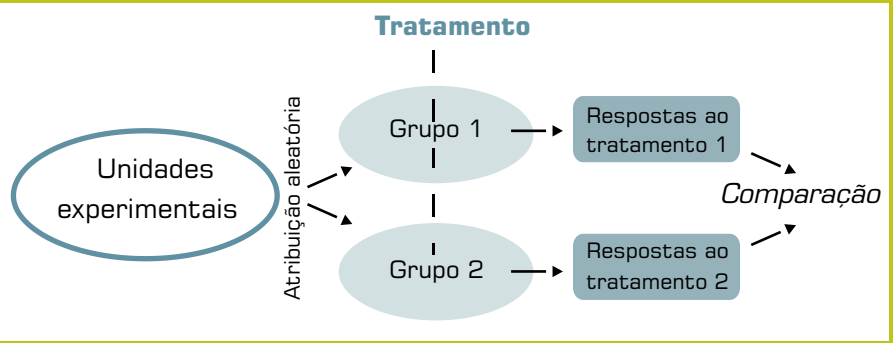
Metade de um grupo de 22000 médicos (homens) foram escolhidos aleatoriamente para tomar uma aspirina todos os dias. A outra metade dos médicos tomou um placebo, que tinha o mesmo aspecto e sabor da aspirina. Depois de vários anos 239 médicos do grupo que tomou placebo, contra 139 do grupo que tomou aspirina, tiveram ataques cardíacos. Esta diferença é suficientemente grande para evidenciar o efeito da aspirina na prevenção dos ataques cardíacos.

Unidades experimentais, tratamento, variável resposta, variáveis explanatórias.

Unidades experimentais são os objectos sobre os quais incide a experimentação e a quem é aplicado uma condição experimental específica, a que chamamos tratamento. Variável resposta é a variável cujo comportamento pretendemos estudar. As variáveis explanatórias são as variáveis que explicam ou causam mudanças na variável resposta.

No estudo considerado anteriormente temos:

- Unidades experimentais – 22000 médicos
- Tratamentos – aspirina ou placebo
- Variável explanatória – se o indivíduo tomou aspirina ou placebo
- Variável resposta – se o indivíduo teve ou não ataque cardíaco.



Sem a comparação de tratamentos os resultados de experimentações em medicina e em ciências do comportamento, duas áreas onde estes métodos são largamente utilizados, poderiam ser muito influenciados pela selecção dos indivíduos, o efeito do placebo, etc. O resultado poderia vir enviesado. Um estudo não controlado de uma nova terapia médica é quase sempre enviesado no sentido de dar ao tratamento um maior sucesso do que ele tem na realidade (efeito placebo).

Exemplo 1.2.2.2 (Moore, 1997) - Um tratamento utilizado durante vários anos para tratar úlceras do estômago consistia em pôr o doente a aspirar, durante uma hora, uma solução refrigerada que era bombeada para dentro de um balão. Segundo o Journal of the American Medical Association, uma experimentação levada a efeito com este tratamento permitiu concluir que o arrefecimento gástrico reduzia a secreção de ácido, diminuindo a propensão para as úlceras. No entanto, veio-se a verificar mais tarde com um planeamento adequado, que a resposta dos doentes ao tratamento foi influenciada pelo efeito placebo – efeito confounding. O que acontece é que há doentes que respondem favoravelmente a qualquer tratamento, mesmo que seja um placebo, possivelmente pela confiança que depositam no médico e pelas expectativas de cura que depositam no tratamento. Num planeamento adequado feito anos mais tarde, um grupo de doentes com úlcera foi dividido em dois grupos, tratando-se um com a solução refrigerada e o outro grupo com um placebo, constituído por uma solução à temperatura ambiente. Os resultados desta experimentação permitiram concluir que dos 82 doentes sujeitos à solução refrigerada - grupo de tratamento, 34% apresentaram melhoras, enquanto que dos 78 doentes que receberam o placebo - grupo de controlo, 38% apresentaram melhoras.

Num planeamento experimental, uma vez identificadas as variáveis e estabelecido o protocolo dos tratamentos, segue-se uma segunda fase que consiste na atribuição de cada unidade experimental a um tratamento. Esta segunda fase deve ser regida pelo princípio da aleatoriedade. Este princípio tem como objectivo fazer com que os grupos que vão ser comparados, tenham à partida constituição semelhante, de forma que as diferenças observadas na variável resposta possam ser atribuídas aos efeitos dos tratamentos. Assim, a atribuição de cada indivíduo ao grupo de tratamento ou de controlo é feita de forma aleatória. Combinando a comparação com a aleatoriedade, podemos esquematizar da seguinte forma o tipo de planeamento mais simples:

Ao comparar os resultados temos de ter presente que haverá sempre alguma diferença que se tem de atribuir ao facto de os grupos não serem perfeitamente idênticos e algumas diferenças que se atribuem ao acaso. O que se pretende é averiguar se as diferenças encontradas não serão “demasiado grandes” para que se possam atribuir a essas causas, ou seja, verificar se não tendo em linha de conta a diferença entre os tratamentos, a probabilidade de obter as diferenças observadas não seria extremamente pequena. Se efectivamente esta probabilidade for inferior a um determinado valor (de que falaremos mais tarde) dizemos que a diferença é estatisticamente significativa, sendo de admitir que foi provocada pelos diferentes tratamentos.

Convém ainda observar que numa experimentação os indivíduos seleccionados para cada grupo não devem saber qual o tipo de tratamento a que estão a ser sujeitos, nem o investigador que está a conduzir a experimentação e a medir os resultados deve saber qual o tipo de tratamento que cada indivíduo seguiu. Temos o que se chama uma experimentação duplamente cega. Esta precaução é uma forma de evitar o enviesamento, quer nas respostas, quer nas medições (um médico ao observar o efeito de um tratamento que provoque, por exemplo, uma mancha vermelha na pele, pode estar condicionado na interpretação da gravidade dessa mancha se souber qual o tratamento a que o doente foi sujeito).

Em muitas situações os investigadores têm de se cingir aos estudos observáveis, já que não é possível conduzir uma experimentação controlada. Por exemplo, para estudar o efeito do tabaco no cancro do pulmão, o investigador limita-se a observar grupos de indivíduos que fumam ou não, não podendo ser ele próprio a seleccionar um conjunto de indivíduos e depois pô-los aleatoriamente a fumar tabaco ou um placebo.

No capítulo seguinte abordaremos de forma introdutória o estudo de alguns planos de amostragem, já que um estudo conveniente do planeamento das experiências, assim como da definição da amostra adequada para o estudo em vista contêm, por si só, matéria suficiente para ser objecto de várias disciplinas num curso de Estatística, nomeadamente as disciplinas de Planeamento de Experiências e Amostragem.

1.3 - Técnicas de amostragem aleatória

Seguidamente apresentaremos alguns dos planeamentos mais utilizados para seleccionar amostras aleatórias. Dos vários tipos de planeamento utilizados, destacam-se os que conduzem a amostras aleatórias simples, amostras aleatórias com reposição, amostras sistemáticas e amostras estratificadas.

1.3.1 - Amostragem aleatória simples (sem reposição) e amostragem aleatória com reposição

O plano de amostragem aleatória mais básico é o que permite obter a amostra aleatória simples:

Amostra aleatória simples

Dada uma população, uma amostra aleatória simples de dimensão n é um conjunto de n unidades da população, tal que qualquer outro conjunto dos $\binom{N}{n}$ conjuntos diferentes de n unidades teria igual probabilidade de ser seleccionado.

Se uma população tem dimensão N e se pretende uma amostra aleatória simples de dimensão n, esta amostra é recolhida aleatoriamente de entre todas as

$$\binom{N}{n} = \frac{N!}{n!\,(N-n)!} = = \frac{N(N-1)(N-2)...(N-n+1)}{n(n-1)(n-2)...1}$$

amostras distintas que se podem recolher da população. Isto implica que cada amostra tenha a mesma probabilidade $\binom{N}{n}^{-}$ de ser seleccionada.

Uma amostra destas pode ser escolhida sequencialmente da população, escolhendo um elemento de cada vez, sem reposição, pelo que em cada

selecção cada elemento tem a mesma probabilidade de ser seleccionado. Um esquema de amostragem aleatória simples, conduz a que cada elemento da População tenha a mesma probabilidade de ser seleccionado para a amostra. No entanto existem outros esquemas de amostragem em que cada elemento tem igual probabilidade de ser seleccionado, sem que cada conjunto de n elementos tenha a mesma probabilidade de ser seleccionado. É o que se passa com a amostragem aleatória sistemática, de que falaremos adiante.

Amostragem com reposição

Na amostragem com reposição, sempre que um elemento é seleccionado, ele é reposto na população, antes de seleccionar o seguinte, ao contrário do que acontece na amostragem sem reposição. Intuitivamente conseguimos apercebermo-nos de que se a dimensão da população for "grande", quando comparada com a dimensão da amostra, estes dois tipos de amostragem podem ser considerados de certo modo equivalentes, já que a probabilidade de seleccionar o mesmo elemento duas vezes é "muito pequena".

Dada uma população de dimensão N, referir-nos-emos a uma amostra aleatória de dimensão n, com reposição, como um conjunto de n unidades da população, tal que qualquer outro conjunto dos N^n conjuntos diferentes de n unidades, teria igual probabilidade de ser seleccionado

A probabilidade de cada uma das amostras ser seleccionada é igual a 1/Nn.

Exemplificamos a seguir um processo de obter uma amostra aleatória simples.

Exemplo 1.3.1.1 – Consideremos a população constituída pelos 18 alunos de uma turma do 10.º ano de uma determinada Escola Secundária, em que a característica de interesse a estudar é a altura média desses alunos. Uma maneira possível de recolher desta população uma amostra aleatória, seria escrever cada um dos indicadores (n.º do aluno, nome, …) dos elementos da população num quadrado de papel, inserir todos esses bocados de papel numa caixa e depois seleccionar tantos quantos a dimensão da amostra desejada.

A recolha tem de ser feita sem reposição pois quando se retira um papel (elemento da população), ele não é reposto enquanto a amostra não estiver completa (com a dimensão desejada). Qualquer conjunto de números recolhidos desta forma dará origem a uma amostra aleatória simples, constituída pelas alturas dos alunos seleccionados (desde que se tenha o cuidado de cortar os bocadinhos de papel todos do mesmo tamanho, para ficarem semelhantes, e de os baralhar convenientemente). A partir de cada amostra, pode-se calcular o valor da estatística média, que será uma estimativa do parâmetro a estudar – valor médio da altura dos alunos da turma. Obter-se-ão tantas estimativas, quantas as amostras retiradas.

Chama-se a atenção para o facto de nesta altura não se poder dizer qual das estimativas é "melhor", isto é, qual delas é uma melhor aproximação do parâmetro a estimar, já que esse parâmetro é desconhecido (obviamente que nesta população tão pequena seria possível estudar exaustivamente todos os seus elementos, não sendo necessário recolher nenhuma amostra - este exemplo só serve para ilustrar uma situação)!

1.3.1.1 – Números aleatórios

O processo que acabámos de descrever não é prático se a população a estudar tiver dimensão elevada. Neste caso, um dos processos de seleccionar uma amostra aleatória simples consiste em utilizar uma tabela de números aleatórios.

Dígitos aleatórios

Uma tabela de dígitos aleatórios é uma listagem dos dígitos 0, 1, 2, 3, 4, 5, 6, 7, 8 ou 9 tal que:

- qualquer um dos destes dígitos tem igual possibilidade de figurar em qualquer posição da lista;

- a posição em que figura cada dígito é independente das posições dos outros dígitos.

Apresenta-se a seguir um extracto de uma tabela de números aleatórios (Moore, 1997). O facto de os dígitos se apresentarem agrupados 5 a 5 é só para facilidade de leitura.

Linha							
101	19223	95034	05756	28713	96409	12531	42544
102	73676	47150	99400	01927	27754	42648	82425
103	45467	71709	77558	00095	32863	29485	82226
104	52711	38889	93074	60227	40011	85848	48767
105	95592	94007	69971	91481	60779	53791	17297
106	68417	35013	15529	72765	85089	57067	50211
107	82739	57890	20807	47511	81676	55300	94383
108	60940	72024	17868	24943	61790	90656	87964
109	36009	19365	15412	39638	85453	46816	83485

A partir da tabela de dígitos aleatórios podem-se obter números aleatórios de 2 dígitos – qualquer par dos 100 pares possíveis 00, 01, ...98, 99, tem igual probabilidade de ser seleccionado, de 3 dígitos - qualquer triplo dos 1000 triplos possíveis 000, 001, ...998, 999, tem igual probabilidade de ser seleccionado, etc., tomando os dígitos da tabela 2 a 2, 3 a 3, etc., a partir de uma linha qualquer e percorrendo-a da esquerda para a direita.

Para seleccionar uma amostra de uma população utilizando a tabela procede-se em duas etapas:

- atribui-se um número a cada elemento da população. Esta atribuição terá de ser feita com as devidas precauções, de forma a que cada número tenha o mesmo número de dígitos, para ter igual probabilidade de ser seleccionado;
- a partir da tabela escolhe-se uma linha ao acaso e começa-se a percorrê-la da esquerda para a direita, tomando de cada vez os dígitos necessários.

Exemplo 1.3.1.1 (cont) - Considerando a população do exemplo anterior, constituída por 18 elementos, vamos numerá-los com os números 01, 02, 03, ..., 17, 18 (podia ser utilizado qualquer outro conjunto de 18 números de 2 dígitos). Para seleccionar uma amostra de dimensão 4 fixamo--nos numa linha qualquer da tabela, por exemplo a linha 107 e começamos a seleccionar os números de dois dígitos, tendo-se obtido:

82	73	95	78	90	20	80	74	75	11	81
67	65	53	00	94	38	31	48	93	60	94
07	20	24	17	86	82	49	43	61	79	09

Tivemos de ler 33 números, dos quais só aproveitámos 4, pois os outros não correspondiam a elementos da população.

Como obter uma tabela de números aleatórios?

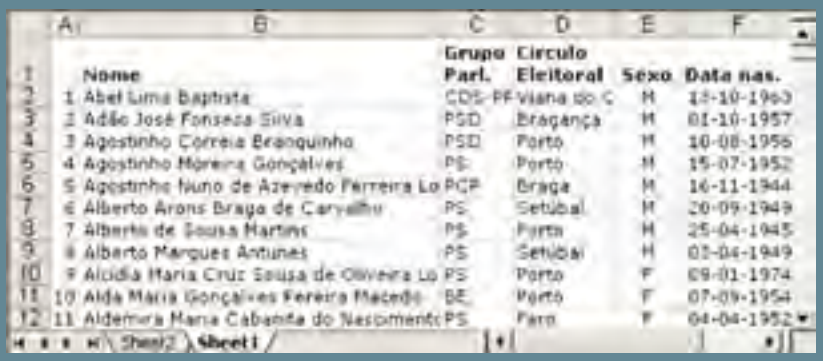
Um processo poderá consistir em meter numa caixa 10 bolas numeradas de 0 a 9 e fazer várias extracções de uma bola, tantas quantas os dígitos que se pretendem para constituir a tabela. De cada vez que se faz uma extracção, lê-se o número da bola, aponta-se e repõe-se a bola na caixa - extracção com reposição. Com este processo qualquer dígito tem igual probabilidade de ser seleccionado. Além disso a saída de qualquer um dos dígitos em qualquer momento, é independente dos dígitos que já saíram anteriormente.

Além das tabelas de números aleatórios também existe a possibilidade de utilizar o computador para os gerar ou uma simples máquina de calcular. Este é o processo mais utilizado hoje em dia, mas convém ter presente que os números que se obtêm são pseudo-aleatórios, já que é um mecanismo determinista que lhes dá origem, embora se comportem como números aleatórios (passam numa bateria de testes destinados a confirmar a sua aleatoriedade). No exemplo seguinte vamos utilizar o computador, mais precisamente o programa Excel, para fazer a selecção de uma amostra aleatória simples e de uma amostra aleatória com reposição.

1.3.1.2 - Utilização do Excel para recolher uma amostra aleatória simples e uma amostra aleatória com reposição

No exemplo seguinte, apresentamos uma forma simples de utilizar o Excel para seleccionar uma amostra aleatória simples e uma amostra aleatória, com reposição, de uma População finita, de que se tenha uma listagem dos elementos.

Exemplo 1.3.1.2 – Considere a população constituída pelos 230 deputados da actual (X) legislatura e que se encontra em Anexo. Para obter esta tabela fomos ao "site" da Assembleia da Republica, onde está uma lista ordenada com o nome de todos os deputados (coluna B), o respectivo grupo parlamentar (coluna C) e o círculo eleitoral (coluna D). Este exemplo vai-nos servir para introduzir alguns conceitos importantes, pelo que fomos completar esta lista com a idade dos deputados, acedendo à página de cada um e recolhendo a informação sobre a data de nascimento (coluna F). Nas situações de interesse, que surgem na vida real, não se vai recolher a informação sobre determinada característica, para a população toda, mas unicamente para os elementos seleccionados para a amostra. Inserimos ainda uma coluna com identificação do sexo (coluna E). Apresentamos a seguir uma pequena parcela desse ficheiro, a que chamámos Deputados.xls. Este ficheiro tem uma primeira coluna (coluna A), onde é indicado o número do deputado, quando estes estão ordenados por ordem alfabética:



	A	B	C	D	E	F
1		Nome	Grupo Parl.	Circulo Eleitoral	Sexo	Data nas.
2	1	Abel Lima Baptista	CDS-PP	Viana do C	M	13-10-1963
3	2	Adão José Fonseca Silva	PSD	Bragança	M	01-10-1957
4	3	Agostinho Correia Branquinho	PSD	Porto	M	10-08-1956
5	4	Agostinho Moreira Gonçalves	PS	Porto	M	15-07-1952
6	5	Agostinho Nuno de Azevedo Ferreira Lo	PCP	Braga	M	16-11-1944
7	6	Alberto Arons Braga de Carvalho	PS	Setúbal	M	20-09-1949
8	7	Alberto de Sousa Martins	PS	Porto	M	25-04-1945
9	8	Alberto Marques Antunes	PS	Setúbal	M	03-04-1949
10	9	Alcídia Maria Cruz Sousa de Oliveira Lo	PS	Porto	F	09-01-1974
11	10	Alda Maria Gonçalves Pereira Macedo	BE	Porto	F	07-09-1954
12	11	Aldemira Maria Cabanita do Nascimento	PS	Faro	F	04-04-1952

Sheet2 Sheet1

Como dissemos anteriormente, vamos utilizá-lo para trabalhar alguns conceitos importantes, tais como:

1. Obtenção de uma amostra aleatória simples e de uma amostra aleatória, com reposição, utilizando o Excel
2. Estatística e parâmetro
3. Variabilidade amostral
4. Precisão

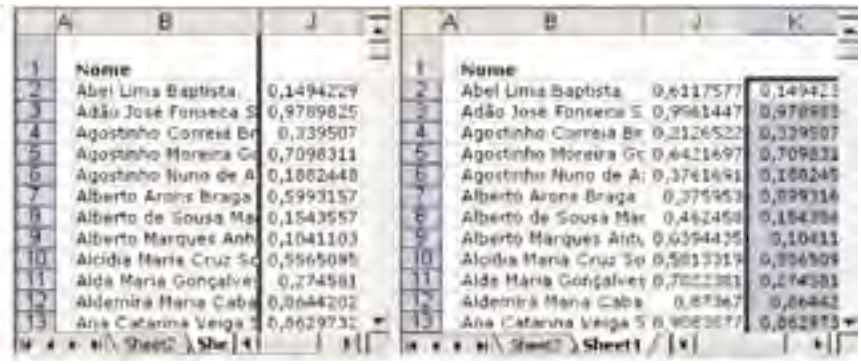
1. Obtenção de uma amostra aleatória simples e de uma amostra aleatória, com reposição, utilizando o Excel

Amostra aleatória simples

1º passo - Utilizando a função RAND(), atribuir um número aleatório, entre 0 e 1, a cada deputado. Para isso basta inserir a função na célula J2 e replicá-la tantas vezes, quantos os deputados (ou seja, 230 vezes):

Tools - Options - View - Formulas

Ok: Uma vez que a função RAND() é uma função volátil, isto é, muda quando se recalcula a folha, no caso de pretendermos ficar com os valores gerados convém ir ao Edit e fazer um Paste Special - Values, como se indica a seguir:

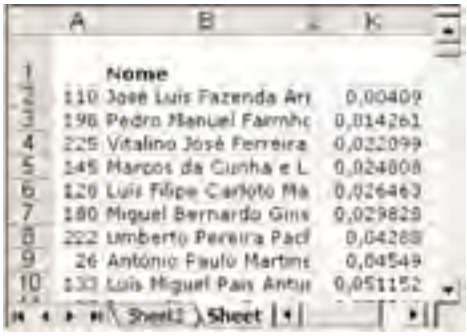


Colámos os valores na coluna K e fizemos o Save. Repare-se que os valores que estavam inicialmente na coluna J foram alterados, dando origem a novos valores (devido ao facto da função RAND() ser volátil, como referimos anteriormente);

2º passo – Ordenar o ficheiro, utilizando como critério a coluna K;

3º passo – Como pretendemos uma amostra de dimensão 10, seleccionar os primeiros 10 deputados do ficheiro ordenado:

Os deputados seleccionados foram os números 110, 198, 225, 145, 128, 180, 222, 26 e 133.

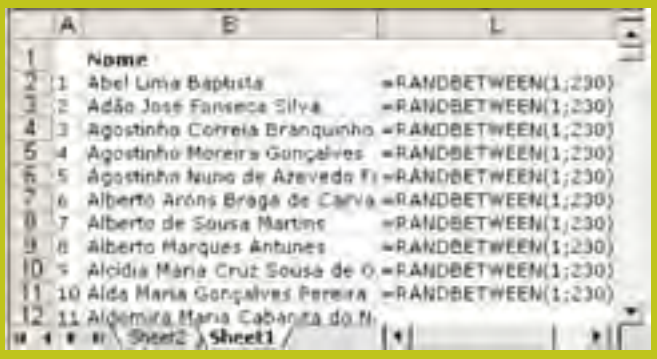


Nota: Embora os números anteriores sejam referidos como aleatórios, convém ter presente que os números que se obtêm são pseudo-aleatórios, já que é um mecanismo determinista que lhes dá origem. No entanto comportam-se como números aleatórios (passam uma bateria de testes destinados a confirmar a sua aleatoriedade) e daí a sua utilização como tal.

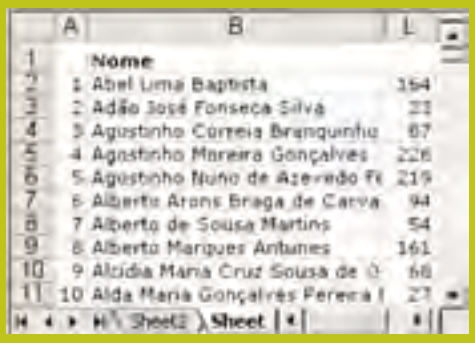
Amostra aleatória com reposição

a) Utilize a função RANDBETWEEN(), para obter números pseudo-aleatórios entre 1 e 230, para simular a extracção de uma amostra aleatória, da população dos deputados.

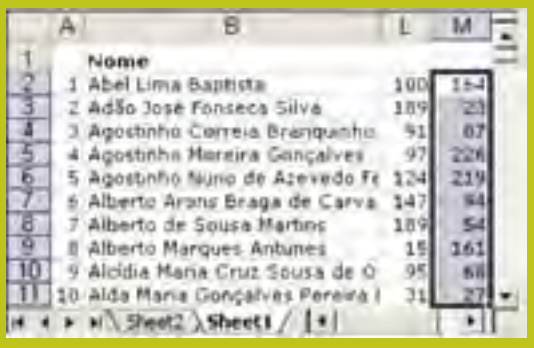
Esta função devolve um número pseudo-aleatório entre os limites especificados nos argumentos. Como pretendemos seleccionar uma amostra de dimensão 10, replicamos a fórmula anterior por 10 células, na coluna L, como se apresenta a seguir:



A amostra obtida é constituída pelos deputados com os 10 números nas células L2, …, L11:



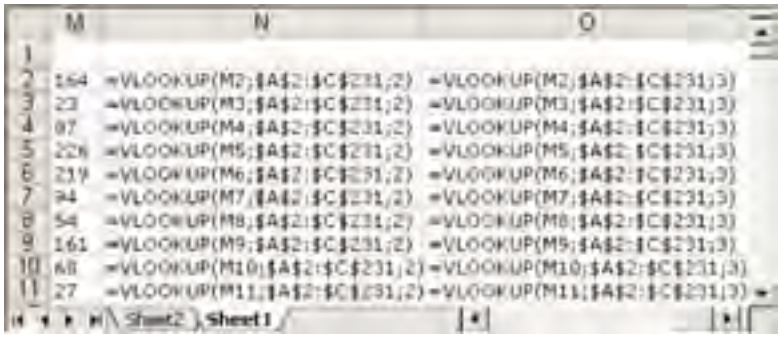
Uma vez que a função RANDBETWEEN é uma função volátil, isto é, muda quando se recalcula a folha, para ficar com os valores gerados fomos ao ***Edit - Paste Special - Values***, como se indica a seguir:



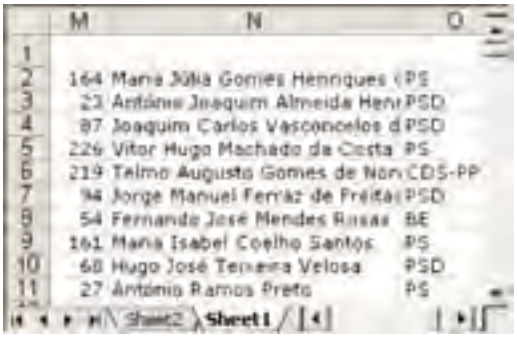
Colámos os valores na coluna M e fizemos o Save. Repare-se que os valores que estavam inicialmente na coluna L foram alterados, dando origem a uma nova amostra (devido ao facto da função RANDBETWEEN ser volátil, como referimos anteriormente):

b) Da tabela dos deputados, seleccione o nome e o grupo parlamentar dos deputados cujo número seja um dos elementos da amostra obtida anteriormente.

Para seleccionar o nome e o grupo parlamentar dos deputados correspondentes aos 10 números obtidos, vamos utilizar uma função do Excel, a função VLOOKUP, do seguinte modo:



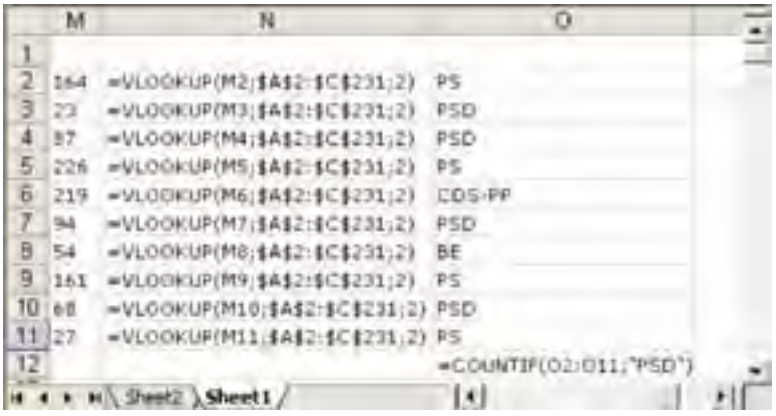
Esta função vai à tabela dos deputados, constituída pelas células (A2:C231) seleccionar o nome (2ª coluna da tabela seleccionada) e o Grupo Parlamentar (3ª coluna da tabela seleccionada) correspondente ao número que está na coluna M, obtendo-se a seguinte amostra:



2. Parâmetro e Estatística.

c) Calcule a percentagem de deputados do grupo parlamentar PSD, na amostra obtida.

Vamos começar por utilizar a função COUNTIF, que inserimos na célula O12, e que conta o nº de células, de entre um conjunto especificado de células, que satisfazem determinado critério, sendo este critério, no caso presente, o de serem iguais a "PSD":



Obtivemos o valor 4 para a frequência absoluta de deputados do PSD. Como o nº de deputados da amostra era 10, a percentagem de deputados do grupo parlamentar do PSD, na amostra é de 40%. Este valor é uma estatística – característica numérica da amostra. Utiliza-se como estimativa do parâmetro "percentagem de deputados do PSD na população em estudo" – característica numérica da população.

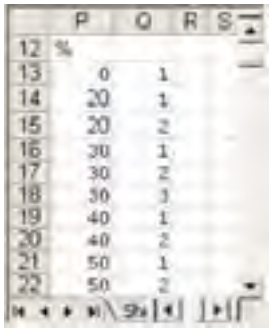
3. Variabilidade amostral

d) Repita 10 vezes o processo descrito nas alíneas anteriores e registe numa tabela os resultados obtidos.

Gerámos 10 amostras e obtivemos os seguintes resultados para a estatística - percentagem de deputados PSD, em cada uma das amostras:

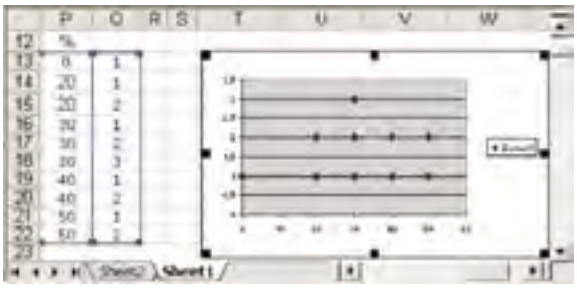
Amostra	% PSD
1	40%
2	20%
3	30%
4	50%
5	20%
6	30%
7	40%
8	50%
9	0%
10	30%

Repare-se na variabilidade apresentada nos resultados obtidos para as diferentes amostras. Os 10 valores obtidos para a percentagem de deputados do PSD existentes em cada uma delas, representam outras tantas estimativas para a verdadeira proporção de deputados existentes na População. Iremos ilustrar esta variabilidade, representando os valores num diagrama de pontos, utilizando uma opção gráfica do Excel, o Scatter. Para obter a representação gráfica pretendida, é necessário começar por construir uma tabela adequada:



Para construir esta tabela, pode-se utilizar a seguinte metodologia: consideram-se duas colunas, onde na primeira coluna se representam todos os elementos do conjunto de dados, pela ordem em que aparecem, e na segunda coluna indica-se a frequência absoluta com que cada elemento surge no conjunto de dados, à medida que se vai percorrendo a coluna, de cima para baixo. Por exemplo, ao lado do primeiro elemento que é o 60%, indicamos um 1, mas a segunda vez que aparece o 60%, indicamos um 2, etc. Se, à partida, dispuséssemos de uma tabela de frequências, para construir esta nova tabela, bastaria repetir cada elemento da amostra, tantas vezes quantas a sua frequência absoluta.

Na folha do Excel, seleccionam-se as duas colunas e no menu ***Chart*** selecciona-se ***Scatter*** e o primeiro subtipo desta representação. Obtém-se o diagrama de pontos com o seguinte aspecto:

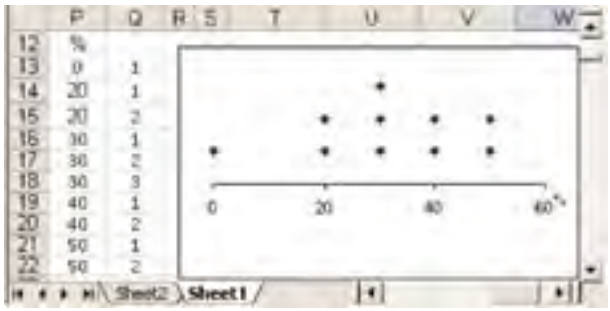


Trabalhámos "esteticamente" esta representação, seguindo os seguintes passos:

Seleccionar:

- Legenda e carregar no botão Delete;
- As linhas e carregar no botão Delete;
- O fundo cinzento e carregar no botão Delete;
- O eixo dos YY e carregar no botão Delete;

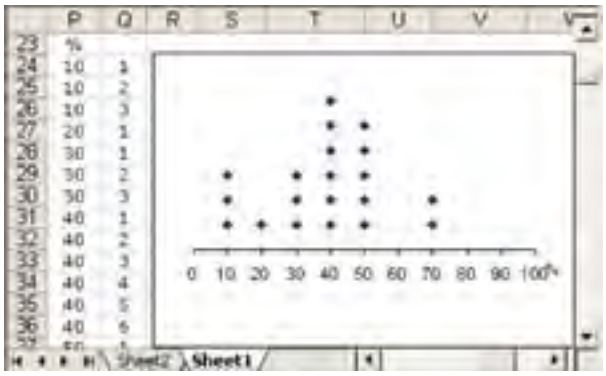
Temos finalmente a seguinte representação:



Da representação gráfica anterior começamos a adivinhar que a distribuição das estimativas apresenta um padrão com uma certa simetria relativamente ao valor de 30%.

e) Considere agora 20 amostras de dimensão 10, calcule para cada uma o valor da estatística em estudo, e construa o diagrama de pontos dos valores obtidos.

Seleccionámos 20 amostras de dimensão 10, calculámos a percentagem de deputados do PSD em cada uma delas e com os resultados obtidos construímos a seguinte representação:

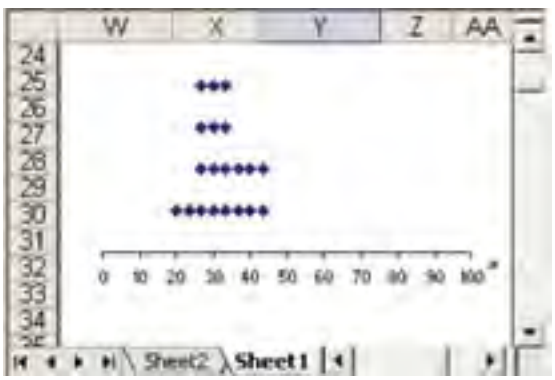


Esta representação é mais elucidativa e reforça a ideia avançada anteriormente, de que o valor do parâmetro em estudo – percentagem de deputados do PSD, se deve situar entre os valores 30% e 40%. Tendo em conta que a verdadeira percentagem de deputados do PSD na população é 32,6%, apesar de o valor apresentado pela estatística variar de amostra para amostra – variabilidade amostral, estes valores apresentam uma distribuição que nos dá informação sobre o parâmetro, já que essa distribuição se localiza ou está centrada em torno do parâmetro.

4. Precisão

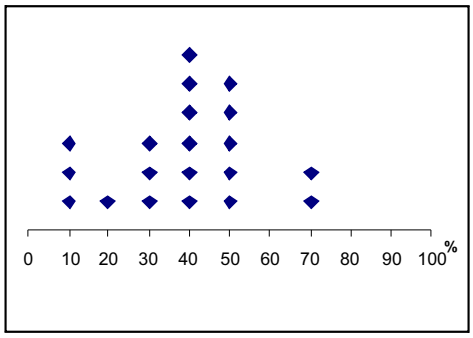
f) Considere agora 20 amostras de dimensão 30, calcule para cada uma o valor da estatística em estudo, e construa o diagrama de pontos dos valores obtidos. Compare a representação obtida, com a que obteve na alínea e).

Seguimos um processo idêntico ao da alínea e), com a particularidade de as dimensões das amostras terem dimensão 30, em vez de 10. Com as percentagens de deputados do PSD existentes em cada uma delas construímos a seguinte representação gráfica:

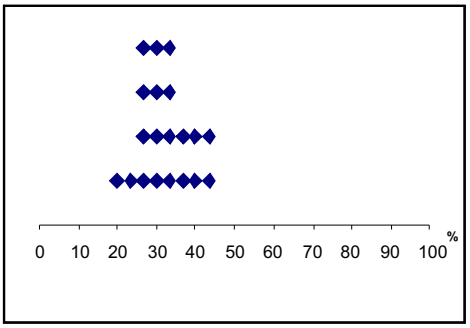


Comparando as duas representações obtidas quando se consideram amostras de dimensão 10 ou de dimensão 30, verificamos que a variabilidade apresentada pelos valores da estatística - percentagem de deputados do PSD, no caso das amostras de maior dimensão, é inferior à apresentada pela estatística no caso das amostras de menor dimensão, como se vê na figura seguinte:

Percentagem de deputados do PSD em amostras de dimensão 10



Percentagem de deputados do PSD em amostras de dimensão 30



A precisão de um estimador é caracterizada pela variabilidade apresentada pelas diferentes estimativas, obtidas quando se consideram diferentes amostras. Quanto menor for a variabilidade apresentada pelas diferentes estimativas, maior é a precisão apresentada pelo estimador.

De um modo geral, diz-se que uma estatística é um "bom" estimador de um certo parâmetro, se a distribuição dos valores dessa estatística, calculados para diversas amostras, revelar uma localização em torno do parâmetro e apresentar pequena variabilidade. Em alguns casos essa análise pode fazer-se do ponto de vista teórico. No entanto, hoje em dia, cada vez se recorre mais à simulação para decidir se um estimador é ou não, um "bom" estimador do parâmetro de interesse.

Observação: Este exemplo que acabámos de apresentar tem como objectivo apresentar alguns conceitos importantes, como o da variabilidade e das propriedades de um estimador. Efectivamente, neste caso, já que temos informação sobre o grupo parlamentar de cada deputado, não teria muito sentido ir recolher uma amostra para obter a percentagem de deputados em cada grupo parlamentar. Repare-se, no entanto, que se o que estivesse em estudo fosse "ter uma ideia" sobre o número médio de filhos dos deputados portugueses e suas idades, já faria sentido recolher uma amostra, pois para obter a informação desejada não seria necessário interrogar todos os deputados e só se interrogariam os seleccionados para a amostra.

1.3.2 - Amostragem aleatória sistemática

Na prática o processo de seleccionar uma amostra aleatória simples de uma população com grande dimensão, não é tão simples como o descrito anteriormente. Se a dimensão da população for grande o processo torna-se muito trabalhoso. Então uma alternativa é considerar uma amostra aleatória sistemática – os elementos são escolhidos de uma maneira regular percorrendo a lista.

Amostra aleatória sistemática

Dada uma população de dimensão N, ordenada por algum critério, se se pretende uma amostra de dimensão n, escolhe-se aleatoriamente um elemento de entre os k primeiros, onde k é a parte inteira do quociente N/n. A partir desse elemento escolhido, escolhem-se todos os k-ésimos elementos da população para pertencerem à amostra.

A amostra aleatória sistemática não é uma amostra aleatória simples, já que nem todas as amostras possíveis de dimensão n, têm a mesma probabilidade de serem seleccionadas.

1.3.2.1 - Utilização do Excel para recolher uma amostra aleatória sistemática

No exemplo seguinte, apresentamos uma forma simples de utilizar o Excel para seleccionar uma amostra aleatória sistemática de uma População finita, de que se tenha uma listagem dos elementos.

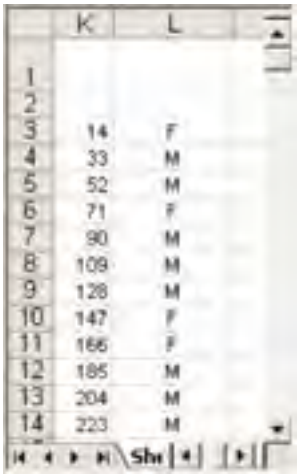
Exemplo 1.3.2.1 – Considere novamente o ficheiro Deputados.xls, que contém o nome, filiação partidária, sexo e data de nascimento dos 230 deputados da actual legislatura e que se encontra em Anexo. Utilizando o processo de amostragem sistemática, obtenha uma amostra de 12 deputados, registando para cada um deles o sexo. Temos uma população de dimensão 230 e pretendemos obter uma amostra de dimensão 12. Vamos utilizar a seguinte metodologia:

Passo 1 – Dividindo 230 por 12 e retendo a parte inteira, obtemos o valor 19.

Passo 2 – Dos primeiros 19 elementos da lista ordenada dos deputados, vamos seleccionar um elemento ao acaso. Vimos na secção anterior que basta utilizar a função Randbetween(1;19), que inserimos na célula K3. A utilização desta função devolveu-nos o deputado número 14.

Passo 3 – A amostra será constituída pelos deputados números 14, 33, 52, 71, 90, 109, 128, 147, 166, 185, 204, 223, que obtivemos adicionando sucessivamente 19, até obtermos 12 elementos (células K3:K14).

Passo 4 -Utilizando a função VLOOKUP(K3;$A$3:$E$232;5), replicada pelas 12 células L3:L14, obteve-se finalmente a informação solicitada, constituída pelo sexo dos 12 deputados seleccionados para a amostra:



1.3.3 – Amostragem estratificada

Pode acontecer que a população possa ser dividida em várias subpopulações ou estratos, mais ou menos homogéneos, relativamente à característica a estudar. Nesta situação existe uma técnica importante e apropriada, que é a amostragem por estratificação. Apresentamos de seguida um exemplo em que privilegiaremos a exemplificação da técnica, em detrimento da apresentação em Excel, uma vez que o tipo de amostragem utilizado, se resume a uma amostragem aleatória simples, já exemplificada anteriormente.

Exemplo 1.3.3.1 (Ted Hodgson and John Borkowski in Getting the Best from Teaching Statistics)

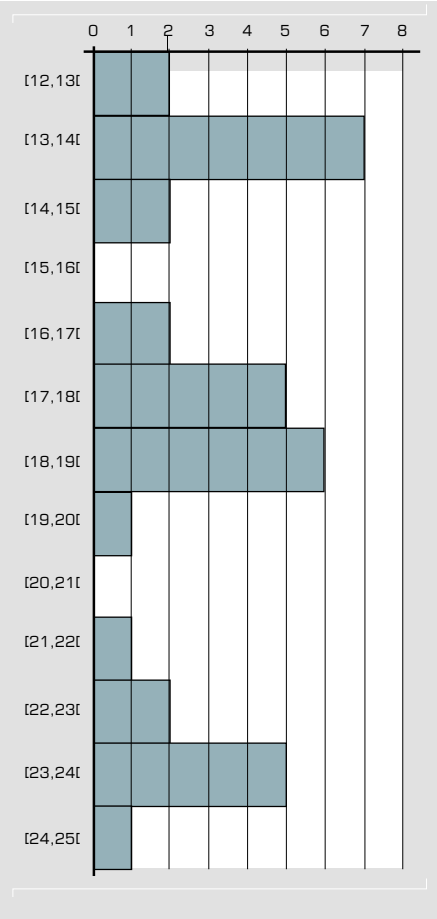
– Consideremos uma população constituída por 40 cartões numerados (20 vermelhos e 20 pretos) de acordo com a seguinte tabela:

Nº	Freq.	Cor
6	4	Ver
7	4	Ver
8	4	Ver
9	4	Ver
10	4	Ver
26	4	Preto
27	4	Preto
28	4	Preto
29	4	Preto
30	4	Preto

A média dos números inscritos nesta população de 40 cartões é de 18 – valor médio da característica populacional em estudo.

Pretende-se, através de uma amostra, obter alguma indicação sobre a média dos números inscritos nos cartões (a qual neste exemplo fictício é conhecida). Colocam-se os cartões num saco e pede-se a cada aluno da turma que retire uma amostra de 4 cartões – amostra aleatória simples, e que calcule a média dos números dos cartões seleccionados. Numa turma de 34 alunos, obtiveram-se os seguintes resultados:

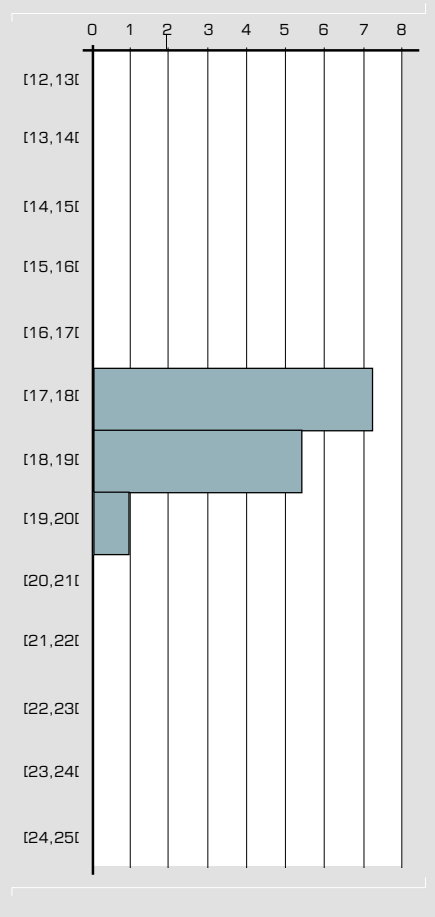
Amostra nº					Média
1	26	7	10	6	12,25
2	10	26	9	6	12,75
3	29	6	7	10	13
4	6	8	9	29	13
5	6	9	8	30	13,25
6	9	8	7	29	13,25
7	7	7	30	9	13,25
8	9	9	10	26	13,5
9	9	8	8	30	13,75
10	9	10	8	29	14
11	10	9	29	9	14,25
12	6	27	6	26	16,25
13	7	7	26	27	16,75
14	28	8	6	26	17
15	7	6	29	26	17
16	6	29	26	8	17,25
17	9	6	26	29	17,5
18	26	9	8	28	17,75
19	7	10	26	29	18
20	27	6	30	9	18
21	6	29	28	10	18,25
22	8	29	26	10	18,25
23	6	8	30	30	18,5
24	26	9	30	10	18,75
25	8	11	28	30	19,25
26	26	27	6	27	21,5
27	30	26	27	6	22,25
28	8	26	29	28	22,75
29	10	26	26	30	23
30	29	6	30	27	23
31	28	9	30	26	23,25
32	27	26	30	10	23,25
33	30	10	29	26	23,75
34	29	30	7	30	24



Esta distribuição não nos ajuda muito a dizer qual a estimativa para o valor médio da população (média dos números inscritos). Gostaríamos de ter obtido para a amostra, cujos elementos são as diferentes médias, uma distribuição com pouca variabilidade, para podermos argumentar que a média destes elementos era uma "boa" estimativa para o parâmetro em estudo, ou seja, o valor médio dos números inscritos nos cartões (Ver secção seguinte).

Diz-se então aos alunos que estamos perante duas subpopulações, a de cartões vermelhos e a de cartões pretos, embora não seja esta a característica em estudo e sobre a qual seria importante haver diferença entre os estratos ou subpopulações. De qualquer modo aqueles são informados que poderá haver diferenças relativamente à característica de interesse e que um processo de amostragem adequado levaria em conta essas diferenças.

Amostra nº					Média
1	6	7	27	28	17
2	8	9	26	27	17,5
3	8	6	28	28	17,5
4	7	8	29	26	17,5
5	9	9	26	26	17,5
6	6	9	29	27	17,75
7	8	10	26	27	17,75
8	10	6	27	28	17,75
9	9	9	28	26	18
10	6	8	28	30	18
11	10	8	27	28	18,25
12	10	7	28	29	18,5
13	9	9	27	29	18,5
14	8	9	29	29	18,75
15	9	10	28	29	19



Procede-se então a uma selecção da amostra, de forma a obter 2 cartões vermelhos e 2 cartões pretos – estes valores devem reflectir a dimensão dos estratos (que no nosso exemplo são iguais). Os resultados obtidos foram os seguintes:

A partir dos dados obtidos para as amostras, confirma-se que efectivamente temos dois estratos distintos, relativamente à característica de interesse – um estrato com cartões com números mais pequenos e outro estrato com cartões com números maiores.

Estes resultados mostram que as médias das amostras estratificadas estão consistentemente próximas do valor médio da população (o qual só deve ser dito aos alunos depois das simulações serem feitas), podendo-se assim observar que a estratificação conduziu a um aumento da precisão.

1.3.4 – Estimador centrado e não centrado. Precisão

Uma vez escolhido um plano de amostragem aleatório, ao pretendermos estimar um parâmetro, pode ser possível utilizar várias estatísticas (estimadores) diferentes. Por exemplo, quando pretendemos estudar a variabilidade presente numa População, que pode ser medida pela variância populacional σ^2, sabemos que podemos a partir de uma amostra, obter duas estimativas diferentes para essa variância, a partir das expressões

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \quad \text{ou} \quad s'^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

Quais as razões que nos podem levar a preferir s^2 em vez de s'^2 ?

Um critério que costuma ser aplicado é o de escolher um "bom" estimador como sendo aquele que é centrado e que tem uma boa precisão. Escolhido um plano de amostragem, define-se:

Estimador centrado

Um estimador diz-se centrado quando a média das estimativas obtidas para todas as amostras possíveis que se podem extrair da População, segundo o esquema considerado, coincide com o parâmetro a estimar. Quando se tem um estimador centrado, também se diz que é não enviesado.

A média das estimativas calculadas a partir da expressão s^2 acima considerada, coincide com a variância.

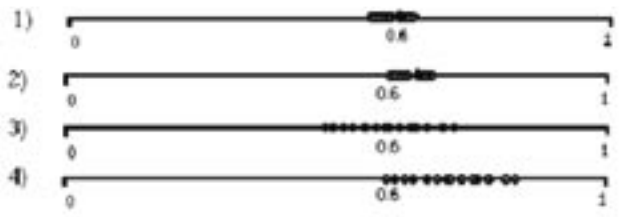
Para se evitar o enviesamento, é necessário estarmos atentos, primeiro na escolha do plano de amostragem e depois na escolha do estimador utilizado para estimar o parâmetro desconhecido. O facto de utilizarmos um estimador centrado, não nos previne contra a obtenção de más estimativas, se o plano de amostragem utilizado sistematicamente favorecer uma parte da População (isto é, fornecer amostras enviesadas).

Precisão

Ao utilizar o valor de uma estatística para estimar um parâmetro, vimos que cada amostra fornece um valor para a estatística que se utiliza como estimativa desse parâmetro. Estas estimativas não são iguais devido à variabilidade presente na amostra. Se, no entanto, os diferentes valores obtidos para a estatística forem próximos, e o estimador for centrado, podemos ter confiança de que o valor calculado a partir da amostra recolhida (na prática recolhe-se uma única amostra) está próximo do valor do parâmetro (desconhecido).

A falta de precisão juntamente com o problema do enviesamento da amostra são dois tipos de erro com que nos defrontamos num processo de amostragem (mesmo que tenhamos escolhido um "bom" estimador). Não se devem, contudo, confundir. Enquanto o enviesamento se manifesta por um desvio nos valores da estatística, relativamente ao valor do parâmetro a estimar, sempre no mesmo sentido, a falta de precisão manifesta-se por uma grande variabilidade nos valores da estatística, uns relativamente aos outros. Por outro lado, enquanto o enviesamento se reduz com o recurso a amostras aleatórias, a precisão aumenta-se aumentando a dimensão da amostra.

Exemplo 1.3.4.1 - Suponhamos que ao pretender estudar a percentagem de eleitores que votariam favoravelmente num candidato à Câmara de determinada cidade, se recolhia uma amostra de 300 eleitores, dos quais 175 responderam que sim. Considerando como estimador, a proporção de elementos na amostra apoiantes do candidato, então uma estimativa para a proporção pretendida seria 0.58. Se considerássemos outra amostra de 300 eleitores, suponhamos que o valor obtido para o número de sim's tinha sido 183. Então a estimativa obtida seria 0.61. A repetição deste processo 15 vezes permitiria obter 15 valores para o estimador, que seriam outras tantas estimativas do parâmetro a estimar -percentagem de eleitores da cidade, potenciais apoiantes do tal candidato. Representando num eixo os valores obtidos e admitindo que o verdadeiro valor do parâmetro era 0.60, poderíamos deparar-nos com várias situações:



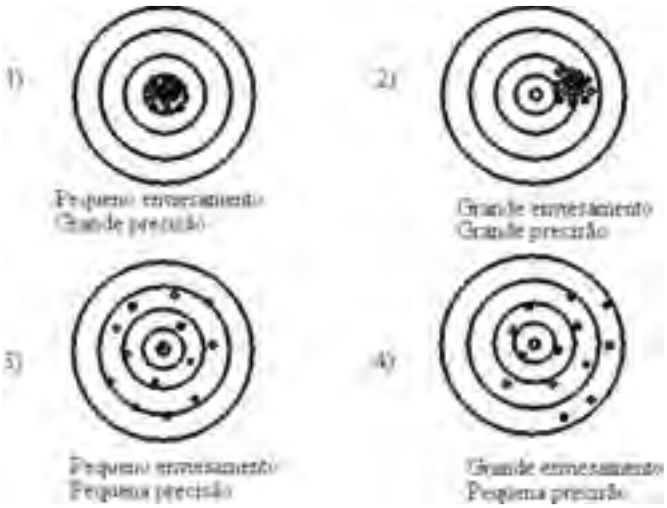
1) reflecte um pequeno ou ausência de enviesamento, pois os valores para a estatística (proporções obtidas a partir das amostras) situam-se para um e outro lado do valor do parâmetro, e verifica-se ainda a existência de uma pequena variabilidade entre os resultados obtidos para as várias amostras, que se traduz em grande precisão.

2) embora se mantenha a precisão, existe um grande enviesamento, pois os valores da estatística situam-se sistematicamente para a direita do valor do parâmetro. Presume-se que o esquema de amostragem não seja aleatório, pelo que as amostras só reflectem parte da População.

3) voltamos a ter uma situação de pequeno enviesamento, mas de pequena precisão devido à grande variabilidade apresentada pelos valores da estatística. Presumimos que as amostras não têm a dimensão suficiente, de forma a garantir uma melhor precisão.

4) a falta de precisão da situação 3) é acompanhada de um grande enviesamento.

Como sugere Moore (1996), fazendo analogia com o que se passa com um atirador que aponta várias setas a um alvo, em que procurava atingir o centro do alvo, teríamos



O estudo de um estimador é feito através da sua distribuição de amostragem, ou seja, da distribuição dos valores obtidos pelo estimador, quando se consideram todas as amostras possíveis.

Distribuição de amostragem

Distribuição de amostragem de uma estatística é a distribuição dos valores que a estatística assume para todas as possíveis amostras, da mesma dimensão, da população.

A forma da distribuição de amostragem, permite-nos verificar se esses valores se distribuem de forma tal, que a sua média coincide com o parâmetro a estimar – caso em que o estimador é centrado, e além disso se apresenta grande ou pequena variabilidade – o que faz com que o estimador apresente, respectivamente, menor ou maior precisão.

