Estatística

1º Semestre 2016/2017

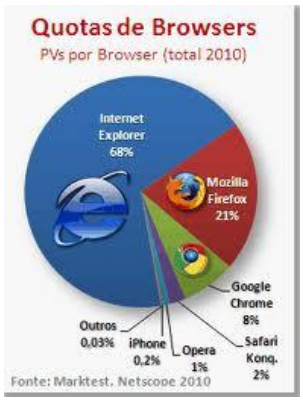
Ana Lúcia Maroco
Ana Meireles
Cláudia Silvestre
Paula Lousão

Estatística

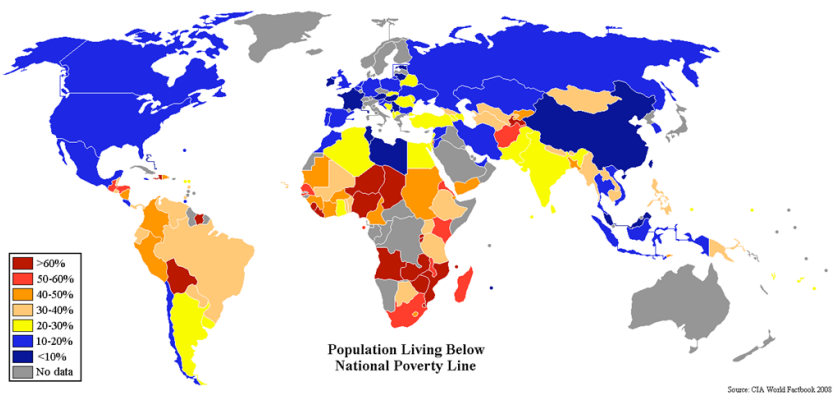
A Estatística no quotidiano

- Quem vai ganhar as eleições?
- Como vai estar o tempo amanhã?
- Estudo do comportamento do mercado de valores de uma bolsa.
- Qual o valor do seguro de vida a pagar?
- Hoje chego a horas à ESCS?
- As estatísticas de um jogo de futebol.

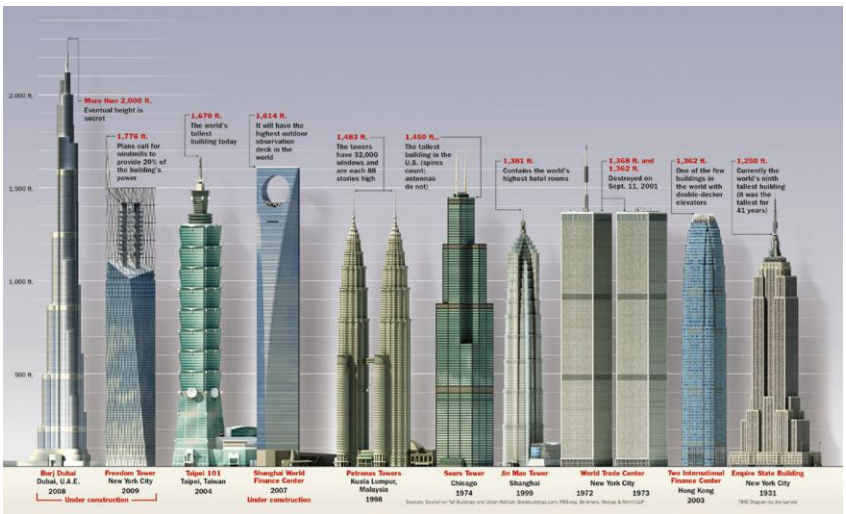
Estatística



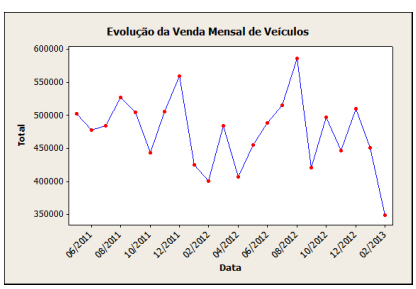
Estatística

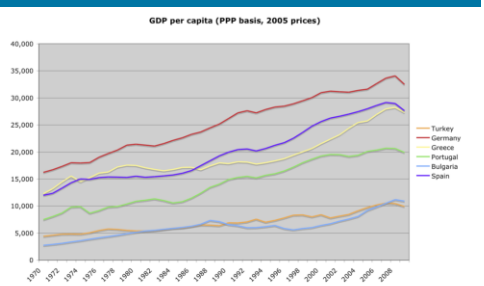


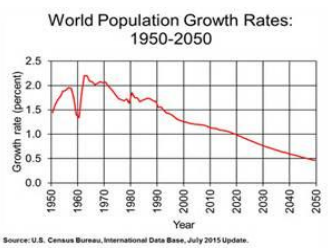
Estatística



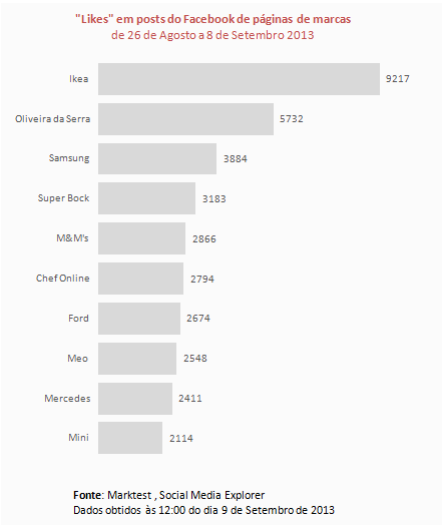
Estatística

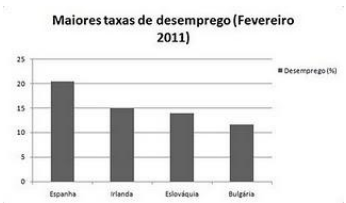


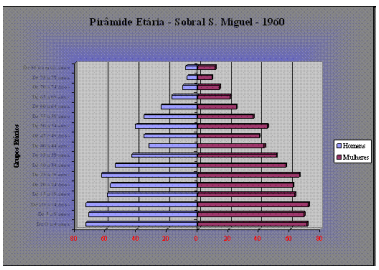




Estatística







Estatística - definição

Estatística - Ramo das matemáticas aplicadas cujos princípios derivam da teoria das probabilidades, que tem por objecto o agrupamento metódico assim como o estudo de séries de factos ou de dados numéricos.

"estatística", in Dicionário Priberam da Língua

Estatística - ciência que se ocupa da recolha e tratamento de informação tendo a capacidade de sintetizar, prever e fazer inferências sobre dados.

Estatística

Censos: a forma mais antiga e direta de conhecer o número de pessoas que, em dado momento, habitam um determinado território, consiste em realizar, literalmente, uma contagem, através duma inquirição exaustiva (habitualmente denominada recenseamento, ou censo) dos indivíduos.

INE – Instituto Nacional de Estatística

Estatística

Pobreza em Portugal – O risco nas famílias (2013)

Publico online, publicado a 30-01-2015
www.publico.pt/sociedade/noticia/portugal-voltou-aos-niveis-de-pobreza-de-ha-dez-anos-1684583

Estatística - Erros

"O problema do mau uso da *matemática* pelos profissionais de comunicação foi identificado há já várias décadas, em especial nos EUA. Apesar disso e da crescente importância que a informação matemática assume na sociedade atual, só atualmente o problema se encontra em estudo no que se refere à imprensa portuguesa."

Susana Simões Pereira, José Manuel Pereira Azevedo, António José de Oliveira Machiavelo

	Jornais diários			
	JN	CM	Público	Total
Notícias com erros (%)	17,9	45,4	35,3	33,4

Tabela 2 - Distribuição das notícias com erros pelos jornais

Foi selecionada uma amostra do conjunto das edições impressas publicadas entre 1 de janeiro de 2013 e 31 de março de 2013.

Estatística - Erros

	Jornais diários			
	JN	CM	Público	Total
Notícias com erros estatísticos*	15 (38,5%)	85 (78,7%)	37 (36,6%)	137 (55,2%)
Notícias com erros gráficos	11 (28,2%)	28 (26,2%)	28 (27,7%)	67 (27,1%)
Notícias com erros numéricos	15(38,5%)	19 (17,8%)	42(41,6%)	76 (30,8%)
Notícias com erros lógicos	0	1(0,9%)	1(1%)	2 (0,8%)

Tabela 4 - Distribuição dos erros quanto ao conteúdo matemático, nas várias notícias de jornais. (*Percentagem calculada sobre o total das notícias com erros no respetivo jornal)

O estudo completo pode ser consultado em:
http://www.lasics.uminho.pt/ojs/index.php/cecs_ebooks/article/view/2261/2178

Estatística - Erros

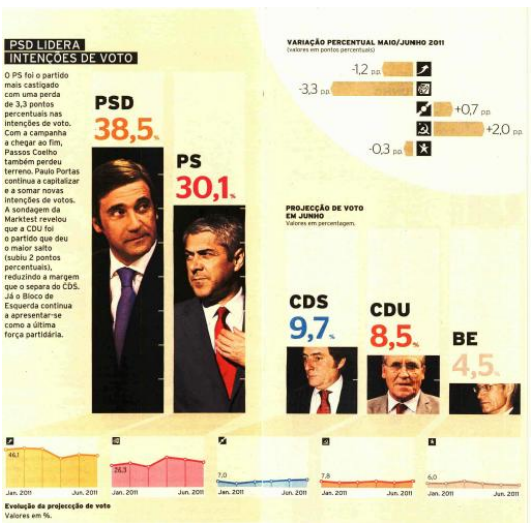
Maria de Lurdes Rodrigues aponta dois estudos recentemente noticiados em Portugal, efectuados com base em dados que o INE fornece ao Eurostat, que apontam - a seu ver "erradamente" - para um decréscimo do emprego qualificado em Portugal na década de 90.

A presidente do OCT, que é também membro do Conselho Superior de Estatística, explicou ao PÚBLICO que o INE efectuou uma renovação da sua amostra a partir de 1998, reconhecendo que a anterior, que tinha como base os Censos de 1991, "encontrava-se envelhecida". Esse facto provocou uma quebra nas séries estatísticas, não sendo por isso legítimo comparar os dados apurados antes de 1997 e depois de 1998, pois resultam de amostras distintas.