A maior parte das vezes não se consegue obter a distribuição de amostragem exacta, mas tem-se uma distribuição aproximada, considerando um número suficientemente grande de amostras da mesma dimensão e calculando para cada uma delas uma estimativa do parâmetro em estudo.

1.3.5 - Qual a dimensão que se deve considerar para a amostra?

Outro problema que se levanta com a recolha da amostra é o de saber qual a dimensão desejada para a amostra a recolher. Este é um problema para o qual, nesta fase, não é possível avançar nenhuma teoria, mas sobre o qual se podem tecer algumas considerações gerais. Pode¬se começar por dizer que, para se obter uma amostra que permita calcular estimativas suficientemente precisas dos parâmetros a estudar, a sua dimensão depende muito da variabilidade da população subjacente. Por exemplo, se relativamente à população constituída pelos alunos do 10º ano de uma escola secundária, estivermos interessados em estudar a sua idade média, a dimensão da amostra a recolher não necessita de ser muito grande já que a variável idade apresenta valores muito semelhantes, numa classe etária muito restrita. No entanto se a

característica a estudar for o tempo médio que os alunos levam a chegar de casa à escola, de forma a obter a mesma precisão que no caso anterior, já a amostra terá de ter uma dimensão maior, uma vez que a variabilidade da população é muito maior. Cada aluno pode apresentar um valor diferente para esse tempo. Num caso extremo, se numa população a variável a estudar tiver o mesmo valor para todos os elementos, então bastaria recolher uma amostra de dimensão 1 para se ter informação completa sobre a população; se, no entanto, a variável assumir valores diferentes para todos os elementos, para se ter o mesmo tipo de informação seria necessário investigar todos os elementos.

Chama-se a atenção para a existência de técnicas que permitem obter valores mínimos para as dimensões das amostras a recolher e que garantem estimativas com uma determinada precisão exigida à partida. Uma vez garantida essa precisão, a opção por escolher uma amostra de maior dimensão, é uma questão a ponderar entre os custos envolvidos e o ganho com o acréscimo de precisão. Vem a propósito a seguinte frase (Statistics: a Tool for the Social Sciences, Mendenhall et al., pag. 226):

"Se a dimensão da amostra é demasiado grande, desperdiça-se tempo e talento; se a dimensão da amostra é demasiado pequena, desperdiça-se tempo e talento".

Convém ainda observar que a dimensão da amostra a recolher não é directamente proporcional à dimensão da população a estudar, isto é, se por exemplo para uma população de dimensão 1000 uma amostra de dimensão 100 for suficiente para o estudo de determinada característica, não se exige necessariamente uma amostra de dimensão 200 para estudar a mesma característica de uma população análoga, mas de dimensão 2000, quando se pretende obter a mesma precisão. Como explicava George Gallup, um dos pais da consulta da opinião pública (Tannenbaum, 1998),: Whether you poll the United States or New York State or Baton Rouge (Louisiana) … you need … the same number of interviews or samples. It´s no mystery really – if a cook has two pots of soup on the stove, one far larger than the other, and thoroughly stirs them both, he doesn´t have to take more spoonfuls from one than the other to sample the taste accurately".

Finalmente chama-se a atenção para o facto de que se o processo de amostragem originar uma amostra enviesada, aumentar a dimensão não resolve nada, antes pelo contrário!

1.3.6 – Outros tipos de erros num processo de aquisição de dados

Além dos problemas relacionados com a amostragem e apontados anteriormente existem ainda outras fontes de erros que não estão relacionadas com o método da recolha da amostra nem com a dimensão da amostra, que são os chamados erros de não amostragem. Se, por exemplo, seleccionarmos uma amostra aleatória simples a partir de uma listagem de elementos que não contenha todos os elementos da população, poderemos obter uma amostra enviesada. Efectivamente, e como já foi referido anteriormente, muitas vezes a recolha da amostra faz-se de uma população que não é a população que se pretende estudar – população alvo ou população objectivo, mas sim de outra população que se pensa representar a primeira – população inquirida. Por exemplo, se se pretende estudar uma determinada característica dos residentes em Lisboa, é comum recolher uma amostra seleccionando aleatoriamente alguns números de telefones da lista telefónica de Lisboa, para representar a população lisboeta. Este processo introduz algum enviesamento, pois existem zonas de Lisboa onde a percentagem de pessoas com telefone é pequena. Além disso, pode acontecer com alguma frequência telefonarem para casa das pessoas quando elas estão ausentes, no trabalho, pelo que a amostra subestimará a percentagem dos lisboetas que trabalham fora de casa.

O exemplo que acabámos de descrever refere-se a um erro de selecção.

Na recolha da informação também se pode ainda verificar que a informação dada não seja verdadeira. Ao responder a um inquérito o inquirido pode sentir-se condicionado pelo inquiridor, face a determinadas perguntas. Isso poderá levá-lo a mentir. Por exemplo ao perguntarem a um indivíduo se ele é racista, ele pode dizer que não, quando na verdade o é.

Finalmente, pode-se ter feito um planeamento adequado da amostra a recolher, mas ao recolher a informação de entre os elementos da amostra, a pessoa encarregada dessa recolha pode ver¬se defrontada com a não resposta. Este problema acontece com frequência quando a amostra é constituída por pessoas, das quais algumas das seleccionadas não são encontradas para darem a informação sobre a variável em estudo, ou então se recusam a responder. Outro problema que pode surgir é devido a erros de processamento que não têm nada a ver com o processo de recolha da amostra, mas que podem influenciar o resultado da estatística, já que esta é calculada com base na informação recolhida. Estes erros surgem com alguma frequência, sendo muitas vezes detectados por serem outliers. Efectivamente, se ao digitar um conjunto de valores correspondentes a pesos de pessoas adultas aparecer 566 quilogramas, ao fazer uma representação gráfica aparecerá este valor como outlier e imediatamente se concluirá que se trata de um problema de processamento: eventualmente ao carregar a tecla do 6 o tempo de apoio foi um pouco maior e apareceram dois 6.

1.4 - Estatística Descritiva e Inferência Estatística

Uma vez recolhida a amostra procede-se ao seu estudo. Este consiste em resumir a informação contida na amostra construindo tabelas, gráficos e calculando algumas características amostrais–estatísticas. Este estudo descritivo dos dados é o objectivo da Estatística Descritiva. Esta fase é a que depende mais da habilidade ou intuição do estatístico (dissemos no início do capítulo que a Estatística além de uma ciência, também é uma arte!). Efectivamente ele vai tentar substituir o conjunto de dados, por um sumário desses dados de forma a realçar a informação que eles contêm. Pense-se o que se passa, por analogia, com um texto comprido e repetitivo em que a pessoa se perde na leitura. Um sumário bem feito do texto, em algumas linhas, dará a informação relevante sobre o texto, que ocupava muito mais linhas. Ao ler o sumário a pessoa fica rapidamente informada sobre o assunto que trata. O mesmo se passa com os dados, sendo necessário que o sumário desses dados seja feito adequadamente de forma a não se perder muita informação, mas também de forma a não sumariar tão pouco que a pessoa seja submergida por tanta informação!

Por exemplo, suponha que perguntou a um aluno se ele foi bom aluno na licenciatura que tirou. Ele responde-lhe com as notas que teve durante os 4 anos que durou a licenciatura:

10	16	11	10	15	17	12	13	17	15	18	14
15	16	12	13	16	11	15	16	12	13	14	14
11	15	17	16	16	13	14	16				

Perante estes dados hesitará um pouco, pois não se vê facilmente qual o tipo de notas que predomina. No entanto se fizer uma representação gráfica muito simples:

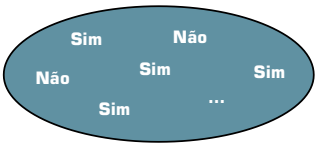
10	**
11	***
12	***
13	****
14	****
15	*****
16	*******
17	***
18	*

Imediatamente concluirá que metade das notas são iguais ou superiores a 15, pelo que se pode considerar um aluno bom. Organizámos os dados através de uma representação gráfica sugestiva, que permitiu realçar a informação desejada. Outro processo seria resumir a informação sob a forma de uma medida que se calculava a partir dos dados (estatística) - a média, que viria igual a 14.2.

Seguidamente, o objectivo de um estudo estatístico, é, de uma maneira geral, o de estimar uma quantidade ou testar uma hipótese, utilizando-se técnicas estatísticas convenientes, as quais realçam toda a potencialidade da Estatística, na medida em que vão permitir tirar conclusões acerca de uma População, baseando-se numa pequena amostra, dando-nos ainda uma medida do erro cometido. A esta fase chamamos Inferência Estatística.

Esta quantificação do erro cometido, ao transportar para a população as propriedades verificadas na amostra, é feita utilizando a Probabilidade. Efectivamente, é nesta fase do processo estatístico que temos necessidade de entrar com este conceito, para quantificar a incerteza associada aos procedimentos aqui considerados. Repare-se que ao transportar para a população uma propriedade verificada na amostra não podemos dizer que essa propriedade é verdadeira porque não a verificamos em todos os elementos da população, mas também não podemos dizer que é falsa, pois a propriedade foi verificada por alguns elementos da população - a mostra. Assim, estamos numa situação entre o que é verdadeiro e falso, caracterizada por uma incerteza, a qual é medida com a utilização da probabilidade.

Exemplo 1.4.1 -O Senhor X, candidato à Câmara da cidade do Porto, pretende saber, qual a percentagem de eleitores que pensam votar nele nas próximas eleições. Havendo algumas limitações de tempo e dinheiro, a empresa encarregada de fazer o estudo pretendido decidiu recolher uma amostra de dimensão 1000, perguntando a cada eleitor se sim ou não pensava votar no Senhor X. Como resultado da amostragem obteve-se um conjunto de sim's e não's, cujo aspecto não é muito agradável, pois à primeira vista não conseguimos concluir nada:



Procede-se à redução dos dados, resumindo a informação sobre quantos sim's se obtiveram, chegando-se à conclusão que nas 1000 respostas, 635 foram afirmativas. Então dizemos que a percentagem de eleitores que pensam votar no candidato, de entre os inquiridos, é de 63.5%. A função da Estatística Descritiva acabou aqui! (Se toda a População tivesse sido inquirida, este estudo descritivo dar-nos-ia a informação necessária para o fim em vista).

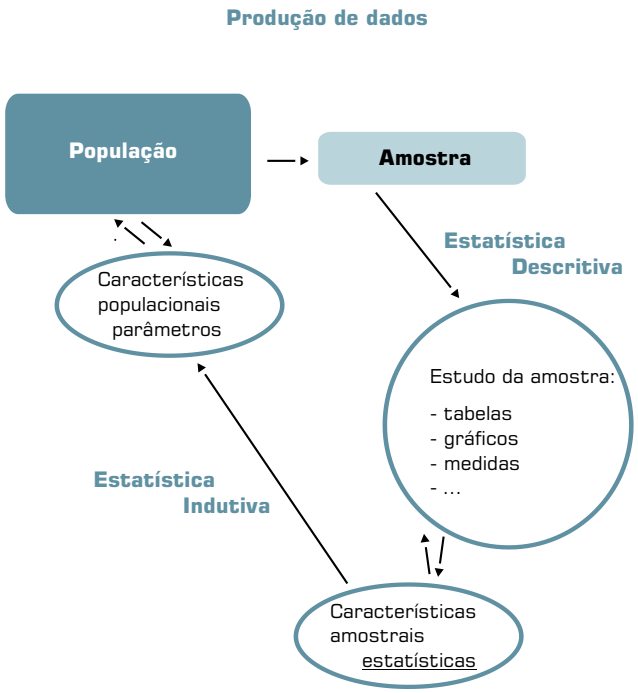
Poderemos agora inferir que 63.5% dos eleitores da cidade do Porto pensam votar no Senhor X? A resposta a esta pergunta nem é sim, nem não, mas talvez. É agora que temos necessidade de utilizar o conceito de Probabilidade, para quantificar a incerteza associada à inferência. Assim, existem processos de inferência estatística que, do resultado obtido a partir da amostra, nos permitirão concluir que o intervalo [60.5%, 66.5%] contém o valor exacto para a percentagem de eleitores da cidade que pensam votar no Senhor X, com uma confiança de 95%.

Observação - A confiança de 95% deve ser entendida no seguinte sentido: se se recolherem 100 amostras, cada uma de dimensão 1000, então poderemos construir 100 intervalos; destes 100 intervalos esperamos que 95 contenham o verdadeiro valor da percentagem (desconhecida) de eleitores da cidade do Porto, que pensam votar no

candidato. Como ao fazer um estudo só se recolhe uma amostra, não sabemos se a nossa é uma das que deu origem a um dos intervalos que continha o parâmetro. Estamos confiantes que sim!

Recorde-se a forma como as previsões são dadas, em noite de eleições, sob a forma de intervalos. Por vezes a guerra de audiências faz com que estas previsões tenham pouco sentido, por apresentarem intervalos com uma tão grande amplitude que a sua precisão, como estimativas das percentagens pretendidas, é muito pequena. Esta situação prende-se com o facto de as amostras utilizadas para a construção dos intervalos terem uma dimensão muito reduzida, havendo assim muito pouca informação disponível (recorde-se o que dissemos anteriormente sobre o processo para aumentar a precisão). No entanto, à medida que a noite vai avançando, os intervalos vão diminuindo de amplitude, estando esta diminuição da amplitude relacionada com a dimensão da amostra que entretanto vai aumentando, até finalmente estarem todos os votos contados. Nesta altura, os intervalos reduzem-se a pontos, que são as percentagens pretendidas - a amostra é constituída por toda a população.

O seguinte esquema pretende resumir as diferentes etapas que normalmente são seguidas num procedimento estatístico:



No esquema anterior a necessidade de utilizar o conceito de probabilidade faz-se sentir ao passarmos das propriedades estudadas na amostra para as propriedades na população, sendo aqui precisamente que vai ser necessário invocar o princípio da aleatoriedade.

Chama-se a atenção para que a compreensão do processo estatístico permitir-nos-á interpretar melhor as notícias que, frequentemente, se lêem nos jornais ou ouvem na televisão. Por vezes alguns estudos sobre os mesmos assuntos, apresentam resultados contraditórios! Isto acontece nomeadamente no estudo de certos aspectos do comportamento humano, utilizando testes psicológicos, ou no estudo de certas doenças utilizando cobaias. Muitas das inferências feitas são imperfeitas, a maior parte das vezes por terem como base dados imperfeitos.

2. Representação e redução de dados. Tabelas e gráficos

2.1- Introdução

Num módulo anterior de Estatística, já foram apresentados vários processos de organizar os dados, de forma a realçar as características principais e a estrutura subjacente da população de onde esses dados foram retirados.

Quer estejamos perante uma variável de tipo discreto ou contínuo, o processo de organizar a informação consiste em, de um modo geral, começar por construir tabelas de frequência e proceder a representações gráficas adequadas. Vamos seguidamente utilizar o Excel na construção dessas tabelas de frequência.

2.2 – Utilização do Excel na obtenção de tabelas de frequência

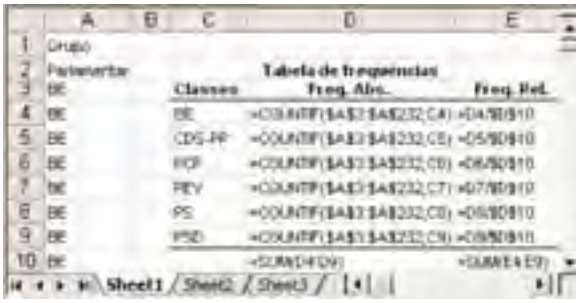
Vamos exemplificar a utilização do Excel na construção de tabelas de frequência a partir do ficheiro Deputados.xls, apresentado no capítulo anterior.

2.2.1 – Tabela de dados qualitativos ou quantitativos discretos

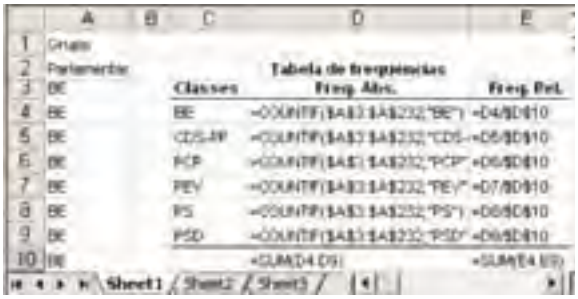
O procedimento para a construção das tabelas de frequência é idêntico, quer tenhamos um conjunto de dados qualitativos ou quantitativos discretos, já que as classes que se consideram são as diferentes categorias ou valores que surgem, respectivamente, no conjunto de dados. A seguir apresentamos a construção destas tabelas utilizando a função COUNTIF. Numa secção posterior veremos a sua construção utilizando a metodologia das PivotTables.

Exemplo 2.2.1 – Considere o ficheiro Deputados.xls. Obtenha uma tabela de frequência para a variável Grupo Parlamentar.

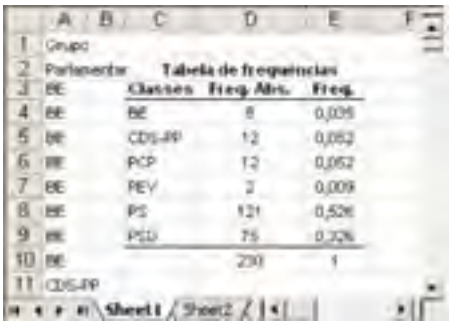
Começámos por copiar a coluna correspondente ao Grupo parlamentar para um novo ficheiro. Ordenámos os elementos por ordem crescente e inserimos na coluna Classes os diferentes elementos do conjunto de dados. Utilizámos de seguida a função COUNTIF (CONTAR.SE) para obter as frequências absolutas de deputados de cada um dos grupos parlamentares:



ou



As fórmulas apresentadas anteriormente, deram origem à seguinte tabela:



2.2.2 – Tabela de dados quantitativos contínuos

Como se viu no módulo anterior de Estatística, no caso de dados contínuos o processo de construção das tabelas é um pouco mais elaborado, já que a definição das classes não é tão imediata. De um modo geral as classes são intervalos com a mesma amplitude, fechados à esquerda e abertos à direita ou abertos à esquerda e fechados à direita. Em certos casos não é conveniente que as classes tenham a mesma amplitude, o que em si não é um problema para a construção da tabela de frequências, mas que implica alguma complicação na construção do histograma associado, quando pretendemos utilizar Excel.

Vamos utilizar ainda o ficheiro Deputados.xls para estudar a variável Idade, que é uma variável quantitativa contínua.

Exemplo 2.2.2 – Utilizando a informação contida no ficheiro Deputados.xls, construa uma tabela de frequências para a variável Idade.

Vamos dividir esta tarefa em duas partes: uma primeira parte consistirá na definição das classes e uma segunda parte no cálculo das frequências.

Copie a coluna "Data de nascimento" para um ficheiro novo com 230 elementos que ocupam as células A2:A231. Para obter a idade em 31/12/2007, podemos utilizar a seguinte metodologia:

- **Passo 1** – Inserir na célula B1 a data 31/12/2007;
- **Passo 2** – Colocar o cursor na célula B2 e introduzir a expressão: =$B$1-A2;
- **Passo 3** – Replicar esta função através das células B3 a B231;
- **Passo 4** – Se no passo anterior se obteve uma coluna de datas, formatar essa coluna com o Format General, por exemplo. Obtém-se a idade em dias;
- **Passo 5** – Para obter a idade em anos, colocar o cursor na célula C2 e introduzir a seguinte função: = INT(B2/365), a qual devolve o maior inteiro contido no quociente (n.º de dias do deputado)/(n.º de dias do ano).
- Replicar esta função através das células C3 a C231.

Definição das classes:

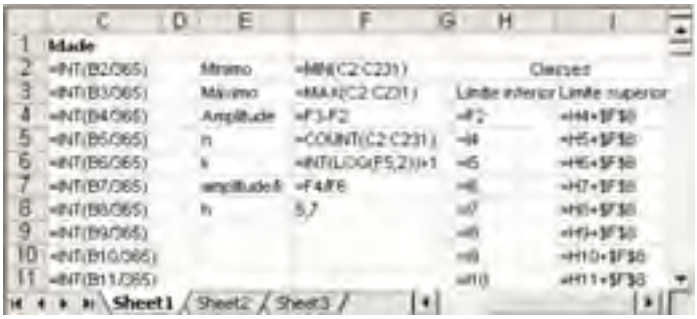
a) Determinar a amplitude da amostra, subtraindo o mínimo do máximo;

b) Dividir essa amplitude pelo número K de classes pretendido. Existe uma regra empírica que nos dá um valor aproximado para o número K de classes e que consiste no seguinte: para uma amostra de dimensão n, considerar para K o menor inteiro tal que 2K≥n. Uma expressão equivalente para obter K, consiste em considerar K=INT(LOG(n;2))+1 ou K=ROUNDUP(LOG(n;2);0), em que a fun-ção ROUNDUP(x;m), devolve um valor de x, arre-dondado por excesso, com m casas decimais;

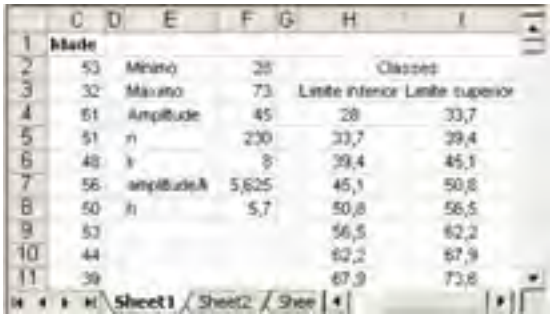
c) Calcular a amplitude de classe h, dividindo a amplitude da amostra por K e tomando para h um valor aproximado por excesso do quociente anteriormente obtido;

d) Construir as classes C1, C2, …, Ck. Vamos considerar como classes os intervalos [mínimo, mínimo + h[,[mínimo + h, mínimo + 2h[, …, [mínimo + (k-1)h, mínimo + kh[. Uma alternativa a este procedimento seria considerar as classes abertas à esquerda e fechadas à direita, da seguinte forma:]max – Kh, max – (K-1)h],]max – (K-1)h, max – (K¬2)h],]max – h, max].

Estes passos são representados na figura seguinte:

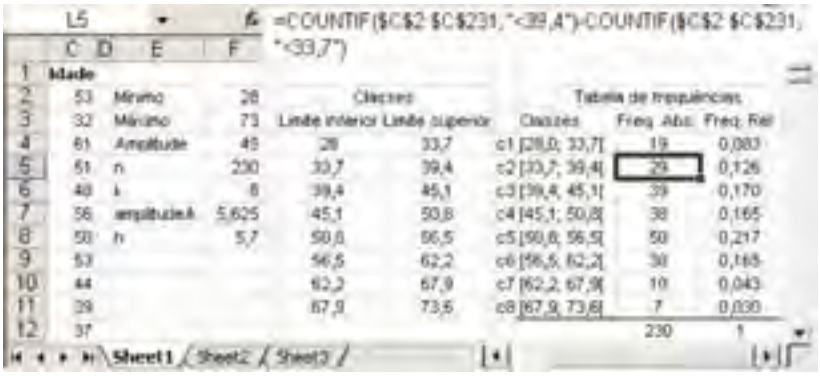


com os seguintes resultados:



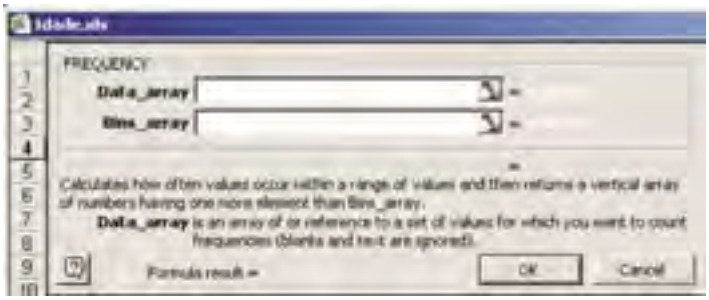
Cálculo das frequências:

Para obter as frequências absolutas, vamos utilizar a função COUNTIF do seguinte modo:



As frequências das classes c1, c3..., c8, são obtidas de forma idêntica à de c2, mudando os limites das classes.

2.2.3 - Construção de uma tabela de frequências utilizando a função Frequency do Excel



O Excel tem uma função, que é a função Frequency(Data_array;Bins_array), que calcula o número de elementos da variável - cujos valores se encontram na Data_array, existentes nas classes - cujos limites se encontram em Bins_array. Este vector Bins_array é constituído por um conjunto de k valores b1, b2, ..., bk, formando (k+1) classes, tais que:

- A 1ª classe é dada por (-∞, b1], isto é, conterá todos os elementos ≤b1;
- A 2ª classe é dada por]b1, b2];
- A 3ª classe é dada por]b2, b3];
- A késima classe é dada por]bk-1, bk];
- A (k+1)ésima classe é dada por]bk, +∞);

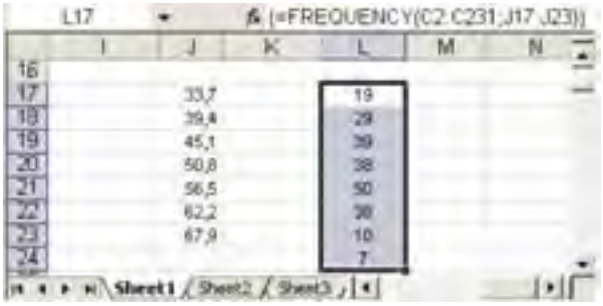
Vamos exemplificar construindo uma tabela de frequências para a variável idade.

Definição das classes:

Considerando as classes definidas em 2.2 e tendo em atenção o que dissemos anteriormente sobre as classes para a utilização da função Frequency, o nosso conjunto de valores para o Bins_array, será constituído por {33,7; 39,4; 45,1; 50,8; 56,5; 62,2; 67,9}; Para utilizar a função Frequency(Data_array;Bin_array), procede-se do seguinte modo:

- Definir a coluna de separadores ou limites das classes, que constituirá o Bins_array;
- Seleccionar tantas células em coluna, quantas as classes consideradas para a tabela de frequências (não esquecer que o número de classes é superior em uma unidade ao número de separadores, pelo que o número de células seleccionadas deverá ser, neste caso, de 8);
- Introduzir a função Frequency, considerando como primeiro argumento o conjunto de células onde se encontram os dados a agrupar, chamado de Data_array, e como segundo argumento as células que constituem o Bins_array;
- Carregar CTRL+SHIFT+ENTER.

Na figura seguinte apresentamos o resultado deste procedimento:



Verifique que os valores devolvidos pela função Frequency, nas células L17: L24, são iguais às frequências obtidas anteriormente e apresentadas na tabela de frequências já construída. Esta situação nem sempre se verifica, nomeadamente se os limites das classes fossem números inteiros, já que agora as classes são consideradas fechadas à direita e abertas à esquerda. Assim, alguns valores da amostra que anteriormente não pertenciam a determinadas classes, poderiam agora pertencer.

2.3 – Utilização do Excel na representação gráfica de dados

De forma idêntica à que fizemos para a construção das tabelas de frequências, vamos também considerar separadamente o caso da variável em estudo ser de natureza qualitativa ou quantitativa discreta, ou de natureza quantitativa contínua.

2.3.1 – Variáveis qualitativas ou quantitativas discretas. Diagrama de barras

Neste caso vimos que a construção da tabela de frequências se resume, de um modo geral, a considerar como classes as diferentes categorias ou valores que surgem na amostra. Uma representação gráfica adequada para estes dados, é o diagrama de barras, que já foi introduzido no módulo de Estatística.

Diagrama de barras – Representação gráfica que consiste em marcar num sistema de eixos coordenados, no eixo dos xx, pontos representando as categorias ou os valores considerados para as classes na tabela de frequências, e nesses pontos barras verticais de altura igual à frequência absoluta ou à frequência relativa.

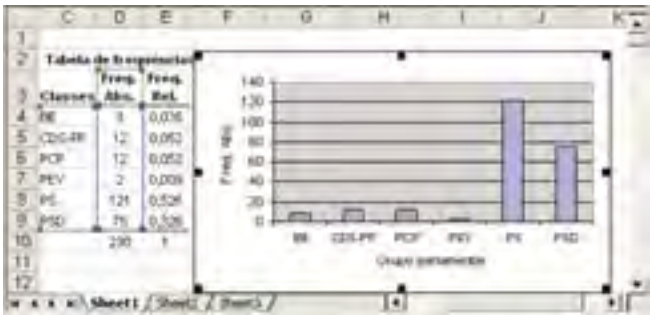
2.3.1.1 - Variável de tipo qualitativo

Exemplo 2.3.1 - Vamos exemplificar a construção de um diagrama de barras de uma variável qualitativa, considerando a tabela de frequências construída em 2.2.1, para estudar a variável Grupo Parlamentar, do ficheiro Deputados.xls:

Tabela de frequências		
Classes	Freq. Abs.	Freq. Rel.
BE	8	0,035
CDS-PP	12	0,052
PCP	12	0,052
PEV	2	0,009
PS	121	0,526
PSD	75	0,326
	230	1

A metodologia seguida para construir o diagrama de barras, consiste em, na folha Excel, que contém a tabela:

- Seleccionar as células que contêm as classes e as frequências absolutas (por exemplo);
- Seleccionar, no menu, o ícone Chart ;
- Na caixa de diálogo que aparece, seleccionar a opção Column;
- Clicar no botão Next, duas vezes, para passar dois passos, até aparecer uma caixa de diálogo, que apresenta várias opções: Em Legend, desactivar a legenda e em Titles, acrescentar o título no eixo dos Y's e no eixo dos X's, como se apresenta a segui, e carregar em Finish:



2.3.1.2 - Variável de tipo quantitativo discreto

2.3.1.2.1 – Diagrama de barras

No caso de dados discretos, para construir a tabela de frequência consideram-se como classes os diferentes valores que surgem na amostra. Estes valores devem ser apresentados, na tabela de frequência, ordenados.

Exemplo 2.3.2 – Suponhamos que para uma amostra de 30 deputados da actual legislatura, se tinha recolhido a informação sobre o número de filhos, tendo-se obtido os seguintes valores:

2, 1, 2, 3, 0, 0, 1, 1, 4, 1, 2, 1, 0, 0, 0, 2, 3, 1, 1, 6, 3, 1, 3, 2, 0, 1, 2, 0, 2, 3

Resuma os dados numa tabela de frequências e construa o diagrama de barras associado.

Introduzimos os dados numa folha de Excel, a que chamámos Filhos.xls e a seguir procedemos do seguinte modo:

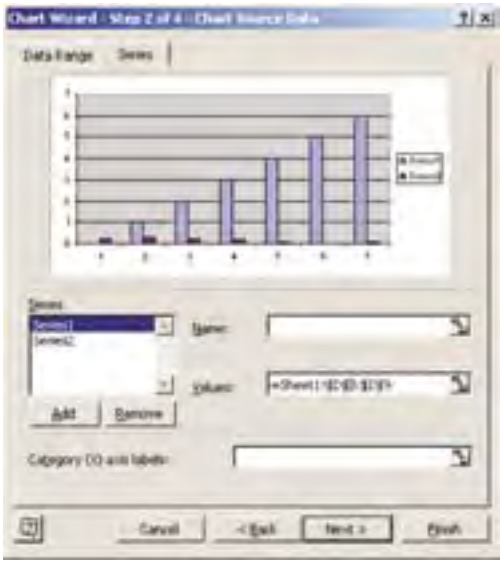
1ª parte – Procedimento para a construção da tabela de frequências:

- Seleccionar as células que contêm os dados e ordená-los utilizando o botão da barra de Excel;
- Considerar para classes os diferentes valores que surgem na amostra. Se faltar algum valor entre o máximo e o mínimo, considerá-lo também na tabela de frequências, se a seguir se pretende construir um diagrama de barras;
- Utilizando a função COUNTIF, determinar as frequências absolutas das classes consideradas no ponto anterior; calcular a partir destas, as frequências relativas:

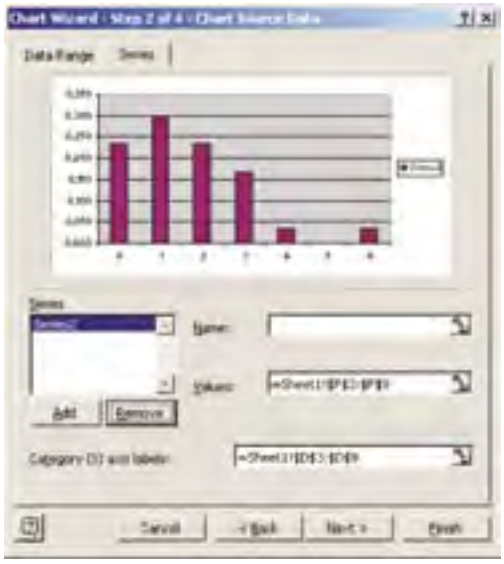
Tabela de frequências		
Classes	Freq. Abs.	Freq. Rel.
0	7	0,233
1	9	0,300
2	7	0,233
3	5	0,167
4	1	0,033
5	0	0,000
6	1	0,033
	30	

2ª parte – Procedimento para a construção do diagrama de barras:

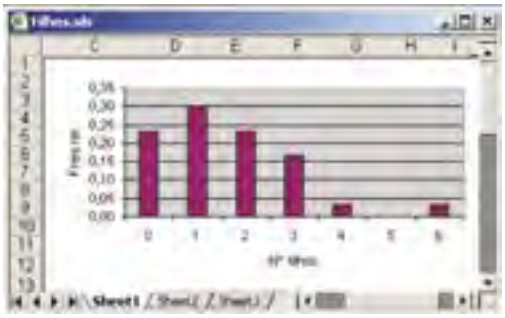
- Seleccionar as células que contêm as classes e as que contêm as frequências relativas (para seleccionar as células que contêm as frequências relativas, como não são adjacentes às que contêm as classes, depois de seleccionar estas, tem que se pressionar a tecla CTRL e com ela pressionada, seleccionar aquelas);
- Seleccionar na barra de ferramentas a opção Chart e a seguir a opção Column, tal como se fez para os dados de tipo qualitativo;
- Seleccionar Next e de seguida Series, como se apresenta a seguir:



- Copiar a Series1, dada pelos valores =Sheet1!$D$3:$D$9, que constituem as classes, para Category (X) axis labels e remover Series1 de Series:



- Seleccionar Next. Nas Chart Options seleccionar Legend e retirar a selecção de Show Legend. Seleccionar Titles e colocar títulos adequados. Carregar em Finish:

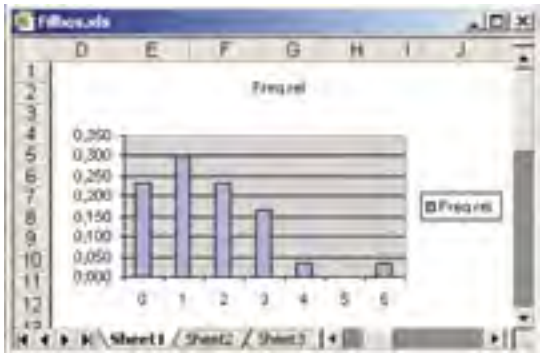


Nota (Esta nota foi sugerida pela leitura do artigo de Neville Hunt – Charts in Excel, in Teaching Statistics, Volume 26, Number 2, pags 49-53): Como vimos na descrição que acabámos de fazer para a construção de um diagrama de barras em Excel, o procedimento está longe de ser natural, já que o que seria de esperar era que, uma vez seleccionada a opção Column, nos surgisse a representação gráfica anterior, aparte pequenas alterações de "cosmética".

Suponhamos, no entanto, que na última classe tínhamos considerado 6+, para significar 6 ou mais filhos. Então, ao fim dos dois primeiros passos da descrição anterior temos a representação gráfica pretendida. O facto é que agora o Excel interpretou as classes como categorias e fez a representação esperada.

Suponhamos ainda, que em vez de modificarmos o 6 para 6+, apagamos o conteúdo de D2:

Seleccionando agora as células D2 a D9 e F2 a F9 e novamente no Chart a opção Column, então a representação que se obtém é, imediatamente, a seguinte:



Depois de apagar a legenda e inserir os títulos de forma conveniente, temos a representação final do exemplo anterior, sem grandes complicações.

2.3.1.2.2 – Função cumulativa

A função cumulativa é uma função definida para todo o valor real x, e que para cada x dá a soma das frequências dos valores da amostra menores ou iguais a x. Quando temos uma variável de tipo discreto, a função cumulativa é uma função em escada, isto é, é uma função que cresce por degraus, mudando de degrau nos pontos em que a frequência é diferente de 0, e em que a altura do degrau é igual à frequência respectiva. Vamos exemplificar a sua construção com o exemplo apresentado na secção anterior para a construção do diagrama de barras.

Exemplo 2.3.2 (cont) – Construa a função cumulativa para os dados do número de filhos da amostra dos 30 deputados.

Retomando a tabela de frequências do exemplo 2.3.2, vamos acrescentar uma coluna com as frequências relativas acumuladas:

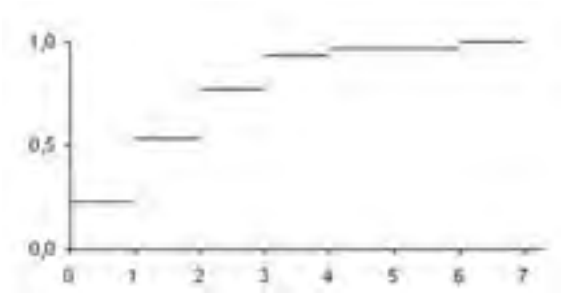
Tabela de frequências

Classes	Freq. Abs.	Freq. Rel.	Freq.rel.acum.
0	7	0,233	0,233
1	9	0,300	0,533
2	7	0,233	0,767
3	5	0,167	0,933
4	1	0,033	0,967
5	0	0,000	0,967
6	1	0,033	1,000
	30		

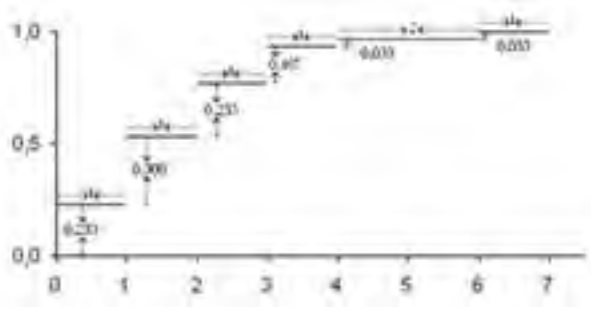
A função cumulativa há-de ser tal que:

- Para valores de x< o, será nula;
- Para valores de o ≤ x < 1, será igual a 0,233;
- Para valores de 1 ≤ x < 2, será igual a 0,533;
- Para valores de 2 ≤ x < 3, será igual a 0,767;
- Para valores de 3 ≤ x < 4, será igual a 0,933;
- Para valores de 4 ≤ x < 6, será igual a 0,967;
- Para valores de x ≥ 6, será igual a 1;

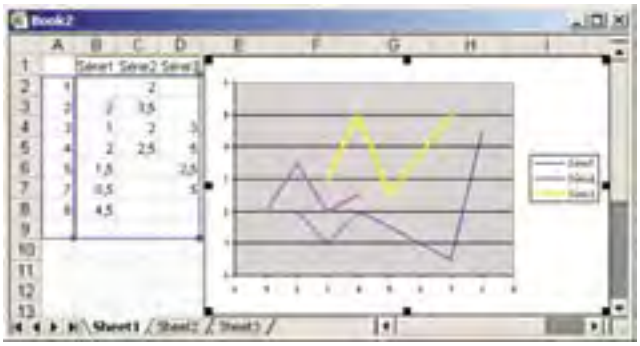
O Excel não dispõe de uma representação imediata para a função anterior, pelo que temos de utilizar um pequeno artifício. Suponhamos, para já, que por algum processo tínhamos conseguido construir o gráfico da função cumulativa, que tem o seguinte aspecto:



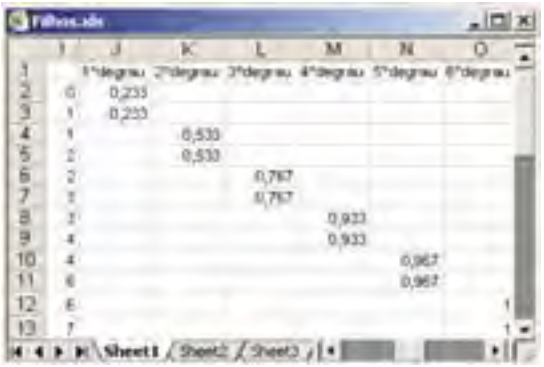
Esta função é constituída por 6 degraus, em que a altura do degrau é, em cada ponto, igual à frequência relativa respectiva e a dimensão do patamar é igual à diferença entre os pontos consecutivos, com frequência relativa diferente de zero:



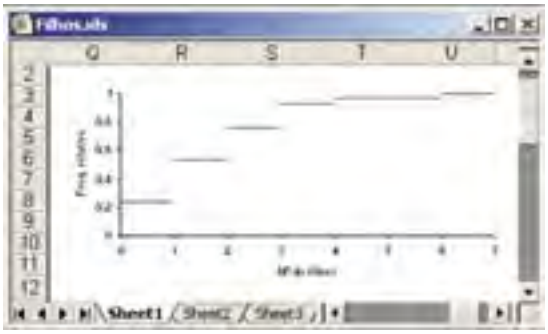
O Excel dispõe de uma representação gráfica, o Scatter (Diagrama de dispersão), em que no último subtipo apresentado para as opções, une os pontos, por ordem crescente das abcissas, simultaneamente de tantas séries (conjuntos de pontos) quantas as desejadas. Exemplifiquemos com os pontos da seguinte tabela, em que pretendemos representar 3 conjuntos de dados a que chamámos Série1, Série2 e Série3:



Vamos utilizar esta função Scatter para construir os sucessivos degraus da função cumulativa, em que cada degrau corresponde a uma série - união de dois pontos, e em que temos tantas séries a representar, quantos os degraus. Assim, o artifício está em representar, numa tabela do Excel, os degraus pretendidos através das coordenadas dos pontos, como exemplificamos a seguir:



Agora basta seleccionar as células I2 a O13 e fazer o diagrama de dispersão, como indicado anteriormente. Proceda como na construção do diagrama de barras, para retirar a legenda e acrescentar títulos:



2.3.2 – Variáveis quantitativas contínuas

2.3.2.1 – Histograma

2.3.2.1.1 – Tabela de frequências com as classes com a mesma amplitude

No caso de um conjunto de dados contínuos, já vimos anteriormente a forma de obter a tabela de frequências. Como se viu, as classes são intervalos e a representação gráfica adequada é o histograma, já apresentado no módulo de Estatística:

Histograma

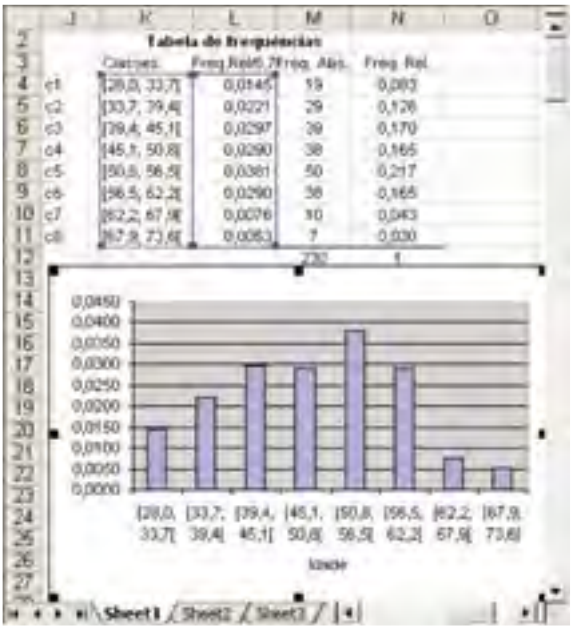
É um diagrama de áreas, formado por uma sucessão de rectângulos adjacentes, tendo cada um por base um intervalo de classe e por área a frequência relativa (ou frequência absoluta). Por conseguinte, a área total coberta pelo histograma é igual a 1 (ou igual a n, a dimensão do conjunto de dados a representar).

Para construir o histograma de forma correcta, isto é, de modo a que as áreas dos rectângulos sejam iguais às frequências, a altura do rectângulo correspondente a determinada classe, deverá ser igual à frequência da classe a dividir pela respectiva amplitude. Contudo, se as classes tiverem todas a mesma amplitude, é usual construir os rectângulos com alturas iguais às frequências relativas (absolutas) das respectivas classes, vindo as áreas dos rectângulos proporcionais e não iguais às frequências. A constante de proporcionalidade é a amplitude de classe. No entanto, se se pretender comparar amostras através de histogramas, embora o histograma não seja a representação mais adequada para a comparação de amostras, deve-se ter o cuidado de os construir da forma indicada inicialmente, e utilizando as frequências relativas, de modo que a área total ocupada por cada um dos histogramas seja igual a 1.

Exemplificamos, de seguida, a construção de um histograma utilizando o Excel.

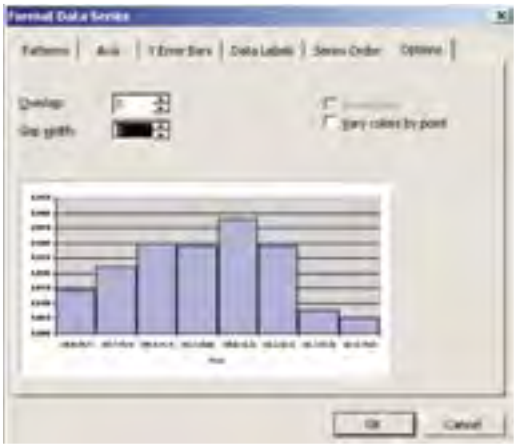
Exemplo 2.3.3 – Considerando a tabela de frequências construída em 2.3 para a variável idade, construa o histograma adequado. Processo utilizado para obter o histograma:

- Acrescentar, à tabela considerada, uma outra coluna com a frequência relativa a dividir pela amplitude de classe (igual a 5,7). No caso presente, inserimos estas células adjacentes às células que contêm as classes. No entanto, não é necessário ter esta preocupação, já que se se pretender seleccionar células não adjacentes, basta seleccionar as células da primeira coluna e se a coluna seguinte não for adjacente, começar por carregar a tecla CTRL e com ela pressionada seleccionar, então, as células pretendidas;
- Seleccionar as células de K4 a L11 (que contêm as classes e as frequências relativas a dividir pela amplitude de classe);
- Proceder como em 3.1 para construir um diagrama de barras, para obter a figura que se apresenta a seguir;

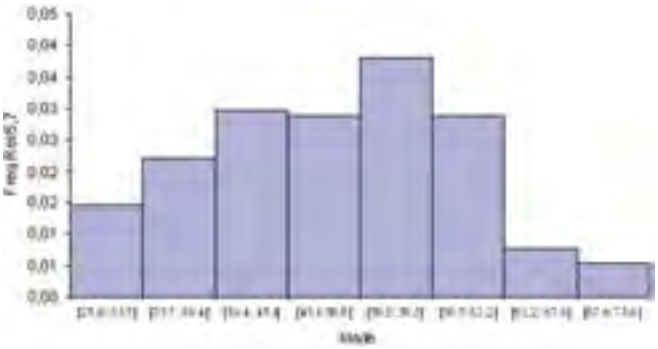


Para obter o histograma, já que o que se nos apresenta na figura anterior não é um histograma pois não tem as barras adjacentes, terá de:

Clicar duas vezes sobre as barras, de forma a que apareça o menu Format Data Series ou Format data Points.; Seleccionar Options e em Gap Width seleccionar 0;OK:



Finalmente pode-se melhorar esteticamente o histograma, diminuindo o número de casas decimais nos valores apresentados no eixo dos YY, retirando as linhas, etc.

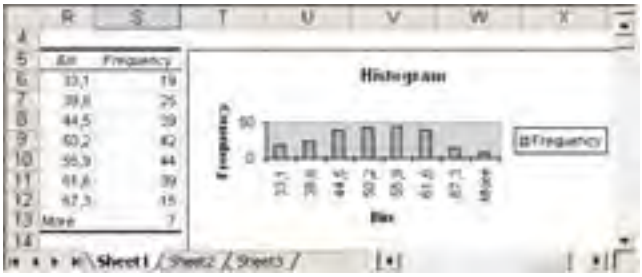


2.3.2.1.2 – Função Histogram

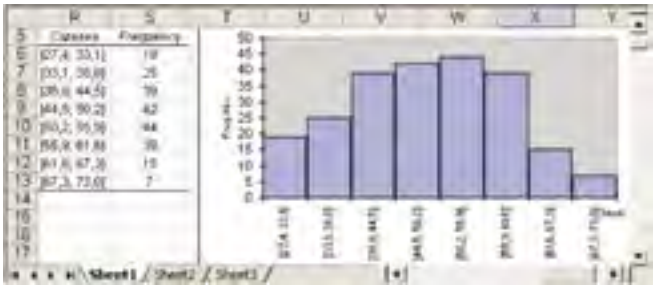
No Excel existe uma função, idêntica à função Frequency, a função Histogram, a que se acede seleccionando Tools-Data-Analysis-Histogram-Ok. Vamos exemplificar a sua utilização para o conjunto de dados da variável Idade, anteriormente considerado:

- Definir a coluna de separadores ou limites de classes, que constituirá o Bin Range: No nosso caso contruímos as classes subtraindo a amplitude de classe sucessivamente ao máximo, obtendo os valores {33,1, 38,8, 44,5, 50,2, 55,9, 61,6, 67,3} (tal como para a função Frequency, as classes são fechadas à direita e abertas à esquerda), que colocámos nas células P4:P10;
- Seleccionar Tools-Data-Analysis-Histogram-Ok:

- Em Input Range, indicámos o local dos dados e seleccionámos ainda a opção Chart Output e clicámos OK. Como resultado obtivemos o seguinte:



- Substituímos os limites das classes pelos intervalos das classes e arranjámos convenientemente o gráfico, já que a representação que se obtém, ao contrário do que é indicado no título, não é um histograma:



Nota: Ao considerar a função Histogram, tem a possibilidade de não indicar os separadores de classe, deixando vazio o espaço denominado Bin Range, uma vez que serão considerados, por defeito, classes. Contudo, não aconselhamos que se deixe esta escolha ao Excel, uma vez que, por exemplo, a primeira classe que é considerada, é constituída pelos valores menores ou iguais ao mínimo, o que não tem qualquer sentido.

2.3.3.1.3 - Tabela de frequências com as classes com amplitudes diferentes

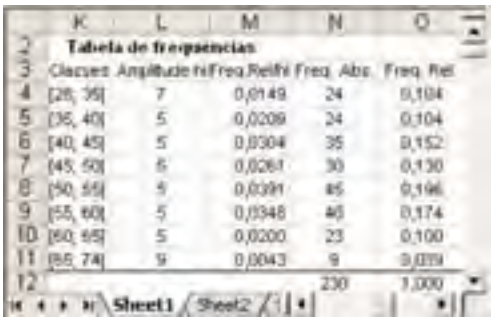
Por vezes a organização e redução de um conjunto de dados contínuos, através de uma tabela de frequências, pressupõe que os intervalos, que constituem as classes, tenham limites escolhidos pelo utilizador, sem obedecerem a um critério estritamente resultante da aplicação de uma regra matemática. É o caso, por exemplo, da variável idade, em que poderá ser interessante escolher determinadas classes etárias.

Tendo em conta a definição de histograma, como sendo um diagrama de áreas, constituído por uma série de rectângulos adjacentes, em que a área de cada rectângulo é igual ou proporcional à frequência de classe, no caso de a tabela de frequências não apresentar as classes todas com a mesma amplitude, já o histograma não se pode reduzir a um diagrama de barras, em que as barras tenham a mesma amplitude e as alturas sejam iguais às frequências.

Não sendo o Excel um software de Estatística, não apresenta uma solução imediata para a construção do histograma nestas condições, sendo necessário recorrer a um artifício. Exemplificaremos a seguir a aplicação de uma técnica possível para a resolução do problema, recorrendo à representação gráfica Scatter.

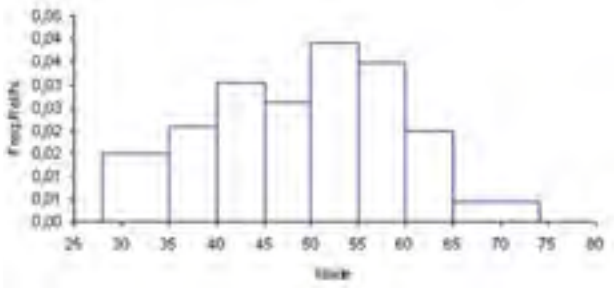
Exemplo 2.3.4 – Consideremos ainda a variável idade dos deputados. Organize os dados segundo uma tabela de frequências, considerando as seguintes classes **[28, 35[, [35, 40[, [40, 45[, [45, 50[, [50, 55[, [55, 65[, [65, 75[, [75, 78].**

A construção da tabela de frequências pode ser feita utilizando a função Frequency, como vimos na secção anterior. No entanto, vai ser necessário acrescentar uma nova coluna onde, para cada classe, se considera a frequência relativa (ou absoluta) a dividir pela amplitude de classe. Será esta coluna que irá fornecer as alturas dos rectângulos que constituirão o histograma. Com esta precaução, garantimos que as áreas destes rectângulos são iguais às frequências relativas (ou absolutas). Apresenta-se a seguir a tabela de frequências obtida, segundo a descrição anterior:



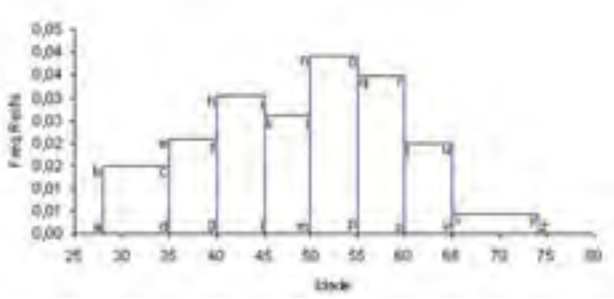
	K	L	M	N	O
2	Tabela de frequências				
3	Classes	Amplitude hi	Freq.Relhi	Freq. Abs.	Freq. Rel.
4	[28, 35[	7	0,0149	24	0,104
5	[35, 40[	5	0,0209	24	0,104
6	[40, 45[	5	0,0304	35	0,152
7	[45, 50[	5	0,0261	30	0,130
8	[50, 55[	5	0,0391	45	0,196
9	[55, 60[	5	0,0348	40	0,174
10	[60, 65[	5	0,0200	23	0,100
11	[65, 74[	9	0,0043	9	0,039
12				230	1,000

O histograma correspondente a esta tabela de frequências, com cuja construção não nos vamos preocupar para já, terá o seguinte aspecto:

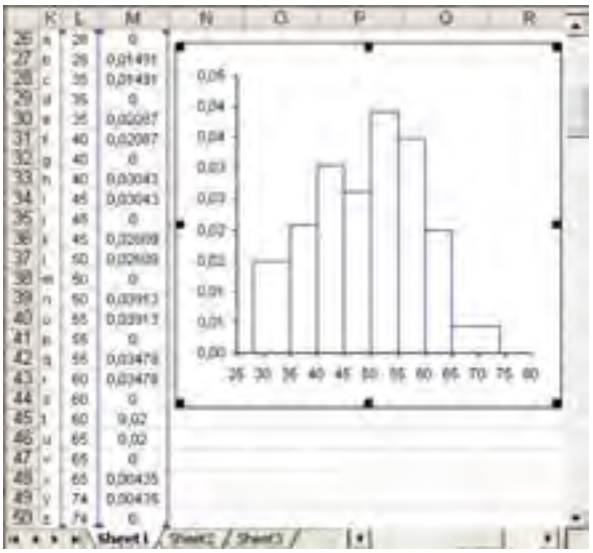


Temos um histograma correctamente construído, em que as áreas dos rectângulos são iguais às frequências relativas, ocupando o histograma uma área total igual a 1.

Na figura anterior, vamos marcar alguns pontos com letras:



Repare que se unir o ponto **a** com **b**, de seguida com c, até esgotar todos os pontos, obtém o histograma. Então, para obter a representação gráfica desejada, basta construir uma tabela, numa folha de Excel, com as coordenadas dos pontos que pretendemos unir e utilizar a representação Scatter, tal como foi feito para representar a função cumulativa em 3.1.2.2:

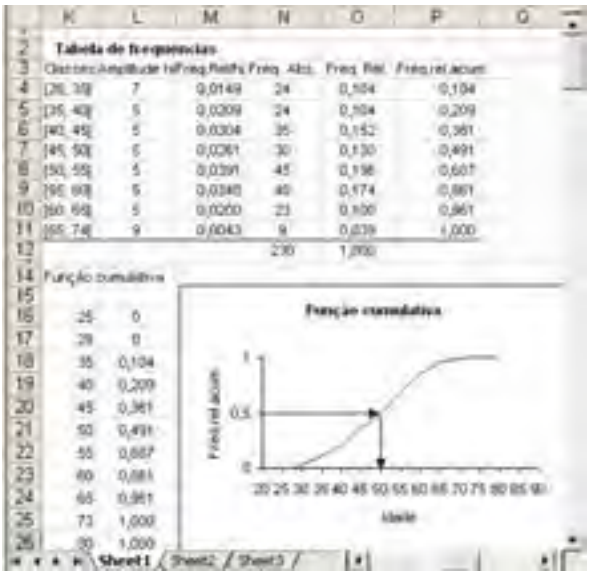


2.3.2.2 – Função cumulativa

Para representar graficamente as frequências acumuladas, considera-se a função cumulativa, que se obtém utilizando a seguinte metodologia:

- Antes do limite inferior da 1ª classe, l1, a frequência acumulada é nula, pelo que se traça um segmento sobre o eixo dos xx, até esse ponto;
- No limite inferior da 2ª classe, l2, a frequência acumulada é a frequência da classe anterior, f1. Admitindo que a frequência se distribui uniformemente no intervalo de classe, unimos os pontos de coordenadas (l1,0) e (l2, f1);
- No limite inferior da 3ª classe, l3, a frequência acumulada é a soma das frequências das duas classes anteriores, (f1+f2). Então unimos os pontos de coordenadas (l2, f1) e (l3, (f1+f2));
- Quando chegarmos à última classe, temos a garantia que a frequência acumulada, correspondente ao seu limite superior, é igual a 1, pelo que nesse ponto marcamos 1 e continuamos com um segmento de recta paralelo ao eixo dos xx.

Exemplo 2.3.4 (continuação) – Construa a função cumulativa, a partir da tabela de frequências apresentada no exemplo 2.3.4. Para obter a função cumulativa, basta acrescentar à tabela de frequências uma nova coluna com as frequências relativas acumuladas. De seguida utiliza-se a representação Scatter, para unir os pontos, tais como foram definidos nas indicações dadas, anteriormente, para a construção da função cumulativa:



Da maneira como foi construída, a função cumulativa tem algumas propriedades importantes, nomeadamente:

- Está definida para todo o x real (na representação gráfica anterior escolhemos arbitrariamente o valor da abcissa igual a 25 para começar a construir a função cumulativa);
- É sempre não decrescente;

- Só assume valores no intervalo [0, 1];
- Permite obter informação sobre qual o valor da abcissa a que corresponde determinada frequência acumulada.

Vamos explorar um pouco mais esta última propriedade.

Suponhamos que se pretendia saber, a partir da representação gráfica da função cumulativa, obtida para o exemplo anterior, qual o valor aproximado para a idade a que corresponde uma frequência relativa acumulada de 50%. De acordo com a figura, este valor deve estar na classe [50, 55[.

Uma vez que se admite que a frequência se distribui uniformemente sobre a amplitude de classe, isto é a frequência 0,196 (=0,687-0,491) distribui-se uniformemente sobre o intervalo de amplitude 5, através da resolução de uma equação de proporcionalidade, obtém-se o valor que andávamos à procura:

$$\frac{0{,}196}{0{,}009}=\frac{5}{x} \qquad x=\frac{0{,}009\times 5}{0{,}196}=0{,}22$$

onde 0,009=0,5-0,491. Então o valor pretendido é 50 + 0,22 = 50,22 anos, ou seja 50 anos.

Ao valor obtido anteriormente, a que corresponde uma frequência acumulada de 50%, chamamos mediana. A mediana, que já foi objecto de estudo no módulo de Estatística, divide a distribuição das frequências em duas partes iguais.

Recordamos que a técnica utilizada permitiu¬nos obter um valor aproximado para a mediana, cujo valor exacto só poderia ter sido determinado a partir dos dados originais, antes de proceder ao agrupamento. Aliás, veremos mais à frente a determinação desta e de outras medidas, utilizando o Excel.

Se em vez de pretendermos determinar o valor a que corresponde a percentagem de 50%, procurássemos os valores a que correspondem as percentagens de 25% ou 75%, obteríamos os chamados quartis, respectivamente 1.º e 3.º quartil, e a metodologia utilizada para os determinar a partir da função cumulativa seria idêntica à utilizada para determinar a mediana.

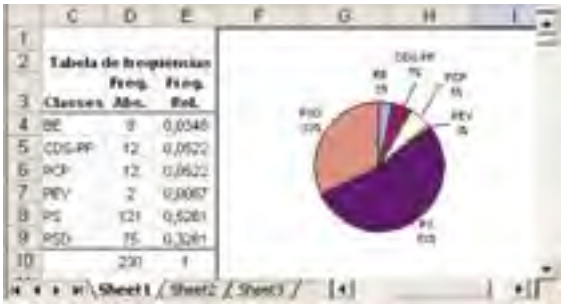
2.3.3 – Outras representações gráficas

Além das representações gráficas consideradas anteriormente, em que destacamos o diagrama de barras para dados discretos e o histograma para dados contínuos, existem ainda outras representações que podem ser utilizadas para dados qualitativos ou quantitativos – diagrama circular, ou dados quantitativos – caule-e-folhas e diagrama de extremos e quartis. Todas estas representações já foram objecto de estudo no módulo de Estatística, pelo que privilegiaremos aqui a forma de os construir utilizando o Excel.

2.3.3.1 – Diagrama circular

Esta representação, utilizada essencialmente para dados qualitativos, é constituída por um círculo, em que se apresentam vários sectores circulares, tantos quantas as classes consideradas na tabela de frequências da amostra em estudo. Os ângulos dos sectores são proporcionais às frequências das classes. A representação deste diagrama, em Excel, é imediata, apresentando várias modalidades.

Exemplo 2.3.5 – Apresente sob a forma de um diagrama circular a distribuição dos deputados do ficheiro Deputados.xls segundo o grupo parlamentar. Esta variável já foi objecto de estudo num exemplo anterior, de forma que recorremos à tabela de frequências já calculada, para obter a representação gráfica pretendida. Seleccionam-se as células com as classes e as respectivas frequências absolutas ou relativas e no menu Chart seleccionassem Pie, a modalidade desejada:



2.3.3.2 – Caule-e-folha

Esta representação, como se sabe, é uma representação que se pode considerar entre a tabela e o gráfico, uma vez que são apresentados os verdadeiros valores da amostra, mas de forma sugestiva, que faz lembrar um histograma. Antes de abordarmos a forma de construir um caule¬e-folhas utilizando o Excel, vamos apresentar um exemplo, que nos poderá ajudar a compreender os passos necessários para essa construção.

Exemplo 2.3.6 – Consideremos a seguinte amostra constituída pela idade de 30 deputados,escolhidos aleatoriamente da tabela de deputados do ficheiro Deputados.xls:

63	59	31	51	51	61	42
65	48	63	57	43	54	42
52						
51	57	34	38	44	61	60
56	66	63	52	47	33	46
52						

Nesta representação considerámos 4 caules e o intervalo entre caules sucessivos é de 10 unidades. No caule 3 pendurámos todas as folhas deste caule e o mesmo foi feito com todos os outros caules. É como se tivéssemos considerado as classes [30, 40[, [40, 50[, [50, 60[e [60, 70[para agrupar os dados. Suponhamos que em vez de considerar estas classes, de amplitude 10, estávamos interessados em considerar classes de amplitude 5, a saber [30, 35[, [35, 40[, [40, 45[, [45, 50[, [50, 55[, [55, 60[, [60, 65[e [65, 70[. Então a representação anterior teria o seguinte aspecto:

3	1 3 4
3	8
4	2 2 3 4
4	6 7 8
5	1 1 1 2 2 2 4
5	6 7 7 9
6	0 1 1 3 3 3
6	5 6

Qualquer que seja a representação considerada, qualquer caule tem sempre a possibilidade de ter penduradas o mesmo número de folhas. No exemplo anterior , no primeiro sub caule 3 (ou 4, ou 5, ou 6) aparecem penduradas as folhas 0, 1, 2, 3 e 4, enquanto que no segundo sub caule 3 (ou 4, ou 5, ou 6) aparecem penduradas as folhas 5, 6, 7, 8 e 9). Uma outra possibilidade seria considerar classes de amplitude 2, fazendo cada caule dividido em 5 sub caules e cabendo a cada sub caule 2 folhas (repare-se com a analogia com a construção do histograma, em que considerámos as classes com igual amplitude).

A esta amplitude de classe é usual chamar comprimento de linha.

Não existe no Excel uma representação imediata para a construção de um caule-e-folhas, pelo que vamos utilizar um processo desenvolvido por Neville Hunt (Hunt, 2001), para o Excel:

• **1º passo** – Insira os dados na coluna C, começando na célula C2; se não estiverem ordenados, ordene-os por ordem crescente;

• **2º passo** – Insira na célula E1 o valor que deseja para o comprimento de linha: 10, 5 ou 2 ou uma potência de 10, destes valores;

• **3º passo** – Na célula A2 escreva a seguinte fórmula = INT(C2/E$1)*E$1 e replique-a tantas vezes quantos os dados inseridos no 1º passo, na coluna C;

• **4º passo** – Na célula B2 escreva o valor 1. Na célula B3 escreva a fórmula = IF (A3=A2; B2+1; 1) e replique a fórmula, tantas vezes quantos os dados inseridos no 1º passo, na coluna C;

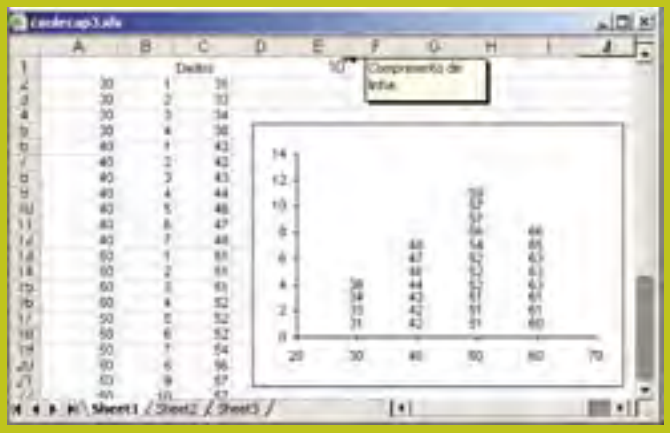
• **5º passo** – Seleccione as células das colunas A, B e C com os resultados obtidos nos passos anteriores e no módulo Chart Wizard (Assistente de Gráficos) escolha Bubble;

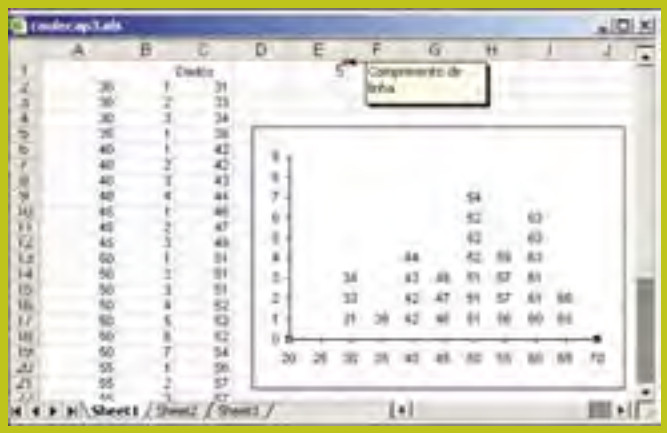
• **6º passo** – Faça um duplo clique numa das bolas representadas e na janela Format data Series (ou clique com o botão direito do rato e seleccione Format data Series) - seleccione Patterns: - Border: None - Area: None - Data Labels: Show bubbles sizes - OK;

• **7º passo** – Faça um duplo clique numa das "Data labels" (ou clique com o botão direito do rato e seleccione Format Data Labels), e na janela Format Data Labels, em Alignment: - Label Position: Centre - OK;

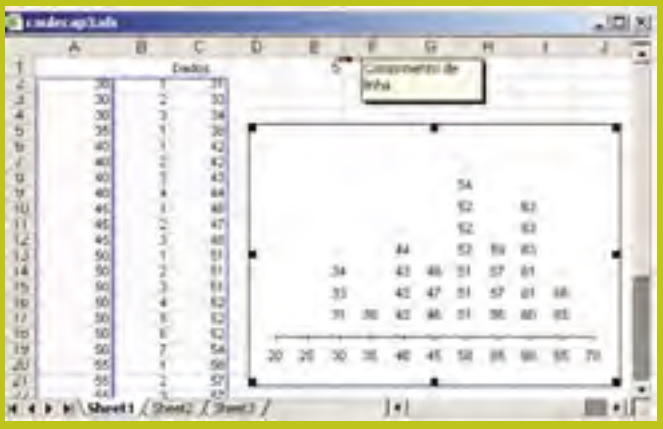
• **8º passo** – Clique numa das linhas horizontais que atravessam o gráfico e apague-as com a tecla Delete. Faça o mesmo ao fundo cinzento, seleccionando-o e carregando na tecla Delete. Apague também a legenda.

• **9º passo** – Formate convenientemente os eixos.





Na folha de Excel, se mudarmos o valor do comprimento de linha para 5, aparece de imediato a seguinte representação (aparte uma formatação adequada do eixo dos xx):



Repare-se que, embora as notações usadas para os caules e as folhas não sejam idênticos aos da representação inicialmente considerada, feita sem o recurso ao Excel, o aspecto gráfico é o mesmo. Para uma maior semelhança, seleccionámos o eixo dos yy e fizemos Delete:

2.3.3.3 – Diagrama de extremos e quartis

Esta representação, muito simples, mas bastante elucidativa ao realçar a informação contida nos dados, no que diz respeito à simetria e variabilidade, pressupõe que se calculem algumas estatísticas necessárias para a sua construção.

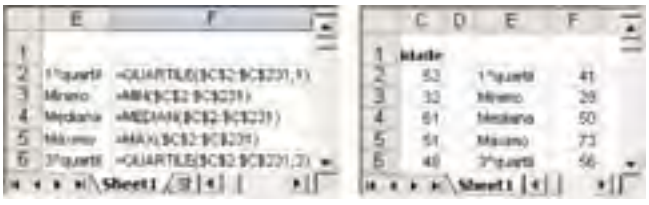
Mais uma vez estamos perante uma representação gráfica cuja construção, por meio do Excel, necessita de alguns "truques". Assim, o primeiro passo para uma dessas construções, consiste em representar, adequadamente, numa folha de Excel, as estatísticas Mínimo, Máximo. 1.º e 3.º quartis e mediana.

Exemplo 2.3.7 – Construa um diagrama de extremos e quartis para a variável idade dos deputados do ficheiro Deputados.xls.

Construção do diagrama de extremos e quartis, em Excel:

1. Utilizando o Excel, começam por se calcular as estatísticas necessárias1, que se apresentam da seguinte forma:

2. Seleccionar as células que contêm as



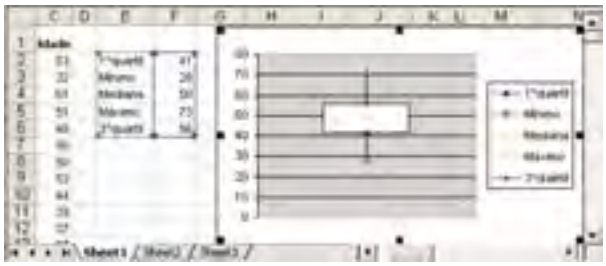
estatísticas, assim como as suas etiquetas: E2 a F6;

3. No módulo Chart Wizard (Assistente de Gráficos) seleccionar:
Line -Seleccionar Line with markers displayed at each data value- Clicar Next -Seleccionar Series in Rows Clicar -Finish

4. Clicar com o botão direito do rato num dos pontos. Seleccionar:

Format Data Series -Seleccionar Options Escolher -High-low lines e Up-down bars; Ajuste à sua escolha Gap width; OK

5. Arranjar "esteticamente" o gráfico:



Esta representação de um conjunto de dados, num diagrama de extremos e quartis, é especialmente indicada para comparação de várias amostras, como se exemplifica a seguir:

Exemplo 2.3.8 – Registou-se o comprimento, em centímetros, das asas de 32 melros-fêmeas e 25 melros-macho, tendo-se obtido os seguintes resultados:

Melro-fêmea -

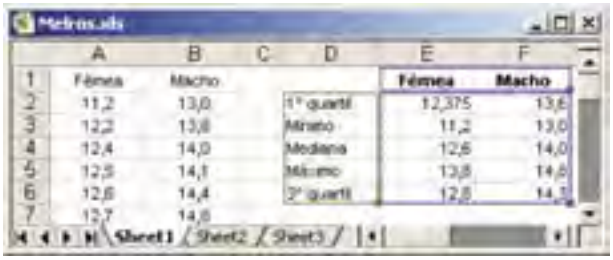
11,2	11,7	12,0	12,1	12,2	12,2	12,3
12,3	12,4	12,4	12,4	12,4	12,5	12,5
12,5	12,5	12,6	12,6	12,7	12,7	12,7
12,8	12,8	12,8	12,8	13,0	13,1	13,1
13,2	13,5	13,6	13,8			

Melro-macho -

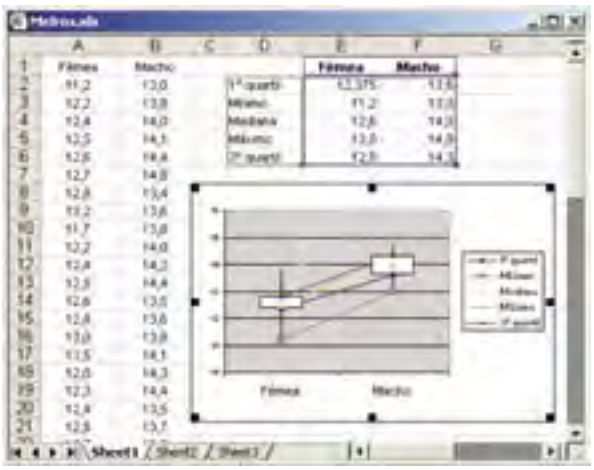
13,0	13,4	13,5	13,5	13,5	13,6	13,6
13,7	13,8	13,8	13,8	13,9	14,0	14,0
14,1	14,1	14,1	14,2	14,3	14,3	14,4
14,4	14,4	14,4	14,8			

Utilizando uma representação adequada, compare os dois conjuntos de dados.

Começámos por introduzir os dados numa folha de Excel, calculando de seguida as características amostrais relevantes para a construção de um diagrama de extremos e quartis:

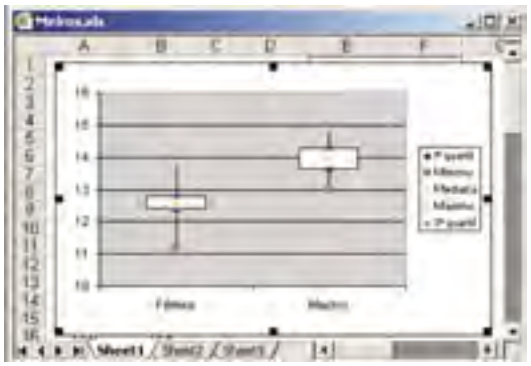


Para proceder à construção do diagrama de extremos e quartis comece por seleccionar as células que contêm os valores das características amostrais, assim como as etiquetas (células D1 a F6), e proceda de acordo com as instruções dadas no exemplo anterior. Depois de formatar convenientemente o eixo dos yy, obterá a seguinte representação:



As linhas a unir as caixas podem ser removidas, seleccionando cada uma, com o botão direito do rato e seleccionando sucessivamente:

Format-Data Series- Patterns-Line: None - Ok



O gráfico anterior é bastante elucidativo na medida em que mostra que o tamanho das asas do melro-macho é, de um modo geral superior ao do melro-fêmea, apresentando ainda uma maior variabilidade.

2.4 – Alguns exemplos

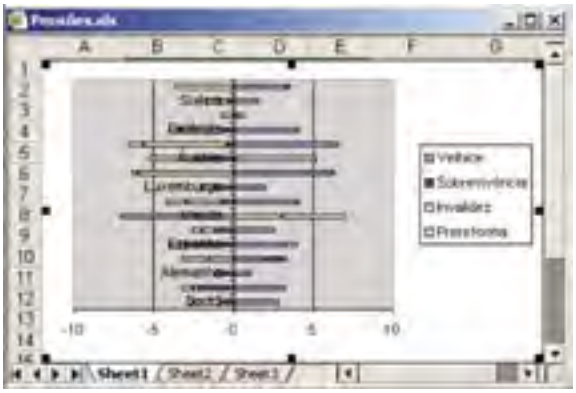
A seguir apresentamos alguns exemplos, sobre a forma de projectos, para os quais podemos utilizar vários tipos de representações gráficas, algumas já referidas anteriormente, outras introduzidas pela primeira vez, mas que apresentam realização imediata com o Excel.

Projecto 1

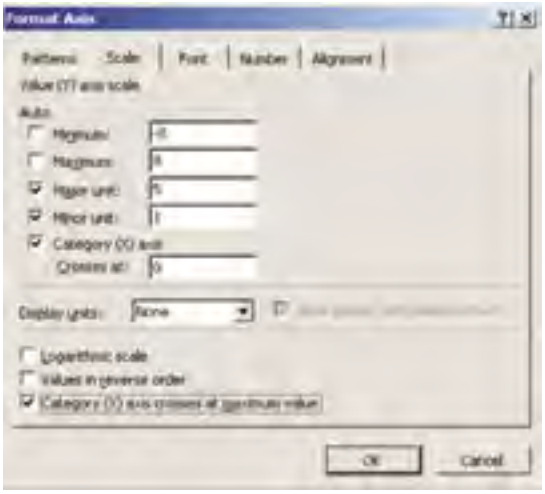
Neste projecto são apresentados alguns dados relativamente à Modificação da Estrutura das Categorias de Pensões entre 1993 e 2001 (em pontos percentuais) (Eurostat – Statistiques en bref – Population et conditions sociales, 8/2004):

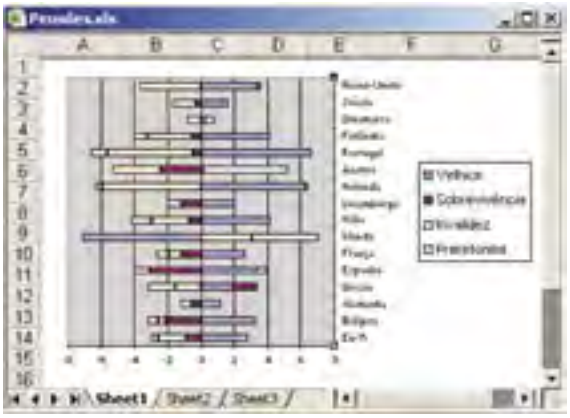
	Velhice	Sobrevivência	Invalidez	Pre-reforma
Eu-15	2,8	-0,8	-1,7	-0,4
Bélgica	3,2	-2,1	-0,4	-0,7
Alemanha	1,1	-0,5	-0,7	0,1
Grécia	1,9	1,5	-1,5	-1,7
Espanha	3,3	-3	-1	0,7
França	2,6	-1,1	-0,8	-0,7
Irlanda	-7,1	0	3,1	4
Itália	4,1	-0,7	-2,2	-1,3
Luxemburgo	2	-1,1	0	-0,9
Holanda	6,2	0,2	-5,9	-0,4
Áustria	0,2	-2,4	-2,9	5,1
Portugal	6,6	-0,4	-5,2	- 1
Finlândia	4,1	-0,5	-2,7	-0,8
Dinamarca	0,3	0	-0,8	0,5
Suécia	1,4	-0,3	-1,4	0,3
Reino-Unido	3,3	0,3	-3,6	0

Uma forma adequada para representar estes dados, é através de um diagrama de barras, nomeadamente barras horizontais, seleccionando na opção Chart o 2º tipo da opção Bar:

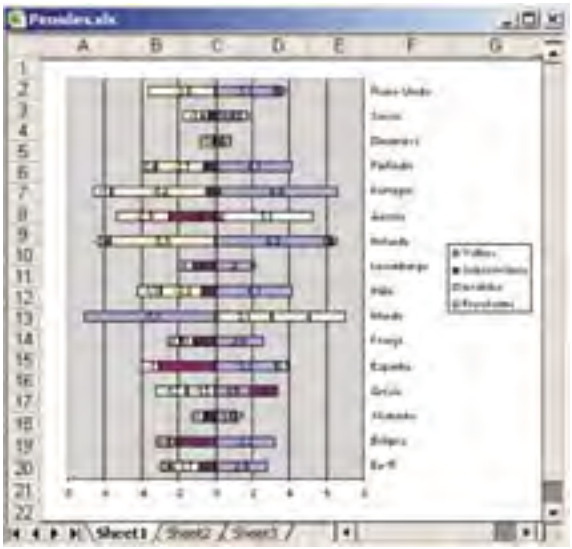


Vamos fazer alguma "cosmética" na representação gráfica anterior, nomeadamente mudando a escala para –8 a 8 e fazendo com que as legendas não se sobreponham ao gráfico:





Podemos ainda acrescentar sobre o gráfico os valores quando houver conveniência em disponibilizar esta informação. Para isso basta seleccionar Chart Options - Data labels - Show Value:

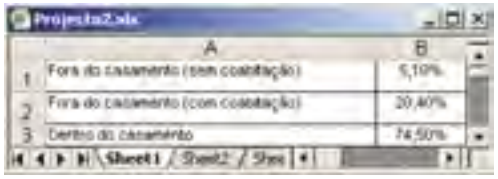


Projecto 2

Entre os dois últimos recenseamentos da população portuguesa, os Censos 91 e os Censos 2001, realizados, respectivamente, em 15 de Abril de 1991 e 12 de Março de 2001, verificou-se que a população residente no território nacional passou de 9.867.147 para 10.356.117 habitantes, a que corresponde um acréscimo de 4.8%. Na generalidade das regiões verificou-se um aumento da população, com excepção das regiões do Alentejo e Madeira. Partindo dos resultados censitários definitivos, estimou-se a população residente em 31 de Dezembro de 2002 em 10.407.500 indivíduos, dos quais 5.030.200 do sexo masculino.

Apresentam-se a seguir algumas tabelas e gráficos com alguns indicadores (www.ine.pt):

1.Nados-vivos segundo a filiação – 2002



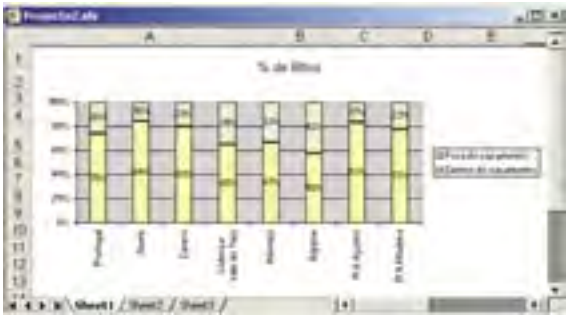
	A	B
1	Fora do casamento (sem coabitação)	5,10%
2	Fora do casamento (com coabitação)	20,40%
3	Dentro do casamento	74,50%

Uma representação adequada para a tabela anterior é o diagrama circular. Assim, vamos seleccionar Chart - Pie - 1ºsubtipo - Next - Next - Data labels - Show label and percent - Finish:

Nados-vivos segundo a filiação, por regiões:

	A	B
6	Portugal	74,50%
7	Norte	83,80%
8	Centro	80,00%
9	Lisboa e Vale do Tejo	64,70%
10	Alentejo	68,70%
11	Algarve	57,60%
12	R.A.Açores	83,10%
13	R.A.Madeira	77,30%

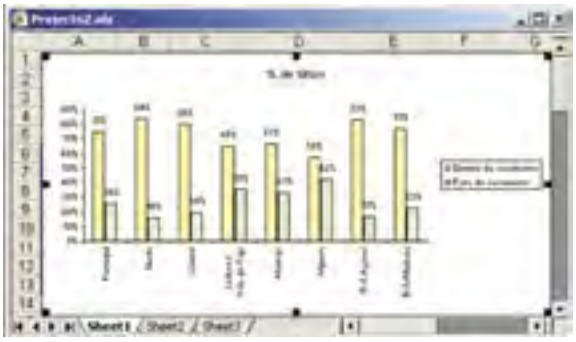
Acrescentámos à tabela anterior uma outra coluna – células C6 a C13, com os filhos fora do casamento e decidimos aqui optar por uma representação em barras verticais. Assim, depois de seleccionar as células A5 a C13, fizemos Chart - Column - 3ºsubtipo- Next - Next - Data labels - Show value - Titles - Chart title - % de filhos - Finish:



Observação: Foi possível optarmos pela representação gráfica anterior, uma vez que os dados das duas características em estudo somavam 100%.

Outra representação possível obtém-se seleccionando Chart - Column - 1ºsubtipo - Next

- Data labels - Show value - Titles - Chart title - % de filhos - Finish:

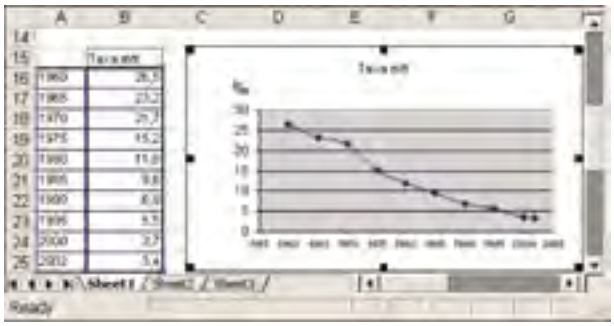


2. Taxa de mortalidade fetal tardia (Taxa mft) (28 ou mais semanas de gestação):

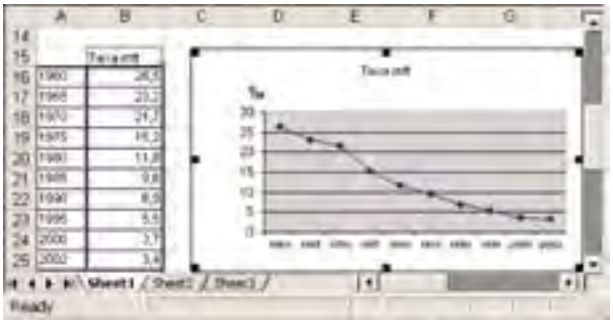
1960	26.5‰
1965	23.2‰
1970	21.7‰
1975	15.2‰
1980	11.8‰
1985	9.6‰
1990	6.9‰
995	5.5‰
2000	3.7‰
2002	3.4‰

Introduzimos a tabela anterior numa folha de Excel e antes de procedermos a uma representação gráfica passámos os pontos para vírgulas e retirámos a permilagem, não reconhecida no Excel.

Seguidamente depois de seleccionar as células A15 a B25, seleccionámos Chart - XY(Scatter) - 2ºsubtipo - Next - Next - Legend:Retirar a selecção de Show Legend - Titles - ‰ em Value(Y) - Finish:



Chamamos a atenção para o facto de ser possível obter uma representação aparentemente semelhante à anterior utilizando a opção Chart - Line - 4ºsubtipo - Next - Next - Legend - Retirar a selecção de Show Legend - Titles - ‰ em Value(Y) - Finish:



Repare-se, no entanto, que a representação anterior não está correcta, pois a variável tempo do eixo dos xx está a ser interpretada como uma variável qualitativa e não quantitativa como deveria ser. Assim, o intervalo entre 1995 e 2000 é igual ao intervalo entre 2000 e 2002, o que obviamente não está correcto.

3. Taxa de mortalidade infantil

1960	77.5‰
1965	64.9‰
1970	58.0‰
1975	38.9‰
1980	24.3‰
1985	17.8‰
1990	10.9‰
1995	7.5‰
2000	5.5‰
2002	5.0‰

A representação gráfica dos dados desta tabela pode ser idêntica à do ponto anterior.

4. Casamentos segundo a forma de celebração

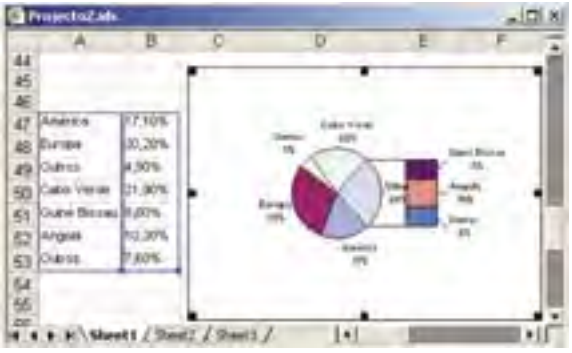
Para esta tabela pode-se usar uma representação gráfica idêntica à usada no ponto 1, para mostrar a percentagem de filhos dentro e fora do casamento.

Unidade %	Civil	Católico
1960	9.2	90.8
1965	11.8	88.2
1970	13.4	86.6
1975	20.0	80.0
1980	25.3	74.7
1985	25.9	74.1
1990	27.5	72.5
1995	31.2	68.8
2000	35.2	64.8
2002	37.5	62.5

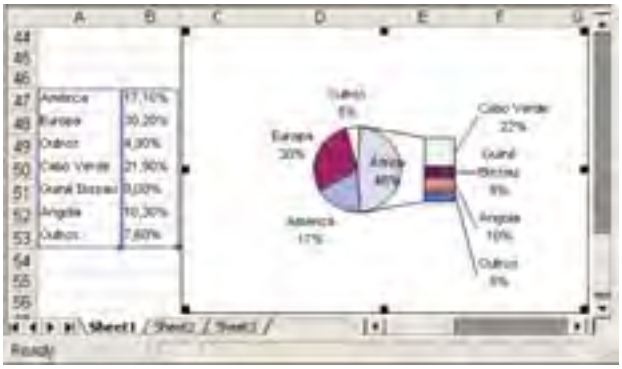
5. População estrangeira com estatuto legal de residente segundo a nacionalidade

América	17,1%	**África**	Angola	10,3%
Europa	30,2%		Cabo Verde	21,9%
África	47,8%		Guiné Bissau	8,0%
Outros	4,9%		Outros	7,6%

Para fazer uma representação destes dados recorremos a um diagrama em Pie (circular), mas num subtipo especial que permite visualizar a forma como África está repartida. Assim considere-se a seguinte tabela em Excel, ocupando as células A47 a B53 e seleccione-se Chart - Pie - 6ºsubtipo - Next - Next - Data labels - Show label and percent - Legend - Retirar a selecção de Show Legend - Finish:



Para incluir Cabo Verde na parte direita do gráfico carregar com o botão direito do rato em qualquer parte do gráfico e seleccionar Format Data Series - Second plot contais the last: 4 - Finish. Finalmente substituir Other (com 48%) por África:

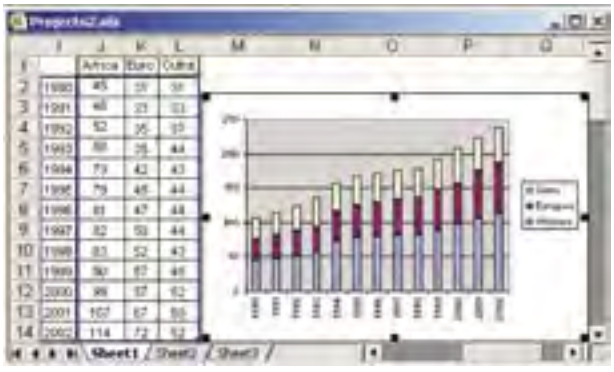


Para representar os dados da tabela seguinte:

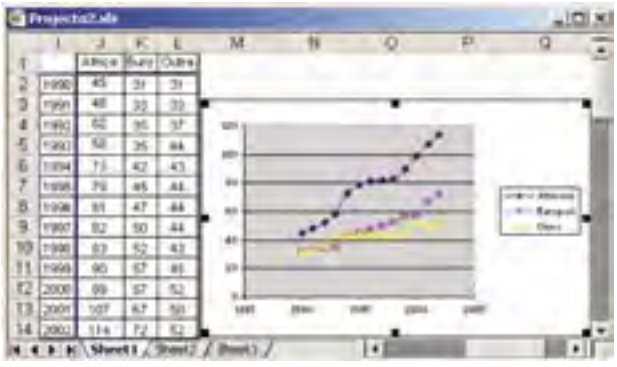
	Africana(1)	Europeia	Outra
1990	45	31	31
1991	48	33	33
1992	52	35	37
1993	58	35	44
1994	73	42	43
1995	79	45	44
1996	81	47	44
1997	82	50	44
1998	83	52	43
1999	90	57	45
2000	99	57	52
2001	107	67	50
2002	114	72	52

(1)Unidade 10^3

Podemos considerar o 2.º subtipo de Column (chama-se a atenção para que neste caso não seria correcto utilizar o 3.º subtipo de Column, uma vez que estamos os dados estão em número absoluto e não em percentagem):



ou o 2.º subtipo de XY(Scatter):



Como vimos há várias representações gráficas para os dados de uma mesma tabela, umas mais sugestivas do que outras. Desde que a representação escolhida esteja correcta, deixa-se a liberdade da escolha ao "artista" que está a organizar e a reduzir os dados.

3. Características amostrais. Medidas de localização e dispersão

3.1 - Introdução

No módulo de Estatística foram apresentadas as medidas ou estatísticas que se utilizam para resumir a informação contida nos dados. Destas medidas, destacam-se as medidas de localização, nomeadamente as que localizam o centro da amostra, e as medidas de dispersão, que medem a variabilidade dos dados.

Neste capítulo não nos debruçaremos sobre as propriedades destas medidas, já apresentadas no módulo referido anteriormente, abordando sobretudo a forma de as calcular, utilizando o Excel. Convém desde já adiantar que este é um trabalho grandemente facilitado pelo facto de existirem funções no Excel que nos dão directamente estas medidas.

Para facilidade de exposição vamos representar a amostra de dimensão n por x1, x2, …, xn onde x1, x2, …, xn representam, respectivamente, os resultados da 1ª observação, da 2ª observação, da n-ésima observação, a serem recolhidas, não pressupondo qualquer ordenação.

3.2 – Medidas de localização

Como medidas de localização, vamos apresentar a média, mediana e quartis.

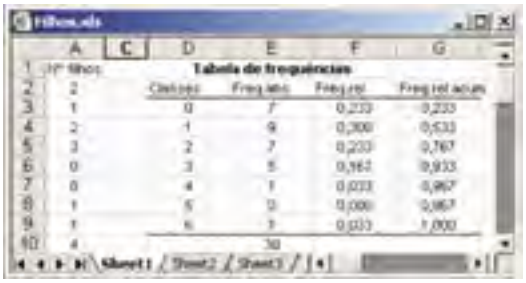
3.2.1 – Média

A média é uma medida de localização do centro da distribuição dos dados. Dada a amostra x1, x2, …, xn, a média representa-se por x e obtém-se adicionando todos os elementos e dividindo o resultado por n. Em Excel, determina-se a média através da função AVERAGE (), que retorna a média aritmética dos seus argumentos, que podem ser números ou endereços de células.

Exemplo 3.2.1 – Retomemos a amostra do exemplo 2.3.2, constituída pelo número de filhos de 30 deputados:

2, 1, 2, 3, 0, 0, 1, 1, 4, 1, 2, 1, 0, 0, 0, 2, 3, 1, 1, 6, 3, 1, 3, 2, 0, 1, 2, 0, 2, 3

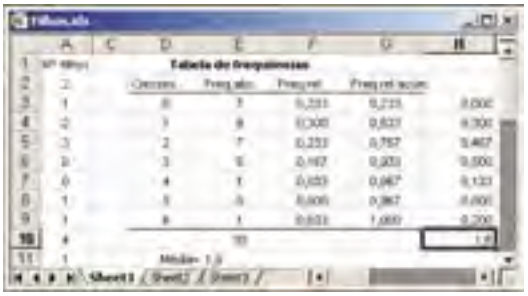
Calcule a média da amostra. Considerámos o ficheiro Filhos.xls, constituído no exemplo 2.3.2, em que os elementos de que pretende calcular a média ocupam as células A2 a A31:



Para calcular a média pretendida, assim como para qualquer outro conjunto de dados de tipo discreto, podemos proceder de dois modos, quer considerando os dados originais, quer agrupados.

1- Cálculo da média, a partir dos dados originais, utilizando a função AVERAGE(): Colocar o cursor na célula onde se pretende colocar a média, por exemplo a célula E11, e inserir a função AVERAGE(A2:A31) – os argumentos desta função são os endereços onde estão os elementos da amostra. Como resultado obtém-se o valor 1,6, que se apresenta na figura seguinte.

2- Cálculo da média, a partir dos dados agrupados: Adicionar à tabela de frequências uma nova coluna com o produto dos valores que constituem as classes, pelas respectivas frequências relativas (Células H3 a H9) e somar os valores obtidos (Célula H10):



No caso de dados discretos, como é o caso anterior, o valor da média é o mesmo, quer seja calculada utilizando os dados originais, quer os dados agrupados (utilizando as frequências relativas), em que as classes do agrupamento são os diferentes valores que surgem na amostra. O mesmo não acontece no caso de dados contínuos, como exemplificamos a seguir.

Exemplo 3.2.2 – Calcule a média das idades dos deputados do ficheiro Deputados.xls.

Para obter a média das idades procede-se como no primeiro caso do exemplo anterior, a partir dos dados originais. Estes dados encontram-se nas células C2 a C231 do ficheiro Idade.xls, Inserindo a função AVERAGE(C2:C231) na célula L13, obtemos o valor de 48,66 anos.

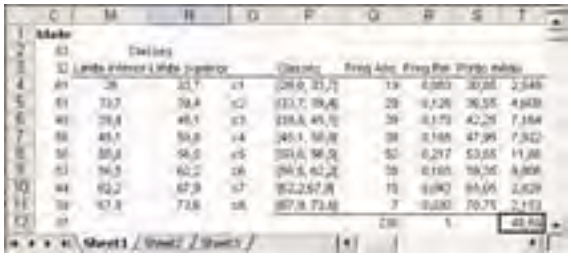
Admitindo que não dispúnhamos dos dados originais, mas apenas de uma tabela de frequências com os dados agrupados, vejamos como obter um valor aproximado para a média.

Reportando-nos ainda ao ficheiro Idade.xls, consideremos a tabela de frequências que serviu para agrupar os dados. Para obter um valor aproximado para a média, procedemos da seguinte forma:

• Adicionar à tabela de frequências uma nova coluna com os pontos médios dos intervalos de classe, que se obtêm fazendo a semi-soma dos limites dos intervalos (células S4 a S11);

• Adicionar à tabela uma nova coluna com os produtos dos pontos médios dos intervalos de classe, pelas frequências relativas respectivas (células T4 a T11);

• Somar os resultados das células T4 a T11 (célula T12):



Repare-se que o valor obtido de 48,69 para a média, é muito próximo do verdadeiro valor obtido com os dados originais.

3.2.2 – Mediana

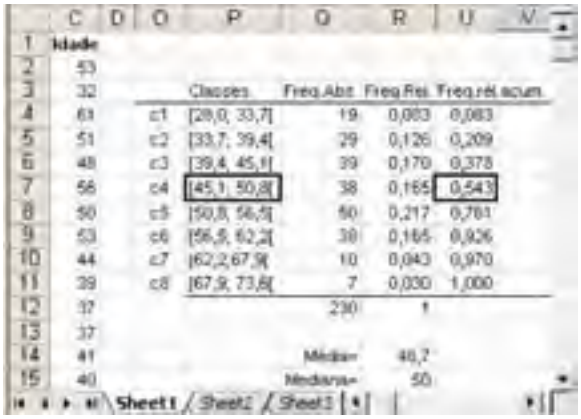
Outra medida de localização do centro dos dados é a mediana. Ordenados os elementos da amostra, a mediana, m, é o valor (pertencente ou não à amostra) que a divide ao meio, isto é, 50% dos elementos da amostra são menores ou iguais a m e os restantes 50% são maiores ou iguais a m. Em Excel, determina-se a mediana através da função MEDIAN(), que retorna a mediana dos seus argumentos, que podem ser números ou endereços de células.

Exemplo 3.2.3 – Calcule a mediana das idades dos deputados. Compare com o valor obtido para a média e diga o que poderia concluir da forma como os dados se distribuem.

Voltando ao ficheiro Idade.xls, utilizado no exemplo anterior, insira na célula R15 a função Median(C2:C231) e obterá como retorno, o valor 50, como se verifica na figura seguinte. O valor obtido para a mediana é ligeiramente superior ao da média, pelo que podemos admitir que a distribuiução é aproximadamente simétrica, com um ligeiro enviesamento para a esquerda.

Se os dados se apresentarem agrupados, já vimos na secção 3.2.2 do capítulo 2, um processo de obter a mediana através da função cumulativa. No entanto, não é necessário construir esta função para obter um valor aproximado para a mediana, pois este pode ser obtido a partir da tabela de frequências, utilizando ainda o processo de interpolação.

Exemplo 3.2.4 – A partir do agrupamento considerado, no exemplo 2.3.3, para a variável idade, calcule um valor aproximado para a mediana. Adicionando à tabela de frequências uma nova coluna com as frequências relativas acumuladas, verificamos que a mediana se encontra na classe [45,1; 50,8[, pois a frequência acumulada de 50% é atingida nesta classe:



	C	D	O	P	Q	R	U	V
1	Idade							
2	53							
3	32			Classes	Freq.Abs	Freq.Rel	Freq.rel.acum	
4	63		c1	[28,0; 33,7[	19	0,083	0,083	
5	51		c2	[33,7; 39,4[	29	0,126	0,209	
6	48		c3	[39,4; 45,1[	39	0,170	0,378	
7	56		c4	[45,1; 50,8[	38	0,165	0,543	
8	50		c5	[50,8; 56,5[	50	0,217	0,761	
9	53		c6	[56,5; 62,2[	38	0,165	0,926	
10	44		c7	[62,2; 67,9[	10	0,043	0,970	
11	39		c8	[67,9; 73,6[	7	0,030	1,000	
12	37				230	1		
13	37							
14	41				Média=	48,7		
15	40				Mediana=	50		

Admitindo que a frequência se distribui uniformemente sobre a amplitude de classe, isto é, a frequência 0,165 se distribui uniformemente sobre o intervalo de amplitude 5,7, resolvendo a equação de proporcionalidade

$$\frac{0{,}165}{0{,}122}=\frac{5{,}7}{x} \qquad x=\frac{0{,}122\times 5{,}7}{0{,}165}=4{,}2$$

onde 0,122=0,5-0,378, obtemos para a mediana o valor aproximado 45,1 + 4,2 = 49,3.

Chamamos a atenção para o seguinte facto: o valor (aproximado) que se obtém para a mediana, depende do agrupamento que se fizer para os dados, pelo que agrupamentos diferentes darão origem a valores diferentes, embora não difiram muito uns dos outros (Lembramos que o valor da mediana apresentado na figura anterior foi obtido a partir dos dados não agrupados). .

3.2.3 – Quartis

Os quartis, 1.º e 3.º, definem-se de forma idêntica à mediana, mas considerando em vez da percentagem de 50%, respectivamente 25% para o 1º quartil, Q1, e 75% para o 3.º quartil, Q3.

Há vários processos para a determinação dos quartis, nem sempre conduzindo aos mesmos resultados. Este facto não é preocupante, pois de um modo geral nas situações que têm interesse em estatística, as amostras têm dimensão suficientemente elevada de forma que os diferentes processos conduzem a valores próximos.

Em Excel a determinação dos quartis faz-se utilizando a função QUARTILE(array;quart):

Repare que a função Quartile(array;quart) tem dois argumentos, em que o primeiro argumento é o endereço das células de que queremos calcular o quartil e o segundo argumento pode tomar vários valores, conforme a medida de localização, de entre as seguintes, que nos interesse calcular:

0 – mínimo

1 – 1º quartil

2 – mediana

3 – 3º quartil

4 – máximo

Assim, esta função, além do 1.º e 3.º quartis, a que estão associadas as percentagens 25% e 75%, respectivamente, ainda calcula a mediana, a que está associada a percentagem de 50% e o mínimo e máximo com percentagens associadas de 0% e 100%.

Exemplo 3.2.5 – Escolha os primeiros 15 elementos da variável Idade, do ficheiro Idade.xls. Obtenha o 1º e 3º quartis. Os primeiros 15 elementos são os seguintes:

53 32 61 51 48 56 50 53 44 39 37 37 41 40 40

Utilizando a função QUARTILE(C2:C16;1) e QUARTILE(C2:C16;3), obtemos Q1=39,5 e Q3=52.

Se utilizar o processo que aprendeu no módulo de Estatística, nomeadamente considerando o 1.º quartil como a mediana da primeira parte da amostra, quando esta é dividida pela mediana, depois de ordenar a amostra e tendo em conta que a mediana é 44, temos para 1.º quartil o

44 48 50 51 53 53 56 61 valor 39, se não considerarmos a mediana como pertencente a nenhuma das partes, ou 39,5 se considerarmos a mediana pertencente às duas partes. Para o 3º quartil obteremos, respectivamente o valor **53** ou **52**, utilizando a mesma metodologia.

Exemplo 3.2.5 (cont) – Repita o exemplo anterior, considerando amostras de dimensão 12 e 13.

Considere agora só os primeiros 12 elementos. Como a mediana é 49, o 1º quartil – mediana da 1ª parte da amostra, será (37+39)/2=38, enquanto que o 3º quartil será (53+53)/2=53.

50 51 53 53 56 61

Utilizando o Excel, os valores que se obtêm são Q1=38,5 e Q3=53.

Considere agora os primeiros 13 elementos. Como a mediana é 48, o 1º quartil – mediana da 1ª parte da amostra, será (37+39)/2=38, enquanto que o 3º quartil será (53+53)/2=53, não considerando a mediana como pertencente a nenhuma das partes. Caso contrário, teremos Q1=39 e Q3=53.

48 50 51 53 53 56 61

Utilizando o Excel, os valores que se obtêm são Q1=39 e Q3=53.

Obervação: Repare que os valores que se obtêm para os quartis, recorrendo ao excel não são iguais aos que se obtiveram sem utilizar o Excel. Efectivamente não existe uniformidade na forma de calcular os quartis, como já havíamos referido anteriormente, embora os resultados obtidos satisfaçam a definição de quartis. Exemplificando com a mediana, repare que pela definição de mediana, quando o número de elementos da amostra é par, podemos considerar para mediana qualquer valor compreendido entre os dois elementos médios da amostra ordenada! Não é costume deixar esta opção ao critério de cada um e considera-se a semi-soma desses elementos médios.

Voltando aos quartis, pode verificar que, no Excel, o 1.º quartil corresponde à observação de ordem (n+3)/4, procedendo-se a uma interpolação, quando necessário (Sugestão – Tente descobrir como é calculado o 3º quartil no Excel).

3.3 – Medidas de dispersão

Continuando na mesma linha de apresentação das medidas de localização, também agora não nos vamos preocupar com as propriedades das medidas de dispersão, pois admitimos que estas já foram estudadas no módulo de Estatística. Debruçar-nos-emos sobre o seu cálculo, utilizando o Excel.

A seguir apresentaremos o cálculo da variância, desvio padrão e amplitude inter-quartil.

3.3.1 – Variância e desvio-padrão

A variância de um conjunto de dados obtém-se fazendo a média dos quadrados dos desvios dos dados, relativamente à média.