Publicado a 23-06-2002

www.publico.pt/noticias/jornal/ha-gato-nas-estatisticas-do-emprego-qualificado-74755

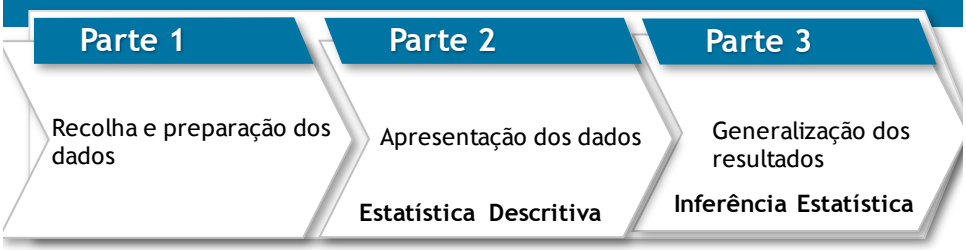
Estatística - Erros



O trabalho de campo foi realizado entre sábado e terça-feira e já incorpora todas as polémicas que, nos últimos dias, têm marcado a campanha eleitoral. Com uma amostra maior (1.208 entrevistas) face ao barómetro habitual (800 entrevistas), os dados da Marktest apontam ainda para um número elevado de "Não sabe/Não responde" (32,4%) e um número reduzido de "Não voto" assumido (4,6%). Os restantes partidos e os votos brancos chegam aos 5,5%. ■

Diário Económico
2 de Junho de 2011

Programa



- Amostragem
- Definição das Variáveis
- Construção da base de dados

- Tabelas de Frequências
- Tabelas de Contingência
- Representações gráficas
- Medidas Descritivas

- Testes Estatísticos:
  - Testes de associação
  - Testes de correlação
  - Testes para a comparação de 2 ou mais grupos
- Intervalos de confiança

Objetivos da Unidade Curricular

- ✓ Conhecer os métodos e as técnicas utilizadas na recolha de dados quantitativos.
- ✓ Ler os resultados de uma sondagem.
- ✓ Utilizar a aplicação SPSS (*Statistical Package for Social Sciences*)
- ✓ Identificar qual o gráfico, tabela ou medida estatística mais adequada aos diferentes conjuntos de dados, bem como construí-los através do SPSS.
- ✓ Fazer uma leitura correta dos diferentes tipos de gráficos, tabelas e indicadores estatísticos.
- ✓ Identificar o teste estatístico mais adequado em cada situação, executá-lo no SPSS e interpretar os seus resultados.
- ✓ Fazer estimação de indicadores sobre uma população.

Avaliação

	Ponderação
AVALIAÇÃO PERIÓDICA	
Avaliação Individual Nota mínima no teste de 8.5 valores.	**60%**
Avaliação em Grupo	**40%**
AVALIAÇÃO POR EXAME	**100%**

Para obter aprovação é necessário que a Média Final ponderada seja superior ou igual a 9.5.

Bibliografia

- Hill, Manuela Magalhães; Hill, Andrew (2000) Investigação por Questionário. Edições Sílabo.
- Maroco, João (2011) Análise estatística com o SPSS Statistics. ReportNumber
- Martinez, Luís Fructuoso; Ferreira, Aristides Isidoro (2008). Análise de dados com SPSS, primeiros passos. Escolar Editora
- Martins, Carla (2011) Manual de Análise de Dados Quantitativos com recurso ao IBM SPSS. Ed. Sílabo
- Murteira, Bento (1993) Análise Exploratória de Dados, Estatística Descritiva. McGraw-Hill.
- Pestana, Maria Helena e Gageiro, João Nunes (2005) Análise de Dados para Ciências Sociais – A Complementaridade do SPSS. Edições Sílabo.
- Reis, Elizabete; Melo, Paulo; Andrade, Rosa e Calapez, Teresa; (1996) Estatística Aplicada. Edições Sílabo.

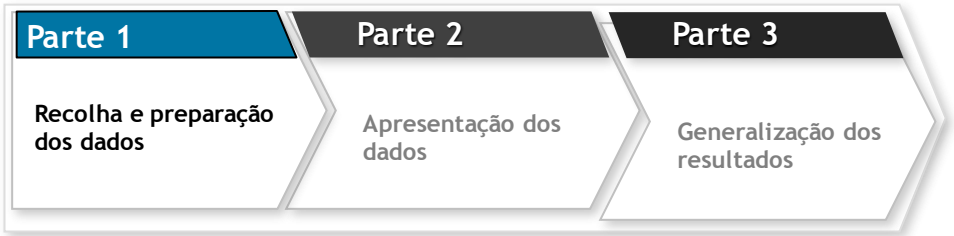
Bibliografia

- Pereira, Alexandre (2008). SPSS Guia Prático de Utilização em Análise de Dados para Ciências Sociais e Psicologia. Sílabo
- Sapsford, Roger ; Jupp, Victor et al (1998) Data Collection and Analysis. Sage Publications Ltd.
- Silva, Cecília Moura da (1994) Estatística Aplicada à Psicologia e Ciências Sociais. McGraw-Hill.
- Vicente, Paula; Reis, Elizabete e Ferrão, Fátima (1998) Sondagens. Edições Sílabo.
- Quantitative Applications in the Social Sciences, Sage Publications:
- Jacoby, William G., Statistical Graphics for Univariate and Bivariate Data, nº 117
- Jacoby, William G., Statistical Grapics for Visualizing Multivariate Data, nº 120

Estatística

Amostragem e Estudos de Opinião

Tipos de Estudos



recolha de informação já existente
- dados secundários → • Estudos de Gabinete ou Estudos Documentais

recolha de nova informação
- dados primários → • Estudos de Campo

Estudos Documentais

Exemplo: dados secundários

destaque
informação à comunicação social

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

Quadro 1 – População residente e presente, famílias, alojamentos e edifícios por NUTS II

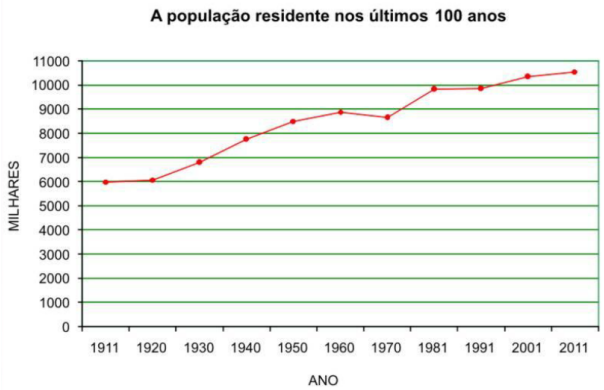
	Censos 2011 (Dados preliminares)						
	População				Famílias	Alojamentos	Edifícios
	Residente		Presente				
	HM	H	HM	H			
Portugal	10555853	5052240	10476291	4980003	4079577	5879845	3550823
Norte	3689713	1769482	3641412	1728877	1341445	1849181	1210720
Centro	2327026	1112257	2301447	1090373	914716	1450268	1113420
Lisboa	2815851	1334637	2783318	1312975	1154904	1486927	450574
Alentejo	758739	367720	749766	361931	306207	472831	384791
Algarve	450484	220183	475220	232885	186456	381026	200481
Açores	246102	121299	245629	121184	82703	110038	98850
Madeira	267938	126662	279499	131778	93146	129574	91987

Estudos Documentais

Exemplo: dados secundários

destaque
informação à comunicação social

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL



Estudos de Campo

Dados Primários

- Acesso a toda a população – Censo ou Recenseamento

XV recenseamento geral da população
CENSOS 2011
V recenseamento geral da habitação

- Acesso a parte da população – Estudo por amostragem

Amostragem e Estudos de Opinião

Exemplos de Estudos por amostragem

- O mercado potencial para um novo produto ou serviço
- Avaliação de produtos ou serviços
- Atitudes dos empregados
- Níveis de satisfação dos clientes
- Opiniões sobre governantes ou políticos
- Intenção de voto

Amostragem e Estudos de Opinião

Lei n.º 10/2000

Regime jurídico da publicação ou difusão de sondagens e inquéritos de opinião

Regula a realização e a publicação ou difusão pública de sondagens e inquéritos de opinião cujo objeto se relacione, direta ou indiretamente, com:

- órgãos constitucionais;
- referendos nacionais;
- regionais ou locais;
- associações políticas ou partidos políticos;
- previsões ou simulações de voto.

Amostragem e Estudos de Opinião

Lei n.º 10/2000
Regime jurídico da publicação ou difusão de sondagens e inquéritos de opinião

Artigo 2.º

Definições

Para os efeitos da presente lei, entende-se por:

a) **Inquérito de opinião**, a notação dos fenómenos relacionados com o disposto no artigo anterior, através de um mero processo de recolha de informação junto de todo ou de parte do universo estatístico;

b) **Sondagem de opinião**, a notação dos fenómenos relacionados com o disposto no artigo anterior, cujo estudo se efetua através do método estatístico quando o número de casos observados não integra todo o universo estatístico, representando apenas uma amostra.

Amostragem e Estudos de Opinião

1. **Métodos de recolha da informação**
   - Entrevista pessoal direta
   - Entrevista telefónica
   - Inquérito postal
   - Inquérito por correio eletrónico ou online

2. **Método de amostragem**
   - Amostragem aleatória
   - Amostragem não aleatória

Métodos de recolha da informação

Entrevista pessoal direta

Vantagens

- Possibilidade de deixar ver, sentir e/ou saborear um produto;
- Possibilidade de encontrar a população-alvo mais facilmente;
- Entrevistas mais longas são, por vezes, toleradas.

Desvantagens

- Geralmente custam mais por entrevista que outros métodos;
- Cada local tem as suas próprias caraterísticas, podendo criar uma amostra não representativa.

Métodos de recolha da informação

Entrevista telefónica

Vantagens

- Contato mais rápido do que com outros métodos (sobretudo com sistema CATI);
- Facilidade de obter contatos aleatórios;
- *Software* CATI permite questionários complexos, por exemplo modificar perguntas com base nas respostas às questões anteriores.

Desvantagens

- *Telemarketing*, a prática massiva de falsas pesquisas levou a uma taxa de recusa mais elevada;
- O horário restrito de disponibilidade da maioria da população ativa;
- Impossibilidade de mostrar produtos por telefone.

Métodos de recolha da informação

Inquérito postal

Vantagens

- Pesquisas por correio estão entre os menos caros;
- O questionário pode incluir fotos - algo que não é possível através do telefone;
- Permitir que o entrevistado responda quando lhe for conveniente.

Desvantagens

- Tempo! Pesquisas por correio demoram muito mais do que outros tipos;
- Em populações de menor escolaridade e alfabetização, taxas de resposta a inquéritos de correio são muitas vezes demasiado pequenas para serem úteis.

Métodos de recolha da informação

Inquérito por correio eletrónico

Vantagens

- A eliminação virtual dos custos de edição;
- Respostas mais precisas às questões sensíveis;
- Velocidade. Podem obter-se vários milhares de respostas em pouco tempo;
- Quase sem custos envolvidos, uma vez que a criação foi concluída;
- Podem anexar-se fotos e arquivos de som.

Métodos de recolha da informação

Inquérito por correio eletrónico

Desvantagens

- Necessidade de uma lista de endereços de *email*;
- Possibilidade de responderem várias vezes ou passar questionários junto aos amigos para responder, caso não haja mecanismos de controle;
- Muitas pessoas não gostam de *email* não solicitados e também podem ser filtrados como SPAM;
- Impossibilidade de generalizar resultados de pesquisas de *email* para as populações inteiras.

Métodos de recolha da informação

Resumo dos métodos de recolha

Velocidade	*Email* e inquéritos online são os métodos mais rápidos, bem como por telefone. Inquéritos postais são os mais lentos
Custo	Entrevista pessoal direta e por telefone são os mais dispendiosos, no polo oposto estão o *email* e os inquéritos online
Utilização de Internet	*Email* e inquéritos online têm evidente vantagem, mas os seus resultados não podem ser generalizados
Habilitações literárias	Pessoas com menos formação raramente respondem a inquéritos postais ou pela internet
Questões Sensíveis	Maior probabilidade de responder a questões sensíveis em inquéritos online
Imagens, sons, sabores	Não possível em inquéritos telefónicos e com limitações nos postais

Métodos de Amostragem

Cuidados a ter na escolha da amostra

Em relação às amostras, deve assegurar-se a sua representatividade relativamente à população de onde foram retiradas.

O objetivo é que os resultados obtidos possam ser próximos dos da população.

Métodos de Amostragem

Tipos de erros:

- Erro Amostral = | estatística – parâmetro |
- Erro de Enviesamento
- Erro de Medida ou Sistemático

Amostragem e Estudos de Opinião

Exemplos de Erro de Enviesamento

Amostra	Provável enviesamento	Razão
Clientes	Favorável	Não seriam clientes se estivessem insatisfeitos, contudo pode ser importante saber porque estão satisfeitos.
Ex-clientes	Desfavorável	Se estivessem satisfeitos não seriam ex-clientes, contudo pode ser importante saber porque ficaram insatisfeitos.
Auto-seleção (Televoto)	Pontos de vista extremos	São sobretudo as pessoas com uma opinião forte sobre determinado assunto que se mobilizam para participar, muitas vezes mais que uma vez.
Horário de trabalho	Reformados, desempregados	A maioria das pessoas que se encontra em casa durante o horário normal de trabalho não está empregada, pelo que o estudo não reflete a opinião da população ativa.
Internet	Cidadão atípico	Limitado a pessoas com acesso a internet, apesar da crescente utilização desta, estas pessoas não são representativas da população em geral, por exemplo em termos de idade, classe social, instrução, etc.

Amostragem e Estudos de Opinião

Sondagem do Literary Digest relativa às eleições presidenciais norte americanas de 1936.

-**Dimensão da amostra:**
-cerca de 2,4 milhões de eleitores

Alf Landon (57%)
Franklin Roosevelt (43%)

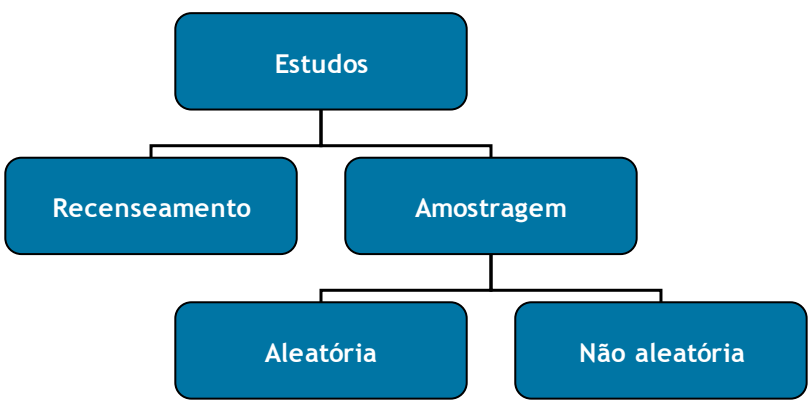
Resultado da eleição:
Alf Landon (40%)
Franklin Roosevelt (60%)

Enviesamento de amostragem: a base de sondagem foi a lista de telefone, de membros de clubes, assinantes da revista, registos de propriedade automóvel., etc.

Enviesamento de não resposta: a proporção de não respondentes era maior entre os democratas.

Amostragem

Estudos Quantitativos



Processos de Seleção

Aleatórios – (probabilísticos)

- Cada elemento da população tem uma probabilidade conhecida de pertencer à amostra;
- É possível determinar a precisão das estimativas;
- Não há interferência nem do entrevistador nem do investigador.

Não aleatórios – (não probabilísticos)

- Não se conhece a probabilidade dos elementos integrarem a amostra nem a precisão das estimativas;
- A seleção dos elementos pode basear-se nos critérios do entrevistador ou do investigador.

Processos Aleatórios

Vantagens

- Cada elemento da população tem uma certa probabilidade (calculável e diferente de zero) de pertencer à amostra;
- É possível determinar a precisão das estimativas;
- É possível determinar matematicamente a dimensão da amostra a recolher em função da precisão e do grau de confiança associado ao estudo;
- Não há interferência do entrevistador nem do investigador (não enviesamento).

Processos Aleatórios

Desvantagens

- Por vezes é necessário a obtenção de uma listagem completa e atualizada de todos os elementos que constituem a população;
  (muitas vezes impossível porque as populações podem ser infinitas, …)
- Custo e tempo necessários para realizar o estudo é muito elevado;
- Taxa de "não respostas" pode ser muito elevada.

Processos Não aleatórios

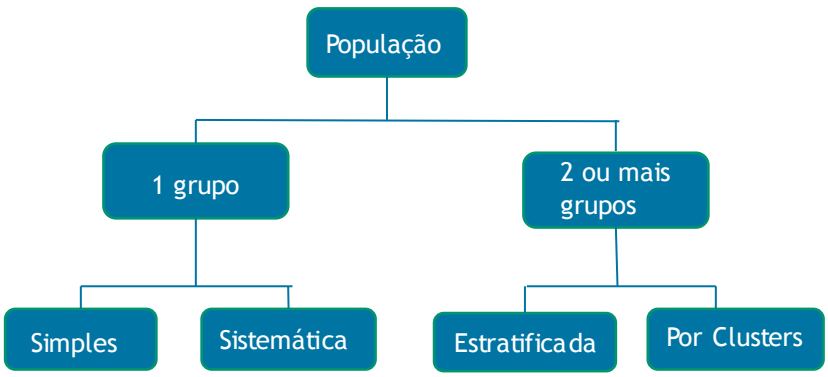
Vantagens

- Não necessita base de referência
- Acessibilidade dos entrevistados
- Tempo
- Custo

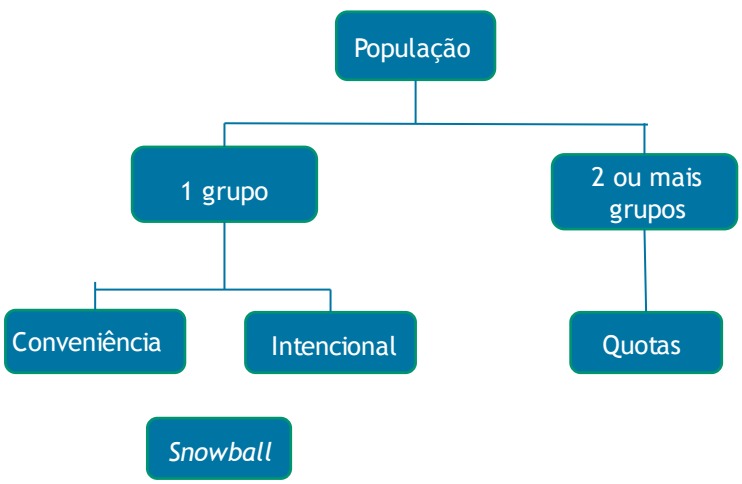
Desvantagens

- Normalmente geram amostras não representativas da população, o que pode causar enviesamentos e interpretações erradas.

Processos aleatórios



Processos não aleatórios

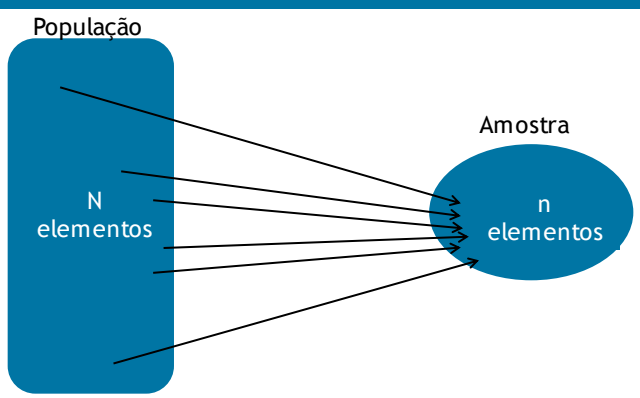


Amostragem Aleatória

Amostragem Aleatória

- Simples
- Sistemática
  - Rotas Aleatórias
- Estratificada
- Por Clusters
  - Multi-etapas

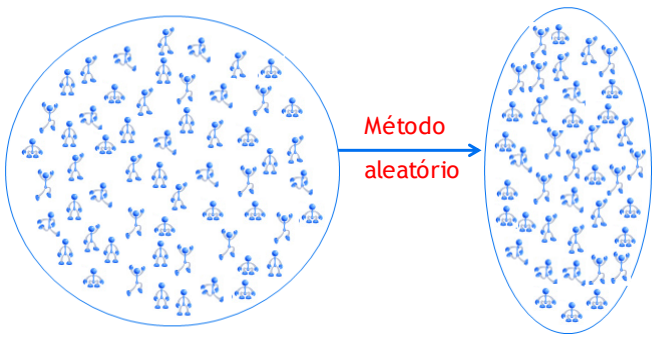
Amostra Aleatória Simples



- Qualquer elemento do universo tem igual probabilidade de ser selecionado (p=n/N).
- Cada uma das combinações possíveis de n elementos dos N tem a mesma probabilidade de ser escolhida.