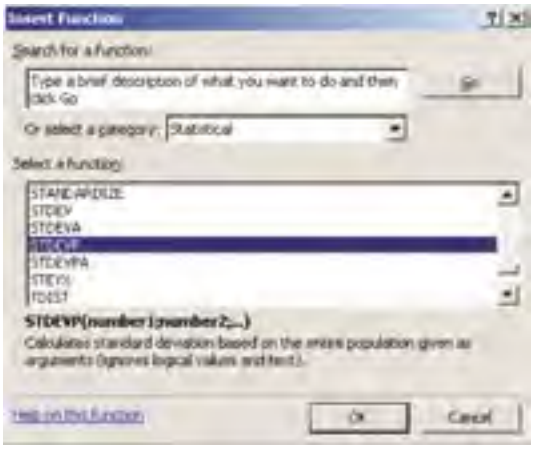
O Excel, tal como as máquinas de calcular, dispõe de duas funções para calcular a variância, conforme estejamos a calcular a variância populacional (parâmetro) ou a variância amostral (estatística). Resumimos no quadro seguinte a situação de estarmos a calcular parâmetros ou estatísticas.

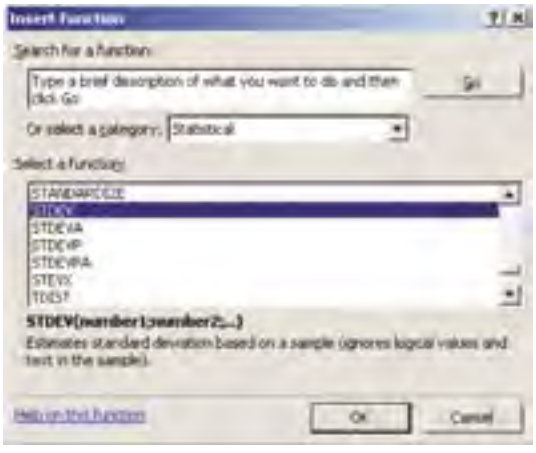
População de **N** elementos	**Amostra** de **n** elementos
$x_1, x_2, \ldots, x_N$	$x_1, x_2, \ldots, x_n$
Valor médio $\mu = \frac{x_1 + x_2 + \ldots + x_N}{N}$	**Média** $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$
Variância populacional $\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \ldots + (x_N - \mu)^2}{N}$	**Variância** amostral $s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}$
Desvio padrão populacional σ	**Desvio padrão amostral** s

Em Excel as funções utilizadas para calcular a variância populacional e amostral, são respectivamente VARP() e VAR(). Como argumento utiliza-se a sequência de números de que se quer calcular a variância, ou o endereço das células que os contêm.

Por exemplo, no caso da população dos deputados, que temos vindo a estudar, temos informação completa sobre a variável Idade, pelo que a fórmula que deve ser utilizada para obter a variância é a VARP, isto é, esta fórmula dá-nos a variância populacional. Se só dispuséssemos da idade de alguns deputados, isto é, uma amostra da população em estudo, então a fórmula a utilizar seria a VAR, que dá a variância amostral. A maneira de calcular as duas variâncias é idêntica, diferindo unicamente no seguinte ponto: enquanto que no caso da variância populacional se divide a soma dos quadrados dos desvios pelo número de parcelas, no caso da variância amostral divide-se a soma dos quadrados dos desvios pelo número de parcelas menos uma.

O desvio padrão obtém-se fazendo a raiz quadrada da variância ou utilizando uma função própria. Como é evidente, existem também duas fórmulas para o calcular, obtendo-se o desvio padrão populacional ou amostral, conforme a fórmula utilizada:



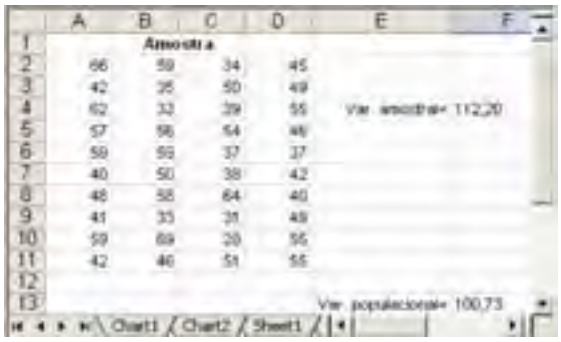


Repare-se que quando se selecciona a função que se quer utilizar, aparece a descrição do que é que a função faz.

Exemplo 3.3.1 – A partir do ficheiro Idade.xls, seleccione uma amostra aleatória simples de dimensão 40. Calcule a variância e o desvio padrão da amostra obtida. Calcule de seguida a variância da população constituída pelas idades dos 230 deputados e compare com a variância da amostra obtida anteriormente.

Utilizando o processo descrito em 1.3.1.2, seleccionámos uma amostra de 40 elementos que posteriormente colocámos nas células A2 a D11, de uma nova folha de Excel. Colocando agora o cursor na célula onde pretendemos colocar a variância, por exemplo na célula F4, inserimos a função VAR (A2:D11) e a função retorna um valor aproximadamente igual a 112, para a variância da amostra.

Para calcular a variância da população das idades, inserimos na célula F5 a função VARP(Sheet1!C2:C231), obtendo-se um valor aproximadamente igual a 101:



Comparando as variâncias, vemos que não são iguais, o que já seria de esperar, uma vez que a variância amostral foi obtida a partir de 40 dos 230 dados e é uma estimativa da variância populacional. Se recolhermos outra amostra, também de 40 elementos, não esperamos obter o mesmo valor para a estimativa. Esperamos sim, obter valores aproximados.

Para calcular o desvio padrão, ou se calcula a raiz quadrada (positiva) do valor da variância, ou se utilizam as funções STDEV() ou STDEVP(), conforme se pretenda o desvio padrão amostral ou populacional. No nosso caso os desvios padrões amostral e populacional vêm, respectivamente, aproximadamente iguais a 10,6 e 10,0.

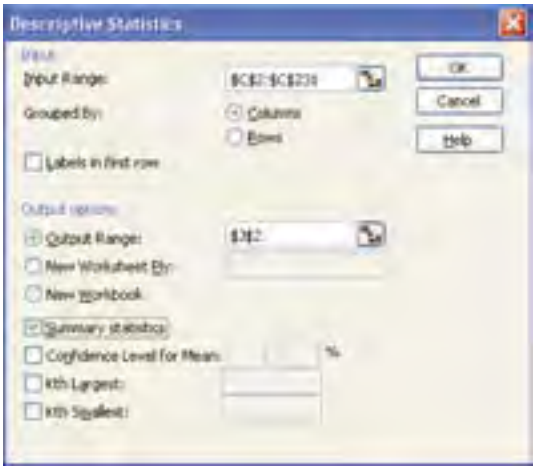
3.3.2 – Amplitude e amplitude interquartis

A amplitude da amostra (não confundir com dimensão da amostra), R, é a medida mais simples para medir a variabilidade, mas tem a grande desvantagem de ser muito sensível à existência na amostra, de uma observação muito pequena ou muito grande. Não existe, no Excel, uma função específica para a calcular, recorrendo-se às funções MAX() e MIN(). Já tivémos, aliás, oportunidade de utilizar estas funções quando necessitámos de calcular a amplitude de um conjunto de dados, para iniciar a construção de um histograma, com classes de igual amplitude. Uma medida mais resistente do que a anterior, é a amplitude interquartis que, como o nome indica, se define como a diferença entre os 1.º e 3.º quartis.

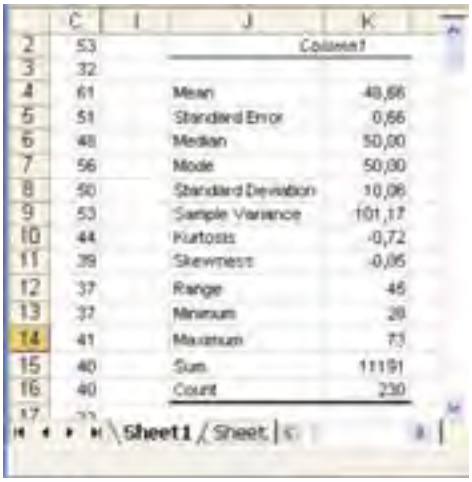
Exemplo 3.3.2 – Calcule a amplitude e a amplitude interquartis da amostra obtida no exemplo anterior. Como os elementos da amostra se encontram nas células A2 a D11, temos:
R = MAX(A2:D11) – MIN(A2:D11) = 69-28 = 41
Recorrendo à terminologia usada quando definimos os quartis, temos: Amplitude interquartis= QUARTILE(A2:D11;3) – QUARTILE(A2:D11;1) = 56,25-39,75=16,5.

3.4 – Função Descriptive Statistics

O Excel dispõe de uma função a que se acede seleccionando Tools - Data Aalysis - Descriptive Statistics - OK



e cujo resultado é o que se apresenta a seguir:



Algumas das funções já são conhecidas das secções anteriores. Chamamos a atenção para o facto de a variância das 230 idades não coincidir com o valor obtido na secção 3.3.1, uma vez que quando se considera um conjunto de dados e se pedem as Estatísticas descritivas, subentende-se que se está perante uma amostra e não da população toda! Por esta razão, a fórmula utilizada para o cálculo da variância é a da variância amostral.

As funções Standard Error, Kurtosis e Skewness saem fora do âmbito estas folhas, pelo que não entraremos em detalhe.

4. Dados bivariados

4.1- Introdução

No módulo de Estatística foi feita referência a dados bidimensionais, de tipo quantitativo. Quando dispomos de uma amostra de dados bivariados, a qual pode ser representada na forma (x1, y1), (x2, y2),..., (xn, yn), apresentamos esta informação através de uma representação gráfica a que se dá o nome de Diagrama de dispersão:

Diagrama de dispersão – É uma representação gráfica para os dados bivariados, em que cada par de dados (xi, yi), é representado por um ponto de coordenadas (xi, yi), num sistema de eixos coordenados.

Já vimos no capítulo 2, a forma de representar, em Excel, dados bivariados, utilizando a opção XY(Scatter). Não apresenta qualquer dificuldade a construção desta representação gráfica, uma vez que basta proceder da seguinte forma:

- Seleccionar as células que contêm os dados, organizados em 2 colunas;
- Carregar no ícone
- seleccionar a opção XY(Scatter) e o sub-tipo pretendido; Formatar convenientemente a representação obtida (retirar a legenda, retirar as linhas de grelha, etc).

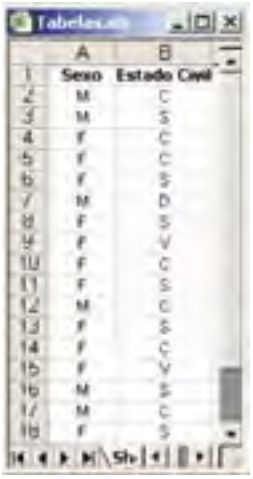
Quando se trata de dados qualitativos, não tem sentido proceder à representação gráfica dos dados através de um diagrama de dispersão. No entanto, é possível organizar essa informação na forma de tabelas de contingência (que aliás também podem ser usadas para dados quantitativos, quer discretos, quer contínuos, depois de proceder à sua discretização).

Vamos, neste capítulo, introduzir uma metodologia que utiliza uma ferramenta do Excel, a PivoTable, que além de permitir construir tabelas de contingência, também pode ser utilizada para proceder a agrupamentos de dados quantitativos.

4.2 – Tabelas de contingência

Suponhamos que estamos interessados em estudar a associação entre variáveis de tipo qualitativo como, por exemplo, sexo e religião. Uma forma de apresentar os dados, é utilizando tabelas de contingência.

Exemplo 4.2.1 – Uma empresa decidiu estudar o seu pessoal quanto ao estado civil e sexo. Representando por M e F as categorias da variável Sexo, e por C (casado(a)), S (solteiro(a)), D (divorciado(a)) e V (viúvo(a)), obteve a seguinte lista: (M,C), (M,S), (F,C), (F,C), (F,S), (M,D), (F,S), (F,V), (F,C), (F,S), (M,C), (F,S), (F,C), (F,V), (M,S), (M,C), (F,S) (Este exemplo é fictício e serve unicamente para introduzir o estudo das tabelas de contingência, pois os casos interessantes em Estatística envolvem amostras de maior dimensão).

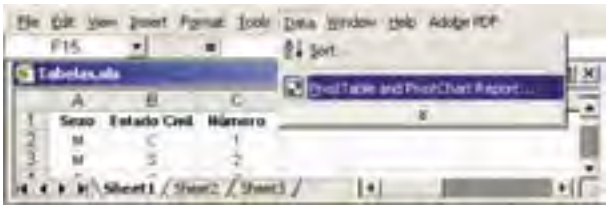


Começámos por introduzir estes dados numa folha de Excel, colocando nas células A1 e B1 os títulos, respectivamente Sexo e Estado Civil, e nas células A2 a A18 a informação sobre o sexo dos 17 elementos e nas células B2 a B18, o respectivo estado civil:

Introduzimos uma coluna auxiliar, a que chamámos Nº, com o número do par, a qual vai ser utilizada para exemplificar a construção de uma tabela de contingência, utilizando as PivotTable.

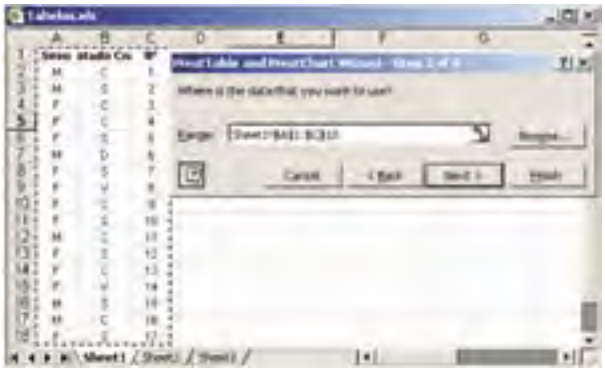
Para criar uma tabela, proceder do seguinte modo:

• No menu Data, clicar em PivotTable and PivotChart Report:

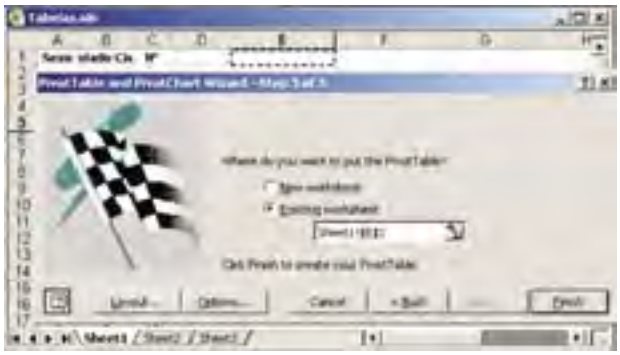


• No passo 1 da PivotTable and PivotTable Wizard, seguir as instruções, e clicar PivotTable à pergunta What kind of report do you want to create?

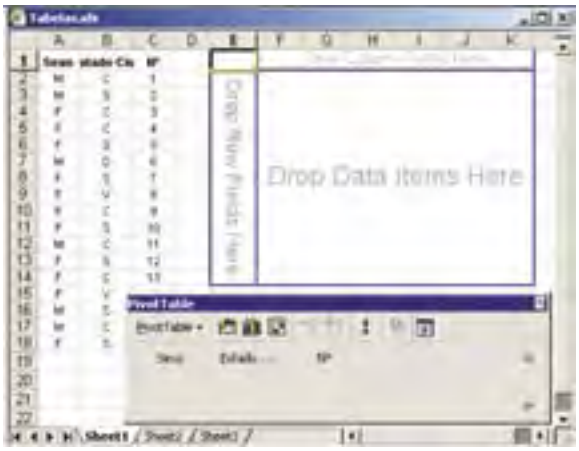
• No passo 2 seguir as instruções, seleccionando os dados que se pretende usar (não esquecer de seleccionar os títulos):



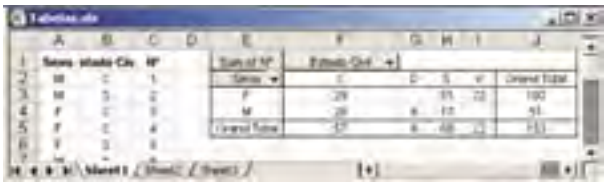
• No passo 3 seleccionar o lugar onde se pretende criar a tabela. Nós optámos por seleccionar a célula E1,



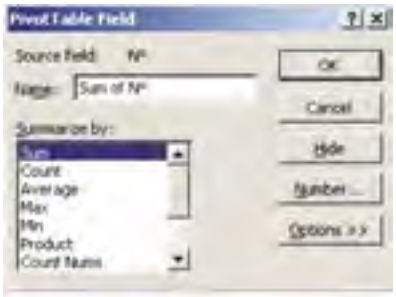
obtendo como resultado:

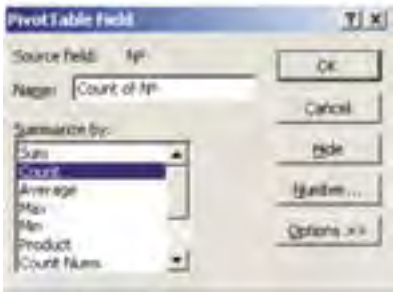


• Arrastar o botão Sexo da barra PivotTable, e colocá-lo (drop it) no campo Row; Arrastar o botão Estado civil da barra PivotTable, e colocá-lo (drop it) no campo Column; Arrastar o botão Nº da barra PivotTable, e colocá-lo (drop it) no campo Data:

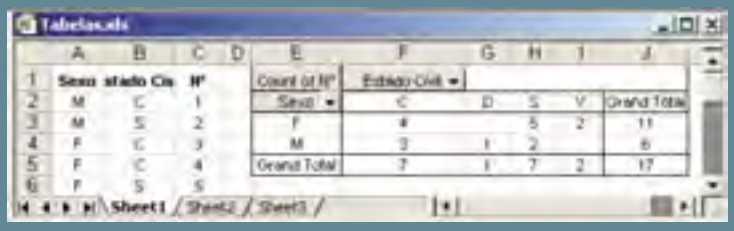


• Esta tabela, que resulta das operações anteriores, não é a que nos interessa, sendo agora necessário clicar 2 vezes no campo Sum of N.º e seleccionar a opção Count:





Finalmente temos a tabela de contingência desejada, que nos dá a distribuição conjunta (em valores absolutos) do par (Sexo, Estado civil), permitindo obter o número de indivíduos que satisfazem simultaneamente cada uma das modalidades (feminino(a),casado(a)), (feminino(a),divorciado(a)), ...(masculino(a),viúvo(a)):



O facto da célula correspondente ao F e D estar vazia, significa que não havia indivíduos do sexo feminino e divorciados. Esta tabela apresenta ainda as distribuições marginais (em valores absolutos) da variável Sexo e Estado civil, respectivamente nas células J3 a J4 e F5 a I5. Efectivamente, através da tabela, pode-se concluir que o número de indivíduos do sexo feminino era 11, enquanto que do sexo masculino eram 6. Analogamente, também podemos tirar conclusões sobre o número de indivíduos em cada modalidade da variável Estado civil.

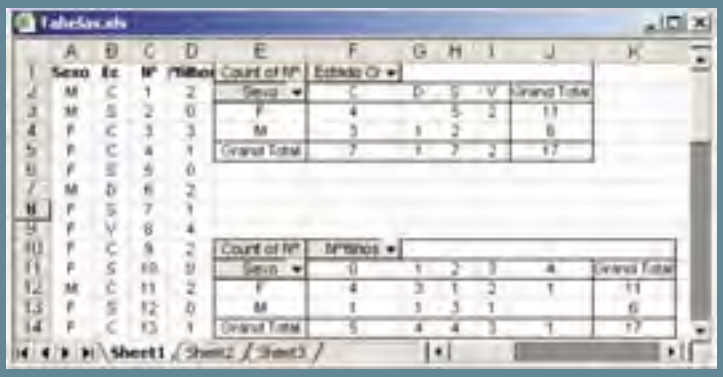
Exemplo 4.2.1 (cont) - Suponhamos que ao recolher a informação, junto de cada indivíduo, sobre o seu estado civil, também se tinha investigado sobre o número de filhos (esta informação é relevante para o serviço de processamento de salários proceder à retenção do IRS). Construa uma tabela de contingência para o par (Sexo, Estado civil).

Inserimos a informação sobre a variável Nº de filhos, e procedemos à construção da tabela de contingência da mesma forma que anteriormente, com as alterações convenientes, nomeadamente:

- No passo 2 seleccionámos as células de A1 a D18;
- No passo 3 seleccionámos a célula E10, para inserir a tabela;
- No passo seguinte arrastámos o botão Sexo da barra PivotTable, e colocámo-lo no campo Row;

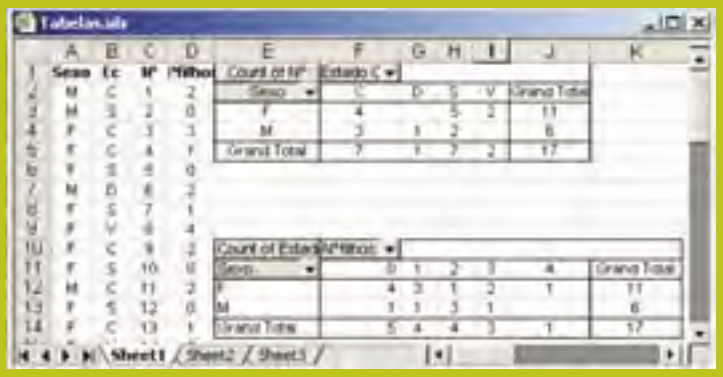
Arrastámos o botão Nº de filhos da barra PivotTable, e colocámo-lo no campo Column; Arrastámos o botão Nº de filhos da barra PivotTable, e colocámo-lo no campo Data;

- Clicámos 2 vezes no campo Sum of Nº e seleccionámos a opção Count:



Nesta 2ª tabela temos a distribuição conjunta do par (Sexo, Nº de filhos).

Exemplo 4.2.1 (cont) – Proceda como no exemplo anterior, excepto no passo seguinte ao passo 3, em que o botão da variável que arrasta para o campo Data é o botão da variável Estado civil. Com este procedimento o resultado é o seguinte:

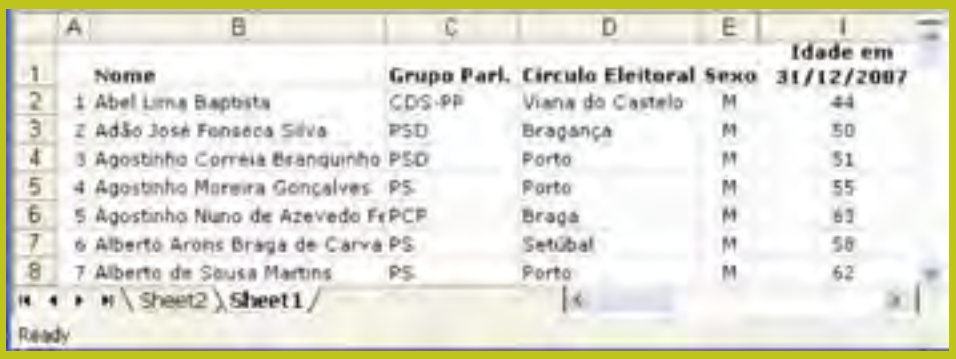


Quando colocámos o botão Estado civil no campo Data, imediatamente obtivemos uma tabela igual à anterior, com as contagens, em vez das somas, já que Count é a opção que está seleccionada, por defeito, quando colocamos no campo Data uma variável não numérica.

4.3 – Utilização das PivotTables para agrupar dados

Quando temos um conjunto de dados, já vimos no Capítulo 2 a forma de proceder ao seu agrupamento. Vamos agora ver, como essa tarefa pode ser feita através da utilização da PivotTable.

Vamos voltar ao ficheiro Deputados.xls (de que apresentamos a seguir uma pequena parte)



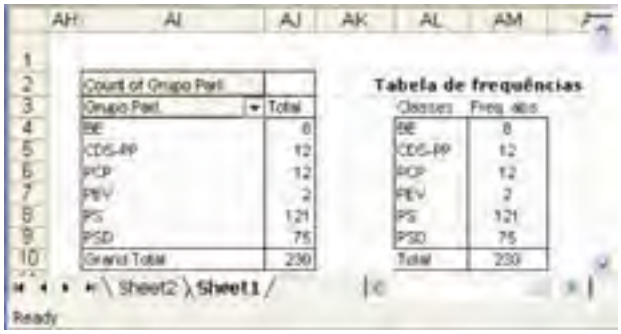
	Nome	Grupo Parl.	Circulo Eleitoral	Sexo	Idade em 31/12/2007
1	Abel Lima Baptista	CDS-PP	Viana do Castelo	M	44
2	Adão José Fonseca Silva	PSD	Bragança	M	50
3	Agostinho Correia Branquinho	PSD	Porto	M	51
4	Agostinho Moreira Gonçalves	PS	Porto	M	55
5	Agostinho Nuno de Azevedo Fe	PCP	Braga	M	61
6	Alberto Arons Braga de Carva	PS	Setúbal	M	58
7	Alberto de Sousa Martins	PS	Porto	M	62

Sheet2 Sheet1 Ready

para exemplificar a construção de uma tabela de frequências de uma variável qualitativa, utilizando a PivotTable.

Exemplo 4.3.1 – Utilizando a PivotTable, proceda ao agrupamento de dados da variável Grupo parlamentar, do ficheiro Deputados.xls.

- No menu Data, clicar em PivotTable and PivotChart Report;
- No passo 1 da PivotTable and PivotTable Wizard, seguir as instruções, e clicar PivotTable à pergunta What kind of report do you want to create?;
- No passo 2 seguir as instruções, seleccionando os dados que se pretende usar (não esquecer de seleccionar os títulos). Neste caso seleccionar as células C1:C231;
- No passo 3 seleccionar o lugar onde pretende criar a tabela. Nós optámos por seleccionar a célula AI2;
- Arrastar o botão Grupo parlamentar da barra PivotTable, e colocá-lo (drop it) no campo Row; Arrastar o botão Grupo parlamentar e colocá-lo (drop it) no campo Data:



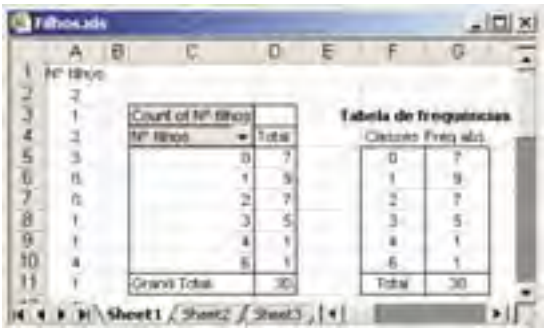
O procedimento anterior conduziu-nos à tabela do lado esquerdo da figura anterior, cujo conteúdo foi copiado para construir a tabela do lado direito, com uma apresentação mais sugestiva.

4.3.2 – Dados de tipo discreto

A organização de dados discretos numa tabela de frequências, utilizando a PivotTable, faz-se do mesmo modo que para os dados de tipo qualitativo. Vamos exemplificar procedendo ao agrupamento da variável N.º de filhos dos dados do ficheiro Filhos.xls.

Exemplo 4.3.2 - Utilizando a PivotTable, proceda ao agrupamento de dados da variável Nº de filhos, do ficheiro Filhos.xls

- No menu Data, clicar em PivotTable and PivotChart Report;
- No passo 1 da PivotTable and PivotTable Wizard, seguir as instruções, e clicar PivotTable à pergunta What kind of report do you want to create?;
- No passo 2 seguir as instruções, seleccionando os dados que se pretende usar (não esquecer de seleccionar os títulos). Neste caso seleccionar as células A2 a A31 (que contêm o n.º de filhos de uma amostra de 30 deputados);
- No passo 3 seleccionar o lugar onde pretende criar a tabela. Nós optámos por seleccionar a célula C3;
- Arrastar o botão N.º de filhos da barra PivotTable, e colocá-lo (drop it) no campo Row; Arrastar o mesmo botão e colocá-lo (drop it) no campo Data;
- Clicar duas vezes no botão Sum of N.º filhos, da tabela, e seleccionar Count:



Obtivemos a tabela do lado esquerdo, a qual foi copiada para o lado direito, com um aspecto mais usual.

4.3.3 – Dados de tipo contínuo

Vamos exemplificar o agrupamento de uma variável de tipo contínuo, utilizando a PivotTable, mas avisamos desde já, que se os dados não forem inteiros, o processo não é correcto e tem de ser utilizado com as devidas precauções, como veremos oportunamente. O processo que vamos utilizar foi sugerido por um artigo de Neville Hunt, na revista Teaching Statistics (Volume 25, Number 2, Summer 2003).

Começaremos por abordar a situação de termos uma variável contínua, mas em que os dados são inteiros.

1ª Parte – Dados em formato de inteiro

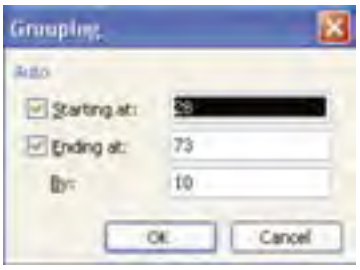
Exemplo 4.3.3 – Considere o ficheiro Idade.xls, que contém a idade de 230 deputados. Proceda ao agrupamento em classes, utilizando as PivotTables.

Considere o ficheiro Idade.xls, em que os dados da variável se encontram nas células C2 a C231 e proceda da seguinte forma:

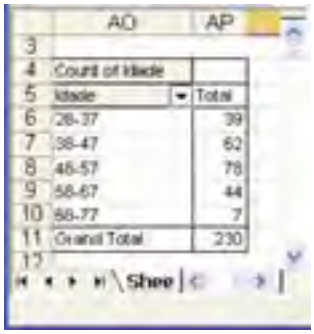
- No menu Data, clique em PivotTable and PivotChart Report;
- No passo 1 da PivotTable and PivotTable Wizard, siga as instruções, e clique PivotTable à pergunta What kind of report do you want to create?;
- No passo 2 siga as instruções, seleccionando os dados que pretende usar. Neste caso seleccione as células C1 a C31 (embora os dados estejam nas células C2 a C231, o título está na C1);
- No passo 3 seleccione o lugar onde pretende criar a tabela. Nós optámos por seleccionar a célula AO4;
- Arraste o botão Idade da barra PivotTable, e coloque-o (drop it) no campo Row; Arraste o mesmo botão e coloque-o (drop it) no campo Data;
- Clique duas vezes no botão Sum of Idade, da tabela, e seleccione Count;

A tabela que aparece depois destas operações, mostra a frequência de cada valor individual (como estamos com dados contínuos, embora inteiros, corremos o risco de termos uma tabela com tantas classes, quantos os dados, todos com frequência igual a 1!). Assim, é necessário proceder a mais algumas operações, para agrupar os dados:

- Clique em algum dos dados da variável Idade e seleccione Data - Group and Outline - Group, que faz surgir o seguinte diálogo:



Por defeito, no diálogo anterior é considerado como "Starting at" e "Ending at" respectivamente, o mínimo e o máximo do conjunto de dados a agrupar. Para "By" é considerado, também por defeito, um valor que dependerá do número de dados e da grandeza desses dados.



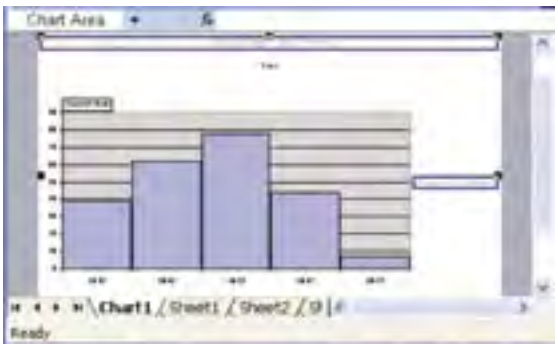
- Clicando em OK, é produzida a seguinte tabela de frequências:

Observação: Repare-se que na construção desta tabela, ao dizer que pretendemos que o agrupamento seja feito By:10, não significa que se adicione 10 ao mínimo para formar a 1ª classe e assim por diante. Neste caso 10 é o número de inteiros que vai do limite inferior de cada classe, até ao limite superior e não significa propriamente amplitude de classe, da forma como é definida, isto é, como sendo a diferença entre os limites do intervalo de classe. Se pretendêssemos classes de amplitude 10, teríamos de ter seleccionado, antes de efectuar o agrupamento, By:11 e obteríamos as classes 28-38, 39-49, 50-60, 61-71 e 72-82.

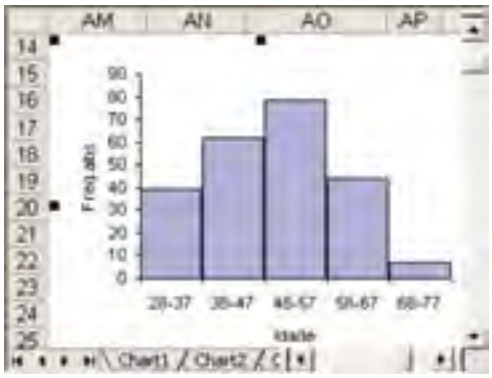
Para construir o histograma associado a esta tabela, basta carregar em alguma parte da tabela e na barra da PivotTable clicar no ícone  .

Por defeito aparece a construção de um gráfico de barras, com intervalos entre as barras, que podem ser removidas por um processo idêntico ao já utilizado, aquando da construção do histograma. Assim:

• Clique com o lado direito do rato numa das colunas e seleccione Format data Series - Options - Gap width:0:



• Finalmente podemos esconder os botões clicando com o lado direito do rato num deles e seleccionando Hide PivotChart Field Buttons e acrescentando de seguida títulos aos eixos:



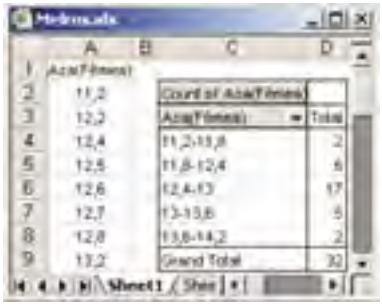
Observação: Para obter o gráfico anterior copiámos a figura obtida numa folha Chart do Excel para uma folha normal (Sheet).

2ª Parte – Dados em formato decimal

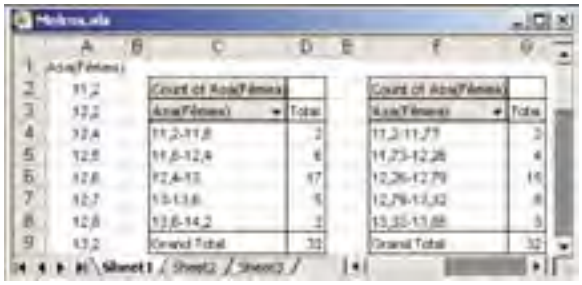
Como vimos na construção das classes da tabela anterior, estas são construídas sem ambiguidade, na medida em que qualquer elemento do conjunto de dados só pode pertencer a uma única classe. O mesmo não acontece se estivermos a trabalhar com dados com casas decimais, como veremos no exemplo seguinte.

Exemplo 4.3.4 – Considere novamente os dados do exemplo 2.3.8, em que se estudou o comprimento, em centímetros, das asas de melros. Proceda ao agrupamento dos dados correspondentes aos melros-fêmea.

Consideremos a amostra constituída pelas 32 medidas das asas de outros tantos melros ¬fêmeas que inserimos numa folha de Excel, ocupando as células A2 a A33, reservando a A1 para o título Fêmea. Construímos uma tabela de frequências, utilizando o processo seguido anteriormente, mas escolhendo para amplitude de classe o valor 0,6. O resultado obtido foi a seguinte tabela:



Como se verifica, ao contrário do que acontecia com a variável Idade, o limite superior de um intervalo é igual ao limite inferior do intervalo seguinte, ficando a dúvida de saber em que classe inserir um elemento igual a um desses limites. Na verdade estes intervalos funcionam como se fossem fechados á esquerda e abertos à direita (excepto a última classe que também é fechada à direita), pelo que um valor igual, por exemplo, a 11,8, será contabilizado na classe 11,8-12,4. Este problema pode ser resolvido, considerando para amplitude de classe um valor decimal, com uma casa decimal a mais dos que os dados. No exemplo anterior, se escolhêssemos como amplitude de classe 0,53, já o problema deixaria de existir, pois não teríamos dúvida em que classe contabilizar qualquer um dos valores do conjunto de dados:



Como diz Neville Hunt no artigo referido anteriormente, página 45, e passamos a citar:
...After reading this article, some teachers will (not unreasonably) decide that Excel is not fit to be used for this type of analysis. However, the universal popularity and availability of Excel are such that students will inevitably try to use it for this purpose at some stage, so it is important that they should be made aware of its limitations and need for vigilance.

Esta citação vem ao encontro daquilo que pensamos e já referimos neste texto, de que o Excel não é um software de Estatística, mas ao nível elementar resolve muitas situações, desde que ao utilizá-lo se saiba o que se pretende. Por exemplo, quando se pretende um histograma, e ao obter um diagrama de barras, é necessário ter presente que, embora o histograma seja construído à custa de barras, estas têm que estar unidas.

5. Introdução à simulação

5.1- Introdução

Pretende-se com este Capítulo, dar a conhecer um instrumento poderoso – a simulação, que sobretudo nas duas últimas décadas, com o desenvolvimento e aperfeiçoamento dos meios computacionais, contribuiu de forma decisiva para o estudo das leis de probabilidade e a obtenção da probabilidade associada a determinados acontecimentos. Veremos assim uma forma de imitar o comportamento aleatório, característico dos fenómenos que têm interesse estudar em Probabilidade, isto é, os fenómenos chamados de aleatórios, por oposição aos determinísticos. Na verdade, essa possibilidade de imitação (simulação), baseia-se no facto de ao realizar uma experiência aleatória, repetidamente e em condições semelhantes, os resultados obtidos mostrarem uma regularidade estatística, que é utilizada para obter estimativas das probabilidades dos acontecimentos associados à experiência em causa. Esta regularidade a longo termo, é a base da interpretação frequencista de Probabilidade. Simulando várias realizações de uma experiência aleatória, é então possível obter as estimativas consideradas anteriormente.

Por exemplo, ao lançar um dado equilibrado repetidas vezes, registando numa tabela de frequências, a frequência relativa da saída de cada face, verifica-se que à medida que o número de lançamentos aumenta, a frequência relativa da saída de cada face tende a estabilizar à volta do valor 0,167 (aproximadamente 1/6).

Embora não tenhamos chamado explicitamente a atenção para o facto, na verdade já utilizámos o conceito de simulação, quando no capítulo 1, utilizámos a função Randbetween do Excel, para "imitar" o comportamento aleatório da extracção de uma amostra, de uma certa população.

Vamos ver de seguida, como por simulação se
podem obter boas aproximações das probabi-lidades de acontecimentos, que teoricamente seriam difíceis, ou mesmo impossíveis de obter.

5.2- Obtenção de probabilidades por simulação

Vamos apresentar exemplos simples, que nos servirão para dar uma ideia da utilização e da potencialidade do método da simulação. Vamos utilizar as funções RAND ou RANDBETWEEN, já utilizadas no capítulo 1, que têm por base o conceito de número aleatório, ou mais propriamente pseudo-aleatório.

Os algoritmos de geração de números pseudo-aleatórios estão concebidos de modo a que ao considerar uma qualquer sequência de números gerados se obtenha aproximadamente a mesma proporção de observações em subintervalos de igual amplitude do intervalo [0,1]. Assim, por exemplo, se se fizer correr o algoritmo 100 vezes, é de esperar que caiam 25 dos números gerados em cada quarto do intervalo [0,1]. Na tabela seguinte está listada uma sequência de 100 NPA's obtida através do gerador RAND do software Excel (Graça Martins, M. E e Loura, L., 2001):

0,842050	0,406320	0,848744	0,810469	0,789583
0,965131	0,676239	0,722927	0,825587	0,702971
0,761648	0,552387	0,079614	0,298300	0,087455
0,359825	0,208420	0,098150	0,818893	0,103532
0,054705	0,102768	0,147229	0,557920	0,996667
0,466613	0,493374	0,150888	0,540352	0,480287
0,814300	0,638416	0,086141	0,007840	0,109918
0,449515	0,090759	0,197460	0,209145	0,713230
0,901502	0,552418	0,466389	0,221584	0,623757
0,862762	0,507097	0,613583	0,389183	0,129629
0,395195	0,415666	0,210044	0,379011	0,302539
0,420519	0,469764	0,05374	0,478208	0,444822
0,124664	0,765629	0,737348	0,696311	0,806147
0,537707	0,451921	0,702749	0,683382	0,377823
0,033277	0,523063	0,908485	0,708764	0,196290
0,024371	0,213326	0,442821	0,983754	0,970551
0,558313	0,283191	0,153907	0,655705	0,995760
0,087859	0,429387	0,735276	0,890680	0,569285
0,069915	0,221549	0,358037	0,578713	0,161851
0,774156	0,039495	0,490216	0,755072	0,753139

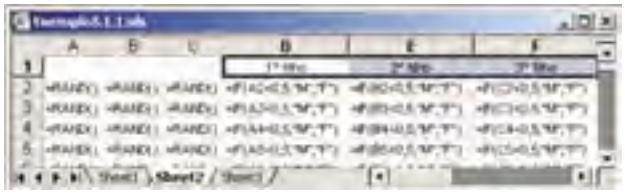
Como se pode verificar por contagem, esta lista inclui 30 números no intervalo [0,0.25], 24 números nos intervalos]0.25,0.5] e]0.5,0.75] e 22 números no intervalo]0.75,1]. Embora haja métodos estatísticos para avaliar se são ou não significativas as diferenças entre estas frequências observadas e as frequências esperadas (25 – 25 – 25 – 25), facilmente a nossa sensibilidade aceita que estes resultados não contradizem o que se esperaria de uma escolha ao acaso de 100 números do intervalo [0,1].

De um modo geral quando falamos em números aleatórios, estamos a referir-nos à obtenção de qualquer real do intervalo [0, 1], de tal forma que a probabilidade de obter um valor de um subintervalo [a, b] de [0, 1], é igual à amplitude desse subintervalo, ou seja (b-a).

Exemplo 5.1.1 (Adaptado do exemplo 6.2.1 de Graça Martins et al, 1999) – Suponha um casal que pretende ter um "casal" de filhos, não desejando mais do que 3 filhos e só tentando o 3.º filho se anteriormente tiver tido ou dois rapazes ou duas raparigas. Qual a probabilidade de ter efectivamente o casalinho?

Admitindo que a probabilidade de nascer rapaz é igual à de nascer rapariga, vamos utilizar a função RAND, para simular um qualquer destes nascimentos, da seguinte forma: Se o resultado da função RAND for inferior a 0,5, simulamos o nascimento de um rapaz – M. Caso contrário simulamos o nascimento de uma rapariga. Numa folha de Excel vamos simular várias repetições da experiência "nascimento de 3 filhos". Poderíamos ter optado por começar por simular o nascimento de dois filhos e só simular o 3.º filho se não houvesse os dois sexos nos dois primeiros filhos. No entanto, este condicionamento da simulação do 3.º filho faz com que cada repetição da experiência dependa do que se obtém anteriormente, o que torna mais demorado o processo da simulação. Assim, simulámos sempre 3 filhos e basta nos dois primeiros haver os dois sexos, para termos como resultado da experiência um sucesso. Assinalamos o sucesso (dois sexos diferentes logo nos dois primeiros filhos ou sexos diferentes nos três filhos) com um 1 – esta notação facilita-nos o cálculo da frequência relativa do nº de sucessos, à medida que repetimos a experiência.

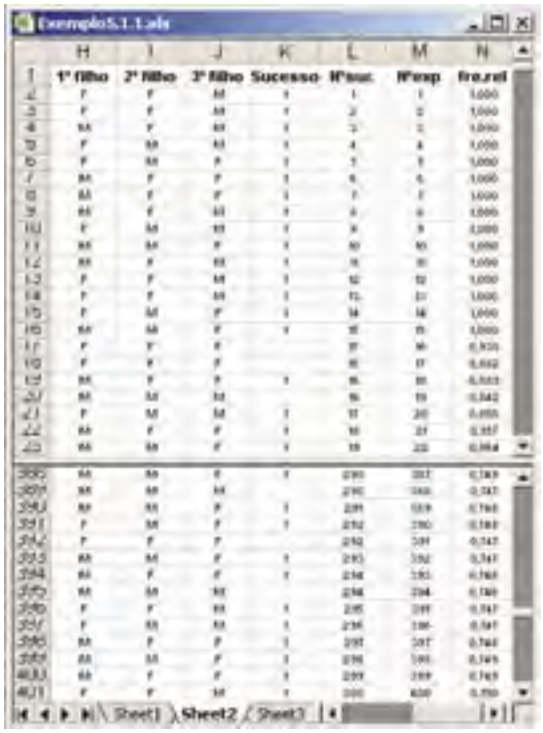
Um procedimento possível para a simulação em causa, pode ser o seguinte:



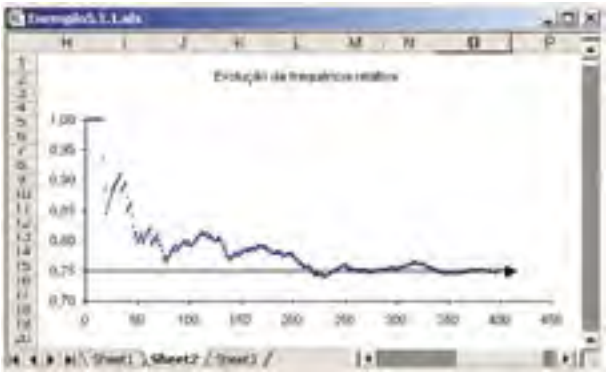
- Inserir a função RAND() nas células A2, B2 e C2 e nas células D2, E2 e F2 a função IF(), como se exemplifica na figura seguinte:
- Replicar (Fill down) as células A2:F2, tantas vezes quantas as vezes que se pretende simular a realização da experiência. Nós replicámos 400 vezes, colocando os resultados nas células A2:F401;
- Copiar (Paste special) os valores das células D2:F401, para as células H2:J401 (Este passo tem como objectivo guardar os valores gerados anteriormente, pois a função RAND() é volátil, como já referimos nos capítulos anteriores);
- Em cada uma das células da coluna K inserir 1 se o resultado da experiência tiver sido sucesso;
- Na coluna L contabilizar o n.º de sucessos acumulados;

- Na coluna M contabilizar o n.º da experiência;
- Na coluna N calcular a frequência relativa de sucesso, à medida que se vão realizando experiências.

O processo anterior é apresentado na figura seguinte. Por uma questão de espaço só apresentamos a parte inicial e a parte final da tabela:



Como se verifica, a frequência relativa estabiliza à volta do valor 0,75, pelo que dizemos que 0,75 é uma estimativa para a probabilidade pretendida (O valor calculado, teoricamente, para esta probabilidade é de 0,75). A título de curiosidade acrescentamos que o resultado da simulação ao fim de 100, 200 e 300 repetições, foi respectivamente 0,790, 0,775 e 0,753. Apresentamos a evolução da frequência relativa na seguinte representação gráfica:



Exemplo 5.1.2 (Ageel, M. I. - Teaching Statistics, Volume 24, Number 2, Summer 2002, pag. 51¬54) – Um segmento de linha de comprimento 1 é partido, aleatoriamente, em três pedaços. Qual a probabilidade de as peças resultantes poderem formar um triângulo?

A resolução deste problema prende-se com uma regra que estabelece que a soma dos comprimentos de dois lados de um triângulo, é superior ao comprimento do outro lado. Vamos resolver este problema fazendo uma série de simulações e calculando a frequência relativa das situações que dão origem a triângulos. Considera-se então uma folha de cálculo e procede-se da seguinte forma:

- Nas células A2 e B2 introduz-se a função RAND(), que devolve um número pseudo-aleatório entre 0 e 1 (equivalente à função RANDBETWEEN(0;1)). Estes números irão representar os pontos P e Q em que uma linha MN de comprimento 1 fica dividida:

- Considera-se para P o menor dos valores obtidos anteriormente, que será o comprimento de MP – célula C2;
- Calcula-se o comprimentos dos segmentos PQ e QN – células D2 e E2, respectivamente:

• Testa-se se 2 quaisquer dos comprimentos obtidos anteriormente é superior ao terceiro comprimento – célula F2;

• Replica-se as células de A2 a F2 até à linha 1001 (1000 réplicas);

• Calcula-se o número de vezes que o teste anterior deu verdadeiro, ou seja TRUE – célula G2, e divide-se por 1000:

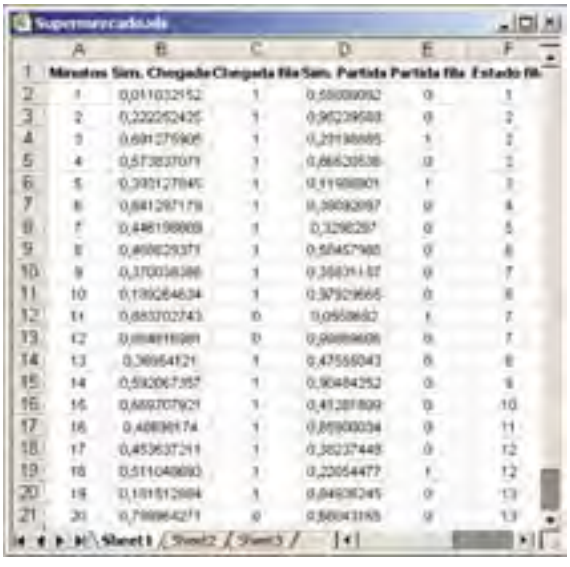
O resultado da simulação anterior deu uma frequência relativa de 0,249, que se pode considerar um valor aproximado para a probabilidade pretendida:

Do mesmo modo que a função RANDBETWEEN, também a função RAND é volátil, pelo que qualquer operação na folha de cálculo modifica os números pseudo-aleatórios considerados para coordenadas dos pontos e consequentemente a estimativa da probabilidade pretendida. Assim, quantas operações forçar na folha anterior, nomeadamente digitar um valor numa das células em branco consiste numa operação, quantas estimativas obterá para a probabilidade pretendida, ou seja, para a probabilidade de conseguir construir um triângulo com as partes de um segmento de recta de comprimento unitário, dividido aleatoriamente em 3 partes.

Exemplo 5.1.3 -Suponha que em cada minuto a probabilidade de alguém chegar à fila de uma caixa de supermercado é de 75%, enquanto que a probabilidade de abandonar a fila, depois de ser servido é de 30%. Ao fim de 20 minutos qual o tamanho que espera para a fila?

Vamos simular a experiência anterior, simulando a chegada de um cliente à fila sempre que o resultado da função RAND for $\leq 0{,}75$ e a saída de um cliente da fila sempre que a função RAND devolver um resultado $\leq 0{,}30$:

Para não correrermos o risco de termos uma fila com um número negativo de pessoas, consideramos a função máximo:

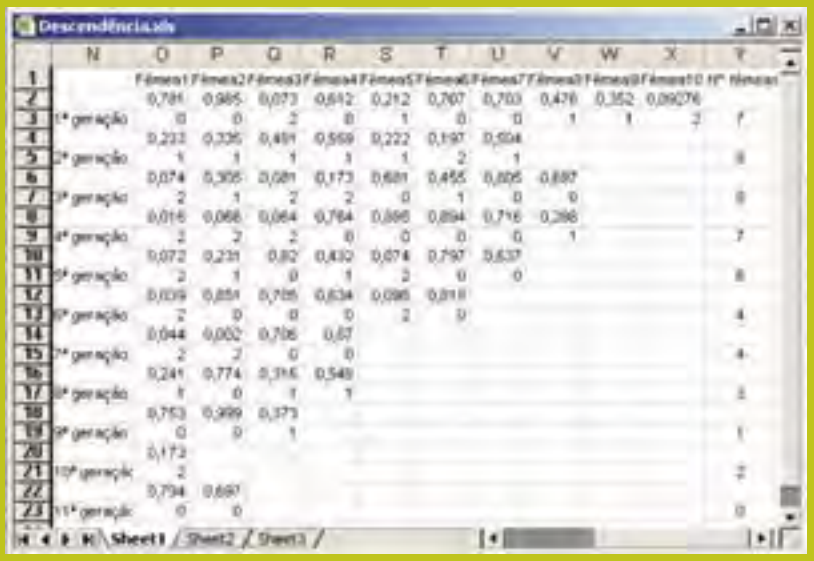


Ao fim de 20 minutos a fila já tem 13 clientes e com tendência para crescer!

Exemplo 5.1.4 – Suponha uma espécie animal em que as fêmeas têm o seguinte comportamento reprodutor:

- 40% morrem antes de deixar descendência
- 40% têm uma fêmea descendente
- 20% têm duas fêmeas descendentes.

Estude o comportamento desta população, nomeadamente se se prevê um crescimento rápido de indivíduos da espécie, a extinção ou uma situação de equilíbrio. Vamos estudar a evolução da população simulando a descendência de 10 fêmeas, ao longo de algumas gerações. Para cada fêmea, geramos um número pseudo-aleatório, cujo resultado será interpretado da seguinte forma: Se o número for inferior a 0,20, a fêmea deixa 2 descendentes fêmeas; Se o número estiver compreendido entre 0,2 e 0,6, a fêmea deixa 1 descendente fêmea; Se o número estiver compreendido entre 0,6 e 1, a fêmea morre sem descendência. Apresentamos a seguir uma simulação da experiência com as 10 fêmeas:



Na tabela anterior considerámos:

- Nas células O2:X2, 10 números pseudo-aleatórios para simular a descendência das 10 fêmeas com que iniciámos a nossa experiência;
- Na célula Y3, o número de fêmeas obtidas ao fim da primeira geração – neste caso 7;
- Nas células O4:U4, 7 números pseudo-aleatórios para simular a descendência das 7 fêmeas obtidas na geração anterior;
- Na célula Y5, o número de fêmeas obtidas ao fim da segunda geração – neste caso 8;
- Repetimos o processo anterior, até não haver descendência de fêmeas.

Como se verifica, a população tem tendência a extinguir-se, pois ao fim da 11.ª geração já não há descendentes das 10 fêmeas com que iniciámos o estudo.

Repita a experiência admitindo que

- 20% morrem antes de deixar descendência
- 40% Têm uma fêmea descendente
- 40% têm duas fêmeas descendentes.

Um outro exemplo interessante e que tem levantado bastante polémica é o seguinte exemplo de decisão estratégica.