Amostra Aleatória Simples

- Numerar os elementos da população de 1 a N;
- Escolher n elementos usando um processo aleatório (lotaria) ou recorrendo a uma tabela de números aleatórios;
- Os elementos que constituem a amostra são os que correspondem aos números escolhidos.



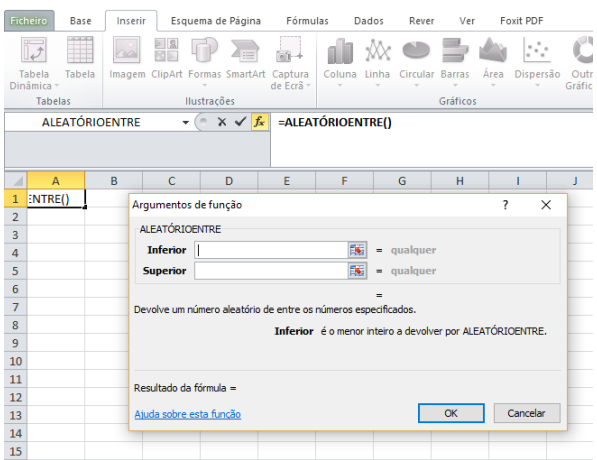
Amostra Aleatória Simples

- Processo moroso e caro, se a amostra for grande;
- Muitas vezes impraticável por exigir a enumeração de todos os elementos da população;
- No caso de uma população pequena pode ser útil e fácil de aplicar se a base de sondagem for credível;
- A seleção dos elementos pode fazer-se através de tabelas de números aleatórios, método da lotaria, etc.

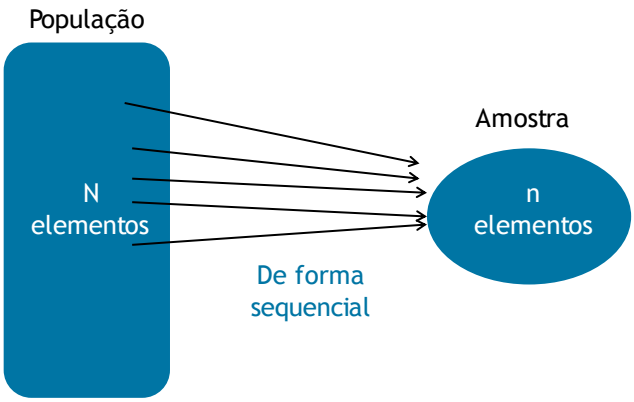
Amostra Aleatória Simples

7	0	0	3	0	0	1	2	9	6	3	9	0	3	5	5	4	0	7	2	2	7	5	9	8	7
4	1	7	9	8	1	3	7	4	1	0	2	7	7	3	0	2	7	3	0	4	6	9	5	3	6
8	4	0	0	6	7	3	7	8	1	0	8	3	0	4	2	4	3	7	1	9	5	6	6	5	6
2	0	1	0	8	3	2	0	2	5	7	4	4	1	5	7	8	4	3	9	9	8	6	2	4	8
9	7	1	5	7	5	5	9	5	6	1	5	4	8	6	6	2	7	3	2	7	8	2	3	6	9
0	0	9	6	7	7	7	7	3	0	0	8	7	3	0	3	0	2	7	3	5	6	3	9	5	4
2	6	2	7	5	7	8	0	1	0	2	6	6	0	6	3	3	5	5	2	5	8	9	5	1	4
5	3	0	2	4	5	6	6	4	5	5	9	7	9	8	8	1	4	3	9	2	3	7	5	3	1
8	5	2	3	9	4	2	2	5	5	9	7	8	3	6	1	2	8	5	9	6	4	6	9	5	8
4	0	6	2	4	7	2	1	9	2	5	0	3	2	9	0	1	7	4	8	0	9	9	6	3	1
0	7	1	6	0	9	0	7	5	7	2	1	8	5	2	1	6	7	8	6	5	6	1	6	7	6
4	2	7	0	1	7	0	9	3	4	3	4	9	1	2	3	3	0	7	4	3	4	3	2	1	5
1	9	5	0	7	6	3	3	8	1	6	7	9	2	2	1	5	1	7	7	2	7	3	5	5	1
2	1	6	6	3	8	8	5	9	0	4	0	3	2	6	8	4	1	3	7	5	1	5	0	2	0
0	4	0	8	0	9	7	0	6	3	3	2	7	5	5	7	9	2	9	6	1	8	4	5	9	3

Amostra Aleatória Simples



Amostra Aleatória Sistemática



Amostra Aleatória Sistemática

1. Calcular o intervalo da amostra: $K = int\left(\frac{N}{n}\right)$
2. Escolher aleatoriamente um número J entre 1 e K;
3. Partindo desse número, adicionar sucessivamente o valor K, ficando, assim, selecionados os elementos J, $J+K$, $J+2K$, $J+3K$, ..., $J+(n-1)K$, perfazendo n elementos.

➲ Processo semelhante ao da amostragem aleatória simples.

Amostra Aleatória Sistemática

Exemplo:

- $N = 200$; $n = 50$ => $K = 4$
- Escolher um número aleatório x : $1 < x < 4$. Por exemplo, **3**
- Os indivíduos selecionados serão:

3º, 7º, 11º, 15º, ..., 195º, 199º

A seleção de um elemento depende do que foi anteriormente selecionado.

Amostra Aleatória Sistemática

Problema:

- Vendas mensais de determinada empresa;
- Calcular o valor de K;
- Quando os elementos têm comportamentos cíclicos tem que se ter atenção ao valor de **K**. Por exemplo, ao selecionar aleatoriamente o número correspondente a um determinado mês, escolher-se-ia sempre o mesmo mês todos os anos => **Enviesamento dos resultados obtidos**.

Amostra: Rotas Aleatórias

- ✓ Caso Particular da Amostragem Aleatória Sistemática;
- ✓ Método útil em sondagens realizadas em localidades. É um meio de orientação para o entrevistador;
- ✓ Base de referência cartográfica.

1. Selecionar aleatoriamente um ponto de partida – **Ponto de amostragem**;
2. Definir critérios de escolha dos elementos – **Definir as regras de orientação para o entrevistador**;
3. Reunir os elementos selecionados para constituírem a amostra.

Amostra: Rotas Aleatórias

A partir de uma determinada rua, virar na 1ª à direita e depois na 2ª rua à esquerda.
Quando se vira à direita entrevistam-se os moradores das casas ímpares; quando se vira à esquerda entrevistam-se os moradores das casas pares.

Amostra: Rotas Aleatórias

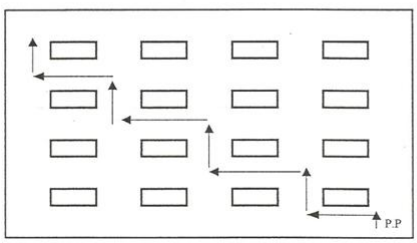
Exemplo:

Seleção de edifícios

O entrevistador situa-se no ponto de partida indicado na capa de série , onde nunca realizará Entrevista de recenseamento do lar.

De frente para o ponto de partida, caminhará para a sua esquerda e virará a seguir na primeira rua à esquerda, depois na primeira rua à direita, depois à esquerda e assim sucessivamente .

Neste percurso deverá fazer entrevista de recenseamento em todos os prédios onde o número de polícia termine no valor de X correspondente a essa série. Se os prédios não tiverem número de polícia considerará um prédio de X em X , do lado direito da rua se X é par ou do lado esquerdo se X é impar..

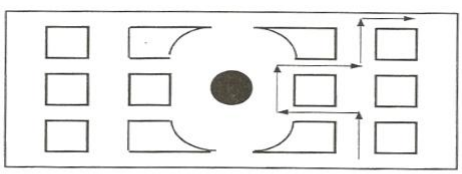


Amostra: Rotas Aleatórias

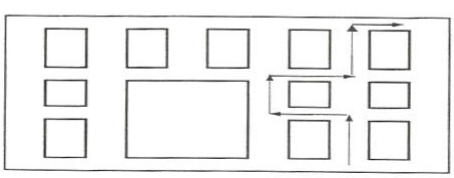
Exemplo:

Seleção de edifícios (cont.)

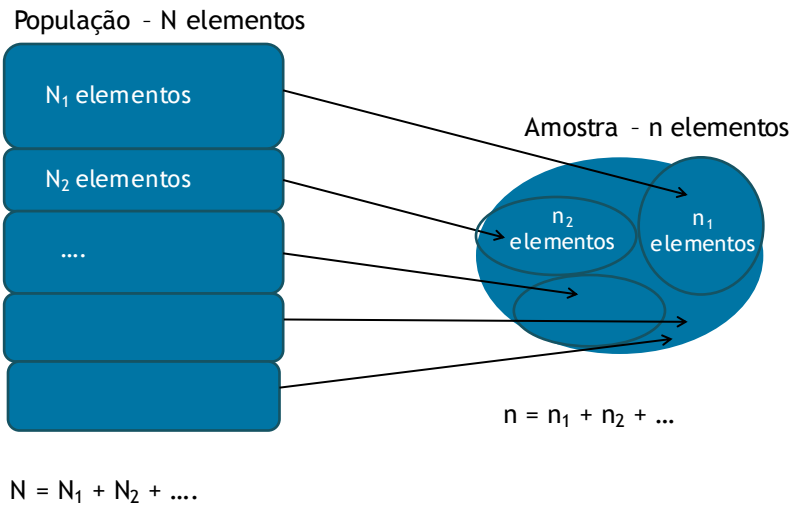
Praceta de configuração circular:



Praceta de configuração rectangular:



Amostragem Aleatória Estratificada



Amostragem Aleatória Estratificada

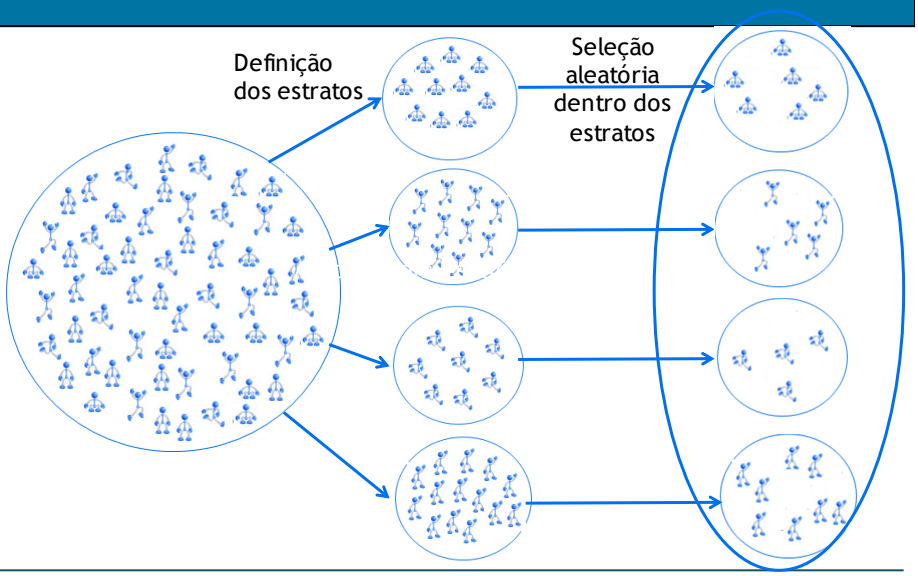
Numa população com ***N*** elementos, identificam-se ***L*** sub-grupos (estratos) com N_i elementos cada. Selecionam-se n_i elementos em cada estrato em vez de recorrer à população como um todo.

Etapas:

1. Definir os estratos;
2. Escolher aleatoriamente os elementos de cada estrato;
3. Reunir os elementos selecionados em cada estrato para constituírem a amostra

Estudo sobre caraterísticas mais importantes num carro.

Amostragem Aleatória Estratificada

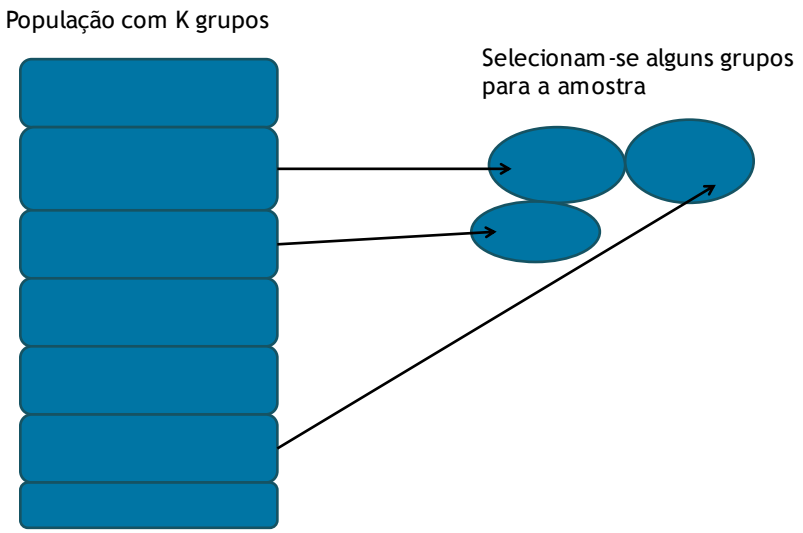


Amostragem Aleatória Estratificada

Caraterísticas:

- Os estratos devem ser exaustivos (cobrir toda a população) e mutuamente exclusivos (não haver elementos que pertençam a mais do que um estrato);
- A divisão da população deve ser feita de acordo com as caraterísticas da população em estudo;
- O ideal é que a variabilidade dentro dos estratos seja pequena e grande entre os estratos;
- A amostragem estratificada pode ser proporcional ou não proporcional.

Amostragem Aleatória por *Clusters*



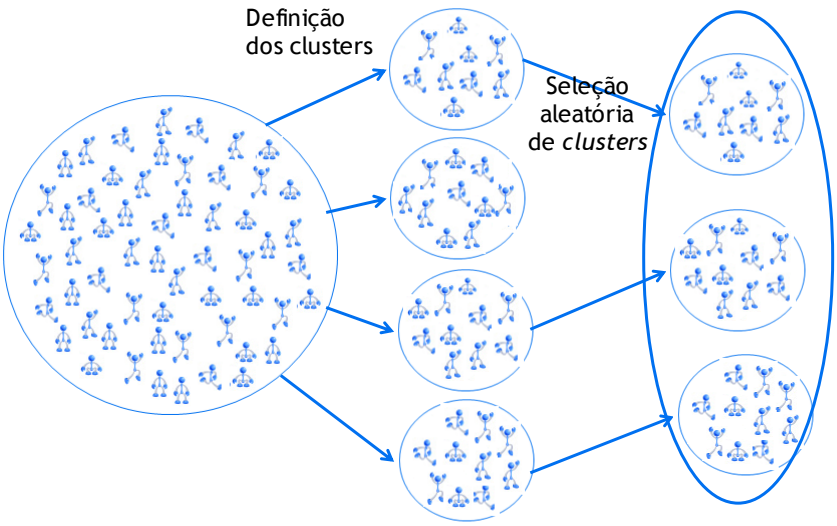
Amostragem Aleatória por *Clusters*

Cluster é um grupo de unidades elementares da população.
Selecionam-se os grupos de elementos e não elementos individuais.

Etapas:

1. Especificar os *Clusters* (Conglomerados);
2. Escolher aleatoriamente um conjunto de *clusters*;
3. Construir a amostra com todos os elementos de cada *cluster* selecionado.

Amostragem Aleatória por *Clusters*



Amostragem Aleatória por *Clusters*

Exemplos:

Aplicação (Exemplo)	*Cluster* ou Unidade Amostral Primária	Unidade Elementar
Conhecer hábitos de consumo de alcoól dos estudantes duma escola secundária	Turma	Aluno
Estimar o tempo médio de espera para atendimento numa consulta	Centro de Saúde	Utentes
Estimar o montante de vendas para o próximo ano	Zona Geográfica de Vendas	Vendedor

Amostragem Aleatória por *Clusters*

Caraterísticas:

- Os *clusters* devem ser exaustivos (cobrir toda a população) e mutuamente exclusivos (não haver elementos que pertençam a mais do que um cluster);
- Cada *cluster* deve refletir a variabilidade da população;
- Custo geralmente mais baixo do que nos casos anteriores.

Amostragem Aleatória Multi-etapas

Extensão do conceito de amostragem por *clusters*.
Selecionam-se aleatoriamente vários *clusters* e, dependendo dos casos, selecionam-se sub-*clusters* até se obterem os elementos individuais.

Etapas:

1. Definir os *Clusters*;
2. Escolher aleatoriamente um conjunto de *clusters*. De acordo com o número de etapas que se considerarem, vão-se definindo e selecionando os novos *clusters* até se obterem as unidades elementares;
3. Construir a amostra com todos os elementos de cada *cluster* selecionado.

Amostragem Aleatória Multi-etapas

Exemplo:

Numa sondagem sobre estudantes de ensino secundário a nível nacional:

1. Selecionar localidades;
2. Selecionar escolas secundárias;
3. Selecionar turmas dentro das escolas secundárias;
4. Entrevistar todos os alunos das turmas selecionadas.

Caraterísticas:

- Método geralmente económico.

Amostragem Não Aleatória

Amostragem Não Aleatória

- Intencional
- *Snowball*
- Por Conveniência
- Por Quotas

Amostragem Intencional

Os elementos são intencionalmente selecionados geralmente por se pensar que têm caraterísticas representativas da população.

Exemplo:

Num estudo sobre o futuro da Televisão pública, escolher uma amostra de especialistas no sector audiovisual.

Caraterísticas:

- Método geralmente usado em estudos exploratórios;
- Obtenção de amostras de dimensão reduzida;
- Impossibilidade de se conseguir uma amostra aleatória;
- Conseguir deliberadamente uma amostra enviesada.

Amostragem *Snowball*

- Usada quando a população alvo é muito pequena.
- Caso particular da amostragem intencional:
- Consiste em ir pedindo aos inquiridos para indicarem novos elementos para a amostra.

Exemplo:

Populações com caraterísticas específicas:

Deficientes, emigrantes, imigrantes, etc.

Amostragem por Conveniência

- Os elementos da amostra são escolhidos porque estão disponíveis na altura ou no local do estudo.
- Apesar de ser uma técnica suscetível de provocar enviesamento nos resultados, é útil se o interesse do estudo for captar ideias ou se se pretender fazer uma exploração prévia sobre algum assunto.

Exemplos:

- Pessoas que passam no local onde o entrevistador faz inquéritos.
- Inquéritos telefónicos feitas por estações de TV ou Rádio em que se convidam as pessoas a dar a sua opinião sobre determinado assunto.

Amostragem por Quotas

- Amostragem estratificada não aleatória;
- Usa os mesmos critérios da estratificação;
- O objetivo da divisão por quotas é garantir a representatividade das caraterísticas da população;
- A proporção de elementos com determinada caraterística na amostra deve ser aproximadamente igual à proporção de elementos com a mesma caraterística na população;
- A população é dividida em subgrupos segundo uma característica de interesse e selecionam-se amostras não aleatórias em cada subgrupo.

Etapas:

1. Definir as quotas;
2. Escolher os elementos respeitando as quotas;
3. Reunir os elementos selecionados para constituírem a amostra.