Exemplo 5.1.5 (Graça Martins, M. E. e Loura, L., 2001) - Num concurso é dada a escolher ao concorrente uma de 3 portas. Atrás de uma delas está um carro e atrás de cada uma das outras duas está uma ovelha. O concorrente escolhe uma das portas (sem a abrir) e o apresentador, que sabe exactamente qual é a porta que esconde o carro, abre, de entre as duas portas que restam, uma onde está uma ovelha. Nesse momento pergunta ao concorrente se deseja ou não trocar a porta que escolheu pela outra porta que ainda está fechada. O primeiro pensamento que ocorre é que não há qualquer vantagem em trocar, pois temos agora apenas duas portas e o carro tanto pode estar atrás de uma como da outra. No entanto, se se calcular teoricamente a probabilidade do concorrente ganhar o carro, trocando de porta, verifica-se que esta é igual a 2/3. Para os mais reticentes uma simulação talvez os faça reconsiderar a sua posição inicial. Não há qualquer dúvida de que ao escolher uma porta ao acaso a probabilidade de ela esconder o carro é igual a 1/3.

Para simular o decorrer de 100 destes concursos vamos então considerar que o concorrente escolheu a boa porta sempre que o valor do número pseudo-aleatório (NPA) estiver entre 0 e 1/3. Nestes casos, quando ele trocar de porta, ficará com a "ovelha" mas, em compensação, ficará com o carro em todos os outros casos (se ele tiver escolhido inicialmente a "ovelha", a porta que resta terá obrigatoriamente o carro pois o apresentador encarregou-se de eliminar a outra porta que também tinha "ovelha"!...)

Eis o resultado da simulação obtida a partir de 100 números pseudo-aleatórios gerados numa folha de Excel:

NPA	O que ganha não trocando	O que ganha trocando	NPA	O que ganha não trocando	O que ganha trocando	NPA	O que ganha não trocando	O que ganha trocando
0,842	Ovelha	Carro	0,406	Ovelha	Carro	0,849	Ovelha	Carro
0,965	Ovelha	Carro	0,676	Ovelha	Carro	0,723	Ovelha	Carro
0,762	Ovelha	Carro	0,552	Ovelha	Carro	0,080	Carro	Ovelha
0,360	Ovelha	Carro	0,208	Carro	Ovelha	0,098	Carro	Ovelha
0,055	Carro	Ovelha	0,103	Carro	Ovelha	0,147	Carro	Ovelha
0,467	Ovelha	Carro	0,493	Ovelha	Carro	0,151	Carro	Ovelha
0,814	Ovelha	Carro	0,638	Ovelha	Carro	0,086	Carro	Ovelha
0,450	Ovelha	Carro	0,091	Carro	Ovelha	0,197	Carro	Ovelha
0,902	Ovelha	Carro	0,552	Ovelha	Carro	0,466	Ovelha	Carro
0,863	Ovelha	Carro	0,507	Ovelha	Carro	0,614	Ovelha	Carro
0,395	Ovelha	Carro	0,416	Ovelha	Carro	0,210	Carro	Ovelha
0,421	Ovelha	Carro	0,470	Ovelha	Carro	0,054	Carro	Ovelha
0,125	Carro	Ovelha	0,766	Ovelha	Carro	0,737	Ovelha	Carro
0,538	Ovelha	Carro	0,452	Ovelha	Carro	0,703	Ovelha	Carro
0,033	Carro	Ovelha	0,523	Ovelha	Carro	0,908	Ovelha	Carro
0,024	Carro	Ovelha	0,213	Carro	Ovelha	0,443	Ovelha	Carro
0,558	Ovelha	Carro	0,283	Carro	Ovelha	0,154	Carro	Ovelha
0,088	Carro	Ovelha	0,429	Ovelha	Carro	0,735	Ovelha	Carro
0,070	Carro	Ovelha	0,222	Carro	Ovelha	0,358	Ovelha	Carro
0,774	Ovelha	Carro	0,039	Carro	Ovelha	0,490	Ovelha	Carro
0,810	Ovelha	Carro	0,709	Ovelha	Carro	0,713	Ovelha	Carro
0,826	Ovelha	Carro	0,984	Ovelha	Carro	0,624	Ovelha	Carro
0,298	Carro	Ovelha	0,656	Ovelha	Carro	0,130	Carro	Ovelha
0,819	Ovelha	Carro	0,891	Ovelha	Carro	0,303	Carro	Ovelha
0,558	Ovelha	Carro	0,579	Ovelha	Carro	0,445	Ovelha	Carro
0,540	Ovelha	Carro	0,755	Ovelha	Carro	0,806	Ovelha	Carro
0,008	Carro	Ovelha	0,790	Ovelha	Carro	0,378	Ovelha	Carro
0,209	Carro	Ovelha	0,703	Ovelha	Carro	0,196	Carro	Ovelha
0,222	Carro	Ovelha	0,087	Carro	Ovelha	0,971	Ovelha	Carro
0,389	Ovelha	Carro	0,104	Carro	Ovelha	0,996	Ovelha	Carro
0,379	Ovelha	Carro	0,997	Ovelha	Carro	0,569	Ovelha	Carro
0,478	Ovelha	Carro	0,480	Ovelha	Carro	0,162	Carro	Ovelha
0,696	Ovelha	Carro	0,110	Carro	Ovelha	0,753	Ovelha	Carro
0,683	Ovelha	Carro						

Como se verifica, nas 100 realizações simuladas deste concurso o concorrente ganharia o carro em 67 dessas realizações, se se decidisse por trocar de porta!...

Lista de algumas funções usadas no Excel:

And() — E()

Devolve verdadeiro se todos os argumentos forem verdadeiros e devolve falso se algum dos argumentos for falso

Average() — Media()

Calcula a média dos valores existentes num conjunto de células

Count() — Contar()

Conta as células com valores numéricos, incluindo datas e fórmulas cujos resultados são numeros

Counta() — Contar.val()

Conta todas as células não vazias

Countblank() — Contar.vazio()

Conta as células vazias

Countif() — Contar.se()

Conta as ocorrências verificadas num conjunto de célula, que obedecem a um critério

Frequency() — Frequência

If() — Se()

Executa uma de duas acções possíveis, em função do resultado da condição

Int() — Int()

Devolve a parte inteira de um número

Max() — Maximo()

Devolve o maior valor de um conjunto de células

Min() — Minimo()

Devolve o menor valor de um conjunto de células

Mod() — Resto()

Devolve o resto de uma divisão

Or() — Ou()

Devolve verdadeiro se um dos argumentos for verdadeiros e devolve falso se todos os argumentos forem falsos

Pie

Product() Produto()

Multiplica os valores de um conjunto de células, ignorando as células vazias e/ou com texto

Rand() Aleatório()

Devolve um número pseudo-aleatório (no intervalo (0,1))

Randbetween() Aleatórioentre()

Devolve um número pseudo-aleatório no intervalo especificado

Round() Arred()

Devolve um número arredondado, na posição indicada

Rounddown() Arred.para.baixo()

Devolve um número arredondado, por defeito, na posição indicada

Roundup() Arred.para.cima()

Devolve um número arredondado, por excesso, na posição indicada

Scatter

Stdev

Stdevp

Sum() Soma()

Soma os valores de um conjunto de células

Sumif() Soma.se()

Soma as ocorrências verificadas num conjunto de células que obedecem a um critério

Anexo -
Ficheiro de Deputados da X Legislatura

	Nome	Grupo Parl.	Círculo Eleitoral	Sexo	Data nas.
1	Abel Lima Baptista	CDS-PP	Viana do C	M	13-10-1963
2	Adão José Fonseca Silva	PSD	Bragança	M	01-10-1957
3	Agostinho Correia Branquinho	PSD	Porto	M	10-08-1956
4	Agostinho Moreira Gonçalves	PS	Porto	M	15-07-1952
5	Agostinho Nuno de Azevedo Ferreira Lopes	PCP	Braga	M	16-11-1944
6	Alberto Arons Braga de Carvalho	PS	Setúbal	M	20-09-1949
7	Alberto de Sousa Martins	PS	Porto	M	25-04-1945
8	Alberto Marques Antunes	PS	Setúbal	M	03-04-1949
9	Alcídia Maria Cruz Sousa de Oliveira Lopes	PS	Porto	F	09-01-1974
10	Alda Maria Gonçalves Pereira Macedo	BE	Porto	F	07-09-1954
11	Aldemira Maria Cabanita do Nascimento Bispo Pinho	PS	Faro	F	04-04-1952
12	Ana Catarina Veiga Santos Mendonça Mendes	PS	Setúbal	F	14-01-1973
13	Ana Isabel Drago Lobato	BE	Lisboa	F	28-08-1975
14	Ana Maria Cardoso Duarte da Rocha Almeida Pereira	PS	Porto	F	16-08-1967
15	Ana Maria Ribeiro Gomes do Couto	PS	Lisboa	F	19-04-1961
16	Ana Maria Sequeira Mendes Pires Manso	PSD	Guarda	F	30-03-1956
17	António Alfredo Delgado da Silva Preto	PSD	Lisboa	M	18-11-1958
18	António Alves Marques Júnior	PS	Porto	M	03-07-1946
19	António Bento da Silva Galamba	PS	Lisboa	M	11-11-1968
20	António Carlos Bivar Branco de Penha Monteiro	CDS-PP	Lisboa	M	31-05-1968
21	António Edmundo Barbosa Montalvão Machado	PSD	Porto	M	09-12-1952
22	António Filipe Gaião Rodrigues	PCP	Lisboa	M	28-01-1963
23	António Joaquim Almeida Henriques	PSD	Viseu	M	05-05-1961
24	António José Ceia da Silva	PS	Portalegre	M	11-04-1963
25	António José Martins Seguro	PS	Braga	M	11-03-1962
26	António Paulo Martins Pereira Coelho	PSD	Coimbra	M	27-04-1958
27	António Ramos Preto	PS	Lisboa	M	19-01-1956
28	António Ribeiro Cristóvão	PSD	Castelo Br	M	07-07-1939
29	António Ribeiro Gameiro	PS	Santarém	M	14-08-1970
30	Armando França Rodrigues Alves	PS	Aveiro	M	22-10-1949
31	Arménio dos Santos	PSD	Lisboa	M	22-11-1945
32	Artur Jorge da Silva Machado	PCP	Porto	M	20-05-1976
33	Artur Miguel Claro da Fonseca Mora Coelho	PS	Lisboa	M	04-07-1952
34	Bernardino José Torrão Soares	PCP	Lisboa	M	15-09-1971
35	Bruno Ramos Dias	PCP	Setúbal	M	19-10-1976
36	Carlos Alberto David dos Santos Lopes	PS	Leiria	M	06-06-1965
37	Carlos Alberto Garcia Poço	PSD	Leiria	M	12-02-1957
38	Carlos Alberto Silva Gonçalves	PSD	Europa	M	20-10-1961
39	Carlos António Páscoa Gonçalves	PSD	Fora da Eu	M	09-02-1952
40	Carlos Jorge Martins Pereira	PSD	Braga	M	15-02-1973
41	Carlos Manuel de Andrade Miranda	PSD	Viseu	M	03-09-1953
42	Cláudia Isabel Patrício do Couto Vieira	PS	Viseu	F	16-10-1967
43	David Martins	PS	Faro	M	05-01-1976
44	Diogo Nuno de Gouveia Torres Feio	CDS-PP	Porto	M	06-10-1970
45	Domingos Duarte Lima	PSD	Bragança	M	20-11-1955
46	Duarte Rogério Matos Ventura Pacheco	PSD	Lisboa	M	25-11-1965
47	Elísio da Costa Amorim	PS	Aveiro	M	14-05-1953
48	Emídio Guerreiro	PSD	Braga	M	23-05-1965
49	Esmeralda Fátima Quitério Salero Ramires	PS	Faro	F	23-10-1955
50	Feliciano José Barreiras Duarte	PSD	Leiria	M	19-04-1966
51	Fernanda Maria Pereira Asseiceira	PS	Santarém	F	18-04-1961
52	Fernando dos Santos Antunes	PSD	Coimbra	M	19-09-1949
53	Fernando dos Santos Cabral	PS	Guarda	M	10-05-1956
54	Fernando José Mendes Rosas	BE	Setúbal	M	18-04-1946
55	Fernando Manuel de Jesus	PS	Porto	M	04-06-1950

56	Fernando Mimoso Negrão	PSD	Setúbal	M	29-11-1955
57	Fernando Santos Pereira	PSD	Braga	M	27-05-1960
58	Francisco Anacleto Louçã	BE	Lisboa	M	12-11-1956
59	Francisco José de Almeida Lopes	PCP	Setúbal	M	29-08-1955
60	Francisco Miguel Baudoin Madeira Lopes	PEV	Lisboa	M	12-01-1975
61	Glória Maria da Silva Araújo	PS	Porto	F	04-01-1976
62	Guilherme Henrique Valente Rodrigues da Silva	PSD	Madeira	M	16-07-1943
63	Helena Maria Moura Pinto	BE	Lisboa	F	05-09-1959
64	Heloísa Augusta Baião de Brito Apolónia	PEV	Setúbal	F	26-06-1969
65	Henrique José Praia da Rocha de Freitas	PSD	Lisboa	M	13-03-1961
66	Hermínio José Sobral Loureiro Gonçalves	PSD	Aveiro	M	30-12-1965
67	Horácio André Antunes	PS	Coimbra	M	05-03-1946
68	Hugo José Teixeira Velosa	PSD	Madeira	M	18-04-1948
69	Hugo Miguel Guerreiro Nunes	PS	Faro	M	12-06-1963
70	Isabel Maria Batalha Vigia Polaco de Almeida	PS	Leiria	F	22-10-1953
71	Isabel Maria Pinto Nunes Jorge	PS	Braga	F	10-02-1953
72	Jacinto Serrão de Freitas	PS	Madeira	M	16-02-1969
73	Jaime José Matos da Gama	PS	Lisboa	M	08-06-1947
74	Jerónimo Carvalho de Sousa	PCP	Lisboa	M	13-04-1947
75	Joana Fernanda Ferreira Lima	PS	Porto	F	18-11-1963
76	João Barroso Soares	PS	Lisboa	M	29-08-1949
77	João Bosco Soares Mota Amaral	PSD	Açores	M	15-04-1943
78	João Cândido da Rocha Bernardo	PS	Aveiro	M	24-09-1955
79	João Carlos Vieira Gaspar	PS	Lisboa	M	22-05-1937
80	João Guilherme Nobre Prata Fragoso Rebelo	CDS-PP	Lisboa	M	02-02-1970
81	João Guilherme Ramos Rosa de Oliveira	PCP	Évora	M	09-07-1979
82	João Miguel de Melo Santos Taborda Serrano	PS	Lisboa	M	15-04-1964
83	João Nuno Lacerda Teixeira de Melo	CDS-PP	Braga	M	18-03-1966
84	João Pedro Furtado da Cunha Semedo	BE	Porto	M	20-06-1951
85	João Raul Henriques Sousa Moura Portugal	PS	Coimbra	M	01-10-1977
86	Joaquim Barbosa Ferreira Couto	PS	Porto	M	01-05-1951
87	Joaquim Carlos Vasconcelos da Ponte	PSD	Açores	M	06-06-1956
88	Joaquim Ventura Leite	PS	Setúbal	M	15-08-1950
89	Joaquim Virgílio Leite Almeida Costa	PSD	Braga	M	13-10-1943
90	Jorge Fernando Magalhães da Costa	PSD	Porto	M	12-01-1959
91	Jorge Filipe Teixeira Seguro Sanches	PS	Castelo Br	M	30-07-1965
92	Jorge José Varanda Pereira	PSD	Braga	M	28-10-1966
93	Jorge Manuel Capela Gonçalves Fão	PS	Viana do C	M	04-11-1957
94	Jorge Manuel Ferraz de Freitas Neto	PSD	Porto	M	03-01-1957
95	Jorge Manuel Gouveia Strecht Ribeiro	PS	Porto	M	07-09-1943
96	Jorge Manuel Monteiro de Almeida	PS	Vila Real	M	20-09-1954
97	Jorge Tadeu Correia Franco Morgado	PSD	Aveiro	M	02-07-1971
98	José Adelmo Gouveia Bordalo Junqueiro	PS	Viseu	M	28-06-1953
99	José Alberto Rebelo dos Reis Lamego	PS	Lisboa	M	05-01-1953
100	José António Freire Antunes	PSD	Porto	M	25-01-1954
101	José Augusto Clemente de Carvalho	PS	Lisboa	M	18-12-1948
102	José Batista Mestre Soeiro	PCP	Beja	M	17-01-1948
103	José Carlos Bravo Nico	PS	Évora	M	11-09-1964
104	José Carlos Correia Mota de Andrade	PS	Bragança	M	25-11-1955
105	José de Almeida Cesário	PSD	Fora da Eu	M	20-07-1958
106	José Eduardo Rego Mendes Martins	PSD	Viana do C	M	09-02-1969
107	José Eduardo Vera Cruz Jardim	PS	Lisboa	M	02-01-1939
108	José Helder do Amaral	CDS-PP	Viseu	M	08-06-1967
109	José Honório Faria Gonçalves Novo	PCP	Porto	M	24-10-1950
110	José Luís Fazenda Arnaut Duarte	PSD	Viseu	M	04-03-1963
111	José Manuel de Matos Correia	PSD	Lisboa	M	08-05-1963
112	José Manuel Ferreira Nunes Ribeiro	PSD	Aveiro	M	18-04-1969
113	José Manuel Lello Ribeiro de Almeida	PS	Porto	M	18-05-1944
114	José Manuel Pereira da Costa	PSD	Faro	M	12-05-1959

115	José Mendes Bota	PSD	Faro	M	04-08-1955
116	José Paulo Ferreira Areia de Carvalho	CDS-PP	Porto	M	29-05-1967
117	José Pedro Correia de Aguiar Branco	PSD	Porto	M	18-07-1957
118	José Raúl Guerreiro Mendes dos Santos	PSD	Porto	M	11-07-1959
119	Jovita de Fátima Romano Ladeira	PS	Faro	F	16-02-1957
120	Júlio Francisco Miranda Calha	PS	Portalegre	M	17-11-1947
121	Leonor Coutinho Pereira dos Santos	PS	Lisboa	F	02-03-1947
122	Lúcio Maia Ferreira	PS	Porto	M	26-03-1950
123	Luís Afonso Cerqueira Natividade Candal	PS	Aveiro	M	02-03-1971
124	Luís Álvaro Barbosa de Campos Ferreira	PSD	Viana do C	M	26-11-1961
125	Luís António Pita Ameixa	PS	Beja	M	13-10-1960
126	Luís Emídio Lopes Mateus Fazenda	BE	Lisboa	M	08-10-1957
127	Luís Filipe Alexandre Rodrigues	PSD	Setúbal	M	05-02-1966
128	Luís Filipe Carloto Marques	PSD	Setúbal	M	17-07-1963
129	Luís Filipe Montenegro Cardoso de Morais Esteves	PSD	Aveiro	M	16-02-1973
130	Luís Manuel Gonçalves Marques Mendes	PSD	Aveiro	M	05-09-1957
131	Luís Maria de Barros Serra Marques Guedes	PSD	Lisboa	M	25-08-1957
132	Luís Miguel Morgado Laranjeiro	PS	Braga	M	13-08-1965
133	Luís Miguel Pais Antunes	PSD	Leiria	M	20-08-1957
134	Luís Miguel Pereira de Almeida	PSD	Coimbra	M	07-08-1970
135	Luís Pedro Russo da Mota Soares	CDS-PP	Lisboa	M	29-05-1974
136	Luísa Maria Neves Salgueiro	PS	Porto	F	02-01-1968
137	Luiz Manuel Fagundes Duarte	PS	Açores	M	06-10-1954
138	Manuel Alegre de Melo Duarte	PS	Lisboa	M	12-05-1936
139	Manuel António Gonçalves Mota da Silva	PS	Braga	M	01-05-1972
140	Manuel Filipe Correia de Jesus	PSD	Madeira	M	16-12-1941
141	Manuel Francisco Pizarro de Sampaio e Castro	PS	Porto	M	02-02-1964
142	Manuel José Mártires Rodrigues	PS	Faro	M	22-08-1949
143	Manuel Luís Gomes Vaz	PS	Bragança	M	05-10-1951
144	Manuel Maria Ferreira Carrilho	PS	Viseu	M	09-07-1951
145	Marcos da Cunha e Lorena Perestrello de Vasconcel	PS	Beja	M	23-08-1971
146	Marcos Sá Rodrigues	PS	Lisboa	M	05-04-1976
147	Maria Antónia Moreno Areias de Almeida Santos	PS	Coimbra	F	14-02-1962
148	Maria Celeste Lopes da Silva Correia	PS	Lisboa	F	08-10-1948
149	Maria Cidália Bastos Faustino	PS	Castelo Br	F	11-04-1947
150	Maria Custódia Barbosa Fernandes Costa	PS	Lisboa	F	20-06-1939
151	Maria de Belém Roseira Martins Coelho Henriques d	PS	Lisboa	F	28-07-1949
152	Maria de Fátima Oliveira Pimenta	PS	Viana do C	F	09-02-1963
153	Maria de Lurdes Ruivo	PS	Porto	F	05-11-1958
154	Maria do Rosário da Silva Cardoso Águas	PSD	Vila Real	F	21-02-1961
155	Maria do Rosário Lopes Amaro da Costa da Luz Carn	PS	Aveiro	F	14-10-1948
156	Maria Helena da Silva Ferreira Rodrigues	PS	Vila Real	F	07-05-1955
157	Maria Helena Passos Rosa Lopes da Costa	PSD	Lisboa	F	06-04-1953
158	Maria Helena Terra de Oliveira Ferreira Dinis	PS	Aveiro	F	22-06-1965
159	Maria Hortense Nunes Martins	PS	Castelo Br	F	21-09-1966
160	Maria Irene Marques Veloso	PS	Lisboa	F	07-12-1945
161	Maria Isabel Coelho Santos	PS	Porto	F	12-02-1968
162	Maria Jesuína Carrilho Bernardo	PS	Europa	F	25-11-1943
163	Maria José Guerra Gamboa Campos	PS	Porto	F	06-07-1948
164	Maria Júlia Gomes Henriques Caré	PS	Madeira	F	25-10-1954
165	Maria Luísa Raimundo Mesquita	PCP	Santarém	F	10-04-1949
166	Maria Manuel Fernandes Francisco Oliveira	PS	Setúbal	F	17-09-1960
167	Maria Manuela de Macedo Pinho e Melo	PS	Porto	F	26-03-1945
168	Maria Matilde Pessoa de Magalhães Figueiredo de S	PS	Coimbra	F	08-07-1943
169	Maria Odete da Conceição João	PS	Leiria	F	03-01-1958
170	Maria Ofélia Fernandes dos Santos Moleiro	PSD	Leiria	F	21-06-1949
171	Maria Teresa Alegre de Melo Duarte Portugal	PS	Coimbra	F	23-08-1939
172	Maria Teresa Filipe de Moraes Sarmento Diniz	PS	Setúbal	F	18-10-1957
173	Mariana Rosa Aiveca Ferreira	BE	Setúbal	F	03-02-1954

174	Mário da Silva Coutinho Albuquerque	PSD	Santarém	M	19-11-1940
175	Mário Henrique de Almeida Santos David	PSD	Leiria	M	20-08-1953
176	Mário Patinha Antão	PSD	Braga	M	26-06-1945
177	Maximiano Alberto Rodrigues Martins	PS	Madeira	M	30-10-1949
178	Melchior Ribeiro Pereira Moreira	PSD	Viseu	M	23-01-1964
179	Miguel Bento Martins da Costa de Macedo e Silva	PSD	Braga	M	06-05-1959
180	Miguel Bernardo Ginestal Machado Monteiro Albuqu	PS	Viseu	M	01-09-1965
181	Miguel Fernando Cassola de Miranda Relvas	PSD	Santarém	M	05-09-1961
182	Miguel Jorge Pignatelli de Ataíde Queiroz	PSD	Porto	M	21-04-1934
183	Miguel Jorge Reis Antunes Frasquilho	PSD	Guarda	M	12-11-1965
184	Miguel Tiago Crispim Rosado	PCP	Lisboa	M	27-08-1979
185	Nelson Madeira Baltazar	PS	Santarém	M	15-06-1951
186	Nuno André Araújo dos Santos Reis e Sá	PS	Braga	M	02-04-1976
187	Nuno Maria de Figueiredo Cabral da Câmara Pereira	PSD	Lisboa	M	19-06-1951
188	Nuno Mário da Fonseca Oliveira Antão	PS	Santarém	M	31-03-1975
189	Nuno Miguel Miranda de Magalhães	CDS-PP	Setúbal	M	04-03-1972
190	Osvaldo Alberto Rosário Sarmento e Castro	PS	Leiria	M	10-08-1946
191	Paula Cristina Barros Teixeira Santos	PS	Vila Real	F	16-08-1966
192	Paula Cristina Ferreira Guimarães Duarte	PS	Porto	F	11-11-1965
193	Paula Cristina Nobre de Deus	PS	Évora	F	05-03-1970
194	Paulo Artur dos Santos Castro de Campos Rangel	PSD	Porto	M	18-02-1968
195	Paulo Miguel da Silva Santos	PSD	Porto	M	24-03-1971
196	Paulo Sacadura Cabral Portas	CDS-PP	Aveiro	M	12-09-1962
197	Pedro Augusto Cunha Pinto	PSD	Lisboa	M	24-10-1956
198	Pedro Manuel Farmhouse Simões Alberto	PS	Lisboa	M	27-06-1961
199	Pedro Miguel de Azeredo Duarte	PSD	Porto	M	12-07-1973
200	Pedro Miguel de Santana Lopes	PSD	Lisboa	M	29-06-1956
201	Pedro Nuno de Oliveira Santos	PS	Aveiro	M	13-04-1977
202	Pedro Quartin Graça Simão José	PSD	Lisboa	M	18-05-1952
203	Regina Maria Pinto da Fonseca Ramos Bastos	PSD	Aveiro	F	04-11-1960
204	Renato Luís de Araújo Forte Sampaio	PS	Porto	M	03-05-1952
205	Renato Luís Pereira Leal	PS	Açores	M	17-06-1953
206	Ricardo Jorge Olímpio Martins	PSD	Vila Real	M	11-09-1972
207	Ricardo Manuel de Amaral Rodrigues	PS	Açores	M	01-06-1958
208	Ricardo Manuel Ferreira Gonçalves	PS	Braga	M	13-09-1957
209	Rita Manuela Mascarenhas Falcão dos Santos Miguel	PS	Guarda	F	28-07-1974
210	Rita Susana da Silva Guimarães Neves	PS	Lisboa	F	10-05-1976
211	Rosa Maria da Silva Bastos da Horta Albernaz	PS	Aveiro	F	04-09-1947
212	Rosalina Maria Barbosa Martins	PS	Viana do C	F	22-12-1955
213	Rui do Nascimento Rabaça Vieira	PS	Lisboa	M	14-04-1948
214	Rui Manuel Lobo Gomes da Silva	PSD	Lisboa	M	23-08-1958
215	Sandra Marisa dos Santos Martins Catarino da Costa	PS	Setúbal	F	05-03-1977
216	Sérgio André da Costa Vieira	PSD	Porto	M	22-08-1970
217	Sónia Ermelinda Matos da Silva Fertuzinhos	PS	Braga	F	12-01-1973
218	Sónia Isabel Fernandes Sanfona Cruz Mendes	PS	Santarém	F	10-12-1971
219	Telmo Augusto Gomes de Noronha Correia	CDS-PP	Lisboa	M	04-02-1960
220	Teresa Margarida Figueiredo de Vasconcelos Caeiro	CDS-PP	Leiria	F	14-02-1969
221	Teresa Maria Neto Venda	PS	Braga	F	30-08-1953
222	Umberto Pereira Pacheco	PS	Lisboa	M	27-11-1952
223	Vasco Manuel Henriques Cunha	PSD	Santarém	M	23-03-1965
224	Vasco Seixas Duarte Franco	PS	Lisboa	M	27-04-1952
225	Vitalino José Ferreira Prova Canas	PS	Santarém	M	14-07-1959
226	Vítor Hugo Machado da Costa Salgado de Abreu	PS	Braga	M	24-01-1977
227	Vítor Manuel Bento Baptista	PS	Coimbra	M	27-05-1952
228	Vítor Manuel Pinheiro Pereira	PS	Castelo Br	M	16-08-1962
229	Vitor Manuel Sampaio Caetano Ramalho	PS	Setúbal	M	21-07-1948
230	Zita Maria de Seabra Roseiro	PSD	Coimbra	F	25-05-1949

Bibliografia / Outros Recursos

• BARNETT, V. (1997) – Sample Survey: Principles & Methods, Arnold, London.

GRAÇA MARTINS, M.E. et al (1999) – Introdução às Probabilidades e à Estatística, Edição da Universidade Aberta.

GRAÇA MARTINS, M.E. (2005) – Introdução à Probabilidade e à Estatística – Com complementos de Excel. Edição da Sociedade Portuguesa de Estatística.

GRAÇA MARTINS, M.E. et al (2001) – Estatística – 10º ano de escolaridade, Edição do Ministério da Educação – Departamento do Ensino Secundário.

GRAÇA MARTINS, M.E. e Loura, L. (2001) – Matemática para as Ciências Sociais – Anexo para apoio à interpretação do programa.

MOORE, D. (1992) – What is Statistics in Perspectives on Contemporary Statistics, Edição de David Hoaglin e David Moore, The Mathematical Association of America.

MOORE, D. ET AL (1996) – Introduction to the Practice of Statistics, Freeman, New York.

MOORE, D. (1996) – The Basic Practice of Statistics, Freeman, New York.

MOORE, D. (1997) – Statistics – Concepts and Controversies, Freeman, New York.

MURTEIRA, B. (1993) – Análise Exploratória de Dados. Estatística Descritiva, McGraw-Hill.

COMAP, (2000) – For all Practical Purposes: Mathematical Literacy in Todays World, Freeman and Company, New York.

ROSSMAN, A. et al (2001) –Workshop Statistics – Discovery with data, Key College Publishing.

TANNENBAUM. P. et al (1998) – Excursions in modern Mathematics, Prentice Hall. VICENTE, P., REIS, E., FERRÃO, F. (1996) – Sondagens, Edições Sílabo.

Artigos da revista /TEACHING STATISTICS

AGEEL, M.I. – Spreadsheets as a Simulation Tool for Solving Probability Problems, Vol 24, 2, 51

Hodgson, T., and Borkowski, J. - Why Stratify? Vol 20, 1, 68-71. NEVILLE, H. – Handling Continuous Data in Excel, Vol 25, 2, 42-45.

NEVILLE, H. – Charts in Excel, Vol 26, 2, 49-53.

Páginas na Internet

ESCOLA SECUNDÁRIA TOMAZ PELAYO E INSTITUTO NACIONAL DE ESTATÍSTICA PROJECTO ALEA – http://www.alea.pt

INSTITUTO NACIONAL DE ESTATÍSTICA – www.ine.pt/ Tem informação sobre Portugal, ao nível da freguesia.

EUROSTAT – europa.eu.int/comm/eurostat/ Tem informação relativa aos diversos países da Europa.

WORLD HEALTH ORGANIZATION – http://www.who.int/research/en/ Tem informação sobre temas ligados à saúde, para todos os países do mundo.

WORLD IN FIGURES – http://www.stat.fi/tup/maanum/index_en.html Tem informação das mais diversas áreas, tais como população e estatísticas vitais, cultura, religiões, emprego, consumo, etc., relativa a todos os países do mundo.

Representações Gráficas

Ana Alexandrino da Silva

Representações Gráficas

Notas sobre a criação e apresentação de alguns tipos de gráficos

Ana ALexandrino da Silva

Sumário:

1.1. Introdução

Os gráficos encontram-se presentes em quase todos os meios de divulgação de informação, designadamente nos jornais e revistas, nos manuais escolares, nas apresentações públicas e até os nossos relatórios individuais já não passam sem eles.

Contudo, fazer um gráfico ou um mapa que de facto informe e seja, simultaneamente, apelativo, legível e coerente com os dados não é tarefa fácil...

A grande vantagem dos gráficos reside na sua capacidade de contar uma história de forma interessante e atractiva permitindo compreender rapidamente fenómenos que dificilmente seriam percebidos de outra forma. Contudo, tal não implica que este processo seja feito de forma simples, sendo necessário muito trabalho e cuidado.

Existem inúmeras formas de apresentar figurativamente a informação estatística e no caso particular dos gráficos são tantas as possibilidades que houve necessidade de restringir o objecto deste dossiê aos gráficos mais correntes e não proceder a uma abordagem exaustiva.

História dos gráficos

A história dos gráficos estatísticos é relativamente recente. O maior avanço deu-se apenas há cerca de 200 anos, em 1786, graças a William Playfair que inventou a maioria das formas gráficas que conhecemos hoje: o gráfico de barras, o gráfico de linhas baseado em dados económicos e o gráfico circular.

Enquanto no século XIX, se assistiu à criação e disseminação alargada dos gráficos estatísticos na comunidade científica, no século XX houve um aumento exponencial da sua utilização em documentos de divulgação alargada e acessíveis ao grande público.

Desde Playfair muito se avançou na divulgação dos gráficos estatísticos, usados agora um pouco por todo o lado - nas escolas, nos média, etc. mas a maioria dos gráficos actualmente em uso datam desse tempo (século XVIII/XIX).

Com o aparecimento dos computadores retomaram-se os estudos desenvolvidos na área dos gráficos sendo imperativo fazer referência a Edward TUKEY (1977) responsável pela invenção de gráficos indispensáveis na análise exploratória de dados, como sejam a caixa de bigodes e o diagrama de caule e folhas, entre outros.

Reflexões sobre a construção de gráficos

Com a tecnologia existente, a produção de gráficos está ao alcance de todos. Mas é importante ter alguns cuidados.

Neste dossiê serão compilados um conjunto de critérios subjacentes à criação de um gráfico. Este processo inicia-se no momento em que se decide optar por um gráfico e só termina quando o resultado se considera satisfatório.

Com a enchente de gráficos que se vive nos dias de hoje, o leitor tornou-se exigente. A reacção a um gráfico demasiado 'carregado' de informação, pode ser o afastamento, e mesmo que lhe seja dedicado alguma atenção, poucas recordações subsistem. Este distanciamento também pode ser causado por um excesso de elementos gráficos não informativos, originando gráficos apelidados por TUFTE (1983) de lixo gráfico (chart junk).

Antes de mais, deve questionar-se a necessidade de mostrar os dados graficamente. De facto, em certos casos, não fará sentido recorrer a um gráfico quando o objectivo não é dar uma imagem, mas sim fornecer dados concretos, quer

em situações em que apenas se detêm poucos valores como para os casos em que se pretendem divulgar muitos dados.

Outro dos problemas com que se debate quem produz gráficos é a restrição de espaço, obrigando à acumulação de informação num único gráfico ou a um dimensionamento reduzido das imagens, com consequências na sua leitura.

WALLGREN (1996) sintetiza esta fase preparatória em oito perguntas que não podem ser respondidas separadamente:

- Um gráfico é realmente a melhor opção?
- Qual é o público-alvo?
- Qual é o objectivo do gráfico?
- Que tipo de gráfico se deve usar?
- Como deve ser apresentado o gráfico?
- Qual deve ser o tamanho do gráfico?
- Deverá ser usado apenas um gráfico?
- A que meios técnicos se deve recorrer?

Após ter sido seleccionado o modelo de gráfico mais adequado ao contexto respectivo, inicia-se a construção do gráfico propriamente dita.

Quando finalmente se pensa ter obtido o gráfico pretendido, torna-se fundamental proceder a uma análise crítica, no sentido de compreender se esta é a forma mais eficaz de transmitir a mensagem inicial. Um gráfico mal compreendido pode provocar uma interpretação errada. Por outro lado, um gráfico visualmente desagradável pode afastar o leitor, em vez de o informar: "Um mau gráfico é pior do que nenhum gráfico" (WALLGREN, 1996, p. 89).

Na tentativa de encontrar a melhor imagem que satisfaça todos os requisitos iniciais, entra-se num processo iterativo que só termina quando se garante uma elevada legibilidade e pertinência.

Por conseguinte, a adopção do gráfico apenas se pode consumar após serem formuladas, e convenientemente respondidas, as seguintes perguntas:

- O gráfico é fácil de ler?
- O gráfico pode ser mal interpretado?
- O gráfico tem o tamanho e a forma certa?
- O gráfico está localizado no sítio certo?
- O gráfico beneficia por ser a cores?
- A compreensão do gráfico foi testada com alguém?

Formatação do gráfico

A representação gráfica é um tema complexo onde se cruzam áreas tão diversas como a estatística, o desenho e a psicologia. Um gráfico pode representar correctamente as variáveis, conter todos os elementos necessários e não ser, nem atractivo, nem de fácil leitura.

É possível redesenhar um gráfico, através da modificação ou supressão de alguns elementos gráficos, sem que haja perda de informação (TUFTE, 1983). No entanto, muitos dos gráficos divulgados necessitam de uma certa sofisticação a este nível, sendo comum encontrar imagens visualmente semelhantes provenientes do assistente de gráficos do software Excel, que por serem imagens muito vistas, e portanto cansativas, não atraem o leitor.

O Excel permite alguma manipulação visual no leque de gráficos que apresenta. Seguidamente, é apresentado um exemplo de como se pode melhorar a leitura, modificando o aspecto do gráfico.

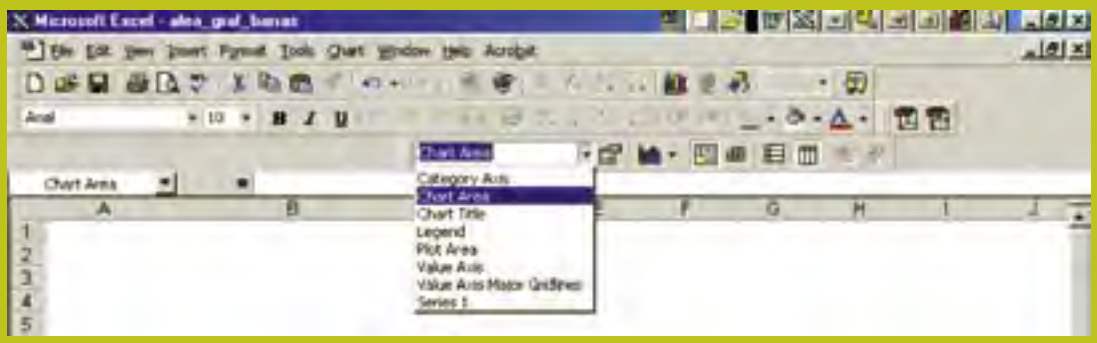
A primeira coisa a ter em conta quando se pretende elaborar um gráfico é a organização dos dados. O tipo de gráfico selecionado é influenciado pela forma como estão dispostos os dados. A melhor forma é dispor os dados numa tabela, com as respectivas identificações, para que estes possam ser utilizados como títulos e legendas do gráfico.

A tabela dos dados:

Qualificação académica da população dos 15-64 anos

Sexo / Qualificação académica	**Masculino**	**Feminino**
Nenhum	7,5%	11,3%
Obrigatório	69,3%	61,5%
Secundário	15,7%	16,7%
Superior	7,5%	10,5%

Possibilidades de formatação de gráficos com o Excel



1. Área do gráfico (chart area)
2. Legenda (legend)
3. Eixo das categorias (category axis)
4. Área do desenho (plot area),
5. Eixo de valores (value axis),
6. Linhas de grelha (gridlines)
7. Série de dados (series)

Descrição do processo de formatação

Partindo do critério de que pelo menos dois terços da área do gráfico devem ser afectados às barras ou, genericamente, à área do desenho, (SCHMID, 1992), ampliou-se o espaço preenchido por estas.

No eixo dos valores foram retiradas as casas decimais e suprimidos alguns valores, apesar de se terem mantido as respectivas linhas de grelha. Poder-se-ia ter deixado apenas o sinal de % junto ao último valor, retirando os sinais de % nos valores 0 e 40. Foi também retirada a linha do eixo e as marcas dos eixos, para além de se ter encurtado a amplitude do intervalo de valores dado que a maior das barras não ultrapassava os 80%.

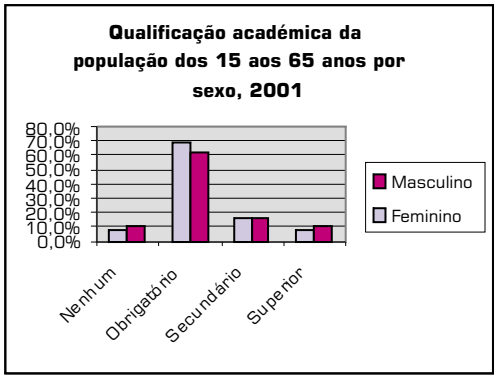
A linha do eixo das categorias apresenta um maior peso visual do que as restantes linhas auxiliares, estando as designações orientadas horizontalmente para facilitar a leitura.

Foram retiradas as molduras do gráfico, da legenda e das barras por se considerar não existir qualquer vantagem em mantê-las, sobrecarregando desnecessariamente a construção gráfica, e posicionou-se a legenda no interior do gráfico para diminuir a distância percorrida pelos olhos entre as componentes e as suas designações. Mudaram-se as cores das barras, aumentou-se a sua grossura e simultaneamente diminuiu-se o espaço entre grupos de barras.

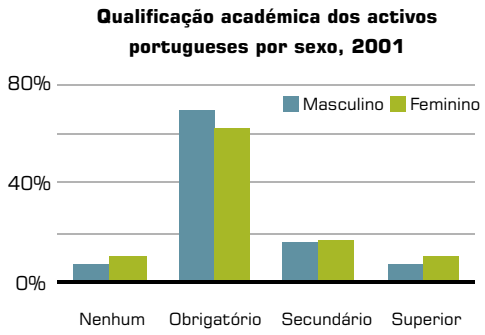
A figura "Depois" não é mais do que a figura "Antes" depois de transformada recorrendo às potencialidades do software.

Figura 1 – Gráfico de barras antes e depois de ser modificado através do Excel

Antes...



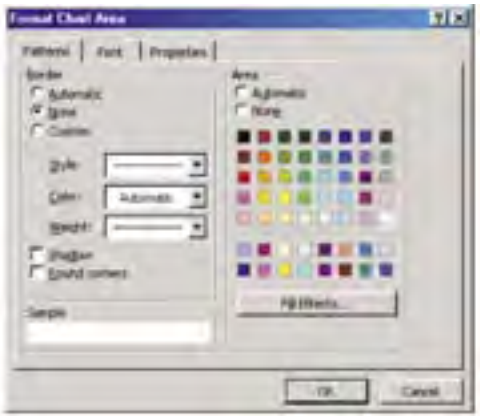
...Depois

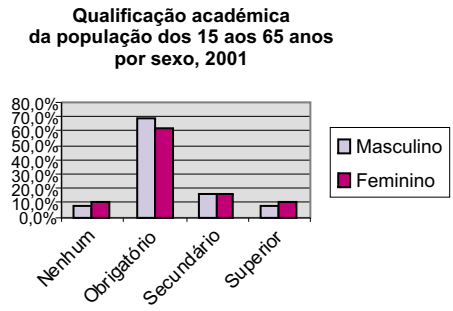


(Re)desenho do gráfico através do Excel

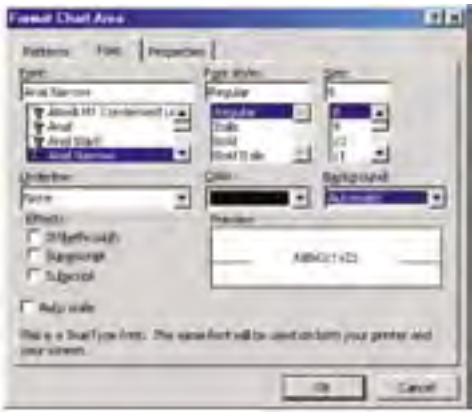
1 – Área do gráfico

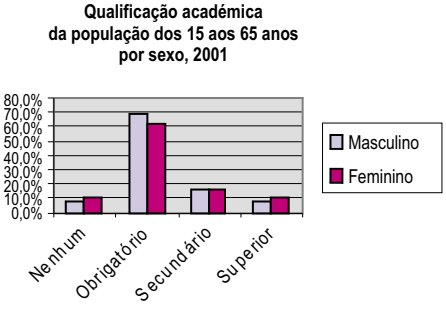
- Gráfico sem moldura e com área a branco...





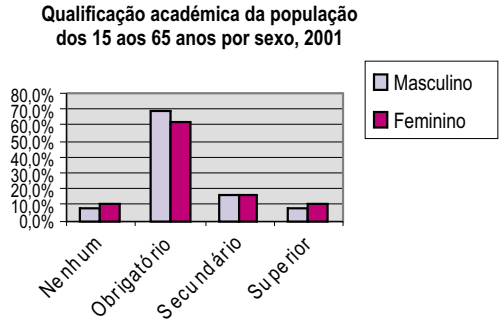
- Gráfico com tipo de letra Arial narrow, tamanho 8...



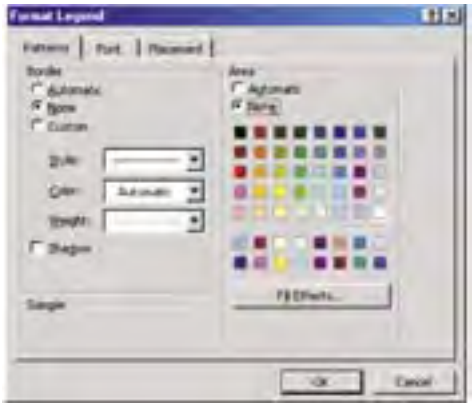


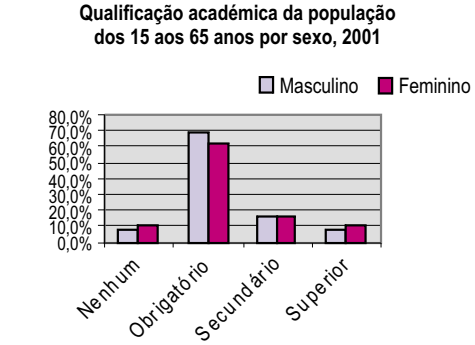
2 – Legenda

• Gráfico com legenda no canto superior direito...



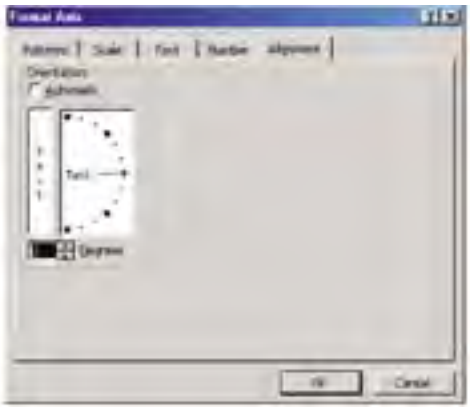
• Gráfico com legenda sem moldura, fundo e símbolos na horizontal...

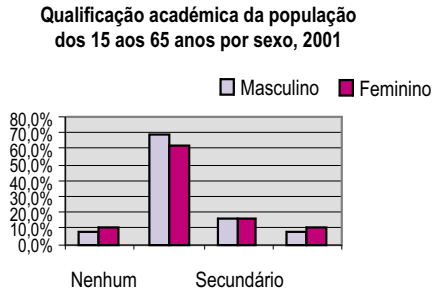




3 - Eixo das categorias

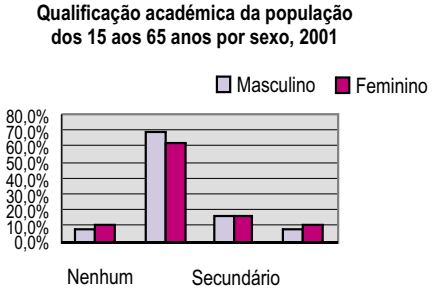
• Gráfico com identificações das categorias na horizontal...



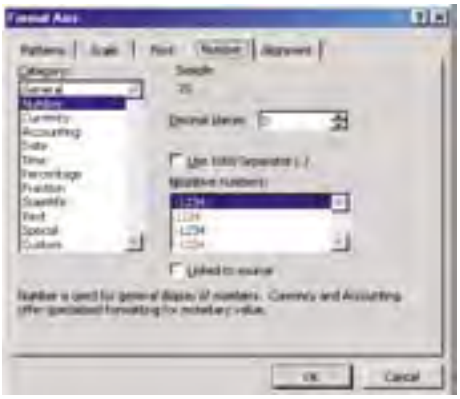


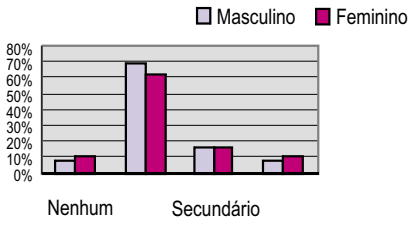
4 - Eixo dos valores

• Gráfico sem linha e tick marks no eixo dos valores...



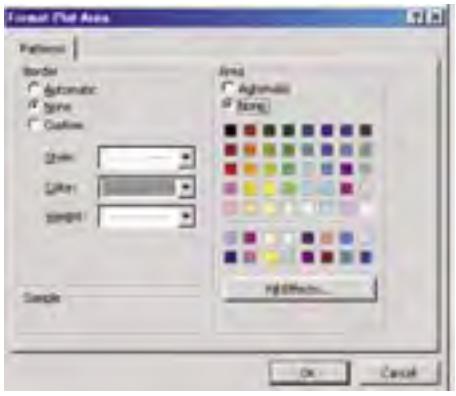
• Gráfico com eixo de valores sem casas decimais...

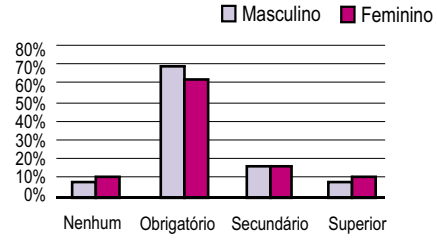




5 - Área do desenho

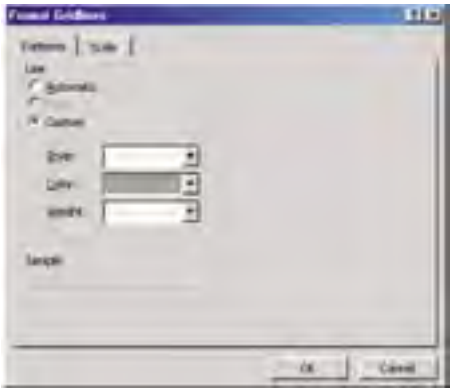
• Gráfico com área de desenho a branco e sem moldura...

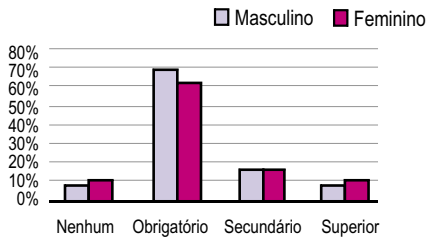




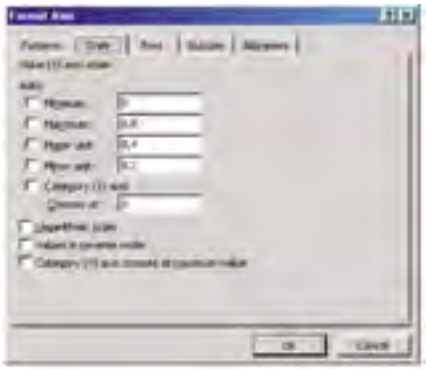
6 - Linhas de grelha

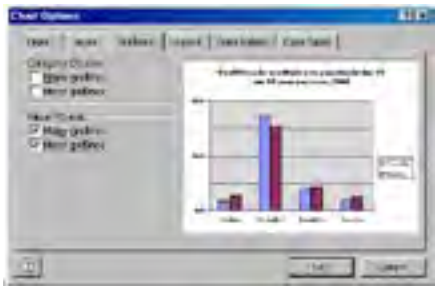
• Gráfico com linhas de grelha a cinzento...

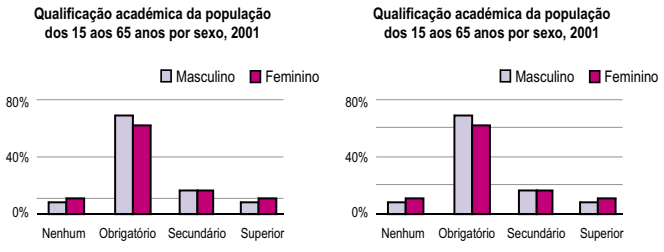




• Gráfico com escala de valores para os dois tipos de linhas de grelha...

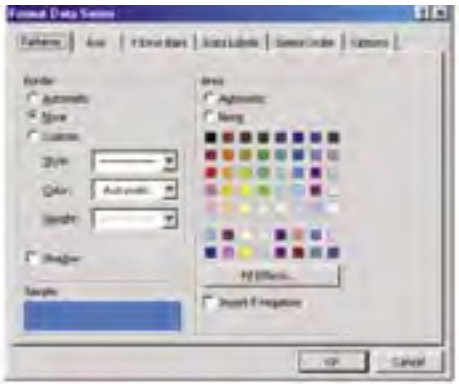


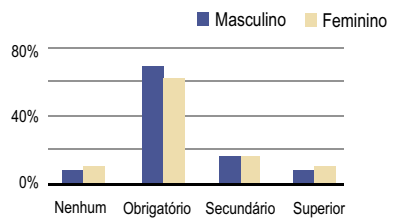




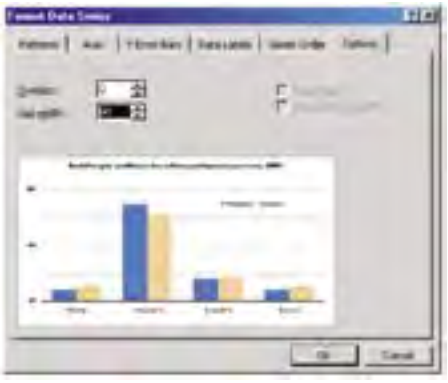
7 - Série de dados

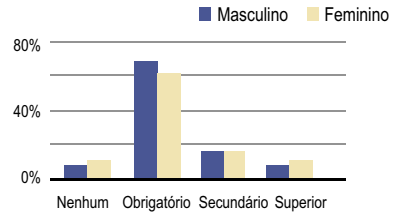
• Gráfico com barras de cor diferente e sem moldura...





• Gráfico com espaço entre as barras alterado...





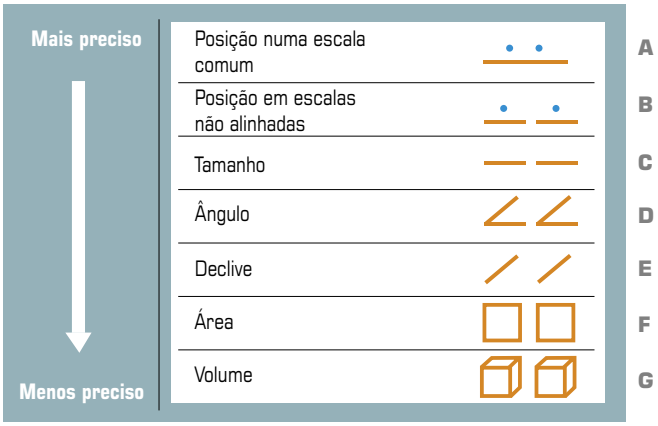
Estudos perceptivos

A percepção gráfica é um dos elementos mais importantes a ter em conta quando se elabora um gráfico, porque permite dar uma fundamentação científica à construção gráfica e sustentar a escolha de uma forma gráfica em detrimento de outra. A leitura das imagens pode ser condicionada pela dificuldade em estimar correctamente os dados representados.

Na fase da construção, a informação é codificada no gráfico através de símbolos, comprimentos, declives dos segmentos de recta, áreas, textura ou cor. Quando um gráfico é analisado, a informação codificada é visualmente descodificada, sendo o processo de descodificação, denominado de percepção gráfica, um factor de controlo na capacidade de um gráfico transmitir informação (CLEVELAND, MCGILL, 1987).

A extracção de informação a partir dos gráficos envolve tarefas perceptivas realizadas pelo sistema visual olho-cérebro. No quadro seguinte, estas tarefas estão ordenadas segundo a sua precisão na extracção de informação quantitativa. Quanto menos precisa for a tarefa preceptiva maior o erro de leitura, ou seja, maior a diferença entre o valor percebido e o valor correcto.

Figura 2 – Avaliação de tarefas perceptivas ordenadas segundo a sua precisão

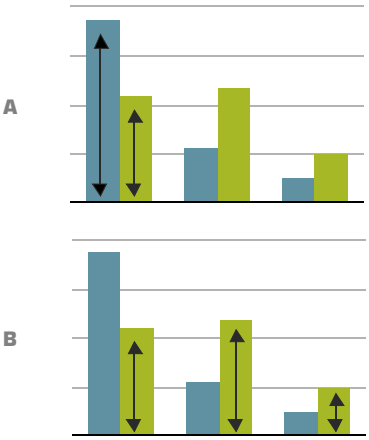


(adaptado de CLEVELAND, MCGILL, 1984, 1987)

Por exemplo, nos gráficos de barras agrupadas, o leitor estima os valores através da posição das barras na mesma escala ou em escalas separadas, consoante a forma de apresentação dos dados.

A comparação entre barras próximas (Figura 3 - A) é melhor do que a comparação entre barras mais afastadas (Figura 3 - B), ou seja, nesta última forma o leitor tem mais dificuldade em estimar os valores.

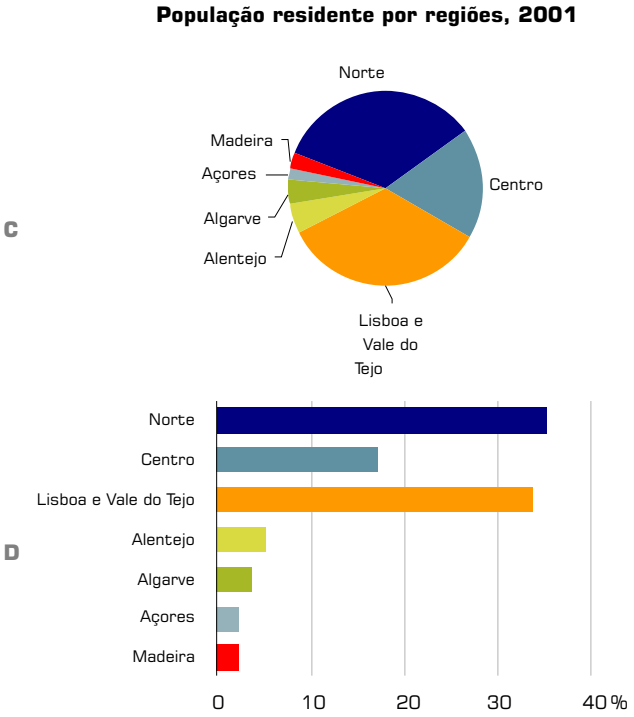
Figura 3 – Exemplos das tarefas A e B



Na comparação entre gráficos de barras e gráficos circulares, os primeiros revelaram-se perceptivamente mais adequados, dado que a estimação dos comprimentos demonstrou ser duas vezes mais precisa que a estimação de ângulos. Veja-se o caso da região Norte e da região de Lisboa e Vale do Tejo. No gráfico circular não se tem a percepção de qual é o maior. Pelo contrário, o gráfico de barras mostra claramente a diferença.

É comum encontrar gráficos a três dimensões em que a profundidade não descreve qualquer variável. Como o volume é o que maiores problemas traz em termos de percepção, não deve ser utilizado.

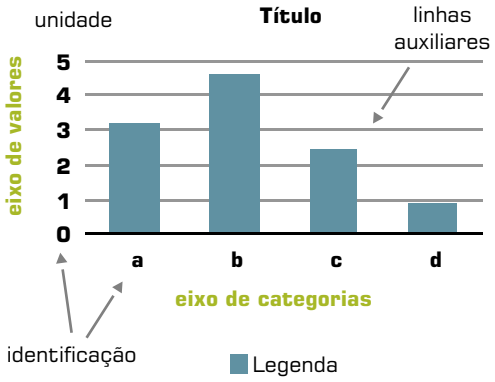
Figura 4 – Exemplos das tarefas C e D



Elementos do gráfico

Os gráficos incorporam o seguinte conjunto de elementos: o título, os eixos de valores e de categorias (baseados no sistema de coordenadas), a legenda, as identificações dos dados e as linhas auxiliares (Figura 5).

Figura 5 – Elementos de um gráfico



Estes elementos são constituídos por símbolos gráficos (pontos, linhas, números, letras, etc.) e sua variação (cor, valor, etc.).

A área do gráfico pode conter todos estes elementos, ou apenas alguns, sistematizados em duas áreas complementares: a 'área do desenho' (plot area) onde está contida a representação gráfica propriamente dita e a 'área exterior' (chart area) onde normalmente estão posicionadas as componentes de auxílio à leitura (título, legenda e identificações).

Área exterior:

Título

O título deve estar presente em qualquer tipo de representação gráfica e ser escrito com vista a orientar o leitor na sua interpretação. Para tal, deve ser redigido por forma a responder às perguntas: O Quê, Onde e Quando. Simultaneamente, deve ser conciso, relevante e claro, ou seja, conter apenas informação essencial para uma interpretação correcta do gráfico. Por exemplo, um gráfico integrado numa publicação temática, relativa a uma dada região ou a um certo período temporal não necessita de incluir sistematicamente a mesma referência regional ou temporal. Sugere-se, igualmente, o posicionamento do título antes do gráfico funcionando como um cabeçalho, centrado horizontalmente (SCHMID, 1992) ou alinhado à esquerda (WALLGREN, 1996).

Identificações (ou rótulos)

Neste conceito genérico enquadra-se toda a informação escrita posicionada na área exterior: as designações dos eixos de valores e categorias, a referência às respectivas unidades e eventuais notas (fontes da informação, esclarecimentos, etc.).

A orientação de todas as palavras deve ser, preferencialmente, horizontal e estar de acordo com o sentido da leitura das palavras escritas na língua, no nosso caso, da esquerda para a direita.

Na maior parte dos gráficos ou tabelas não se justifica uma grande precisão nos dados apresentados.

Um número excessivo de casas decimais (separadas das unidades por uma vírgula), ou mesmo uma casa decimal em valores elevados, envolve um rigor desnecessário e prejudicial à leitura. Para ser mais legível, a formatação de valores acima dos milhares pode ser feita com um espaço em vez de com um ponto ou uma vírgula.

Os valores da escala devem ser expressos em valores arredondados múltiplos de 1, 2 e 5 (ex. 5, 10, 25, 50, 100, etc.). Aconselha-se a que não se apresentem números com mais de 5 dígitos, adaptando, caso seja preciso, a unidade para milhares ou milhões.

Legenda

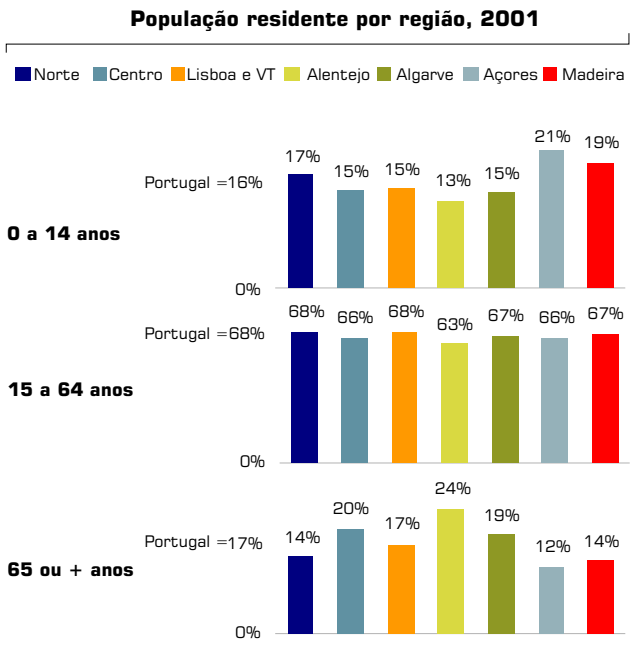
Uma boa legenda deve fazer mais do que simplesmente etiquetar as componentes do gráfico. Deve dizer-nos o que é importante e qual é o objectivo do gráfico: informar o leitor e obrigar quem faz o gráfico a estruturar a informação (CLEVELAND, MCGILL, 1984a).

A legenda é constituída por símbolos e respectivas designações. O preenchimento dos símbolos (cor ou outros) deve ser realizado de modo a que não haja lugar para qualquer confusão visual entre eles e, consequentemente, para que exista uma ligação clara entre os símbolos e a componente representada. As designações, por seu lado, devem ser claras e concisas, deixando para notas adjacentes eventuais esclarecimentos.

Os símbolos devem aparecer na mesma ordem que as respectivas componentes: horizontalmente quando estão lado a lado (Figura 6) e verticalmente quando estão umas sobre as outras (WALLGREN, 1996).

Aconselha-se a manutenção da legenda para gráficos em que as componentes surjam mais do que uma vez (Figura 6).

Figura 6 – Gráfico com uma legenda comum



Note-se que a localização da legenda na área exterior obriga o sistema visual a alternar a procura de informação entre a legenda e o gráfico, dificultando a sua interpretação imediata. Por este facto, é aconselhada sempre que possível a omissão da legenda e o posicionamento das designações junto das respectivas componentes, nomeadamente em gráficos de linhas (ver Figura 8) e circulares.

As designações da legenda podem ser deslocadas da 'área externa' para a 'área do desenho', permitindo não só que o próprio gráfico ocupe menos espaço, mas também diminuir a distância percorrida pelo sistema visual (ver Figura 10, onde essas designações surgem junto às linhas dos dados).

Área do desenho:

Eixo de categorias ou variáveis

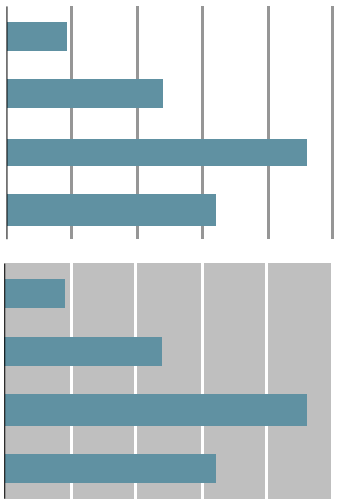
Neste eixo estão posicionadas as variáveis ou categorias que se pretendem retratar. No caso de gráficos que representam séries que evoluem ao longo do tempo, a este eixo estão associados os períodos temporais, em que a cada mês, trimestre, ano ou outro, corresponderá apenas um ponto ou uma barra no gráfico. Esta relação é obviamente unívoca, ou seja, não faz sentido representar numa mesma barra valores anuais e semestrais, ou no eixo anos e décadas, ou no mesmo espaço valores anuais e trimestrais (TUFTE, 1983).

O eixo das categorias deve ser visualmente mais 'pesado' do que as restantes linhas auxiliares (Figura 5) (SCHMID, 1992).

Linhas auxiliares (ou linhas de grelha)

Um dos elementos gráficos visualmente mais monótono são as linhas auxiliares. Devem, por isso ser suprimidas ou abafadas de tal forma que a sua presença se torne implícita. Ainda que possam auxiliar a leitura dos dados, a maioria das linhas auxiliares escuras tem um grande peso visual, encobrindo muitas vezes, o mais importante do gráfico: a informação. Quando forem realmente necessárias deve-se optar por usar uma cor neutra e, no caso particular de um fundo branco, a cor cinzenta (Figura 7).

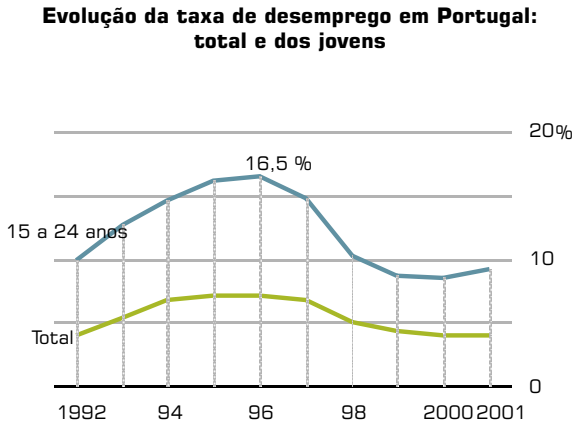
Figura 7 – Linhas auxiliares em fundo branco e de cor



Em certos casos, em particular nas séries temporais, pode ser considerado importante incluir linhas auxiliares verticais como auxilio à leitura de valores, por forma a complementar a leitura evolutiva da série com a leitura de valores em particular (Figura 8).

Evolução da taxa de desemprego em Portugal: total e dos jovens

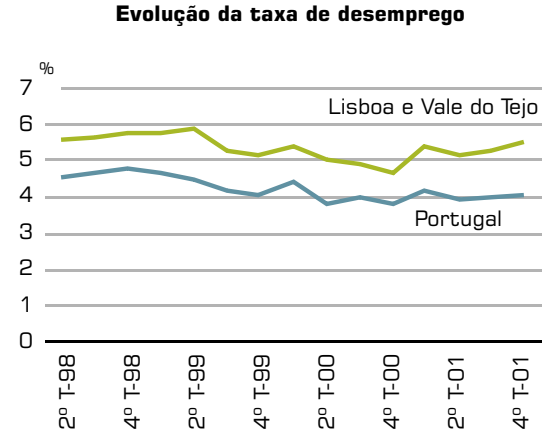
Figura 8 - Linhas auxiliares verticais num gráfico de linhas



Eixo de valores

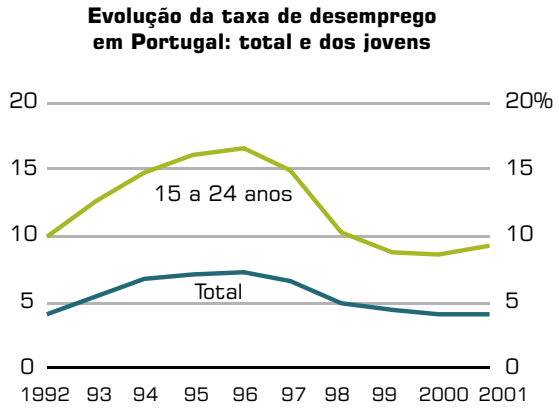
Na maioria dos gráficos de séries temporais, os dados mais recentes estão situados à direita e longe das identificações do eixo dos valores, normalmente localizados à esquerda (Figura 9), fazendo com que o olho humano tenha que se movimentar alternadamente entre os dados e os valores ao longo das margens do gráfico.

Figura 9 – Eixo de valores com identificações à esquerda



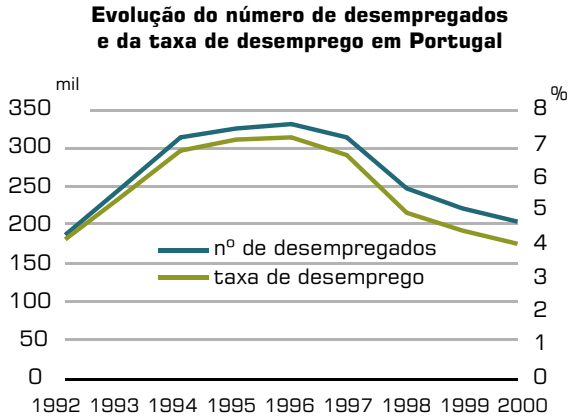
Esta imprecisão na leitura pode ser atenuada posicionando o eixo à direita junto dos dados mais recentes (ver Figura 8), duplicando o eixo (Figura 10), ou posicionando os valores junto das coordenadas respectivas (TUFTE, 1983).

Figura 10 – Gráfico com duplicação do eixo



Os gráficos com dois eixos distintos são normalmente utilizados quando se têm diferentes unidades de medida (Figura 11) ou existem diferenças consideráveis de valores nas categorias de uma variável. Este tipo de gráficos deve ser evitado dado que é normalmente de difícil interpretação e, em muitos casos, bastante confuso (SCHMID, 1992).

Figura 11 – Gráfico com dois eixos distintos



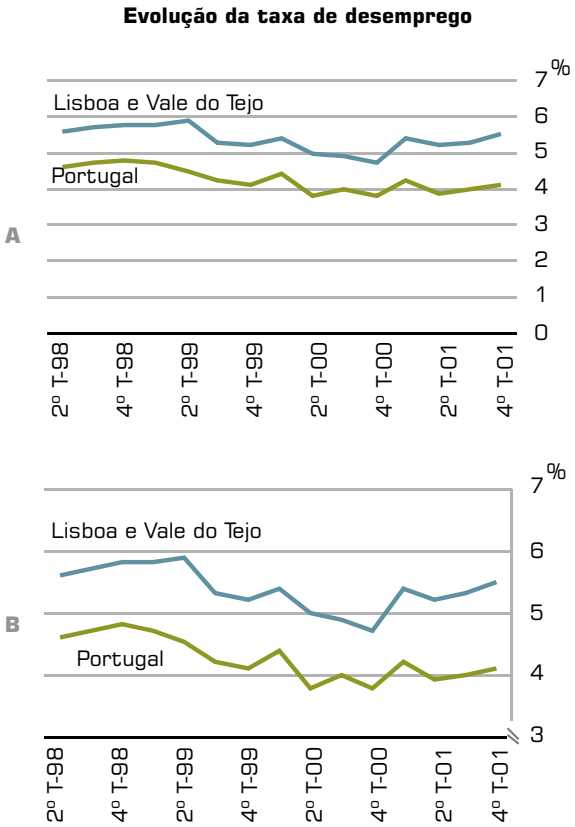
Quebra de escala

Por princípio, deve privilegiar-se a escala completa (com início em zero ou noutro valor de referência) em nome da honestidade na apresentação (Figura 12 - A). Contudo, essa quebra é admissível nos casos em que a informação apresenta pequenas variações, desde que acompanhada por uma simbologia perceptível ao leitor (Figura 12 - B).

Para melhor compreender os dados na fase da análise exploratória não existe qualquer problema em manipular as escalas e extrapolar eventuais variações, mas na fase da divulgação, deve existir algum cuidado para não evidenciar graficamente alterações nos dados que na verdade não ocorreram.

A quebra de escala é um exemplo de como se pode distorcer a mensagem transmitida. Quando o efeito nos dados é significativamente diferente do efeito no gráfico, os valores aparecem visualmente sub ou sobre-avaliados (TUFTE, 1983).

Figura 12 – Gráficos sem e com quebra de escala



Existem dois tipos de leitura possíveis num gráfico com mais de uma série temporal: a comparação vertical em que se confronta a dimensão relativa de uma série face a outra (ex: Portugal tem uma taxa de desemprego cerca de 3/4 da de Lisboa e Vale do Tejo) e a comparação de declives em que é feita uma análise da evolução de ambas as séries.

No caso de se terem duas séries aparentemente constantes, a comparação entre elas apenas pode ser feita na vertical, dado que dificilmente se detectam, visualmente, variações na sua evolução.

Neste caso, a utilização da quebra de escala permite detectar melhor as diferenças nos declives mas a comparação vertical entre as linhas deixa de fazer qualquer sentido (WALLGREN, 1996). É esta a razão pela qual não se devem fazer quebras de escala em gráficos de barras verticais, a comparação vertical entre as barras, após uma quebra de escala, não pode ser feita.

Variáveis visuais

Jacques BERTIN, em Sémiologie graphique (1973, 2ª ed.), foi o primeiro a sistematizar os conhecimentos sobre a aparência visual dos símbolos gráficos, criando uma tipologia com as seguintes variáveis visuais.

Localização – dada através das duas dimensões x,y do plano;

Tamanho – variação em comprimento, largura ou área, estando naturalmente ligado à importância numérica dos dados;

Valor – refere-se à variação (percebida) claro-escuro da cor ou à variação preto-branco;

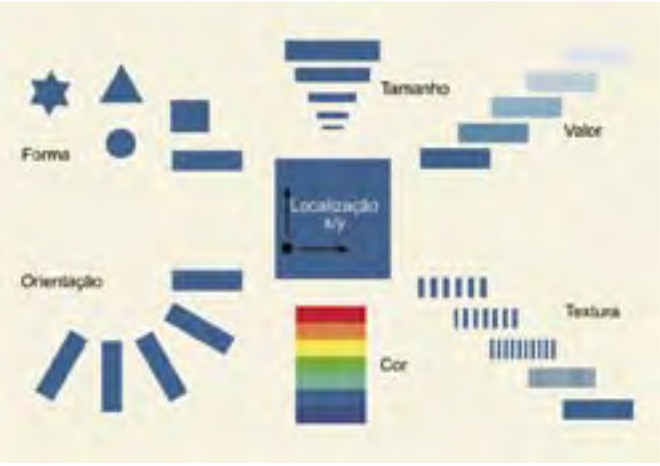
Textura – tamanho e espaçamento dos elementos gráficos que constituem o símbolo (pontos, linhas ou outros), expresso pelo número desses elementos que se repetem por unidade de comprimento;

Cor – sensação pela qual se diferencia entre porções particulares do espectro electro-magnético, isto é, azul, verde, vermelho, etc.;

Orientação – também designada por direcção, corresponde ao ângulo com a linha de leitura;

Forma – pode ser geométrica (como quadrados ou círculos) ou então irregular.

Figura 13 – As variáveis visuais segundo Bertin



1.2. Gráficos de barras

Os gráficos de barras são uma das formas mais populares de representar informação, em parte pela facilidade quer de execução, quer de leitura.

São para apresentar um conjunto de dados e também para comparar vários conjuntos de dados. Devem ser utilizados para representar variáveis discretas ou qualitativas, em termos absolutos ou relativos, ou para comparar categorias de variáveis quantitativas.

Podem, igualmente, representar a evolução de uma variável ao longo do tempo.

Neste tipo de gráficos, o leitor extrai os valores dos dados através da visualização da posição das barras relativamente a uma escala comum (CLEVELAND, MCGILL, 1984).

Normalmente, as barras começam no eixo das categorias, o que facilita a comparação das posições relativas.

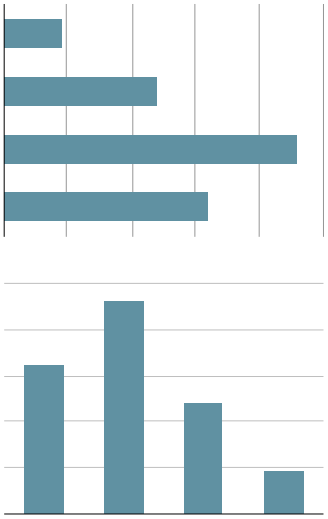
Gráficos de barras simples (verticais ou horizontais)

Num gráfico de barras, as frequências podem ser indistintamente representadas no eixo das abcissas ou das ordenadas, ou seja, as barras podem ser horizontais ou verticais (Figura 14).

Apesar do gráfico de barras verticais ser o mais comum, existem situações em que é preferível optar pela outra disposição. O gráfico de barras horizontais é considerado de leitura mais fácil, quando é expressiva a diferença entre o valor mínimo e o valor máximo da variável.

Num contexto de limitação do espaço disponível para posicionar o gráfico, é igualmente preferível optar pelo gráfico de barras horizontal, uma vez que permite a inclusão de variadas categorias sem aumentar significativamente o espaço ocupado.

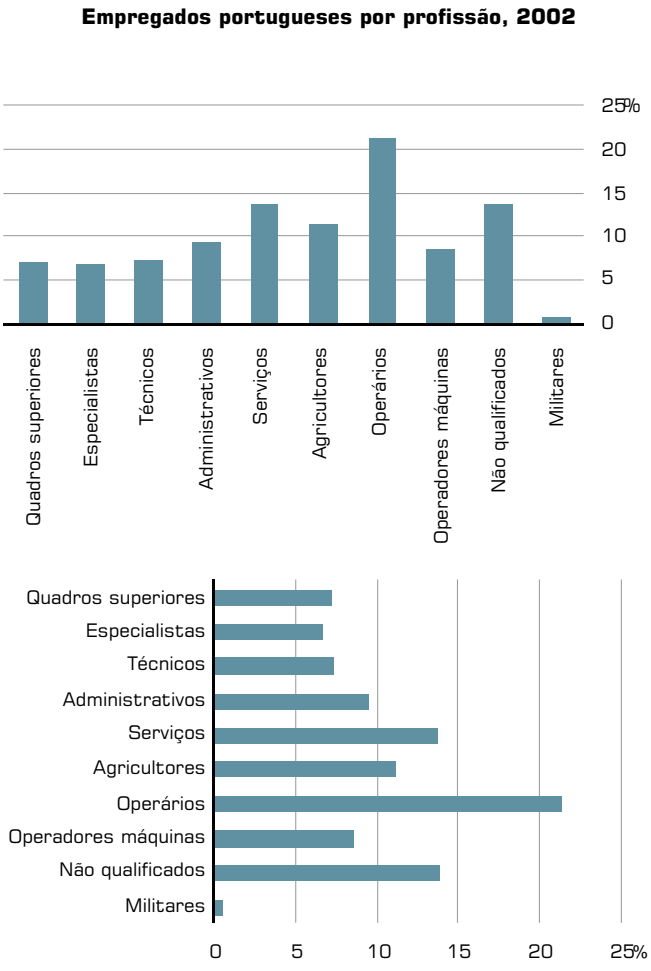
Figura 14 – Gráfico de barras horizontal e vertical



Aconselha-se o gráfico de barras horizontais para variáveis cujas categorias têm designações extensas, dado que nos gráficos de barras verticais o espaço para as designações é curto (Figura 15). Relembre-se que as designações não devem ser abreviadas, nem posicionar-se de forma a dificultar a leitura (verticalmente ou obliquamente) acabando, muitas vezes, por ocupar mais espaço do que o próprio gráfico.

Refira-se também que os gráficos de barras horizontais mostram, de forma mais clara, as diferenças entre os dados uma vez que possuem um eixo dos valores mais amplo. A Figura 15 é exemplo disso: apesar de ambos os gráficos ocuparem a mesma área, provocam efeitos visuais distintos quando se observam as categorias com maior frequência.

Figura 15 – Designações num gráfico de barras vertical e horizontal

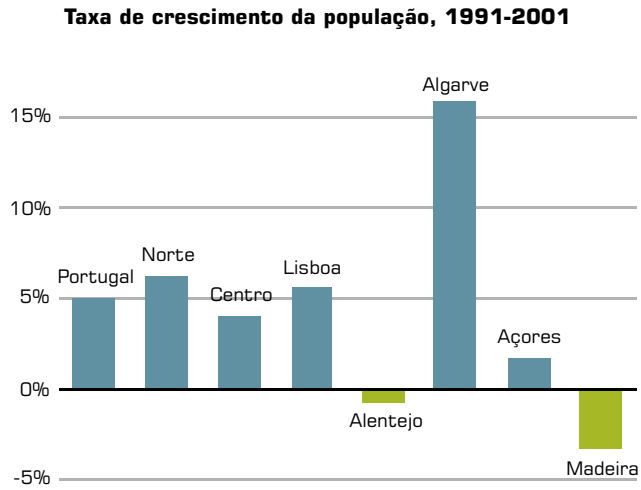


Representação de valores negativos

A representação de valores negativos é desaconselhada em gráficos de barras horizontais, dado que, convencionalmente, aos valores negativos está associada uma barra numa posição descendente (Fig. 16).

De facto, a associação visual entre esquerda e direita e valores negativos e positivos, respectivamente, pode não ser directa para um leitor menos experiente. Por essa razão, devem ser utilizados gráficos de barras verticais quando existem valores negativos.

Figura 16 – Representação de valores negativos



Algumas regras relacionadas com a construção dos gráficos de barras

Escala no eixo dos valores

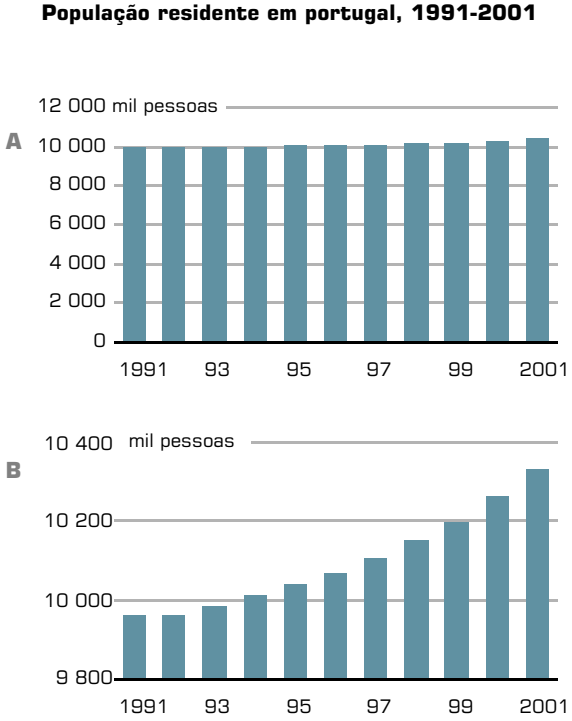
Nos gráficos de barras não é admissível a quebra de escala por deixar de ser possível efectuar comparações verticais entre categorias.

Uma quebra de escala é enganadora, porque mostra visualmente a existência de grandes variações nos dados que, de facto, não existem (Figura 17 A e B).

Olhando para a Figura 17 B, um leitor menos atento poderia dizer que em 1991 existiam cerca de um terço das pessoas de 2001, o que é falso.

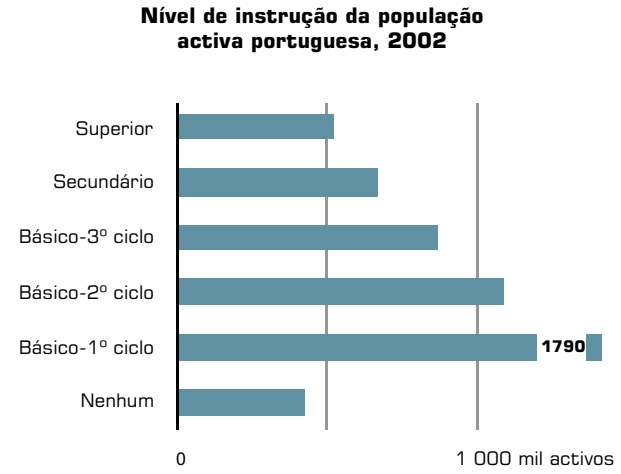
População residente em Portugal, 1991-2001

Figura 17 – Gráfico sem quebra de escala e erradamente com quebra de escala



No entanto, quando uma das barras assume um valor anormal e ocupa muito espaço na imagem, é admissível truncá-la. Tal terá que ser feito de forma clara e compreensível para o leitor, apresentando, por exemplo, o valor respectivo e também uma simbologia que permita compreender que a barra foi interrompida (Figura 18).

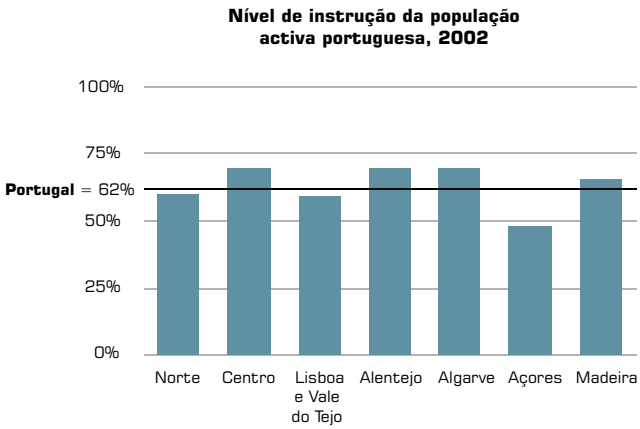
Figura 18 – Gráfico com barra truncada



Pode ser indicado, em certos casos, fazer variar a escala entre 0 e 100 % (Figura 19) para que o leitor possa perceber quanto é que falta em cada barra para atingir os 100%.

Sempre que for possível, é aconselhável comparar as categorias com o total - neste caso Portugal – enriquecendo, desta forma, a leitura do gráfico (Figura 19).

Figura 19 - Gráfico com escala entre 0 e 100%

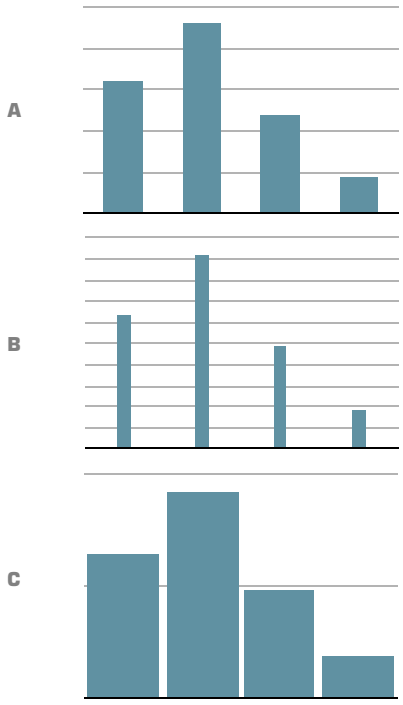


Equilíbrio visual: espaços entre as barras e linhas auxiliares

Os espaços entre as barras devem estar construídos de forma a que não se dificulte a comparação (Figura 20 - B) nem se assemelhe a um histograma (C), sugerindo uma continuidade quando, afinal, a variável representada é discreta. É aconselhado um espaço entre as barras aproximadamente igual ao tamanho das mesmas (A).

As linhas auxiliares existem para ajudar o sistema visual a fazer comparações e ler valores aproximados. Um gráfico com demasiadas linhas auxiliares (B) dá mais peso visual do que deve a estes elementos secundários, sem que daí advenham vantagens significativas ao nível da leitura de valores aproximados. Por outro lado, um gráfico com poucas linhas auxiliares não traz grande valor acrescentado à leitura (C) (WALLGREN, 1996).

Figura 20 – Espaçamento de barras e linhas auxiliares



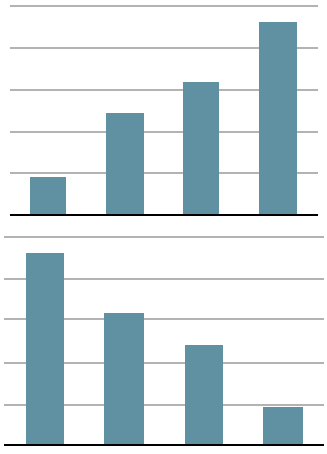
Ordenação

Na representação da informação, por vezes, é importante organizar as categorias por ordem crescente ou decrescente (Figura 21) para melhor compreender certos fenómenos implícitos.

É igualmente comum ordenar alfabeticamente (ou geograficamente) as designações das categorias, nomeadamente nos casos em que se representam países ou outro tipo de unidades administrativas, mas tal nem sempre é a melhor opção.

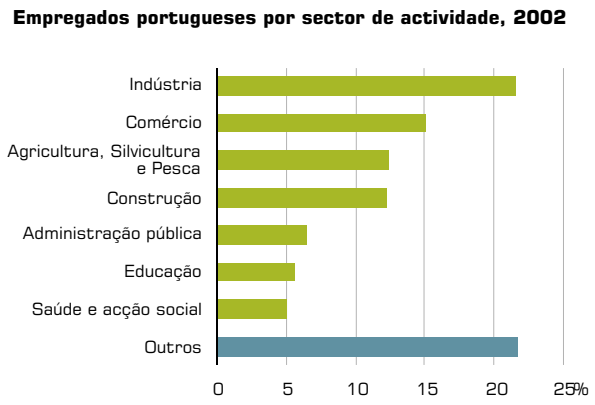
Se o mesmo conjunto de categorias é apresentado em mais do que um gráfico, então a posição relativa de cada categoria deve manter-se, ou seja, as categorias devem aparecer na mesma ordem em todos os gráficos. Da mesma forma, o tamanho e a escala dos gráficos deve ser o mesmo, se o objectivo for a comparação entre eles.

Figura 21 – Gráfico de barras por ordem crescente ou decrescente



Quando as categorias não são todas discriminadas, existindo, por exemplo, uma que reúne as restantes categorias sob a designação de 'Outros', é aconselhável não a incluir na ordenação e reservar-lhe o último lugar (WALLGREN, 1996; SCHMID, 1992) (Figura 22). Caso se utilizem cores para diferenciar as categorias, a categoria 'Outros', por ser a menos importante, deve ter uma cor que não se destaque (ex: cinzento).

Figura 22 – Ordenação das categorias



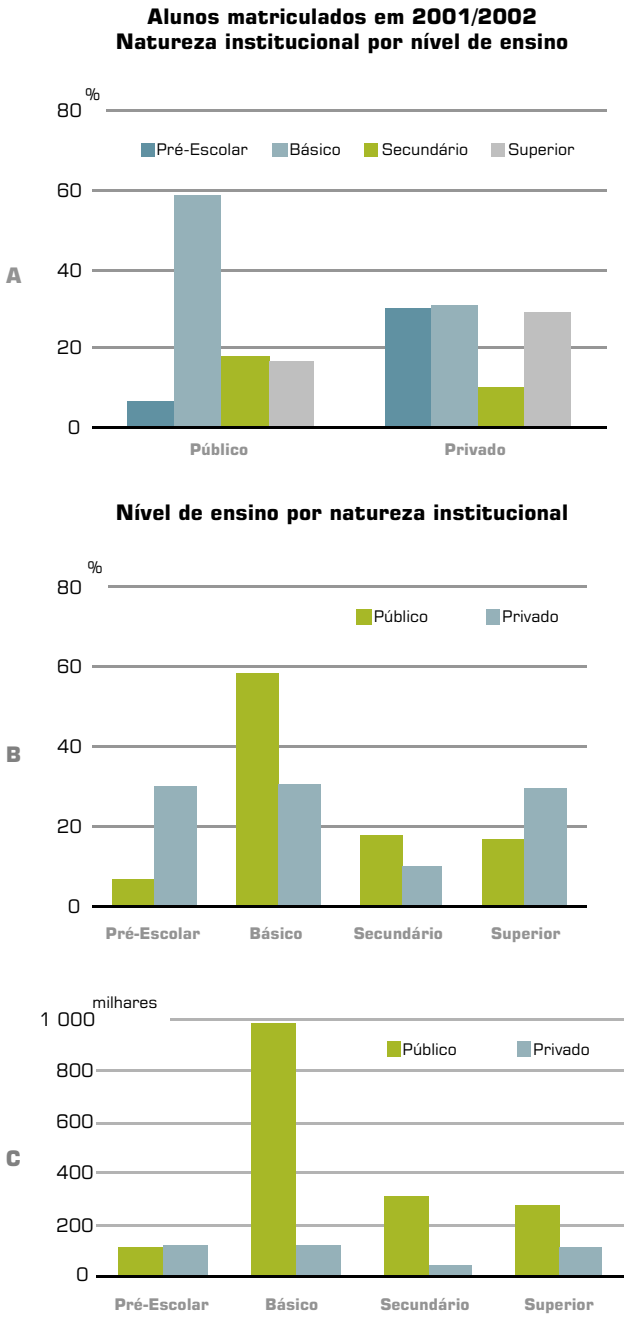
Gráficos de barras agrupadas

Os gráficos de barras agrupadas são utilizados para descrever, simultaneamente, duas ou mais categorias, para uma dada variável discreta, ou quando se pretende realçar o valor das categorias em detrimento do valor total das variáveis (WALLGREN, 1996).

As diferentes categorias são representadas por barras sendo a distinção entre elas feita recorrendo às variáveis visuais (cor ou valor). Os grupos de entidades devem estar separados por um espaço em branco, mas não deve existir qualquer espaço entre as categorias de cada grupo.

Dado que a comparação entre barras adjacentes ao nível da estimação de valores é mais eficaz, em termos perceptivos, do que entre barras mais afastadas, o agrupamento escolhido deve estar de acordo com as categorias a que se pretende dar ênfase. Assim, em termos visuais são comparadas primeiro as categorias que constam da legenda e só depois são relacionadas as desagregações da variável (Figura 23 - A e B).

Figura 23 – Gráfico de barras agrupadas em quatro e duas categorias, em valores relativos e absolutos



As barras podem apresentar indiferentemente valores relativos ou absolutos, consoante o tipo de análise, sendo por vezes de extremo interesse projectar ambos quando existem diferenças significativas (Figura 23 – B e C).

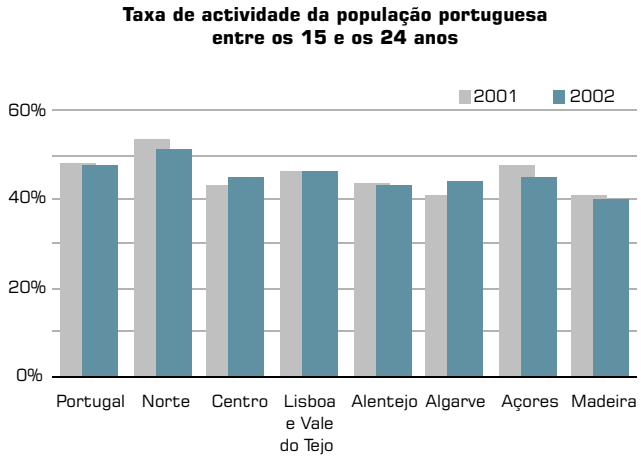
Este processo é tanto menos legível quanto maior for o número de categorias representadas, sendo aconselhável não incluir mais do que três/quatro categorias, por variável, num gráfico.

Nos casos em que existem diversos grupos compostos por variadas categorias, é preferível construírem-se diferentes gráficos em vez de acumular a informação num só.

Sobreposição em gráficos de barras agrupadas

Nos gráficos agrupados, as barras, que representam as categorias de cada grupo, podem tocar-se ou mesmo sobrepor-se (SCHMID, 1992). A sobreposição permite ordenar as categorias para além de poupar espaço e incluir mais informação, Note-se que as barras que se localizam num plano mais distante (e com uma cor menos forte) são percebidas como sendo menos importantes (Figura 24).

Figura 24 - Gráfico de barras agrupadas parcialmente sobrepostas

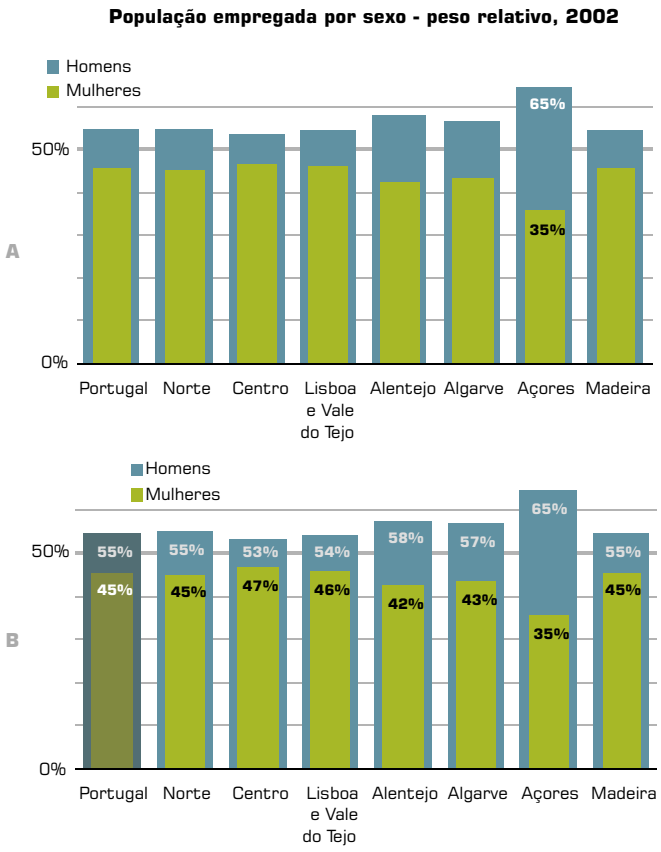


É, igualmente, proposta a sobreposição de barras nos casos em que os valores são sistematicamente menores numa categoria do que na outra (Figura 25). Realçar valores ou acontecimentos é também uma forma de análise dos dados. Por vezes, é importante dar ênfase visual a um determinado valor ou a uma determinada categoria.

Neste caso, e a título de exemplo, tornou-se mais grossa a linha auxiliar referente a 50% dos empregados - a única que tem um valor numérico associado - e deixou-se a leitura dos restantes valores para as linhas auxiliares não numeradas (Figura 25).

Para realçar a categoria referente a Portugal, pode-se utilizar uma moldura (A) ou uma cor mais escura (B). Apenas se apresentam os valores das categorias que se considerem dignas de análise (A - os Açores apresentam a maior diferença entre sexos) em vez de carregar demasiado o gráfico (B).

Figura 25 – Gráficos de barras agrupadas totalmente sobrepostas



Gráficos de barras empilhadas

Recorre-se aos gráficos de barras empilhadas (Figura 26) em situações análogas aos gráficos de barras agrupadas, ou seja, quando o conjunto de dados contém duas ou mais categorias.

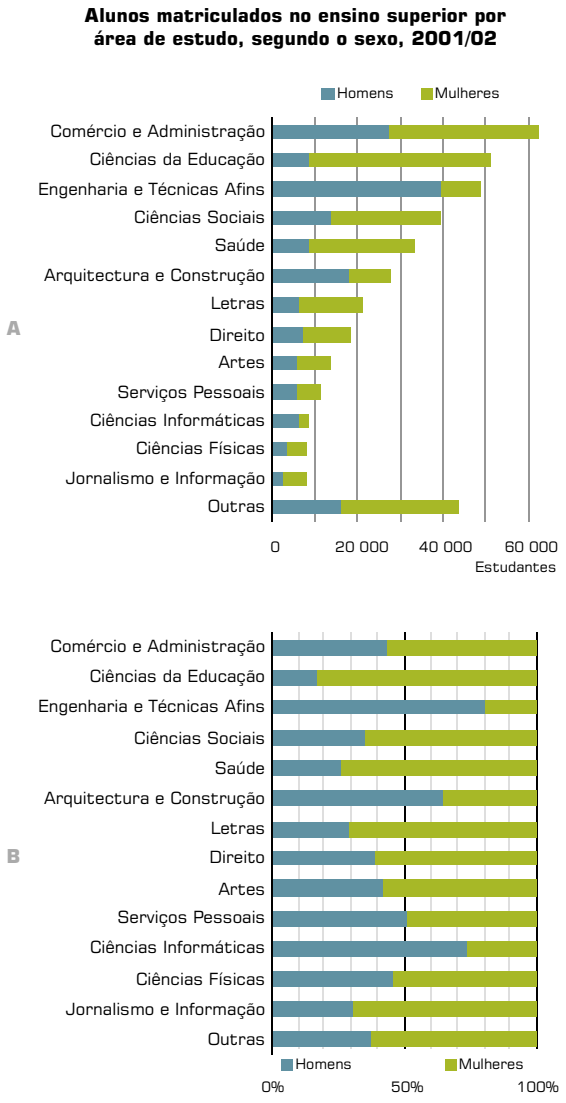
Neste tipo de gráficos, cada barra subdivide-se em pelo menos duas categorias, com distintas cores ou padrões, permitindo mostrar a relação entre cada categoria (Homens/Mulheres) e o respectivo subtotal (ex: Comércio e Administração). As categorias surgem assim posicionadas umas sobre as outras, se for um gráfico de barras vertical (ou lado a lado, se o gráfico for horizontal), sendo que a altura (ou a largura) de cada componente corresponde ao valor absoluto ou relativo da categoria.

O gráfico em valor absoluto (A) adequa-se aos casos em que se pretende evidenciar mais o valor total das variáveis do que das respectivas categorias (WALLGREN, 1996), dado que o todo é apreendido com maior precisão do que as partes. Tal precisão advém de, para o total, ser comparada a posição relativa numa mesma escala, enquanto que na estimação dos valores das categorias são confrontados e ordenados os tamanhos respectivos.

Se o maior objectivo destes gráficos é indicar graficamente a soma total, mais do que estimar visualmente as respectivas categorias, valerá então a pena questionar porque não se opta por representar apenas o total ou então substituir esta por outra forma de representação.

No gráfico em valor relativo (B) apenas se pode estimar o valor das categorias observando o tamanho das barras que lhes correspondem. Alunos matriculados no ensino superior por área de estudo segundo o sexo, 2001/02

Figura 26 – Gráfico de barras empilhadas horizontalmente em valores absolutos e relativos



Com duas categorias torna-se mais fácil estimar os valores, dado que a base e o topo da escala servem de ponto de referência, mas com mais de duas categorias a leitura é consideravelmente mais difícil.

Desvantagem dos gráficos de barras empilhadas

De facto, as primeiras componentes são facilmente comparáveis por começarem junto ao eixo, mas nas seguintes apenas se consegue inferir aproximadamente os valores, sendo tanto mais difícil quanto maior for a variação da primeira categoria (Figura 27).

Por conseguinte, as flutuações e o peso excessivo da primeira categoria podem comprometer a leitura das restantes variáveis representadas. Se a comparação entre categorias com base no tamanho pode envolver erros, não negligenciáveis, entre os verdadeiros valores e os estimados visualmente, a ordenação entre as categorias de uma mesma barra pode até ser incorrectamente realizada, pondo em causa a validade desta forma de apresentação de informação (CLEVELAND, MCGILL, 1984a).

É por esta razão que os gráficos de barras empilhadas devem ser limitados a um conjunto restrito de variáveis e categorias. Em certos casos é preferível substituir por um gráfico de barras agrupadas, porque melhora a estimação dos valores individuais, apesar de não facilitar a comparação entre categorias.

Histograma

Um histograma mostra a distribuição de valores de uma variável contínua através de um gráfico de barras unidas. Contudo, se uma variável discreta apresentar muitos valores distintos, também pode ser representada por um histograma. Normalmente, os histogramas são representados por barras com bases iguais em que a altura (ou o comprimento) varia em função da frequência relativa ou absoluta. De facto, no caso em que os intervalos têm a mesma amplitude, a área depende apenas da altura. Mas, quando as classes têm diferentes dimensões, a área de cada barra já não é proporcional à altura, devendo ser calculada a altura por forma que a área de cada rectângulo seja proporcional à frequência relativa de cada classe. Enquanto no primeiro caso o eixo dos valores transmite a informação alusiva à frequência relativa de cada classe, no segundo caso este eixo não tem qualquer significado sendo o leitor obrigado a comparar áreas para interpretar a informação, o que se revela bastante mais difícil.

Esta forma gráfica permite indicar valores extremos e enviesamentos, demonstrando visualmente se a variável segue uma distribuição normal.

A representação das percentagens permite também comparar conjuntos de dados de diferentes dimensões.