Amostragem por Quotas

Quotas independentes

Género	
Masculino	48
Feminino	52
Total	100

Idade	
15-24	18
25-34	25
35-44	30
Mais de 44	27
Total	100

Amostragem por Quotas

Quotas interrelacionadas

Idade / Género	15-24	25-34	35-44	Mais de 44	Total
Masculino	10	11	12	15	48
Feminino	8	14	18	12	52
Total	18	25	30	27	100

Processos de Amostragem Mistos

- Processos em várias fases
- Combinação entre processos aleatórios e não aleatórios

Dimensão da amostra

Dimensão da Amostra

Amostras não aleatórias:

- Orçamento disponível
- Dimensão utilizada em estudos anteriores com as mesmas caraterísticas
- Dimensão de amostras aleatórias (indicativo)

Dimensão da Amostra

Amostras aleatórias:

- Dimensão da população
- Variabilidade da caraterística de interesse na população
- Precisão do estudo e nível de confiança
- Custo

Dimensão da Amostra

Amostra aleatória Simples:
Erro máximo na estimação de caraterísticas qualitativas

$$e = 1{,}96 \times \sqrt{\frac{P(1-P)}{n}}$$

Dimensão da Amostra

Amostra aleatória Simples:
Estimação de caraterísticas qualitativas

$$n = \frac{P(1-P)}{\left(\frac{e}{1{,}96}\right)^2 + \frac{P(1-P)}{N}}$$

Erro máximo que se pretende

1,96

Para um nível de confiança de 95%

N infinito

$$n = \frac{P(1-P)}{\left(\frac{e}{1.96}\right)^2}$$

Dimensão da Amostra - Exercício

Para um nível de confiança de 95% e erro de estimação de

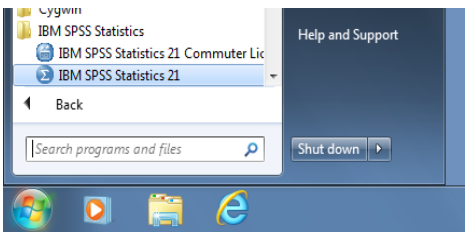
		2%	3%	5%
Partido	Proporção	Amostra	Amostra	Amostra
PSD (38,63%)	0,3863	2277	1012	365
PS (28,05%)	0,2805	1939	861	311
CDS-PP (11,71%)	0,1174	996	442	160
PCP-PEV (7,90%)	0,0790	703	311	112
BE (5,17%)	0,0517	473	210	76
	0,5	2401	1068	385

Introdução ao *software* SPSS

Introdução ao *Software* SPSS

Para iniciar o SPSS clique em

Start
IBM SPSS Statistics
IBM SPSS Statistics 21

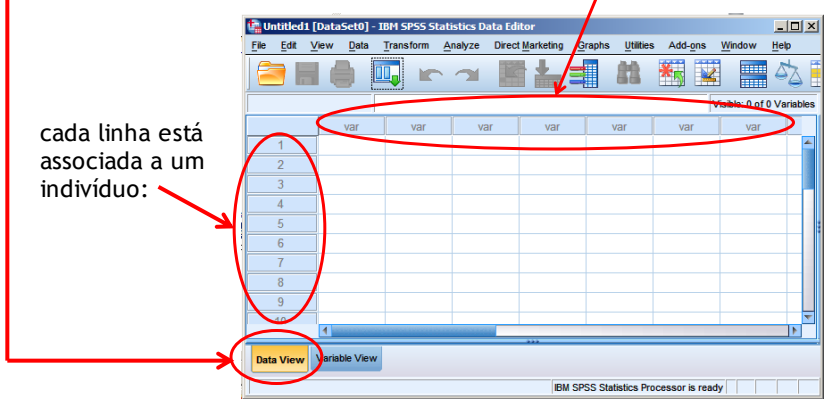


Introdução ao *Software* SPSS
Ambiente de trabalho

Janela de edição de dados:

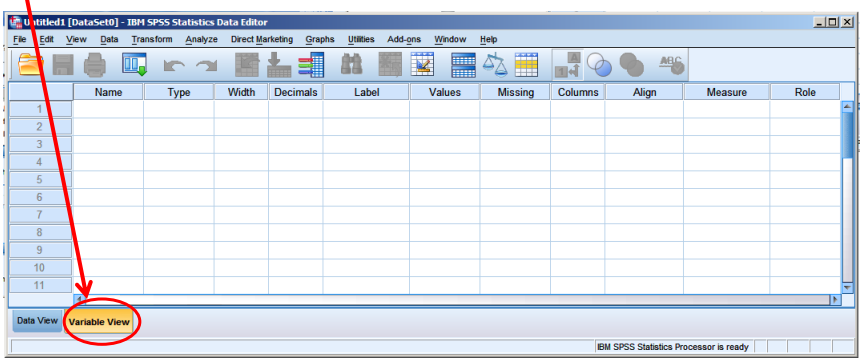
cada coluna representa uma variável

cada linha está associada a um indivíduo:



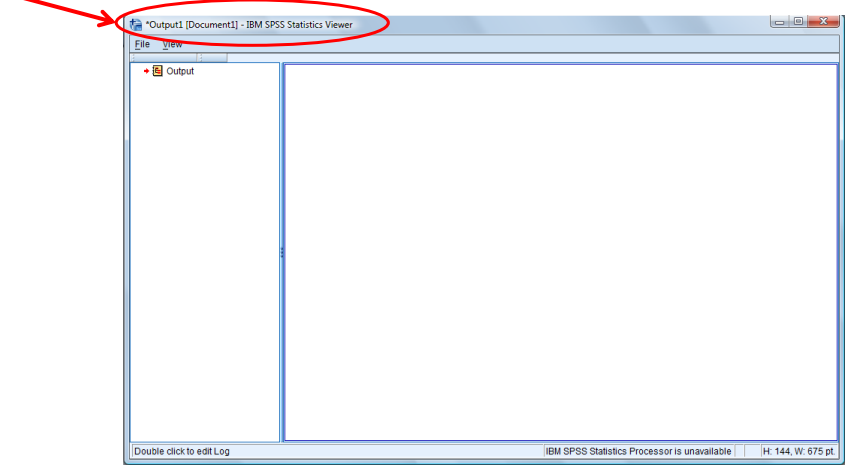
Introdução ao Software SPSS
Ambiente de trabalho

Janela de especificação das variáveis:



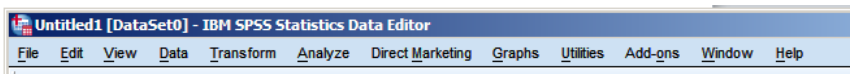
Introdução ao Software SPSS
Ambiente de trabalho

Janela de visualização de resultados:



Introdução ao Software SPSS
Ambiente de trabalho

Barra de Menus:



Barra de Ferramentas:

Ficheiro de Dados - Escalas de Medida

Antes de começar a definir um ficheiro de dados, deve ter em conta a **escala de medida** associada a cada variável.

São diferenciadas pelo tipo de relação que existe entre os objetos

Ficheiro de Dados - Escalas de Medida

Os dados que caraterizam as escalas de medida podem ser do tipo:

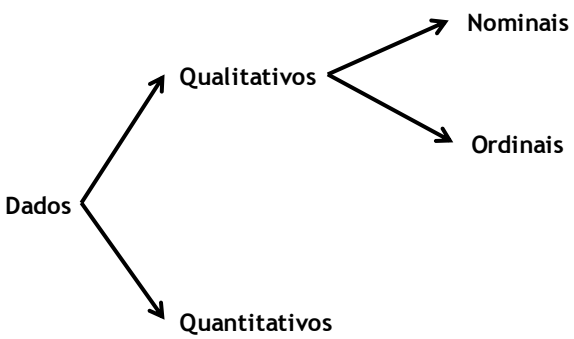
Qualitativo

Representam "**qualidades**" associadas aos indivíduos.
Podem ser representadas por números, mas estes não têm qualquer significado matemático.

Quantitativo

Representam "**quantidades**" associadas aos indivíduos.
São representadas por números, e esses números têm significado matemático (por exemplo, 4 é o dobro de 2).

Ficheiro de Dados - Escalas de Medida



Ficheiro de Dados - Escalas de Medida

Os dados **Qualitativos** podem ser definidos numa escala

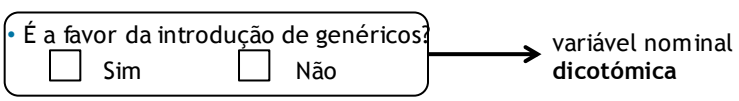
Nominal

- Os dados são identificados pela atribuição de um nome que se designa por **categoria** (classe ou modalidade).
- As categorias são **exaustivas** e **mutuamente exclusivas**.
- Apenas permite caraterizar algum aspeto qualitativo da caraterística que está a ser estudada.
- Não existe qualquer relação de ordem entre as categorias (apenas permite dizer se uma modalidade é igual ou diferente de outra).
- Podem ser representadas por números, mas estes **não têm qualquer significado matemático**.

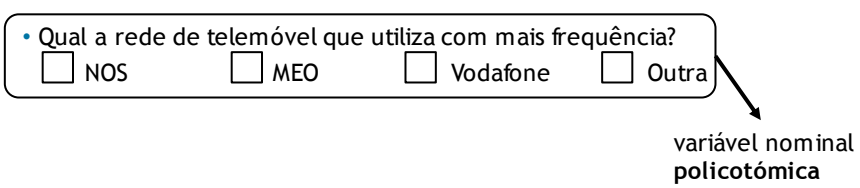
Ficheiro de Dados - Escalas de Medida

Escala Nominal – Exemplos:

- Num inquérito à opinião sobre a utilização de genéricos:



- Num inquérito sobre as diferentes redes de telemóvel usadas:



Ficheiro de Dados - Escalas de Medida

Ordinal

- Apresentam as mesmas caraterísticas que os dados definidos numa escala nominal, mas agora é possível estabelecer uma **relação de ordem** entre as categorias. (melhor/pior, maior/menor, …)

Escala Ordinal – Exemplos:

- É a favor da introdução de genéricos?

Totalmente contra	Contra	Nem Contra nem a favor	A favor	Totalmente a favor
☐	☐	☐	☐	☐

- Está satisfeito/a com os serviços prestados pela sua operadora de telemóvel? (1 – Nada satisfeito, …, 5 – Completamente satisfeito)

1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Ficheiro de Dados - Escalas de Medida

Os dados podem ser definidos numa escala **Quantitativa** quando

- Os valores numéricos associados a esta escala são verdadeiramente **quantitativos**.
- Permitem **quantificar** e **comparar** as diferenças existentes entre as diferentes modalidades, bem como a realização de **operações matemáticas**. (somas, produtos, diferenças, …)
- As diferenças entre dois valores consecutivos são **iguais**.