Figura 27 – Gráfico de barras empilhadas verticalmente e gráfico de barras agrupadas

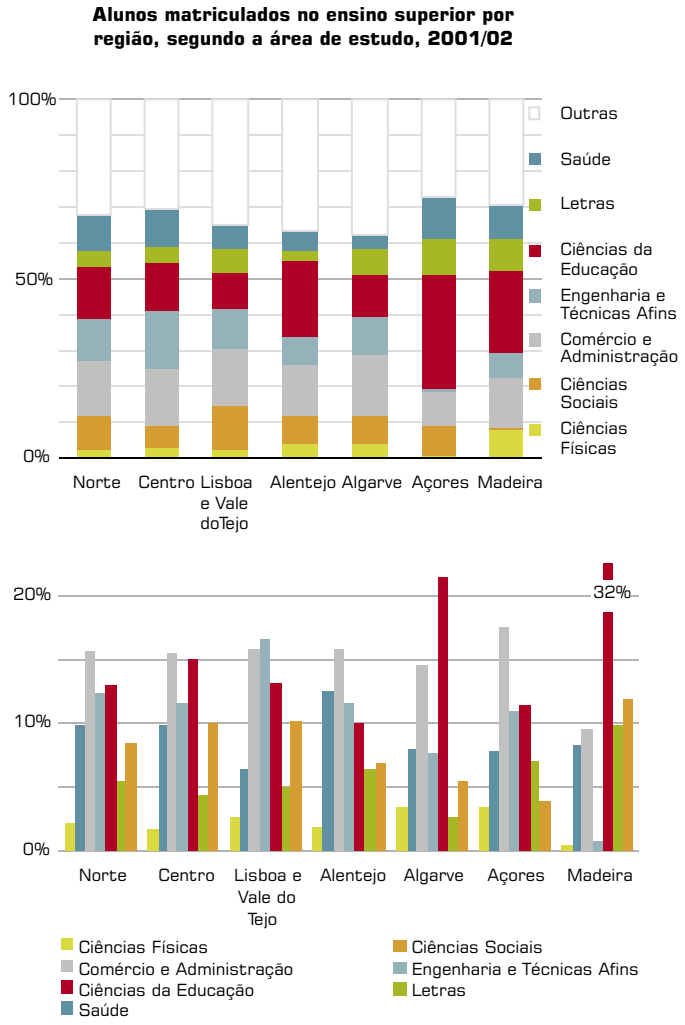
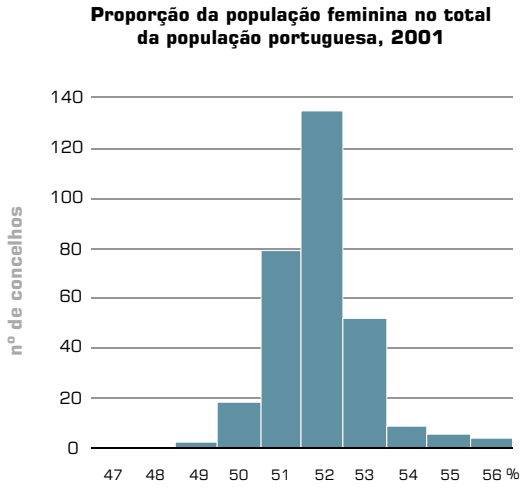


Figura 28 – Histograma



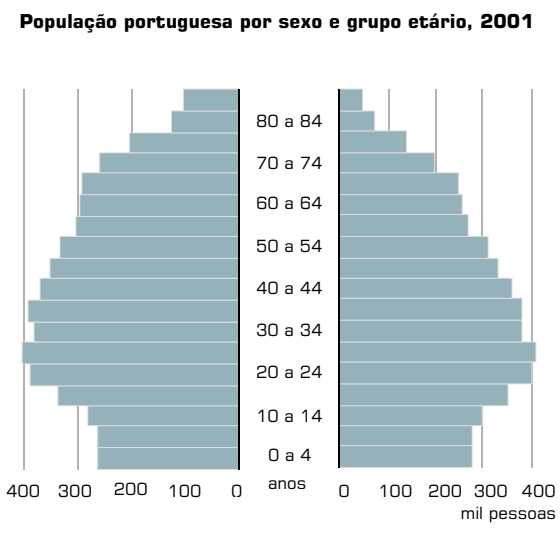
Pirâmide Etária

A pirâmide etária é também um histograma e é muito utilizada em análises demográficas por permitir visualizar numa única imagem a distribuição da população por idades e simultaneamente compará-la entre os dois sexos. A sua representação é feita em dois eixos horizontais (um para os efectivos masculinos e outro para os femininos) podendo esta ser em valores absolutos ou relativos.

As idades encontram-se representadas no eixo vertical, servindo de legenda a ambos os gráficos e são normalmente apresentadas em grupos etários de cinco anos, mas também podem ser representadas ano a ano.

A representação em valores absolutos fornece a dimensão dos dados mas impede qualquer tipo de comparação no espaço ou no tempo, que apenas é possível se os dados forem apresentados em termos relativos (NAZARETH, 1996; INE, DRLVT, 2001). No entanto, esta forma de apresentação pode ser aplicada a outro tipo de informação demográfica (como, por exemplo, o nível de instrução) ou até para representar variáveis contínuas com uma legenda comum (WALLGREN, 1996).

Figura 29 – Pirâmide etária



Séries temporais em Gráficos de barras

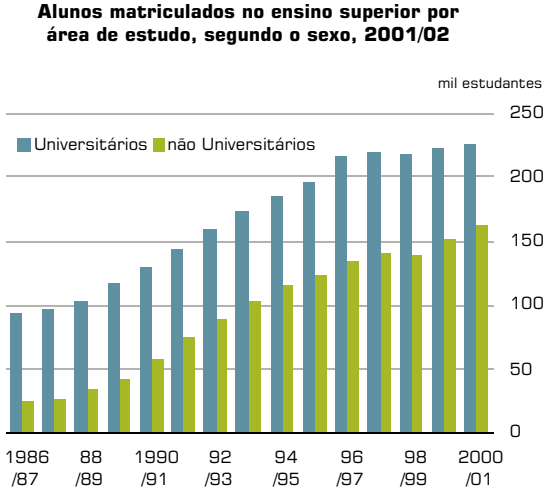
Um gráfico de barras verticais pode ter datas no eixo das categorias, possibilitando a representação de evoluções ao longo do tempo.

Os gráficos de barras podem substituir os gráficos de séries temporais nos casos em que a série de dados é muito curta. São igualmente indicados quando se pretendem fazer comparações verticais de determinadas variáveis num período específico, ou seja, quando se dá importância ao valor da variável em cada período e se pretende sobretudo relacionar quantidades individuais.

Para uma única série de dados, ambas as possibilidades (barras e linhas) são adequadas para mostrar tendências, mas para mais de uma série de dados, os gráficos de linhas são claramente preferíveis (JACOBS, 1997). Por isso, não é aconselhável utilizar os gráficos de barras para representar várias séries de dados. Quando as variáveis assumem valores sistematicamente inferiores ainda é possível acompanhar a sua evolução (Figura 30) mas quando as variáveis se entrecruzam o gráfico torna-se ilegível (Figura 31).

Nos casos em que a informação contida no gráfico é tal que impede uma correcta visualização, deve ser considerada a sua substituição por uma tabela de dados, ou então, a partição em vários gráficos.

Figura 30 – Gráfico de barras com duas séries temporais



1.3. Gráficos de linhas

O gráfico de linhas é indicado para mostrar tendências e evoluções de uma variável contínua por outra variável contínua.

O mais comum é aquele que representa séries temporais (ou cronológicas), em que uma determinada variável contínua é analisada ao longo do tempo. O eixo do y mede a(s) variável(eis) em estudo, enquanto o eixo do x apresenta as unidades temporais dispostas cronologicamente em intervalos iguais de tempo, começando à esquerda com a data mais antiga (Figura 32).

Num gráfico de linhas, ao contrário dos gráficos de barras, as séries podem ser longas.
O objectivo nestes gráficos é comparar os declives das curvas por forma a responder a perguntas do tipo: Em que períodos a variação foi significativa? Quando foram os pontos de inflexão? (WALLGREN, 1996).

Visualmente, para um determinado conjunto de dados, a união dos pontos (pares de coordenadas: x,y), é feita através de uma linha que sugere a continuidade.

Não devem ser incluídas mais do que três linhas por gráfico, caso contrário tornam o gráfico de difícil leitura (SCHMID, 1992; TUFTE, 1983). Quando muitas linhas se sobrepõem (Figura 33), é preferível substituir o gráfico de linhas por vários gráficos.

Figura 31 – Gráfico de séries temporais: barras e linhas

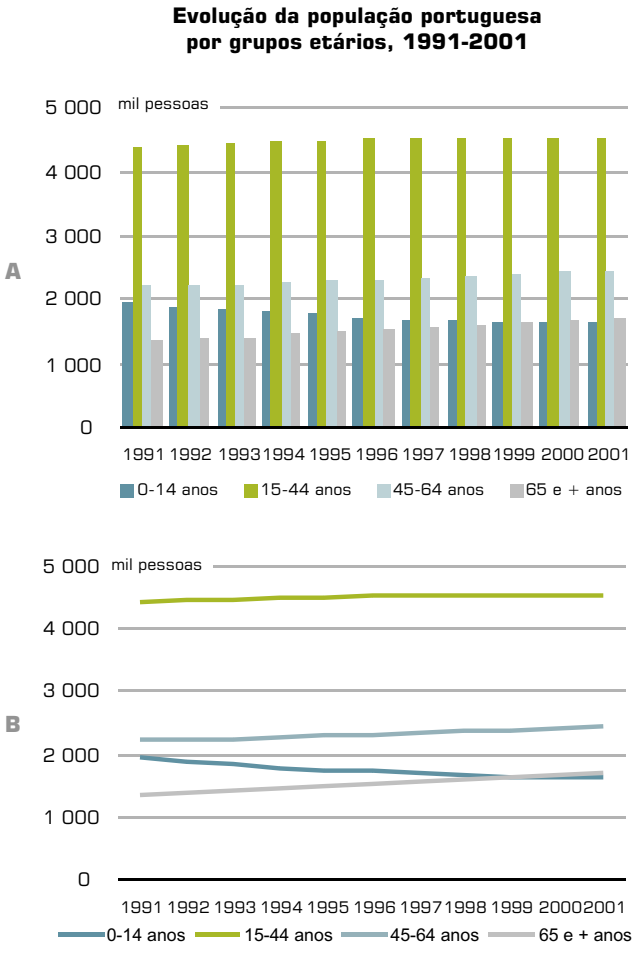
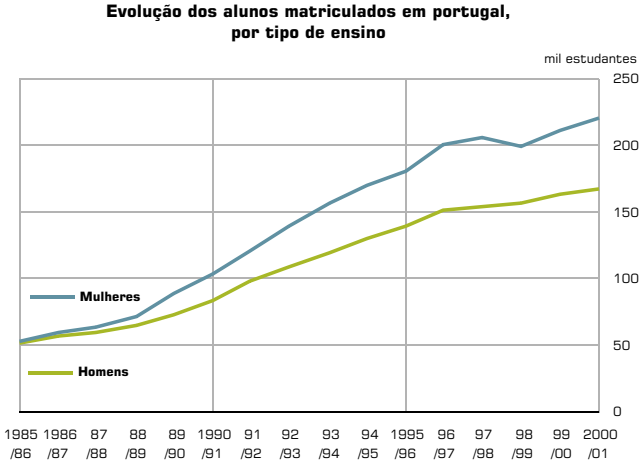
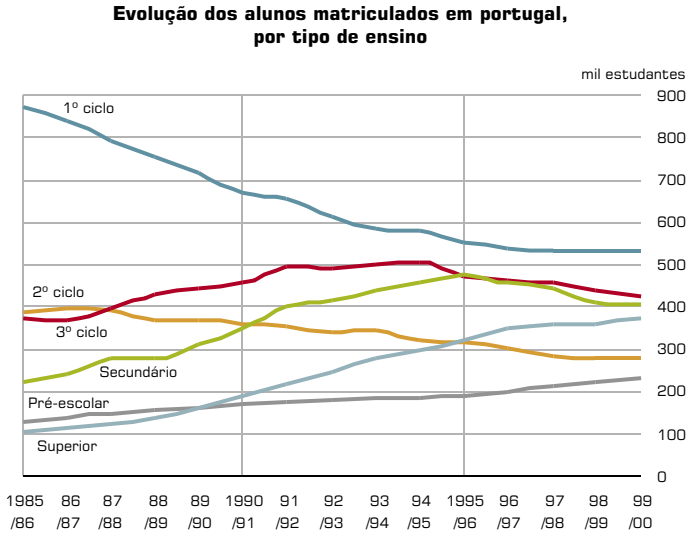


Figura 32 – Gráfico de séries temporais



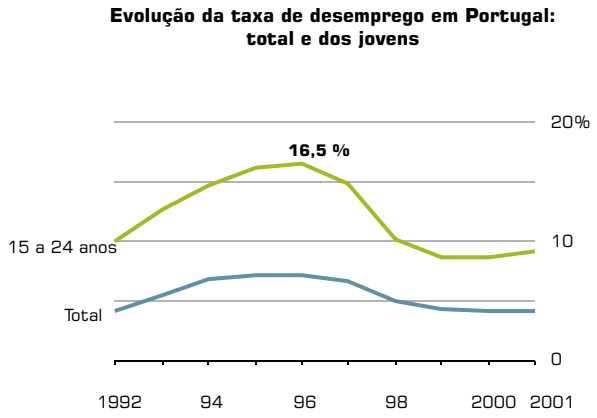
Deve ser usado um estilo de linha diferente para cada gráfico, recorrendo à cor, forma, tamanho ou valor. Mesmo se as linhas se diferenciarem pela cor, pode ser necessário distinguir as linhas de outra forma, para facilitar a interpretação nos casos de impressão a preto e branco ou de reprodução através de fotocópias. Porém, tal opção pode dar uma ordem visual às linhas, não coincidente com a realidade, dado que, por exemplo, uma linha a tracejado é visualmente menos importante que uma linha a cheio.

Figura 33 – Gráfico com demasiadas linhas



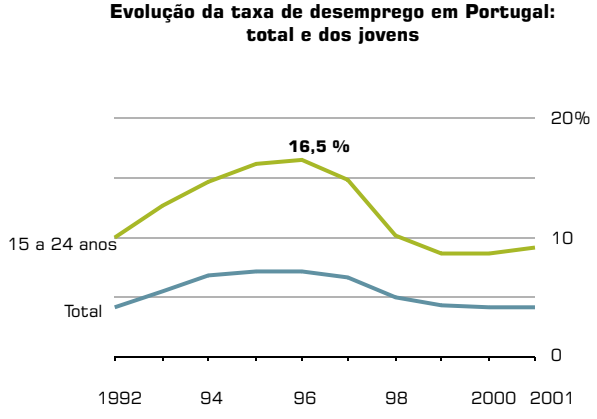
A variável medida no eixo das categorias nos gráficos de linhas não pode ser qualitativa (Figura 34). De facto, neste caso, a evolução da série não têm qualquer significado, ou seja, entre o Algarve e a Madeira não se pode afirmar que existe uma quebra na série de dados, mas apenas que os Açores têm um valor inferior. Também não é possível estimar os valores intermédios entre as categorias da variável, neste caso, não se pode dizer que existem x% de desempregados no Oceano Atlântico (gráfico correcto: Figura 19).

Figura 34 – Gráfico de linhas incorrecto



Os períodos devem estar igualmente espaçados se forem consecutivos e proporcionalmente espaçados se forem descontínuos, ou seja, quando ocorrem intervalos irregulares de tempo é indicado um ajustamento no espaçamento das colunas. Por exemplo, o espaço entre dados de 1998 e 2000 deve ser o dobro do que entre 2000 e 2001 (Figura 35).

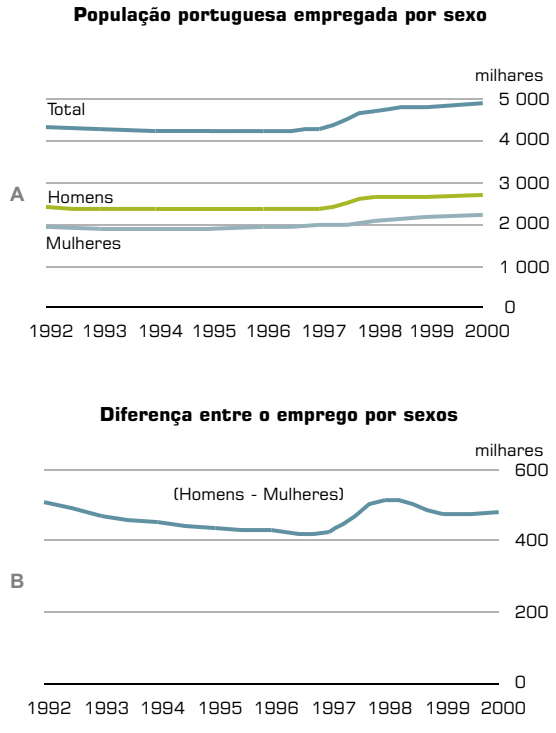
Figura 35 – Espaço entre os valores no eixo das categorias



Quando se pretendem comparar duas curvas que apresentam comportamentos muito semelhantes (Figura 36 - A), é preferível projectar a diferença entre elas, neste caso entre homens e mulheres (Figura 36 - B) em vez das curvas propriamente ditas.

Uma modificação repentina nos dados pode ser encoberta se o gráfico começar depois dessa modificação, mostrando uma estabilidade incorrecta (WAINER, 1984). Pelo contrário, uma alteração pode tornar-se brusca se o gráfico apenas representar aquele período e não o contextualizar, como, por exemplo, em séries com uma sazonalidade forte.

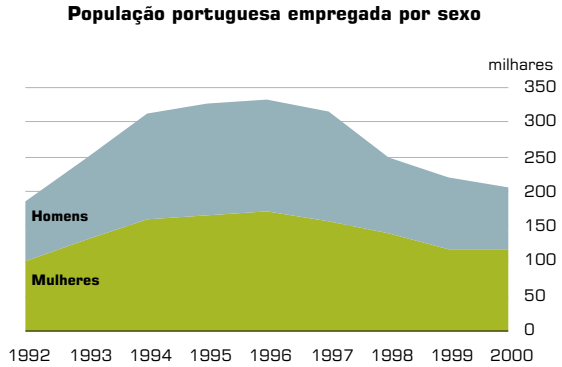
Figura 36 – Comparação de séries paralelas



Gráficos de área

Recorre-se aos gráficos de área quando se pretende visualizar simultaneamente a evolução do total e das respectivas componentes. Tal como nos gráficos de barras empilhados, existem poucas vantagens nesta forma de apresentação dado não ser possível responder de forma imediata a perguntas sobre o crescimento ou decréscimo ao longo do tempo, sobretudo quando a primeira das componentes apresenta oscilações significativas.

Figura 37 – Gráfico de área empilhada



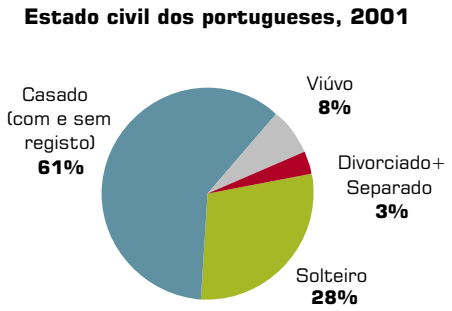
Os gráficos de área são utilizados como alternativa aos gráficos de linhas. No entanto, trazem dificuldades acrescidas quando as áreas se intersectam porque deixa de ser possível seguir a evolução das componentes.

1.4. Gráficos circulares

O gráfico circular tornou-se muito comum em publicações direccionadas para um público alargado, mas tem vindo a ser amplamente contestada pela sua falta de capacidade informativa (WAINER, 1990; TUFTE, 1983; BERTIN, 1977, etc.).

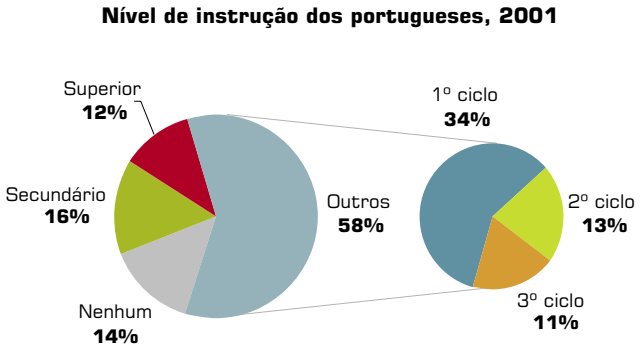
Os gráficos circulares exibem as partes do todo como se fatias de um bolo se tratassem; a isso se deve a denominação inglesa 'pie chart' traduzida em português para queijo ou tarte. Para um determinado período temporal, a variável em análise é projectada num círculo correspondendo a cada componente um ângulo, por forma a que as componentes no seu conjunto perfaçam os 360º (Figura 38).

Figura 38 – Gráfico circular



A sua utilização é desaconselhada quando se pretende comparar mais do que um período temporal, para variáveis que contenham mais de cinco componentes ou quando as componentes têm aproximadamente o mesmo peso, sendo neste caso, preferível substituir o gráfico circular por um gráfico de barras (SCHMID, 1992). Muitas fatias ou fatias demasiadamente estreitas são dificilmente interpretáveis, sendo por isso necessário complementar o gráfico com os valores respectivos (Figura 38) ou associar um subconjunto de valores a outro gráfico circular de tamanho proporcional à quantidade que representa (Figura 39).

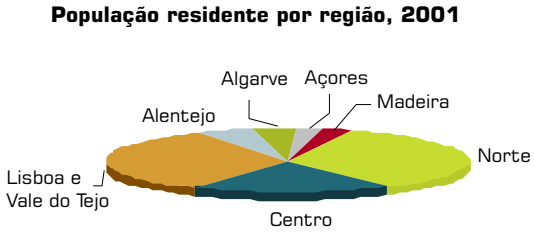
Figura 39 – Gráfico circular subdividido



Assim, a utilização dos gráficos circulares é apenas referida positivamente nos casos em que uma ou duas componentes dominam o total para dar uma ideia genérica dos dados, mas poder-se-á questionar se não será melhor recorrer a uma tabela.

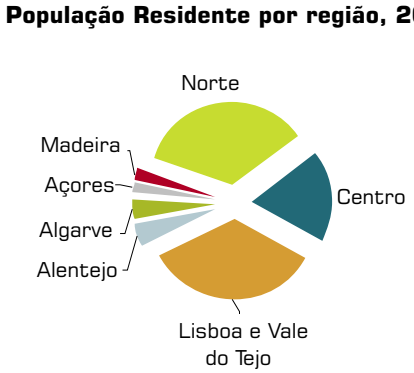
É comum encontrar gráficos circulares distorcidos, ou seja, assumindo formas não circulares, para poupar espaço ou então por razões que a razão desconhece. Tornar uma figura circular numa elipse é altamente enganador, particularmente para os segmentos mais estreitos e deve ser evitado por desvirtuar completamente o gráfico original.

Figura 40 – Gráfico circular distorcido



Outra prática corrente é a separação das fatias movendo-as radialmente para fora, provocando afastamentos desiguais entre fatias díspares. Como para manter as separações iguais é necessário posicionar as fatias de forma não circular, pelo que nenhuma das opções é formalmente correcta (BOUNFORD, 2000).

Figura 41 – Gráfico circular com fatias separadas



1.5. Pictogramas

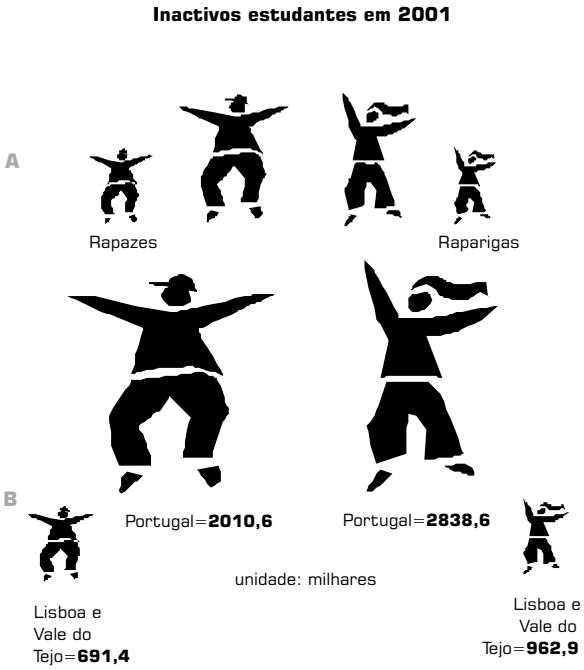
Os pictogramas são gráficos comuns, mas com características decorativas. A sua utilização é indicada numa apresentação superficial em que o contacto com a imagem é breve, nomeadamente, em jornais ou revistas de âmbito alargado ou quando o público-alvo tem um nível educacional médio ou baixo.

Os pictogramas mais usuais são os baseados no critério do tamanho: em que a variação em área do tamanho das formas utilizadas é proporcional à variação da variável representada (Figura 42 - A).

No entanto, é vulgar encontrar imagens, particularmente nos média em que foi aumentada a altura e a largura simultaneamente, e não a área, tornando o desenho desproporcionado e transmitindo uma ideia completamente errada.

Senão veja-se: na Figura 42 – B, Portugal tem 3 vezes mais estudantes do que Lisboa e Vale do Tejo, para ambos os sexos. Assim, a área do boneco referente a Portugal deve ser 3 vezes maior. Por isso, este tipo de apresentação é considerado como um dos mais enganadores (SCHMID, 1992; TUFTE, 1983).

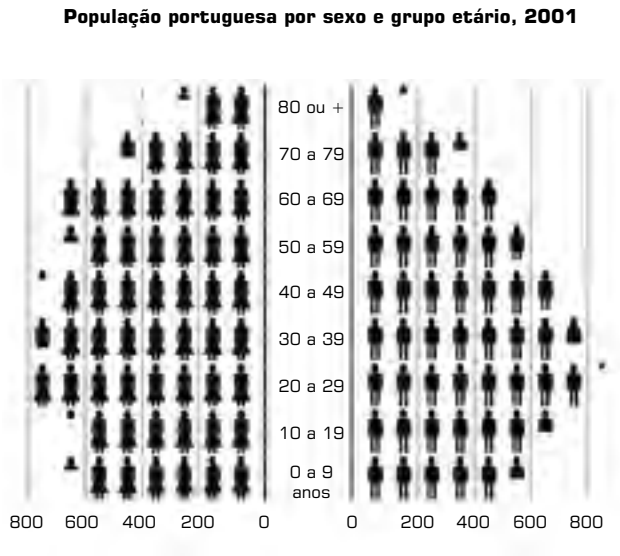
Figura 42 – Pictograma baseado no critério do tamanho



Os pictogramas constituídos por formas unitárias são também bastante utilizados. Neste caso, a cada elemento é atribuído um valor existindo, assim, tantos elementos quanto a dimensão da variável.

A pirâmide etária cujas barras são formadas por elementos que representam pessoas, é um dos mais difundidos. Um dos problemas surge com o tratamento dado às casas decimais. Modley (1952, in SCHMID, 1992) diz que as fracções de símbolos devem ser minimizadas, devendo-se, preferencialmente, arredondar os valores. De facto, é comum encontrar nas pirâmides etárias acima referidas, barras em que o último símbolo é fraccionado, ou seja, que terminam em braços, pernas ou cabeças (Figura 43).

Figura 43 – Pictograma: pirâmide etária



1.6. Ver também ...

Neste dossiê são referidas, sucintamente, algumas das questões mais importantes associadas à representação gráfica, nomeadamente, as que se relacionam com a construção dos gráficos mais conhecidos e utilizados.

A informação utilizada para os gráficos aqui incluídos é bastante actual e pode ser encontrada em www.ine.pt. Todas as figuras, à excepção da última, foram construídas através do software Excel.

Este texto baseia-se na minha dissertação de mestrado intitulada: Representação gráfica e cartográfica da informação estatística e defendida, em Junho de 2003, no ISEGI/ Universidade Nova de Lisboa.

Sobre os gráficos e a estatística existem diversos livros, artigos, web sites, dos quais se destacam os seguintes:

Publicações, livros e artigos em revistas

• BENIGER, James R.; ROBYN, Dorothy L. (1978), "Quantitative graphics in statistics: A brief history", The American Statistician, 32 (1), p. 1-11.

• BERTIN, Jacques (1973) 2.ª ed. (1ª ed. 1967) - Sémiologie graphique. Paris: Gauthier-Villars.

• CHAMBERS, John C.; CLEVELAND, William. S.; KLEINER, Beat; TUKEY, Paul A. (1998) 2ª ed. (1ª ed. 1983) - Graphical methods for data analysis. USA: Chapman & Hall.

• CLEVELAND, William S.; MCGILL, Robert (1987a), "Graphical perception: The visual decoding of quantitative information on graphical displays of data", Journal of the Royal Statistical Society, 150, p. 192-229.

• CLEVELAND, William S.; MCGILL, Robert (1984a), "Graphical perception: Theory, Experimentation, and application to the development of graphical methods", Journal of the American Statistical Association, 82, p. 419-423.

• GRAPHICS GUIDELINES: The theory and practice of presenting statistical data graphically, together with proposals for education of statisticians in appropriate use of graphics for presentation (1994). COMISSION OF THE EUROPEAN COMMUNITIES - EUROSTAT. Kent: White Waghorn Limited.

• HUFF, Darrell (1991) 3ª ed. (1ª ed. 1954) - How to lie with statistics. England: Pinguin Books.

• INE, DRLVT (2001), "As pirâmides de idades", Revista de Estudos Regionais nº 2 (Conceitos e metodologias), Instituto Nacional de Estatística, p. 75-78.

• JACOBS, Bernhard (1997), "Experimental analysis of the graphical presentation of data in line graphs and bar charts in superposition and juxtaposition", http://www.uni-saarland.de/philfak/MZ/graph/gesamtue.html.

• NAZARETH, J. Manuel (1996) - Introdução à demografia - Teoria e prática. Lisboa: Editorial Presença.

• SCHMID, Calvin F. (1992) 2ª ed.; (1983, 1ª ed.) - Statistical graphics - Design principles and practices. Krieger.

• SILVA, Ana A. (2003) - Representação gráfica e cartográfica da informação estatística. Dissertação de mestrado defendida no Instituto Superior de Estatística e Gestão de Informação da Universidade Nova de Lisboa.

• TUFTE, Edward R. (1983) - The visual display of quantitative information. Cheshire-Connecticut: Graphic Press.

• TUKEY John W. (1977) - Exploratory data analysis. USA: Addison-Wesley.

• WAINER, Howard (1990), "Graphical Visions from William PLAYFAIR to John TUKEY", Statistical Science, 5 (3), p. 340-346.

• WAINER, Howard (1984), "How to display data badly", The American Statistician, 38 (2), p. 137-147.

• WALLGREN, Anders; WALLGREN, Britt; PERSSON, Rolf; JORNER, Ulf; HAALAND, Jan-Aage (1996) (English translation from Swedish "Statistikens Bilder - Att Skapa Diagram" Statistics Sweden 1995) - Graphing statistics & data: Creating better charts. California: SAGE Publications.

Páginas na Internet @

American statistical association - Section on Statistical Graphics:

• http://www.amstat-online.org/sections/graphics/

Journal of computational and graphical statistics:

• http://www.amstat.org/publications/jcgs/

Outros:

• http://www.edwardtufte.com/tufte/
(um dos melhores autores sobre esta temática – ver livros)

• http://www.mhhe.com/business/opsci/bstat/vistat.mhtml
(visual statistics)

• http://www.nas.nasa.gov/Groups/VisTech/visWeblets.html
(links sobre visualização científica)

• http://www.bell-labs.com/topic/societies/asagraphics/resources.html
(software, livros, revistas, etc.)

Estatística com R

Pedro Campos
Rita Sousa

com a colaboração de Emília Oliveira

Estatística com R

Uma iniciação para o ENSINO BÁSICO e SECUNDÁRIO

Pedro Campos
 # Rita Sousa

Sumário:

1. Introdução

O R é uma linguagem (e ambiente de computação estatística e construção de gráficos) aberta e gratuita cujo número de utilizadores tem vindo a aumentar consideravelmente. O dossiê começa por apresentar o R, referindo os seus aspectos fundamentais e descrevendo, de seguida, os principais comandos. No capítulo 4 apresenta-se o *R-Commander*, uma ferramenta importante que permite tornar a interface gráfica do R mais apelativa. No final há um conjunto de exercícios resolvidos utilizando o código R.

2. A utilização de software no Ensino da Estatística

O software estatístico que foi sendo introduzido nas últimas décadas trouxe novas formas de explorar a Estatística, proporcionando maior rapidez na resolução de problemas e permitindo a comparação expedita de soluções. Além disso, abriu caminho a um conjunto de utilizadores nos meios académico, empresarial e administrativo que desta forma puderam passar a utilizar a Estatística como uma ferramenta eficaz na resposta aos seus problemas.

No ensino em geral a utilização do computador permitiu introduzir diversas melhorias, pois no contexto escolar usual, "os alunos têm grande dificuldade em aprender novos assuntos cujo significado não vislumbram e que não lhes despertam qualquer interesse" (ver João Pedro da Ponte na Introdução de "A Família em Rede", de Seymour Papert, 1997). O computador e, em particular, o software estatístico permitiram incentivar a participação voluntária do aprendiz no processo educativo, fazendo com que o aluno passe a explorar os dados e a ser cada vez mais o centro desse desafio do ensino/aprendizagem da estatística.

No entanto, apesar de serem reconhecidas as vantagens da utilização do software estatístico, nomeadamente no que respeita ao ensino da estatística, a sua utilização deve ser sempre suportada por um adequado conhecimento das técnicas estatísticas envolvidas ou orientada por quem detenha esses conhecimentos (ALEA, Dossiê Didáctico X – Software Estatístico, Luís Cunha e Helder Alves).

No Dossier Didáctico X (Software Estatístico - Uma introdução a alguns aplicativos, numa abordagem inicial dos dados, Helder Alves, Luís Cunha) foram apresentadas algumas aplicações informáticas (Minitab, SAS, SPSS, Statistica) para a análise estatística de dados, numa abordagem preliminar dos dados, ao nível da estatística descritiva. Neste dossiê, concentramos as atenções no R, um importante e poderoso veículo de análise interactiva de dados que, devido à sua crescente utilização nos meios académico e empresarial, não poderia passar despercebido no contexto do ALEA.

3. O que é o R e para que serve?

O R é uma linguagem e ambiente de computação estatística e construção de gráficos; é considerada uma variante da linguagem S (laboratórios Bell, desenvolvida por John Chambers e seus colegas). Surge pela criação da R Foundation for Statistical Computing, com o objectivo de criar uma ferramenta gratuita e de utilização livre, para análise de dados e construção de gráficos.

O R é compatível com diversas plataformas: UNIX, Windows e MaCOS e permite a ligação a interfaces de diferentes formatos: Excel, Access, SPSS, SAS, SQL Server. Sendo Open Source, permite ao utilizador aceder ou alterar funcionalidades existentes, bem como criar novas funcionalidades para responder aos seus problemas específicos de forma mais eficaz. Tal é possível graças à possibilidade de o R se estender a partir de um crescente conjunto de livrarias (packages) que podem ser acedidas pelo utilizador.

A interacção com o utilizador é baseada numa janela de comandos e exige o recurso a programação, embora existam packages gráficos que permitem a interacção através de menus. Um desses packages é o *R Commander* que será abordado no contexto deste dossiê.

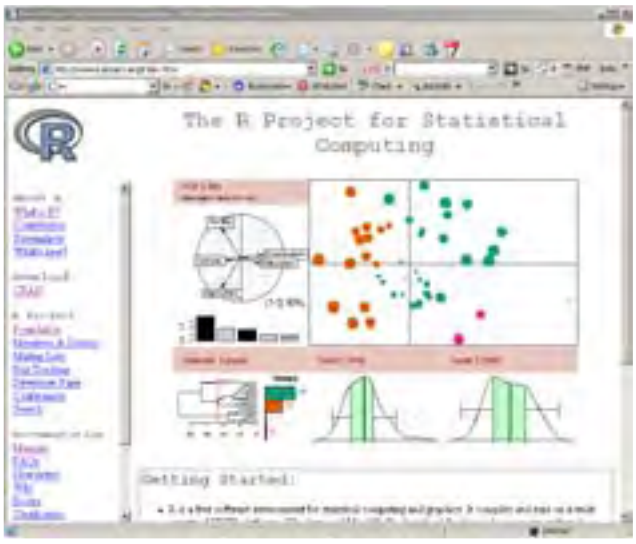
Apesar de existirem muitas facilidades de entreajuda na comunidade de utilizadores do R, esta linguagem não tem suporte técnico assegurado.

4. Primeiros passos

4.1. Instalar o R

A instalação do R é gratuita e pode ser feita directamente a partir da página principal do *R Project for Statistical Computing* em http://www.r-project.org/. A figura seguinte indica o local onde se pode efectuar a importação do R.

Fig. 1 - O download do R é feito a partir da página principal do Projecto R na área CRAN *(Comprehensive R Archive Network)*



Para a importação do R é necessário escolher: um país a partir do qual o ficheiro será transferido, o sistema operativo (MacOS X, Linux, ou Windows), o link base e, finalmente, o programa executável. A última versão à data deste dossiê é: R-2.9.1 -win32.exe .

Após importação deste ficheiro, a instalação é rápida e intuitiva.

4.2. Abrir e Encerrar o R, Ajuda e os Packages

O "prompt"

Ao iniciar o R mostra-se imediatamente a janela de comandos (V. Fig. 2). Esta janela exibe um cursor vermelho em forma de sinal "maior" (>) designado por prompt onde são escritos os comandos. Por exemplo, para se obter o número da versão do R em causa deve-se escrever:

```
> R.version
```

Para sair do R, pode-se utilizar o menu (File/Exit) ou então escrever:

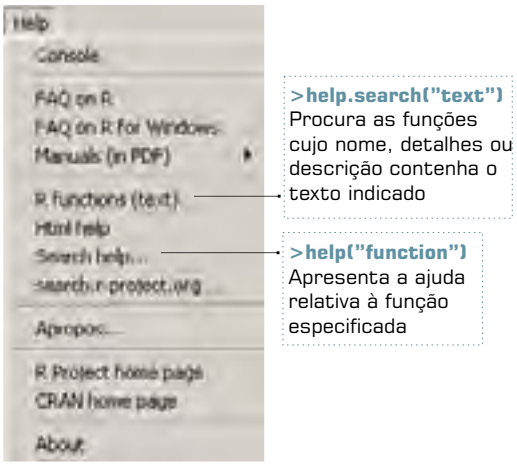
```
> q()
```

Fig. 2 - Janela de comandos do R da versão 2.9.1

Entrar e Sair

Uma das perguntas que surge habitualmente ao abandonar o R é se pretende guardar o espaço de trabalho (workspace). De facto, o R pode guardar no seu workspace o nome e o valor dos objectos criados. Veremos nas secções seguintes como criar esses objectos.

Para qualquer tipo de ajuda (que é muito útil quando se tem uma linguagem como o R) existem muitas opções, sendo a mais intuitiva a que está acessível pelo menu Help da barra de menus. Outra forma muito prática para obter ajuda para qualquer função consiste em digitar help.search("text") em que text representa o que pretendemos pesquisar. Em alternativa, caso se conheça o comando (por exemplo, sum) e haja dúvidas quanto a sua utilização, pode-se digitar help("sum") ou simplesmente ?sum.



>help.search("text")
Procura as funções cujo nome, detalhes ou descrição contenha o texto indicado

>help("function")
Apresenta a ajuda relativa à função especificada

Os Packages

Todos os recursos do R (dados ou funções) estão armazenados em packages. O conteúdo de um determinado package só fica disponível quando este é carregado. O package base (standard) é considerado parte integrante dos recursos do R, sendo carregado automaticamente aquando da instalação do programa. As funções básicas que permitem ao R trabalhar os principais objectos de dados, funções estatísticas e gráficas, já estão disponíveis no package *base*.

Existem funções específicas para extrair informação sobre os packages: por exemplo, para ver os packages que estão instalados no PC deverá escrever o comando library(). Para carregar um determinado package deve usar library ("package").

A instalação dos packages e o seu carregamento (Install package(s) from zip files...) e (load package) devem ser feitos por esta ordem e podem ser executados directamente a partir dos menus do R. Os packages pretendidos podem ser previamente importados em formato zip através do site do R (http://www.cran.r-project.org/) e carregados posteriormente.

>livrary("package")
>require("package")
Mostra todos os packages disponíveis para carregamento

Instalação ou actualização de packages por ligação directa à Internet

Instalação de um package através de um ficheiro ZIP previamente importado dos recursos do R na Internet

4.3. Menus e comandos principais

O R exibe uma barra de ferramentas e um sistema de menus que permite executar algumas operações. Basicamente o menu File permite Gravar e abrir sequências de comandos (scripts), abrir ou gravar espaço de trabalho (workspace), sair do R, etc. Permite ainda, carregar livrarias (packages), que serão descritas mais adiante neste dossiê.

Aceder a ficheiros de programação R ou guardar a informação actual em memória

Uma das opções disponíveis neste menu principal é a ajuda (help). O R dispõe de um completo sistema de ajuda, composto pelas seguintes opções:

- Opções de ajuda sobre a aplicação;
- Opções de ajuda com acessos a partir do browser;
- Opções de ajuda na janela de comandos do R.

Algumas dessas opções podem ser efectuadas directamente no prompt da seguinte forma:

> help("function") ou

>?function

Apresenta a ajuda relativa à função especificada;

> help.start()

Dá acesso a informação auxiliar a partir do browser;

> help.search("text")

Procura as funções cujo nome, detalhes ou descrição contenha o texto indicado;

> apropos("text")

Procura as funções cujo nome contenha o texto indicado.

```
>help("function")
>help.start()
>help.search("text")
>apropos("text")
```

Funcões que permitem obter a ajuda do R

4.4. Regras de sintaxe e Objectos

Primeiras Regras

Uma das regras importantes do R é o facto de ser case sensitive. Por esta razão as letras 'a' e 'A' podem corresponder a diferentes variáveis. Além disso, o R ignora espaços, ou seja, os resultados '8+3' e '8+ 3' dão origem exactamente ao mesmo resultado. Outras regras importantes:

- Podemos agrupar comandos, para serem executados em simultâneo, se estiverem entre chavetas '{ }' e separados por ';';
- O '#' é utilizado para comentários;
- Quando um comando não está completo, o R coloca o sinal de '+' na linha seguinte, permitindo que este seja terminado.

Objectos

No R todos os diferentes conteúdos tais como números, textos, vectores, matrizes, expressões, chamadas funções, etc. são guardados na memória do computador sob a forma de objectos. Todos os objectos têm um nome associado e para armazenamento num objecto usa-se o operador de atribuição, '<-' ou '='. Para visualizar o conteúdo do objecto basta digitar o nome do mesmo.

```
>texto<-"teste"
>texto
[1]"teste"
```

Forma possível de criação de um objecto designado por texto, contendo "teste"

4.5 Introdução de dados com c()

O vector coluna c()

Uma das formas práticas de armazenar valores em R é feita através de objectos denominados vectores. O vector é considerado a estrutura de dados mais simples e consiste numa colecção organizada de elementos. A atribuição é feita a partir da função c( ), cujos argumentos correspondem aos próprios elementos do vector.

A atribuição pode ser feita também por intermédio da função assign( ) que é particularmente útil nas atribuições automáticas, em que desconhecemos os nomes dos objectos.

```
>x<-c(3.5.1.4.5.2.6.7.4.8)
>x
[1] 3.5.1.4.5.0.2.6.7.0.4.8
```

Atribuição de valores ao vector *x*

```
>assign("x",c(3.5,1.4,5,2.6,
  7,4.8)
>x
[1] 3.5  1.4  5.0  2.6  7.0  4.8
```

Atribuição de valores ao vector *x* (alternativa)

Operações com vectores

Uma das vantagens do R é a facilidade na operação com vectores. O vector exemplo, x (composto pelos números 1, 2, 3, 4, 5), pode ser transformado num vector y (que seja igual a 2x+1) desta forma simplificada:

```
> x <- c(1,2,3,4,5)
> y <- 2*x + 1
> y
[1]  3  5  7  9 11
```

De uma forma simples podemos também listar todos os números que sejam superiores a um certo limite, utilizando operadores lógicos. Assim sendo, se pretendermos guardar num outro vector z apenas os valores de y superiores a 3, devemos escrever:

```
> z <- y[y>3]
> z
[1]  5  7  9 11
```

4.6. Importação e exportação de dados

O R dispõe de um conjunto de funções que permitem a importação ou exportação de dados. Para importar ou exportar ficheiros externos, o R dispõe de conjunto de funções que variam de acordo com o formato do ficheiro.

Para ler ficheiros de dados em formato de tabela existem funções mais específicas (dependendo do tipo de ficheiro) e a função read.table que é mais abrangente:

```
> read.table(file,...)
> read.csv(file,...)
> read.csv2(file,...)
> read.delim(file,...)
> read.delim2(file,...)
```

Para saber como se deve usar cada um destes comandos, basta escrever, no R, o nome do comando antecedido de **?**, por exemplo:
>?read.csv

Na importação de ficheiros há alguns parâmetros que é importante definir para garantir a correcta leitura dos dados, tais como:

- sep="\t", para indicação do caracter tabulação como separador entre variáveis;
- dec=",", para indicação do separador decimal;
- header = TRUE, para indicação da existência dos nomes das variáveis na primeira linha.

Ao importar um ficheiro para o R, este deve ficar associado a um objecto. Para tal, o resultado do comando de importação deve ser atribuído ao nome do objecto a que se quer associar. Para importar, através da função read.csv, um ficheiro de texto designado por "ex.csv" e o associar a um objecto Dataset, dever-se-á fazer:

```
> Dataset<-read.csv("C:/.../ex.csv",
  sep="\t",dec=",",header = TRUE)
```

4.7. Primeiros passos na Estatística Descritiva

Análise descritiva

O R dispõe de um conjunto de funcionalidades que permitem fazer uma análise descritiva de dados bastante completa. As medidas descritivas utilizadas e a forma de sumarização da informação deve sempre atender ao tipo de dados de que dispomos, ou seja, às características das variáveis que estamos a analisar. É sabido que para as variáveis quantitativas se podem aplicar, entre outras, medidas de localização e de dispersão[1].

Em resumo, podemos recordar que as Medidas de Localização são medidas que localizam um determinado ponto da distribuição tais como os quartis, o mínimo e o máximo. Quando o ponto em questão corresponde ao centro da distribuição, estas denominam-se por medidas de tendência central (exemplos: média, moda, mediana). As Medidas de Dispersão são as medidas que aferem a variação dos dados em relação ao centro da distribuição (exemplos: variância, desvio padrão, coeficientes de variação e de dispersão). De seguida apresentam-se alguns exemplos simples de utilização das medidas de localização e de dispersão com R.

[1] Geralmente definem-se dois tipos de variáveis: qualitativas e quantitativas. Para saber mais sobre os tipos de dados e sobre as medidas a aplicar em cada caso, consultar o ALEA em 'Noções de Estatística: III – Dados, tabelas e gráficos, disponível em: http://www.alea.pt/html/nocoes/html/cap3_1_i.html

Medidas de Localização

- Média aritmética: mean() calcula a média aritmética simples, para variáveis quantitativas (discretas e contínuas).

```
>a<-c(1,2,3,4,5)
>mean(a)
[1] 3
```

A função mean() calcula a média de uma lista de valores

- Mediana: median() calcula a mediana ou valor central de uma distribuição após ordenação da amostra (é definida pela sua posição na sucessão das observações ou na distribuição de frequências); é também conhecida por percentil 50 ou segundo quartil.

```
>a<-c(1,2,3,4,5)
>median(a)
[1] 3
```

A função median() calcula a mediana de uma lista de valores

- Quantis: quantile() a função calcula os quantis que são estatísticas de ordem que separam a distribuição de acordo com um limite percentual de observações. No caso dos quartis, a distribuição é dividida em quatro partes iguais; estando ordenadas as observações, por ordem crescente, o 1º e o 3º quartis acumulam (até si) 25% e 75% das observações, respectivamente.

```
>a<-c(1,2,3,4,5)
>quantile(a)
0% 25% 50% 75% 100%
1   2   3   4   5
```

A função quantile() calcula os quartis de uma lista de valores

Medidas de Dispersão

- Variância: var() - calcula a variância para uma variável quantitativa.

```
>a<-c(1,2,3,4,5)
>var(a)
[1]2,5
```

A função var() calcula a variância de uma lista de valores

- Desvio padrão: sd() - calcula o desvio padrão de uma variável quantitativa.

```
>a<-c(1,2,3,4,5)
>sd(a)
[1]1.581139
```

A função sd() calcula o desvio padrão de uma lista de uma variável quantitativa

O R dispõe de algumas funções que permitem fazer uma sumarização de dados, essencialmente para variáveis quantitativas (discretas e contínuas). Uma dessas funções é o **summary()**, que calcula para as variáveis quantitativas as seguintes medidas: Mínimo (Min), 1º quartil (1st Qu), Mediana (Median), Média (Mean), 3º quartil (3rd Qu) e Máximo (Max).

```
>a<-c(1,2,3,4,5)
>summary(a)
Min. 1st Qu. Median Mean 3rd Qu. Max.
```

A função summary() calcula algumas estatísticas básicas de uma lista de variáveis.

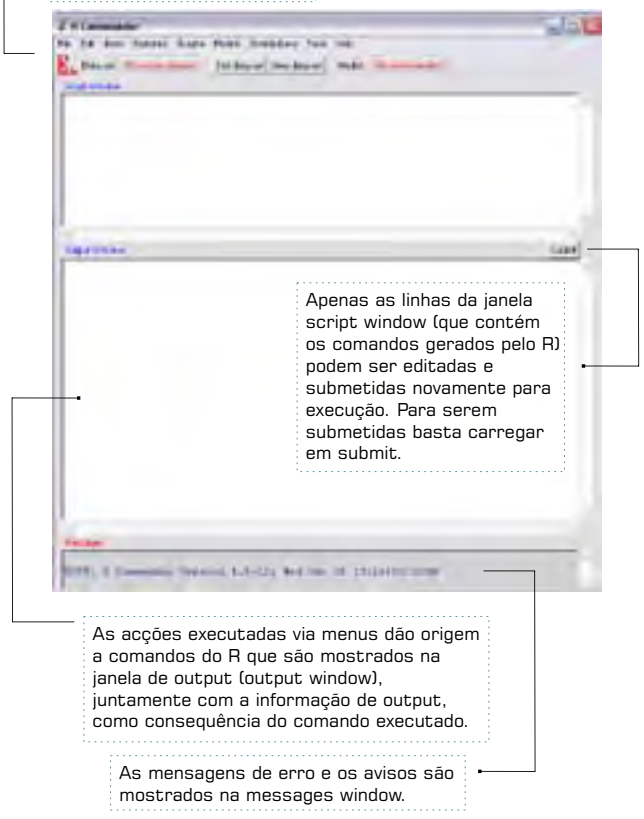
Em resumo, sintetizamos no quadro seguinte os nomes das funções apresentadas, bem como de outras mais específicas, que permitem calcular as respectivas medidas estatísticas no R:

Função	Descrição
table()	Cruzamento de variáveis
mean()	Média aritmética
median()	Mediana
sum()	Soma
summary()	Sumarização de dados
var()	Variância
sd()	Desvio padrão
quantile()	Quartis com descrição
fivenum()	Quartis sem descrição
IQR()	Amplitude inter-quartil
cor()	Coeficiente de correlação

5. *"R Commander"*: um ambiente gráfico

O que é?

Devido ao seu tipo de interface o R torna-se muitas vezes uma ferramenta de utilização pouco amigável. Por essa razão, têm surgido alguns ambientes gráficos que permitem uma utilização do R de uma forma mais intuitiva. O *R-Commander* é uma dessas interfaces gráficas que abre uma janela inicial contendo vários menus e botões de acesso a diferentes procedimentos. Além disso, este ambiente contém uma janela que gera os comandos R que são utilizados em cada procedimento, permitindo assim repetir ou alterar esses comandos. O aspecto geral da janela do *R-Commander* é apresentado de seguida.

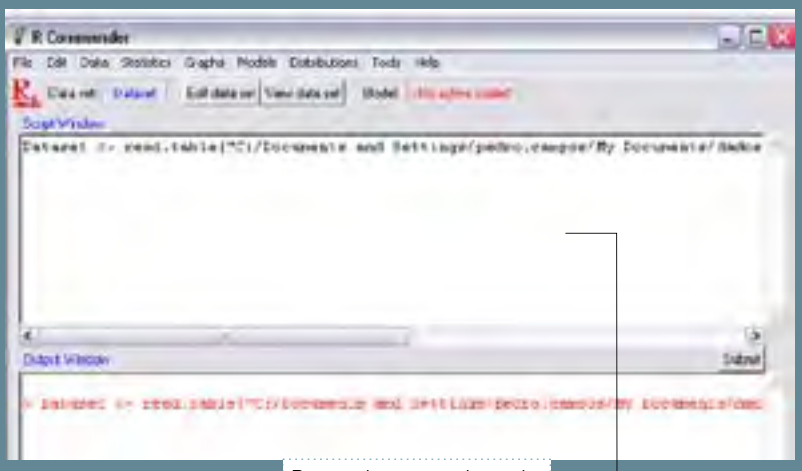


Como se instala?

O *R-Commander* é um package standard (designado por Rcmdr) e os processos de instalação e carregamento fazem-se da mesma forma do que nos outros packages (seguir o procedimento install packages – escolhendo o package Rcmdr e, depois, load package). Existem, por vezes, alguns aspectos a ter em conta durante a instalação: um dos pontos a ter em conta é que o *R-Commander* utiliza alguns "contributed" packages que devem estar instalados para que o *R-Commander* funcione adequadamente [2].

Como funciona?

Um dos primeiros passos a dar depois de entrar no *R-Commander* consiste em activar um conjunto de dados. A partir desse momento, todas as acções serão executadas nesse conjunto de dados. Ao abrir-se um novo conjunto de dados, este passará a ser o conjunto de dados activo. O utilizador pode, em qualquer momento, seleccionar o conjunto que pretende, entre todos os conjuntos de dados que já estiveram activos anteriormente.



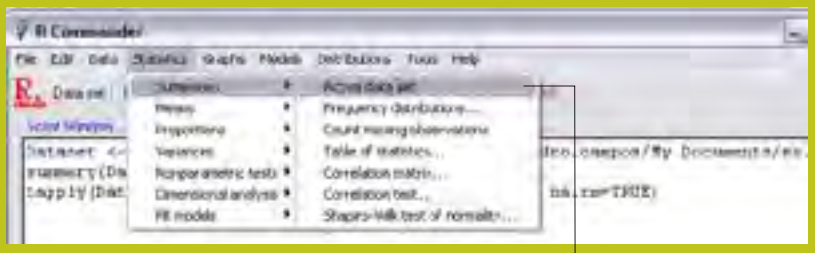
Para activar um conjunto de dados pode-se importar um ficheiro de texto através do menu: (Data/Import Data/ from text file or clipboard

[2] No caso da versão 1.4-2 do *R-Commander* esses packages são: *abind, car, effects, lmtest, multcomp, mvtnorm, relimp, sandwich, strucchange, e zoo.* Além destes packages, deve-se instalar também o package *rgl* no caso de se pretender construir gráficos 3D.

O ficheiro em causa contém dados sobre as peças produzidas numa determinada fábrica de peças para automóveis. Para cada peça produzida dispõe-se de informação sobre:

- **seccao**: secção onde a peça foi produzida (var. qualitativa: valores de 1 a 6);
- **cod**: código da peça (var. qualitativa: valores possíveis: 12, 45, 78, 96);
- **peso**: peso da peça (var. quantitativa);
- **diametro**: diâmetro da peça (var. quantitativa);
- **empregado**: empregado que executou/verificou a peça (var. qualitativa: valores de 1 a 3);
- **tipo**: tipo de aplicação da peça: (var. qualitativa: (c) coluna ou (d) dentro);
- **qualidade**: resultado da verificação: (var. qualitativa: (0) rejeitada ou (1) aprovada).