Escala de medida - Exercício

Para cada uma das bases de dados a seguir apresentadas, identifique a escala de medida associada a cada variável:

Aluno	Género	Grau Escolaridade Enc. Educ.	Nº irmãos	Compe-tência Leitora	Rendi-mento Escolar	Nº médio horas sono
1	Feminino	E. Básico	0	5	103	<7 horas
2	Feminino	E. Básico	1	8	109	7 a 9 horas
3	Feminino	E. Básico	2	10	102	<7 horas
4	Feminino	E. Básico	1	8	109	7 a 9 horas
5	Feminino	E. Básico	1	9	110	>9 horas
6	Masculino	E. Básico	0	15	115	>9 horas
7	Masculino	E. Básico	1	4	106	7 a 9 horas
8	Masculino	E. Básico	2	5	105	<7 horas
9	Feminino	E. Secundário	3	19	114	>9 horas
10	Feminino	E. Secundário	1	15	115	>9 horas
11	Feminino	E. Secundário	0	14	110	>9 horas
12	Feminino	E. Secundário	1	6	108	7 a 9 horas
13	Feminino	E. Secundário	3	18	112	>9 horas
14	Feminino	E. Secundário	0	4	102	<7 horas
15	Masculino	E. Secundário	3	7	106	<7 horas

Aluno	Género	Grau Escolaridade Enc. Educ.	Nº irmãos	Compe-tência Leitora	Rendi-mento Escolar	Nº médio horas sono
16	Masculino	E. Secundário	2	15	109	7 a 9 horas
17	Masculino	E. Secundário	1	10	108	>9 horas
18	Masculino	E. Secundário	0	9	109	>9 horas
19	Masculino	E. Secundário	0	8	108	7 a 9 horas
20	Masculino	E. Secundário	0	11	107	7 a 9 horas
21	Feminino	E. Superior	1	12	109	>9 horas
22	Feminino	E. Superior	1	19	115	>9 horas
23	Feminino	E. Superior	0	14	111	>9 horas
24	Feminino	E. Superior	1	17	115	>9 horas
25	Feminino	E. Superior	2	18	110	>9 horas
26	Feminino	E. Superior	2	12	109	7 a 9 horas
27	Masculino	E. Superior	1	14	108	7 a 9 horas
28	Masculino	E. Superior	1	9	104	<7 horas
29	Masculino	E. Superior	2	5	107	<7 horas
30	Masculino	E. Superior	3	15	102	<7 horas

Competência leitora: definida numa escala de 0-20
Rendimento escolar: Definida numa escala de 0-120

Escala de medida - Exercício

Considere o seguinte conjunto de dados referente à avaliação contínua de 27 estudantes:

LEGENDA:
Genero: Masculino, Feminino
Ano Curso: 1º ano, 2º ano
Presenças: Nº de presenças às aulas (0-12)
TPC: Nº de TPC realizados (0-8)
Teste_1: Classificação no 1º teste (0-20)
Teste_2: Classificação no 2º teste (0-20)

Aluno	Género	Ano Curso	Presenças	T.P.C.	Teste_1	Teste_2
1	Masculino	1	1	0	2,00	0,00
2	Masculino	1	6	4	8,10	5,70
3	Masculino	1	5	0	3,00	9,80
4	Masculino	2	0	0	5,00	3,90
5	Masculino	1	2	0	4,50	12,50
6	Masculino	1	4	4	6,50	11,00
7	Masculino	2	3	0	5,30	8,30
8	Masculino	1	2	2	3,70	5,00
9	Masculino	1	8	0	2,00	3,50
10	Masculino	1	6	2	11,80	11,40
11	Feminino	1	11	8	11,00	19,70
12	Feminino	1	7	2	0,50	3,80
13	Feminino	2	1	0	0,50	5,30
14	Feminino	1	12	8	14,80	15,80
15	Feminino	1	8	2	7,30	12,10
16	Feminino	1	9	8	16,10	18,40
17	Feminino	1	11	0	8,80	12,80
18	Feminino	1	8	4	6,40	17,40
19	Feminino	2	11	6	17,40	14,90
20	Feminino	1	12	8	14,50	15,50
21	Feminino	1	4	0	0,00	0,80
22	Feminino	1	8	6	8,60	7,40
23	Feminino	1	12	8	18,80	20,00
24	Feminino	2	9	0	15,10	12,00
25	Feminino	1	2	0	9,30	3,30
26	Feminino	2	2	4	10,10	11,50
27	Feminino	1	9	2	5,10	10,50

Escala de medida - Exercício

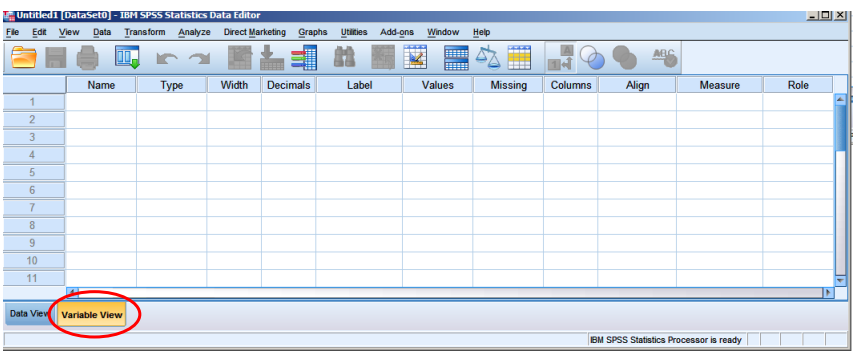
Ind	Género	Idade	N_Filhos	C_Trab	A_Emp	D_Casa
1	Feminino	21	0	Pess	< 3	33
2	Feminino	53	2	Exc	3 a 5	23
3	Feminino	54	2	Med	6 a 8	76
4	Feminino	44	3	Exc	3 a 5	21
5	Feminino	54	1	Med	6 a 8	23
6	Feminino	53	2	Más	> 8	66
7	Feminino	67	1	Med	> 8	41
8	Feminino	28	0	Pess	6 a 8	61
9	Feminino	39	0	Med	< 3	55
10	Feminino	60	1	Boas	3 a 5	75
11	Feminino	50	0	Boas	3 a 5	5
12	Feminino	28	2	Med	3 a 5	35
13	Feminino	52	2	Med	6 a 8	70
14	Feminino	66	0	Más	6 a 8	1
15	Feminino	31	1	Más	< 3	12

Ind	Género	Idade	N_Filhos	C_Trab	A_Emp	D_Casa
16	Masculino	27	1	Boas	< 3	68
17	Masculino	54	0	Pess	> 8	9
18	Masculino	38	1	Pess	6 a 8	30
19	Masculino	59	1	Boas	3 a 5	68
20	Masculino	32	0	Boas	6 a 8	14
21	Masculino	48	0	Med	> 8	27
22	Masculino	42	1	Med	< 3	26
23	Masculino	53	1	Más	3 a 5	15
24	Masculino	55	2	Med	3 a 5	58
25	Masculino	55	1	Med	6 a 8	22
26	Masculino	37	0	Más	< 3	58
27	Masculino	51	2	Más	> 8	19
28	Masculino	36	0	Boas	3 a 5	23
29	Masculino	66	1	Pess	3 a 5	63
30	Masculino	54	3	Pess	6 a 8	19

LEGENDA:
Género:Género do funcionário (Fem, Masc)
Idade: Idade do funcionário (em anos)
N_Filh: Número de Filhos do funcionário
C_Trab: Condições de trabalho (Pess, Más, Médias, Boas, Exc)
A_Emp: Antiguidade na empresa (< 3, 3 a 5, 6 a 8, > 8)
D_Casa: Distância do emprego a casa (em Km)

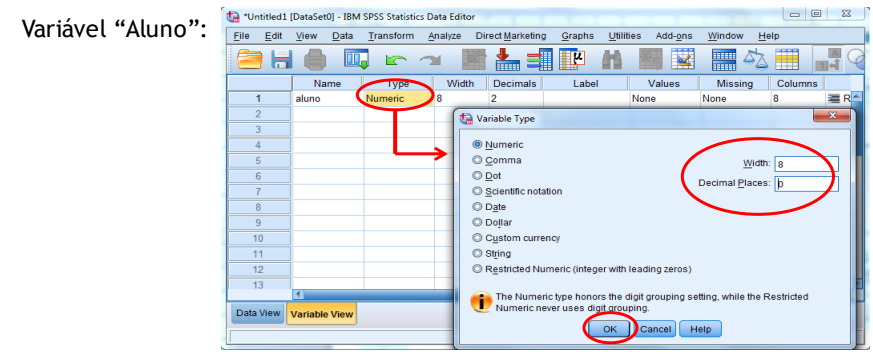
Conceção de um Ficheiro de Dados

Considere novamente o conjunto de dados associado ao desempenho escolar de 30 alunos.
Para inserir estes dados num ficheiro de dados SPSS deve começar por definir as diferentes variáveis na folha de especificação de variáveis:



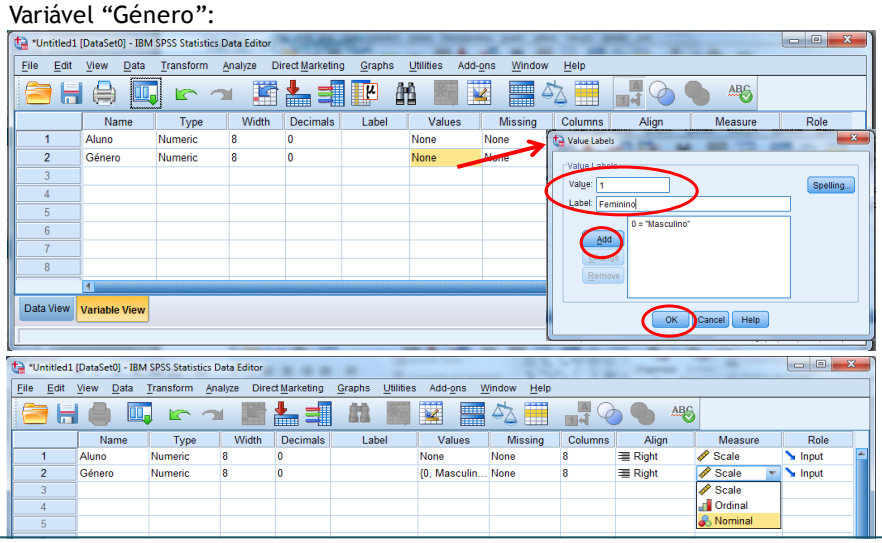
Conceção de um Ficheiro de Dados

Variável "Aluno":