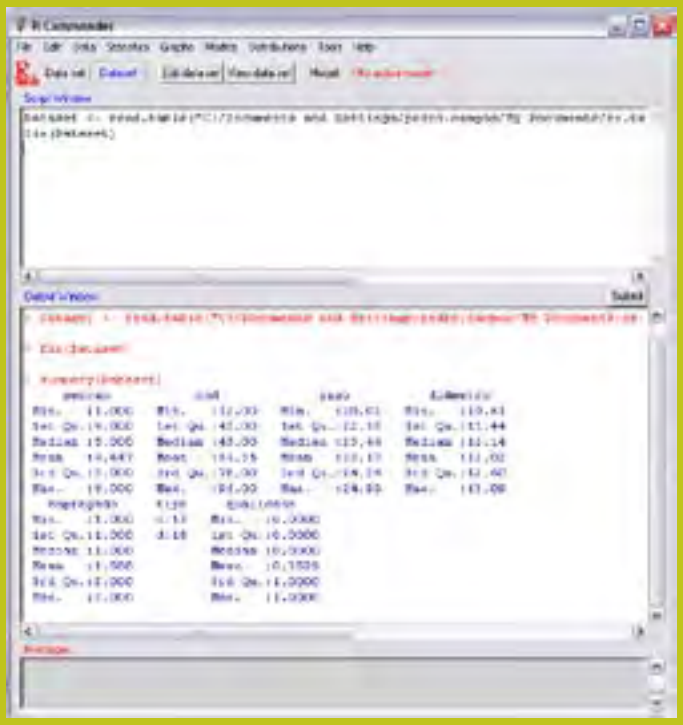
De seguida, no menu *'Statistics/Summary/Active Data Set'* pode solicitar as estatísticas básicas (mínimo, máximo, mediana, quartis) que correspondem à execução do comando summary.



No menu Statistics seleccione a opção Summary/Active Data Set que permite calcular as estatísticas básicas (mínimo, máximo, mediana, quartis), que correspondem à execução do comando summary().

Os resultados encontram-se na figura ao lado (output window). Para cada variável foram calculadas as estatísticas: mínimo, máximo, 1º, 2º e 3º quartis, a média e a mediana. Estes resultados poderiam ter sido obtidos directamente através do comando:

```
>summary(dataset)
```



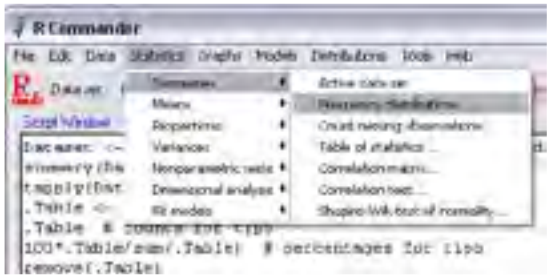
Como neste conjunto de dados existem variáveis de vários tipos, podemos utilizar algumas funcionalidades disponíveis do *R-Commander*, tais como distribuições de frequências, cálculos de estatísticas variadas, representação gráfica, etc. Desenvolveremos esta análise nos próximos capítulos do dossiê.

```
> 100*.Table/sum(.Table)  # percentages for tipo
c       d
47.05882 52.94118
```

6. Análise de Dados

Frequências absolutas e relativas

Prosseguindo com o exemplo anterior, em que dispomos de variáveis de vários tipos (qualitativas e quantitativas), interessa analisar agora as potencialidades do *R-Commander*. Após a primeira sumarização, onde se calcularam as medidas de localização, podemos agora, por exemplo, calcular as frequências absolutas das variáveis qualitativas Para tal, deve-se escolher no menu Statistics a opção 'Summatize/Frequency Distributions'.



O resultado é mostrado na janela output window como sendo a aplicação da função table() da seguinte forma:

```
> .Table <- table(Dataset$tipo)
> .Table  # counts for tipo
c  d
16 18
```

É de notar que a expressão Dataset$tipo é a forma como correctamente nos referimos à variável tipo do conjunto de dados denominado Dataset e que é equivalente a utilizar a expressão Dataset[, "tipo"].

No *R-Commander* mostram-se ainda as frequências relativas associadas a estas frequências absolutas.

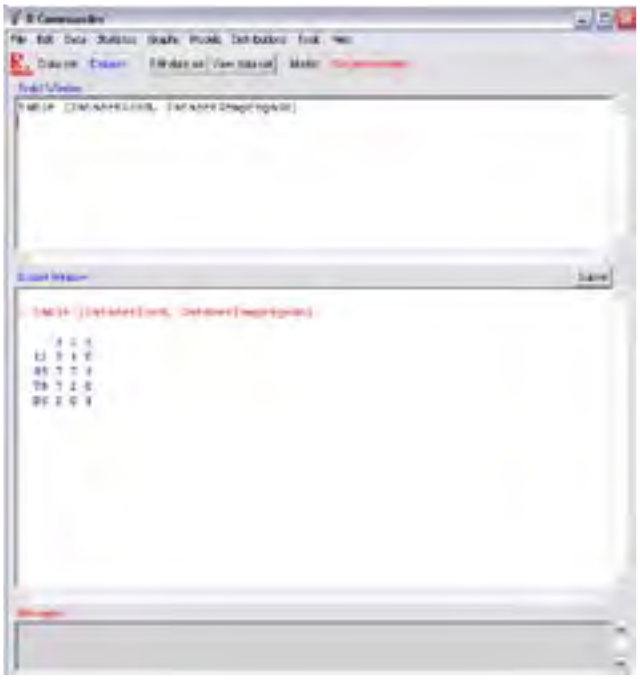
Tabelas de contingência

Podemos também combinar variáveis e calcular tabelas de contingência que resultam das frequências cruzadas entre variáveis qualitativas. Embora não exista um comando directamente acessível, através dos menus do *R-Commander*, pode-se escrever o comando na janela Script Window e carregar no botão Submit para executar o comando. Assim sendo, para podermos, por exemplo, identificar quantas (e quais) as peças que foram feitas por cada empregado, devemos escrever:

```
>table (Dataset$cod, Dataset$empregado)
```

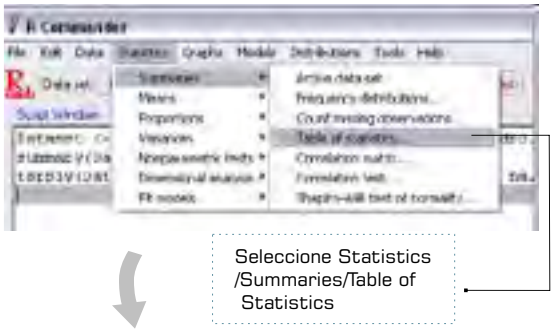
O resultado é o seguinte:

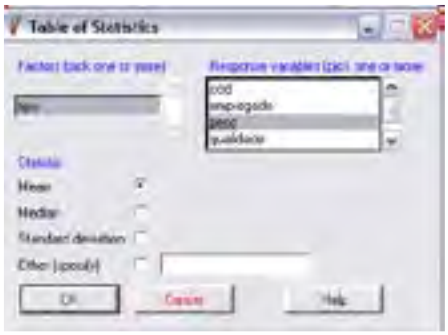
```
   1 2 3
12 3 1 0
45 7 7 4
78 7 2 0
96 2 0 1
```



Medidas de localização e de dispersão:

De seguida podemos também calcular as medidas de localização e de dispersão para uma variável quantitativa, por grupos definidos segundo as modalidades de uma variável qualitativa. Por exemplo, podemos calcular estatísticas sobre o peso das peças produzidas, tendo em conta o tipo de peça. Para tal devemos escolher a opção *'Statistics/Summaries/Table of Statistics'* e, de seguida, escolher como Factor a variável *tipo*. Neste caso, o *tipo* é aqui considerada uma variável independente.





O resultado é a execução do comando tapply que aplica um procedimento à variável quantitativa para grupos distintos (identificados pela variável qualitativa).

```
> tapply(Dataset$peso,
list(tipo=Dataset$tipo), mean,
na.rm=TRUE)

tipo

c        d

26.02323 29.12170
```

Correlação

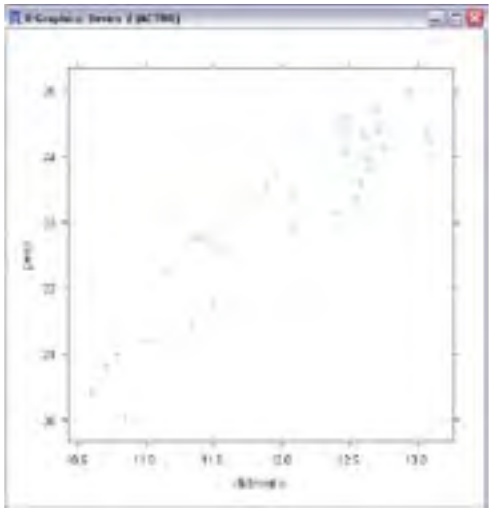
Quando numa base de dados se dispõe de mais do que uma variável, pode fazer sentido calcular o nível ou grau de associação existente entre essas variáveis. Em geral, estes coeficientes medem a força e a direcção (no mesmo sentido ou em sentidos opostos) da relação entre as variáveis. Existem vários tipos de coeficientes de correlação conforme o tipo de variáveis em estudo: qualitativas nominais, qualitativas ordinais, quantitativas, etc. O coeficiente de correlação linear de *Pearson* é um dos mais conhecidos e aplica-se quando as variáveis são quantitativas[3].

Para se perceber que tipo de relação existe entre um par de variáveis, é habitual começar-se por desenhar um diagrama de pontos. Este tipo de representação é muito útil, pois permite realçar algumas propriedades entre os dados, nomeadamente no que diz respeito ao tipo de associação entre as variáveis.

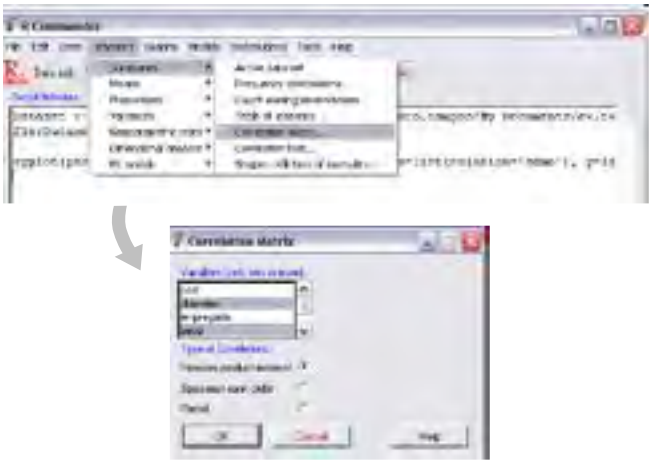
No caso do conjunto de dados em estudo, vamos verificar a relação existente entre as variáveis peso e *diâmetro* das peças. Para tal escolhemos no *R-Commander* a opção *'Graphs/XY Conditioning plot'*.[4]

[3] Embora este coeficiente se aplique especialmente no caso em que as variáveis seguem distribuição Normal, esta restrição é muitas vezes ignorada. Para saber mais sobre o coeficiente de correlação, consulte o curso de Noções de Estatística no ALEA, Capítulo VI – Distribuições Bidimensionais, em http://www.alea.pt/html/nocoes/html/cap6_3_1.html e/ou ActivALEA n.º 4 "Associação entre variáveis quantitativas: O coeficiente de Correlação."

[4] No capítulo 6 deste dossiê pretende-se aprofundar um pouco mais a questão da representação gráfica em R.



Este gráfico sugere a existência de uma relação directa entre as variáveis diâmetro e peso, ou seja, a valores grandes de diâmetro correspondem, de um modo geral, valores grandes de peso e vice-versa. Esta informação pode ser confirmada pelo cálculo do coeficiente de correlação linear de *Pearson* (ou r de *Pearson*). Este procedimento pode ser desencadeado através do menu (ver figura seguinte) e corresponde à execução do comando cor(x,y), em que x e y representam as variáveis em estudo para as quais se pretende calcular o coeficiente de correlação.



Na janela Output Window podemos observar o resultado:

```
> cor(Dataset[,c("diâmetro","peso")],
use="complete.obs")

          diâmetro      peso
diâmetro 1.0000000 0.9166048
peso     0.9166048 1.0000000
```

De facto, podemos notar que a correlação existente entre o diâmetro das peças (x) e o peso das peças (y) é de, aproximadamente, 0.92.

O *R-Commander* dispõe também de outras opções de análise de dados: análise factorial, testes paramétricos e não paramétricos, etc. Estas técnicas não são abordadas no contexto deste dossiê.

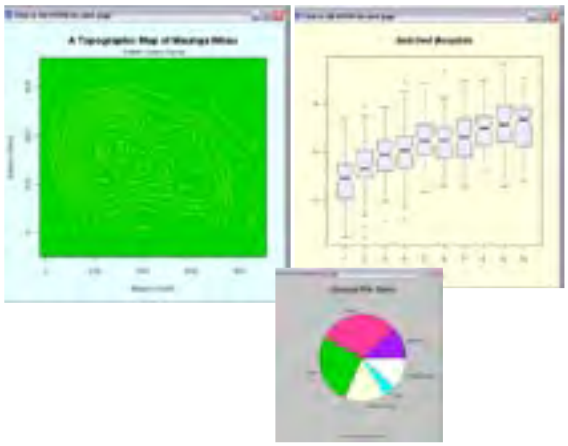
Gestão das variáveis

No *R-Commander* existe a possibilidade de se fazer a gestão do conjunto de dados: acrescentar novas variáveis, novas observações, agregar valores em classes, etc. Esta opção encontra-se disponível através de *'Data/Manage variables in active data set'*.

Para fazer a gestão dos dados recorra à opção 'Data/Manage variables in active data set'.

7. Gráficos

As facilidades gráficas são uma componente importante e muito versátil no ambiente R, sendo possível utilizar essas facilidades numa larga variedade de gráficos estatísticos predefinidos, bem como construir gráficos novos que podem ser formatados e apresentados com grande qualidade.



Os gráficos constituem uma forma de sumariar a informação, sendo que a sua representação gráfica deve ser feita de forma a dar relevo às propriedades importantes dos dados. A construção dos gráficos deve ter em conta o tipo de variáveis que se pretende representar. Na tabela seguinte apresenta-se um resumo do tipo de gráficos, mais comuns, que deve ser feito para cada tipo de variável:

Tipo de variável	Representação gráfica
Qualitativa (ordinal,nominal)	Gráficos de barras, diagramas circulares.
Quantitativa discreta	Gráficos de barras, diagramas circulares, diagramas de dispersão, diagramas de caixas e bigodes, etc.
Quantitativa contínua	Histogramas, diagramas de dispersão, diagramas de caixa e bigodes, etc.

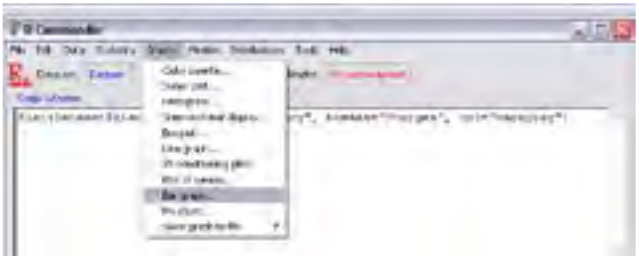
Neste capítulo pretende-se fazer uma visita geral a alguns tipos de gráficos mais conhecidos (gráficos de barras, diagramas circulares, histogramas e gráficos de pontos) e à forma com se podem construir através do *R-Commander*. A apresentação específica de cada gráfico e a sua formatação não são objectivo principal desta abordagem, pelo que deverá consultar as ajudas do R para comandos adicionais.

Apresenta-se, de seguida, a forma como pode fazer alguns destes gráficos tomando por base o mesmo conjunto de dados dos exemplos anteriores.

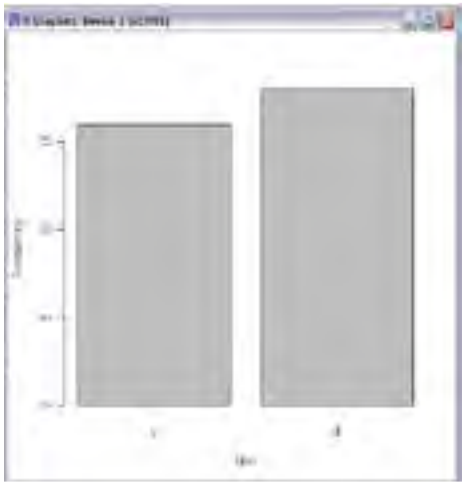
Gráfico de barras e diagramas circulares

O gráfico de barras é uma forma de representação adequada a variáveis qualitativas e quantitativas discretas. No gráfico de barras cada valor associado a uma modalidade da variável é representado através de uma barra cuja altura é proporcional à sua frequência.

De seguida apresentam-se os passos necessários para fazer um gráfico de barras no *R-Commander* para a variável *tipo* (variável qualitativa relacionada com o tipo de aplicação da peça: (c) coluna ou (d) dentro).



Para fazer um gráfico de barras recorra à opção 'Graphs/Bar Graph' e escolha, depois, a variável qualitativa que pretende representar

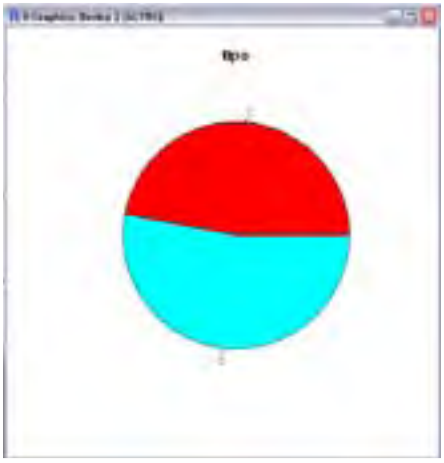


O comando gerado pelo *R-Commander* que permite fazer este gráfico directamente no R é o seguinte:

```
>barplot(table(Dataset$tipo), xlab="tipo",
ylab="Frequency")
```

Para construir um diagrama circular, igualmente adequado a este tipo de dados, o procedimento é idêntico, excepto na opção de gráficos, onde se deve escolher pie chart em vez de bar graph. O comando gerado no R é o seguinte:

```
>pie(table(Dataset$tipo),labels=levels(Dat
aset$tipo),main="tipo",col=rainbow(length
(levels(Dataset$tipo))))
```



Histograma

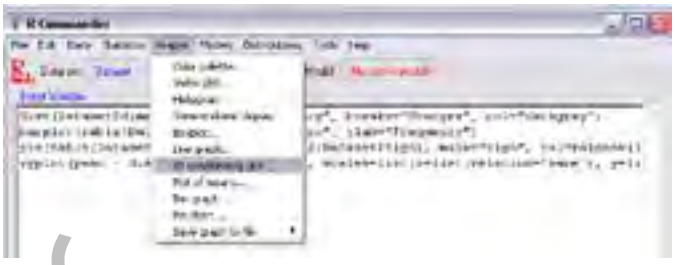
O histograma é uma das formas mais importantes de representar dados quantitativos. Para se fazer um histograma é necessário começar por agrupar as observações em classes e depois representar, para cada classe, uma barra cuja altura seja proporcional ao número de observações. Uma vez que as classes ou intervalos em que os dados são agrupados são contíguas, as barras são apresentadas sem separação. Para fazer um histograma no *R-Commander* considerando a variável diâmetro proceda como se indica na figura:

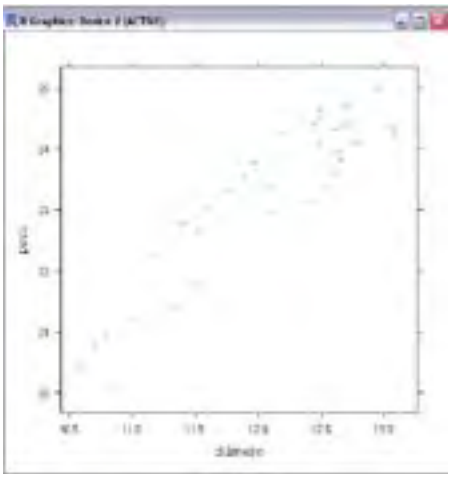
O comando gerado pelo *R-Commander* que permite fazer este gráfico directamente no R é o seguinte:

```
>hist(Dataset$diâmetro,
scale="frequency", breaks="Sturges",
col="darkgray")
```

Diagrama de pontos

Também conhecido por diagrama de dispersão, o gráfico de pontos é muito adequado nos casos em que pretendemos representar duas variáveis quantitativas (discretas ou contínuas), particularmente quando pretendemos analisar a sua correlação.





O comando gerado pelo *R-Commander* que permite fazer este gráfico directamente no R é:

```
> xyplot(peso~diâmetro,auto.key=TRUE,scales=list(x=list(relation='same'), y=list(relation='same')), data=Dataset)
```

8. Exemplos de Aplicação

Este capítulo contém alguns exercícios de aplicação imediata e problemas resolvidos através do R tais como: **"Número de irmãos dos alunos da turma H do 9º ano", "Alturas dos Alunos", "Construir um Triângulo"**, "Uma Corrida Com Dados" e "Resultados de um teste" (este último associado ao programa PISA).

Pensamos que estes exercícios e problemas ajudam a aprofundar os conhecimentos de R apresentados neste dossiê, sendo que, para a sua resolução, se utilizaram conceitos que são usualmente trabalhados no ensino básico e secundário.

Número de irmãos dos alunos da turma H do 9º ano

1	0	1	2	1	1	1	3	0	4	0	1	1
4	2	3	2	1	3	1	2	1	2	1	2	3

Construa:

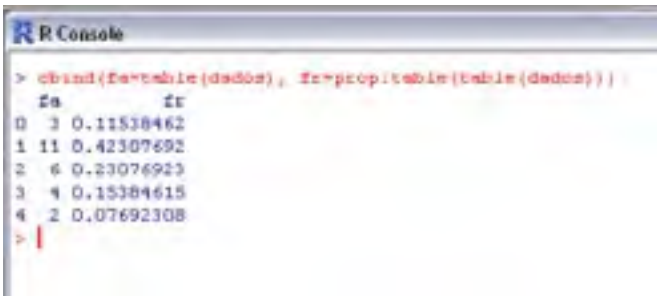
a) a tabela de frequências.

b) o diagrama de barras

Resolução com R:

a) Para construir a tabela de frequências:

```
> cbind(fa=table(dados), fr=prop.table(table(dados)))
```



b) Para construir o diagrama de barras:

```
> barplot(table(dados), main="Número de Irmãos no 9º H", xlab="Número de Irmãos", ylab="Frequência", col=rep("pink",5), ylim=c(0,12))
```

Alturas dos Alunos

Para este exercício, foram registadas as alturas, em centímetros, dos alunos de uma turma do 10º ano:

Altura dos alunos						
150	169	174	155	165	170	172
152	158	163	158	166	158	166
170	171	162	171	161	154	168
161	164	166	164	162	156	167

Construa uma tabela de frequências, agrupando os dados em classes e represente graficamente os dados, utilizando o tipo de gráfico que achar mais conveniente. Faça ainda um diagrama de caule-e-folhas.

Resolução com R:

• O primeiro passo consiste em transmitir os dados ao R. Para tal, podemos criar um ficheiro com estes dados (exercício1.csv) ou lê-los através de um vector.

```
> dados<-read.csv("Exercicio1.csv")
```

ou

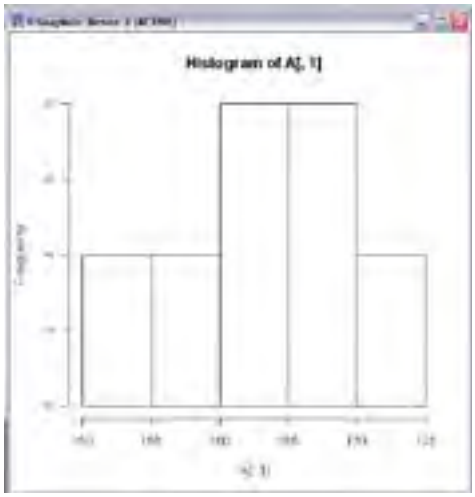
```
>dados<-c(150,169,174,155,165,170,
172,152,158,163,158,166,158,166,170,
171,162,171,161,154,168,161,164,166,
164,162,156,167)
```

• De seguida aplicamos o comando hist.

```
> hist(dados[,1])
```

Para formatar melhor o gráfico, podemos recorrer aos parâmetros do comando hist:

```
> hist(A[,1],breaks="Sturges", col="light blue", xlab="Altura", ylab="Frequência", main="Alturas de Alunos")
```



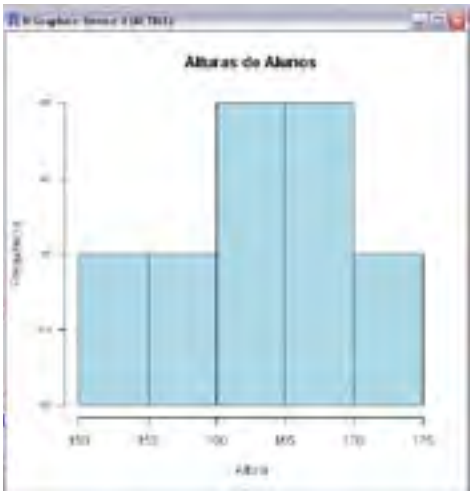
E o resultado é...

A partir do comando do histograma, poderemos construir uma tabela de frequências. Para tal, basta guardar o resultado do comando hist.

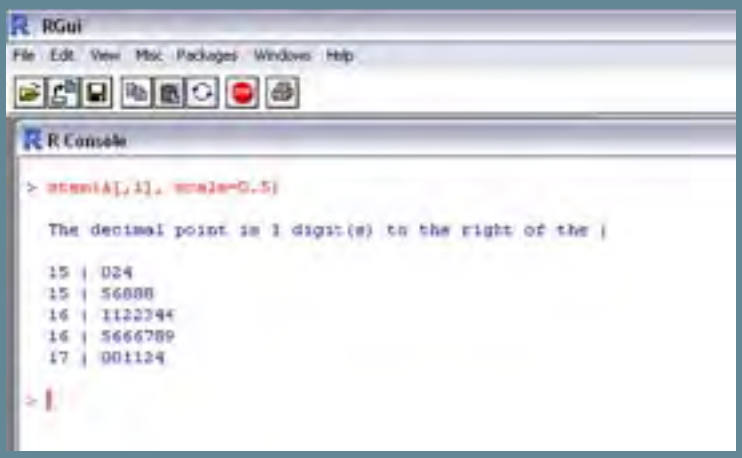
```
> s<- hist(dados[,1])
> s
```

```
$breaks
[1] 150 155 160 165 170 175
$counts
[1] 4 4 8 8 4
$intensities
[1] 0.02857142 0.02857143 0.05714286
0.05714286 0.02857143
```

(...)



Para fazer um diagrama de caule-e-folhas[5] deveremos aplicar o comando *stem*:



Construir um triângulo...

Um segmento de comprimento unitário é dividido em 3 partes, aleatoriamente. Qual a probabilidade de as partes resultantes poderem formar um triângulo?

A resolução deste problema prende-se com uma regra que estabelece que a soma dos comprimentos de dois lados de um triângulo é superior ao comprimento do outro lado.

Nota – Quando se fala em números aleatórios, estamos intuitivamente a pensar em números com uma distribuição uniforme, no intervalo [0,1].

Resolução do problema por simulação no R:

Vamos fazer um determinado número de simulações e calcular a frequência relativa das situações que dão origem a triângulos. Para tal, vamos gerar dois números aleatórios entre 0 e 1 e estes números irão representar os pontos P e Q em que um segmento [MN] de comprimento 1 fica dividido:

Vamos considerar para P o menor dos valores obtidos, que será o comprimento de MP. Calcula-se o comprimento dos segmentos PQ e QN e depois testa-se se dois quaisquer dos comprimentos obtidos são superiores ao terceiro comprimento. Terminado o número de simulações, calcula-se o número das situações que dão origem a triângulos e divide-se pelo número de simulações.

[5] Para saber mais sobre este tipo de gráfico consulte o AELA em: http://www.alea.pt/html/nocoes/html/cap3_2_20.html

Script 1 "Problema do triângulo"

```
cont=0
NumSim=1000
segmentos=array(0,dim=c(NumSim,3))
for (i in 1:NumSim) {
    M=0
    N=1
    A=runif(1,0,1)
    B=runif(1,0,1)
    MP=min(A,B)
    PQ=abs(A-B)
    QN=1-max(A,B)
    if (MP+PQ > QN & MP+QN>PQ & PQ+QN>MP) cont=cont+1
    segmentos[i,1]=MP
    segmentos[i,2]=PQ
    segmentos[i,3]=QN
}
cat("frequência relativa",cont/NumSim)
```

Por exemplo, pedindo 1000 simulações, obteve-se:

Frequência relativa de triângulos: 0.256

Acrescentando ao script anterior, o cálculo do comprimento médio de cada segmento nos casos em que é possível construir um triângulo:

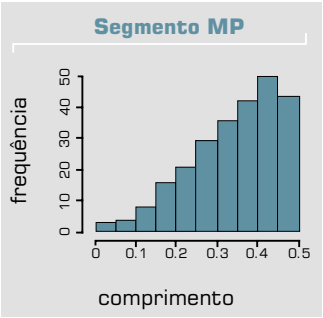
Script 2 "Problema do triângulo"

```
cont=0
NumSim=1000
segmentos=array(0,dim=c(NumSim,3))
for (i in 1:NumSim) {
            M=0
            N=1
            A=runif(1,0,1)
            B=runif(1,0,1)
            MP=min(A,B)
            PQ=abs(A - B)
            QN=1- max(A,B)
            if (MP+PQ > QN & MP+QN>PQ & PQ+QN>MP) {
                        cont=cont+1
                        segmentos[cont,1]=MP
                        segmentos[cont,2]=PQ
                        segmentos[cont,3]=QN
                        par(mfrow=c(2,2))
                        cor1=c("blue")
                        cor2=c("pink")
                        cor3=c("yellow")
                                    }
}
segmentos=segmentos[1:cont,]
hist(segmentos[,1],col=cor1,xlab="comprimento",ylab=
"frequência",main="Segmento MP")
hist(segmentos[,2],col=cor2,xlab="comprimento",ylab=
"frequência",main="Segmento PQ")
hist(segmentos[,3],col=cor3,xlab="comprimento",ylab=
"frequência",main="Segmento QN")

cat("frequência relativa de triângulos",cont/NumSim)
cat("comprimento médio do segmento MP",mean(segmentos[,1]))
cat("comprimento médio do segmento PQ",mean(segmentos[,2]))
cat("comprimento médio do segmento QN",mean(segmentos[,3]))
```

Fizemos nova simulação e obtivemos:

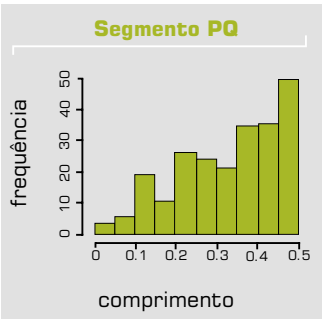
Comprimento médio do segmento MP:

0.3432921



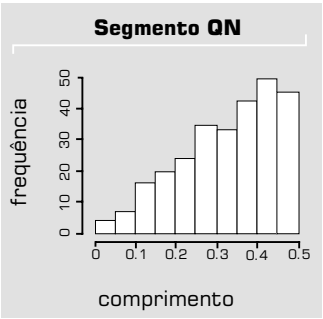
Comprimento médio do segmento PQ:

0.3286406



Comprimento médio do segmento QN:

0.3280673



"Curiosamente" o comprimento médio dos segmentos aproxima-se de 1/3.

Efectuando maior número de simulações, a frequência relativa dos casos em que é possível construir um triângulo aproxima-se de 0,25 e o comprimento médio dos segmentos desses triângulos é um valor próximo de 0,33.

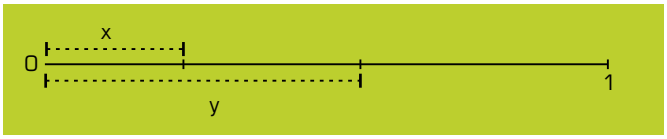
Voltando à simulação no R...

O script elaborado inicialmente pareceu-nos o processo mais indicado para ser explicado aos alunos, mas explorando um pouco mais as potencialidades do R, fizemos um novo script tendo por base o seguinte raciocínio: considere-se duas variáveis aleatórias X e Y (com distribuição uniforme no intervalo [0,1]) e independentes:

- X tem distribuição uniforme no intervalo [0,1]
- Y tem distribuição uniforme no intervalo [0,1]

Quando se seleccionam 2 números, um com distribuição X e outro com distribuição Y, podemos ter uma de duas situações: X<Y ou X>Y.

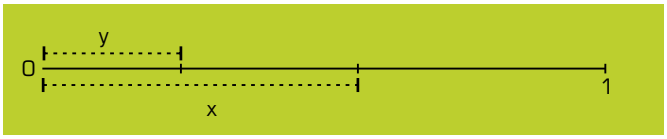
X<Y



Lados do (suposto) triângulo:

X (Y-X) (1-Y)

X>Y



Lados do (suposto) triângulo:

Y (X-Y) (1-X)

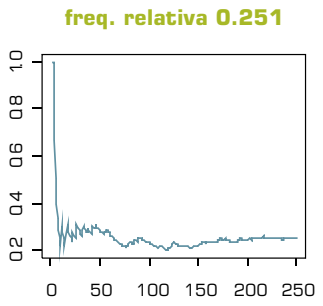
Para que possam, efectivamente, ser os lados de um triângulo, cada lado tem de ser inferior à soma dos outros dois.

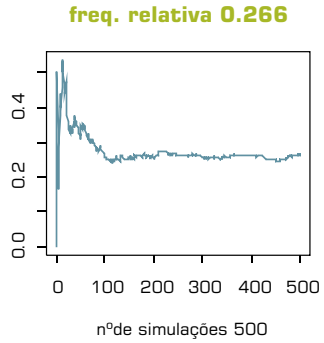
Neste novo script indicamos apenas o número de simulações desejadas e obtemos graficamente a evolução da frequência relativa, dos casos em que é possível construir um triângulo, observando-se em simultâneo a frequência relativa para os quartis do número n de simulações indicadas.

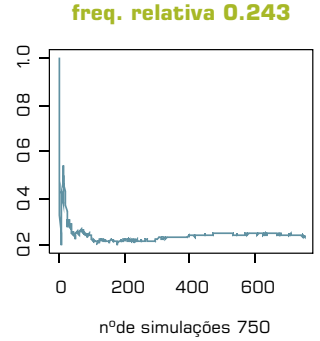
Script 3 "Problema do triângulo"

```
grafico=function(n) {
calculo=function(n) {

x=runif(n,0,1)
y=runif(n,0,1)
cond=((x>1/2 & (x-y)<1/2 & y<1/2) | (x<1/2 & (y-x)<1/2 & y>1/2))
v=round(sum(cond)/n,3)
color=c("blue")
plot(1:n,col=color,cumsum(cond)/ (1:n), type="l",main=paste
("freq. relativa",v ), xlab=paste("nºde simulações", round(quantile(n),
0)[i]),ylab="")

}
par(mfrow=c(2,2))
for (i in 2:5) {calculo(round(quantile(1:n),0)[i])}
}
```

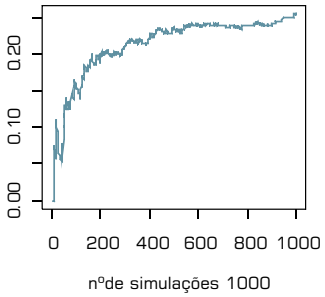
Por exemplo, digitando grafico(1000) (designamos a nossa função no R por grafico), obtivemos os resultados para 1000 simulações, ilustradas na figura seguinte:





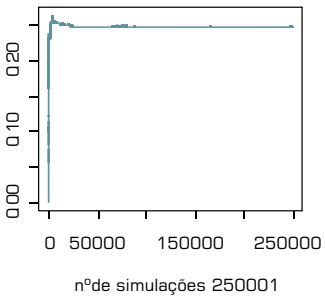


freq. relativa 0.254

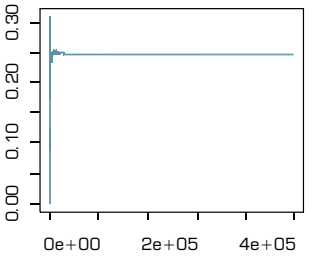


Para 1 000 000 simulações:

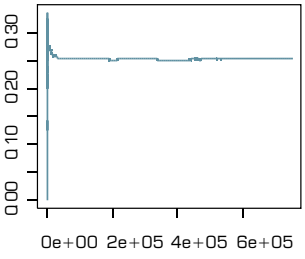
freq. relativa 0.249



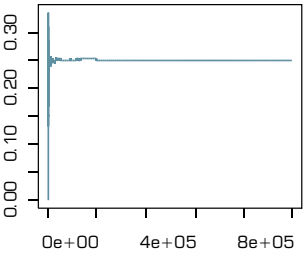
freq. relativa 0.249



freq. relativa 0.25



freq. relativa 0.25



Aumentando o número de simulações, a frequência relativa tende a estabilizar à volta do valor 0,25, o que vem comprovar a definição frequencista do conceito de probabilidade: a probabilidade de um determinado acontecimento é o valor obtido para a frequência relativa com que se observou esse acontecimento, num grande número de realizações da experiência aleatória.

Uma Corrida com Dados

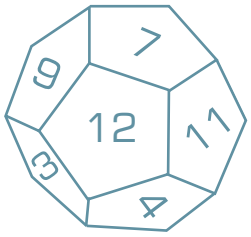
> O Bruno arranjou um dado especial com a forma de um dodecaedro. Tem 12 faces, numeradas de 1 a 12.

> A Tânia tem dois dados normais. São cubos, cada um deles com as faces numeradas de 1 a 6.

Resolveram fazer um jogo. Cada jogada consiste no lançamento dos três dados.

Vão somando os pontos que cada um obtém: o Bruno com o seu dado de 12 faces e a Tânia com os seus dois dados de 6 faces. Ganha quem primeiro chegar aos 100 pontos.

Se por acaso os dois chegarem aos 100 pontos na mesma jogada, ganha quem tiver o total maior. Se esse total for igual para os dois, há empate.



Algum dos jogadores está em vantagem? Ou é o jogo equilibrado?

(Desafios do Público)

Antes da realização das experiências cada elemento do grupo conjecturou sobre quem teria maior probabilidade de vencer, se o Bruno lançando o dodecaedro, se a Tânia lançando dois dados cúbicos. Surgiram opiniões diversas:

• A Tânia obtém, no mínimo, por jogada, dois pontos enquanto que o Bruno pode obter um;

• No dodecaedro a probabilidade de sair **doze** é $\frac{1}{2}$ que é maior que $\frac{1}{36}$, correspondente à probabilidade do mesmo resultado no caso dos dados cúbicos;

• A probabilidade de obter **seis** é maior no lançamento dos dois dados cúbicos, $\frac{5}{36}$, contra $\frac{1}{12}$ no dodecaedro; essa vantagem acentua-se mais no caso da obtenção do valor sete ao qual corresponde as probabilidades $\frac{1}{6}$, nos dados cúbicos, e $\frac{1}{12}$ no outro dado.

Script 1 "Corrida de Dados" em R

```
#Simular um jogo da corrida de dados
L=1
AcumCubico=0
AcumDode=0
while (AcumCubico<100 & AcumDode<100) {
AcumCubico=AcumCubico+round(runif(1,1,6))+round(runif(1,1,6))
AcumDode=AcumDode+(round(runif(1, 1, 12)))
L=L+1
}
if (AcumCubico>AcumDode) print ("Foi o par de dados cubicos")
else if (AcumDode>AcumCubico) print ("Foi o dodecaedro") else
if (AcumCubico==AcumDode) print ("Empate")
print (paste("Total de jogadas", L))
print (paste("Total de pontos dos dados cúbicos", AcumCubico))
print (paste("Total de pontos do dodecaedro", AcumDode))
```

Começamos por elaborar um script para a simulação de um jogo:

Na simulação que realizámos o resultado foi o seguinte: venceu "o par de dados cúbicos", realizaram-se "16 jogadas", sendo o total dos pontos dos dados cúbicos "107" e o total de pontos do dodecaedro "105".

Elaborámos um outro script para simular vários jogos:

Script 2 "Corrida de Dados" em R

```
#Simular vários jogos da corrida de dados
dados=function(n) {
CUBICO=0
DODE=0
EMPATE=0
for (i in 1:n) {
L=1
AcumCubico=0
AcumDode=0
while (AcumCubico<100 & AcumDode<100) {
AcumCubico=AcumCubico+round(runif(1,1,6))+round(runif(1,1,6))
AcumDode=AcumDode+(round(runif(1, 1, 12)))
L=L+1
}
if (AcumCubico>AcumDode) CUBICO=CUBICO+1 else if
(AcumDode>AcumCubico) DODE=DODE+1 else if (AcumCubico=
=AcumDode) EMPATE=EMPATE+1
}
print (paste("Freq.relativa do n.ºde vezes em que os dados
cubicos ganharam", CUBICO/n))
print (paste("Freq.relativa do n.ºde vezes em que o dodecaedro
ganhou", DODE/n))
print (paste("Freq.relativa do n.ºde empates", EMPATE/n))
}
```

Executado o script para simular 100 jogos, digitamos na consola do R "dados (100)" e obtivemos:

- "Freq. relativa do n.º de vezes em que os dados cúbicos ganharam 0.67"
- "Freq. relativa do n.º de vezes em que o dodecaedro ganhou 0.32"
- "Freq. relativa do n.º de empates 0.01"

Se o número de experiências for suficientemente grande, a percentagem de cada resultado estará próxima do valor real da probabilidade (Lei dos Grandes Números).

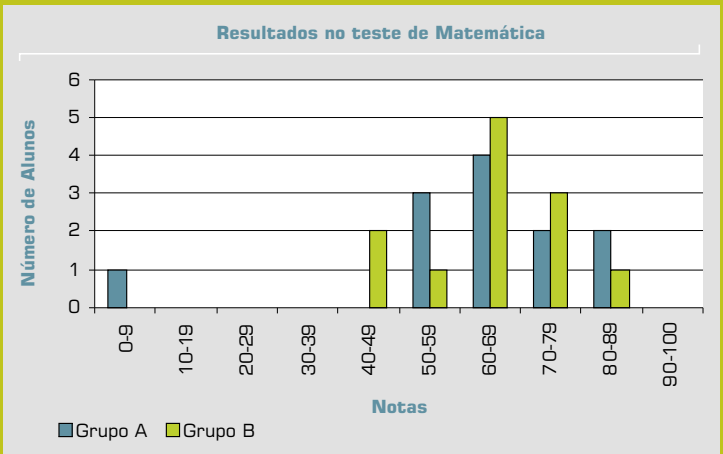
Simulámos no R, um milhão de jogos e ao fim de alguns minutos... obtivemos:

- "Freq. relativa do n.º de vezes em que os dados cúbicos ganharam 0.676556";
- "Freq. relativa do n.º de vezes em que o dodecaedro ganhou 0.304982";
- "Freq. relativa do n.º de empates 0.018462".

Assim, a probabilidade de a Tânia ganhar será aproximadamente 67,7% e a do Bruno 30,5%. A probabilidade de empate é de 1,8%. Claro que estes não são valores exactos... mas estarão próximos dos valores reais.

Resultados de um Teste

O gráfico seguinte mostra os resultados de um teste de Matemática obtidos por dois grupos de alunos, designados por "Grupo A" e "Grupo B". A nota média no grupo A é de 62,0 e no grupo B de 64,5. Os alunos passam neste teste se tiverem uma nota igual ou superior a 50.



Com base nesta informação, o professor concluiu que o grupo B teve melhores resultados neste teste que o grupo A.

Os alunos do grupo A não estão de acordo com o professor. Tentam convencer o professor de que o grupo B não teve necessariamente melhores resultados.

Utilizando a informação dada, apresente pelo menos um argumento matemático que possa ser utilizado pelos alunos do grupo A.

adaptado do Programa para a Avaliação Internacional de Alunos 2003, PISA – Programme for Internacional Student Assessment

Argumentos que podem ser utilizados:

- Há mais alunos que passaram no teste no Grupo A do que no Grupo B (há mais "positivas" no Grupo A do que no Grupo B);
- O Grupo A tem mais alunos com nota igual ou superior a 80 que o grupo B;
- Se ignorarmos o aluno mais fraco do Grupo A, os alunos do Grupo A têm melhores resultados que os do grupo B.

Respeitando a informação dada no problema, consideremos que os resultados obtidos pelos dois grupos foram os seguintes:

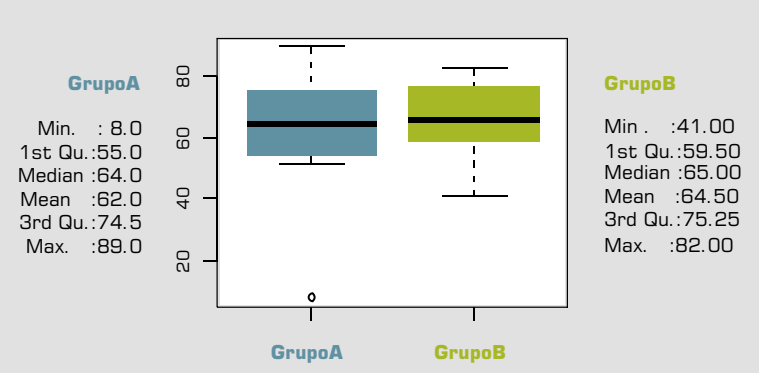
Grupo A: 8, 51, 52, 56, 61, 63, 65, 67, 74, 76, 82, 89

Grupo B: 41, 43, 55, 61, 62, 63, 67, 68, 74, 79, 79, 82

Utilizando o programa R[9] , calculemos as principais estatísticas descritivas destes dois grupos, bem como os respectivos boxplots (caixas de bigodes):

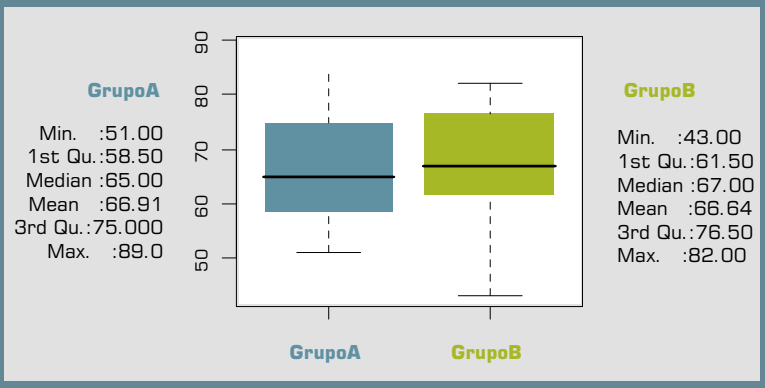
[9] Ver script elaborado no final deste problema.

Note-se que a nota mais baixa do Grupo A, que se afasta significativamente das restantes (outlier), está assinalada com um (ponto). Este valor interfere bastante na média dos resultados do Grupo A. Efectivamente, se retirarmos a nota mais baixa a cada um dos grupos, respectivamente 8 e 41, obtemos:



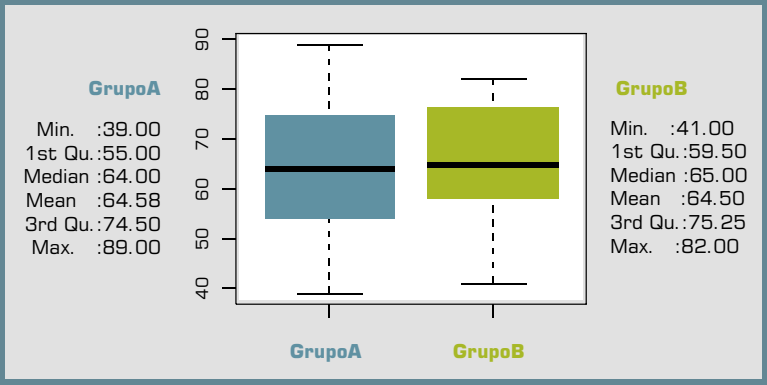
Com esta alteração obtemos uma melhor leitura do gráfico, dada a redução na dispersão dos dados. Confirma-se assim uma subida das estatísticas descritivas, em particular no Grupo A, em que a média das notas do Grupo A supera a média das notas do Grupo B.

Retomando as doze notas iniciais de cada grupo, alteremos agora apenas o menor valor do Grupo A, a nota 8 para 39 (nota mínima, de qualquer modo inferior à nota mínima do Grupo B).

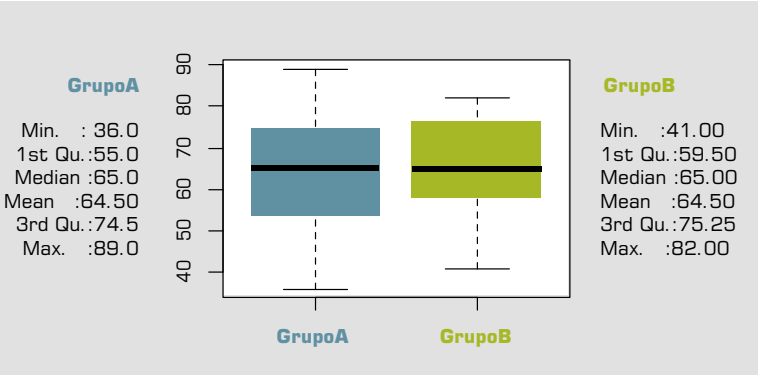


A alteração do valor extremo teve como consequência uma subida significativa da média, mantendo-se, o valor da mediana. Esta situação ilustra bem a maior resistência da mediana a valores extremos relativamente à média.

Apesar da importância destas duas medidas de tendência central, poderemos ter um conjunto de dados diferentes com igual média e mediana, sendo necessário recorrer a outras medidas estatísticas para analisar melhor os dados.



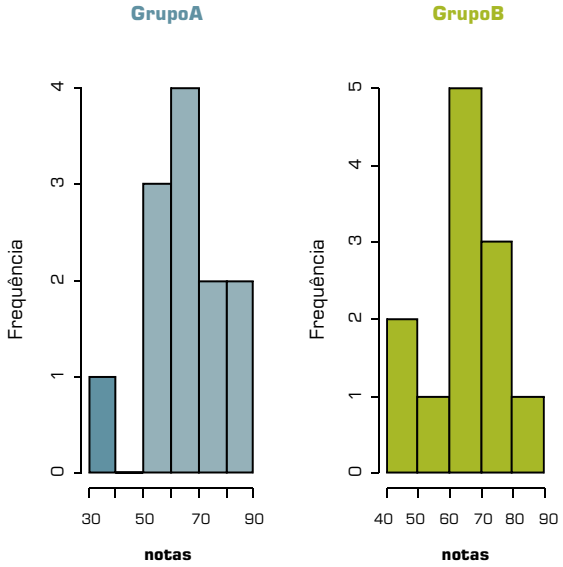
Ainda na situação apresentada, se alterarmos no Grupo A, por exemplo, duas notas: 8 para 36 e 63 para 65, obtemos:



A média e a mediana são iguais, sendo por isso necessário recorrer as outras medidas, por exemplo, de dispersão para analisarmos melhor os dados e concluir, eventualmente qual dos grupos tem melhores resultados.

No Grupo A a amplitude interquartil é superior, bem como o desvio padrão o que significa que neste grupo existe uma maior variabilidade das notas em relação à média.

Os histogramas destes conjuntos de dados apresentam-se a seguir:



Script "Resultados de um teste"

```
notas=data.frame(GrupoA=c(8,51,52,56,61,63,65,67,74,76,
82,89),GrupoB=c(41,43,55,61,62,63,67,68,74,79,79,82))
summary(notas)
par(mfrow=c(2,2))
color=c("red","blue")
boxplot(notas,col=color)

notas2=data.frame(GrupoA=c(51,52,56,61,63,65,67,74,76,
82,89),GrupoB=c(43,55,61,62,63,67,68,74,79,79,82))
summary(notas2)
boxplot(notas2,col=color)

notas3=data.frame(GrupoA=c(39,51,52,56,61,63,65,67,74,
76,82,89),GrupoB=c(41,43,55,61,62,63,67,68,74,79,79,82))
summary(notas3)
boxplot(notas3,col=color)

notas4=data.frame(GrupoA=c(36,51,52,56,61,65,65,67,74,
76,82,89),GrupoB=c(41,43,55,61,62,63,67,68,74,79,79,82))
summary(notas4)
boxplot(notas4,col=color)
sd(notas4$GrupoA)
sd(notas4$GrupoB)
# histogramas do problema Resultados de um teste
par(mfrow=c(1,2))
color=c("red")
hist(notas4$GrupoA,main="GrupoA",xlab="notas",ylab=
"frequência",col.main=color)
color=c("blue")
hist(notas4$GrupoB,main="GrupoB",xlab="notas",ylab=
"frequência",col.main=color)
```

9. Para saber mais: recursos práticos para aprendizagem do R

Publicações

• ALEA, Dossiê X – "Software Estatístico - Uma introdução a alguns aplicativos, numa abordagem inicial dos dados", Helder Alves, Luís Cunha.

• Figueiredo, F., Figueiredo, A., Ramos, A., e Teles, P., Estatística Descritiva e Probabilidades: Problemas Resolvidos e Propostos com Aplicações em R, Escola Editora, 2007.

• Ponte, João Pedro da, Introdução, in Seymour Papert, "A Família em rede", Relógio d'Água, 1997.

• ALEA, Dossiê X – "Software Estatístico - Uma introdução a alguns aplicativos, numa abordagem inicial dos dados", Helder Alves, Luís Cunha.

• L. Torgo (2009), A Linguagem R – Programação para a Análise de Dados, Escola Editora.

• Paul Murrell (2006), R Graphics, Chapman & Hall/CRC, London.

• Peter Dalgard (2002), Introductory Statistics with R, Springer, New York.

WebSites:

• The R Project for Statistical Computing: http://www.r-project.org/index.html

• R Site Search: http://finzi.psych.upenn.edu/search.html

• R mailing lists archive: http://tolstoy.newcastle.edu.au/R/

• The R Commander – A Basic-Statistics GUI for R: http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/

• Tinn-R: http://www.sciviews.org/Tinn-R/