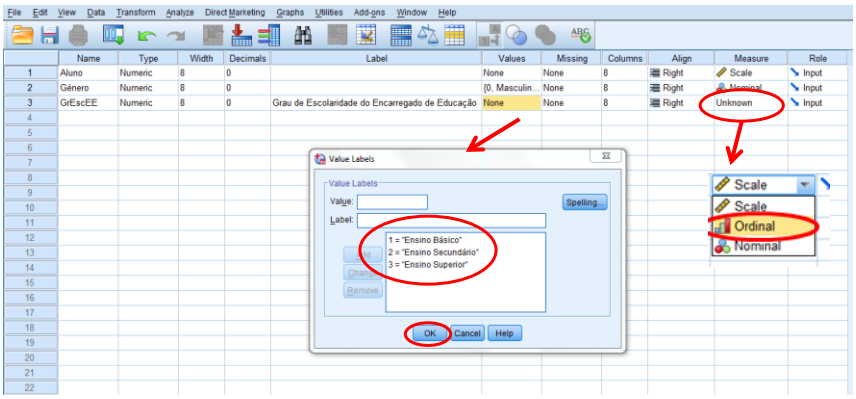
Conceção de um Ficheiro de Dados

Variável "Género":



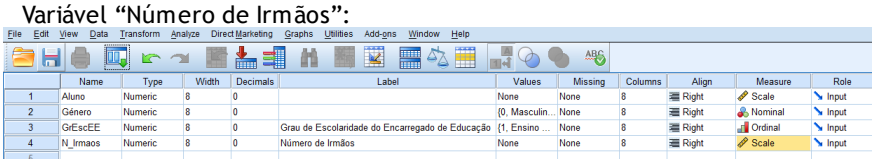
Conceção de um Ficheiro de Dados

Variável "Grau de Escolaridade do Encarregado de Educação":



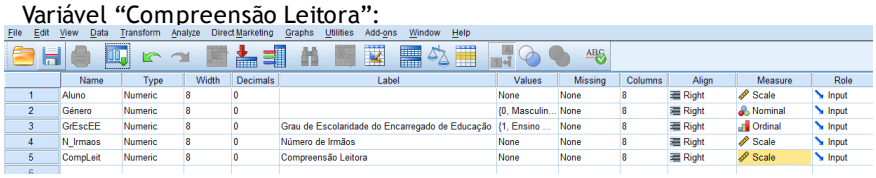
Conceção de um Ficheiro de Dados

Variável "Número de Irmãos":



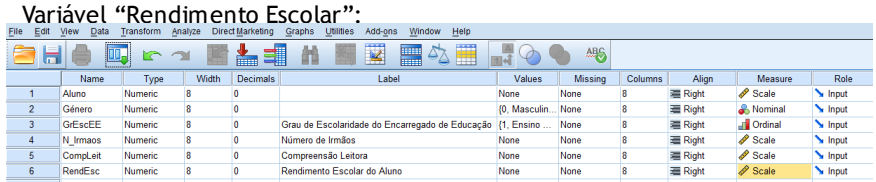
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Aluno	Numeric	8	0		None	None	8	Right	Scale	Input
2	Género	Numeric	8	0		{0, Masculin...	None	8	Right	Nominal	Input
3	GrEscEE	Numeric	8	0	Grau de Escolaridade do Encarregado de Educação	{1, Ensino ...	None	8	Right	Ordinal	Input
4	N_Irmaos	Numeric	8	0	Número de Irmãos	None	None	8	Right	Scale	Input

Variável "Compreensão Leitora":



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Aluno	Numeric	8	0		None	None	8	Right	Scale	Input
2	Género	Numeric	8	0		{0, Masculin...	None	8	Right	Nominal	Input
3	GrEscEE	Numeric	8	0	Grau de Escolaridade do Encarregado de Educação	{1, Ensino ...	None	8	Right	Ordinal	Input
4	N_Irmaos	Numeric	8	0	Número de Irmãos	None	None	8	Right	Scale	Input
5	CompLeit	Numeric	8	0	Compreensão Leitora	None	None	8	Right	Scale	Input

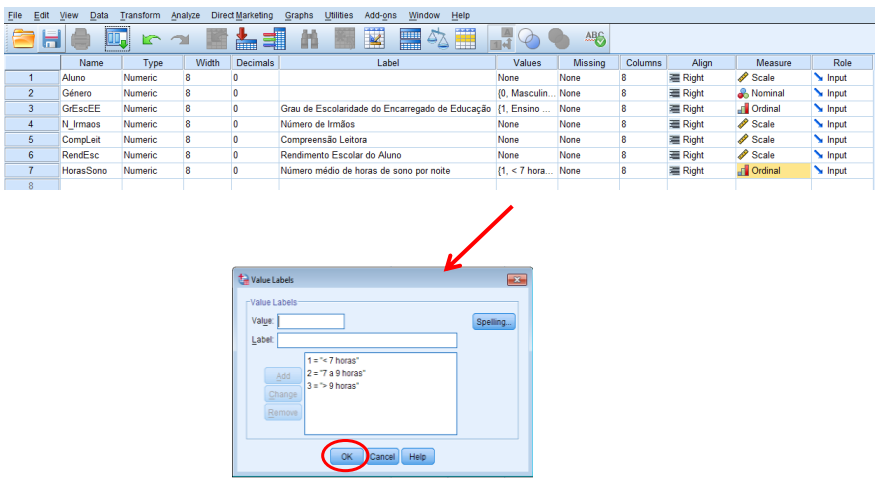
Variável "Rendimento Escolar":



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Aluno	Numeric	8	0		None	None	8	Right	Scale	Input
2	Género	Numeric	8	0		{0, Masculin...	None	8	Right	Nominal	Input
3	GrEscEE	Numeric	8	0	Grau de Escolaridade do Encarregado de Educação	{1, Ensino ...	None	8	Right	Ordinal	Input
4	N_Irmaos	Numeric	8	0	Número de Irmãos	None	None	8	Right	Scale	Input
5	CompLeit	Numeric	8	0	Compreensão Leitora	None	None	8	Right	Scale	Input
6	RendEsc	Numeric	8	0	Rendimento Escolar do Aluno	None	None	8	Right	Scale	Input

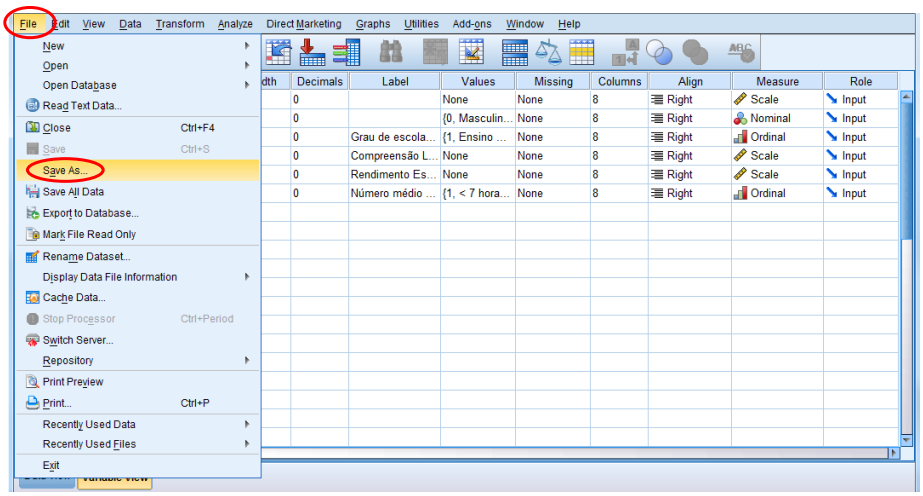
Conceção de um Ficheiro de Dados

Variável "Número médio de horas de sono por noite":

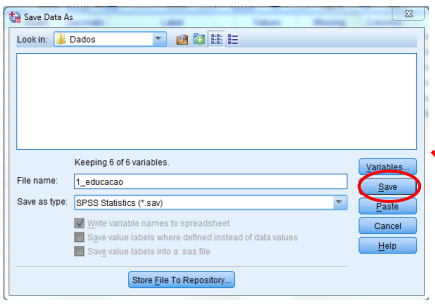


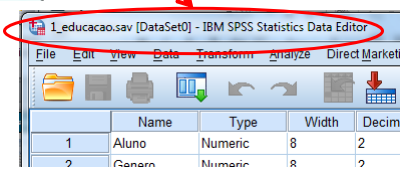
Conceção de um Ficheiro de Dados

Guardar o Ficheiro de Dados:



Conceção de um Ficheiro de Dados

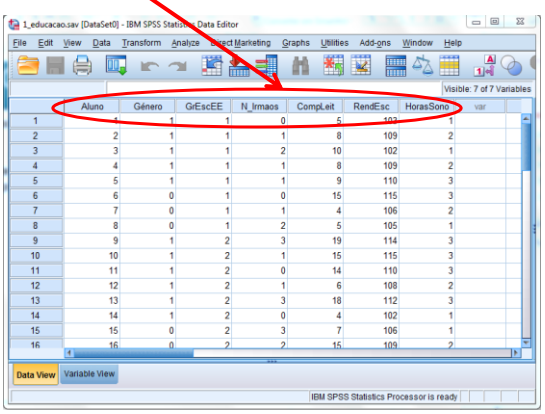




Conceção de um Ficheiro de Dados

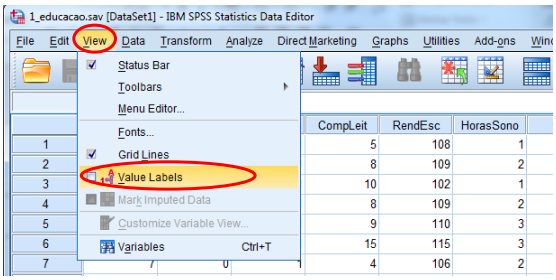
Cada coluna corresponde a uma variável:

Insira os valores observados, considerando que cada linha corresponde aos valores de cada aluno:

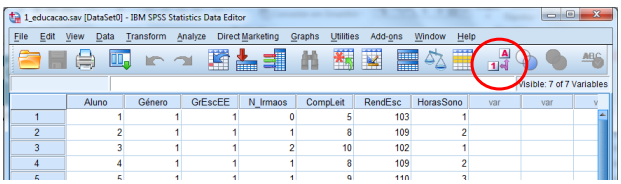


Conceção de um Ficheiro de Dados

Para visualizar as codificações feitas:

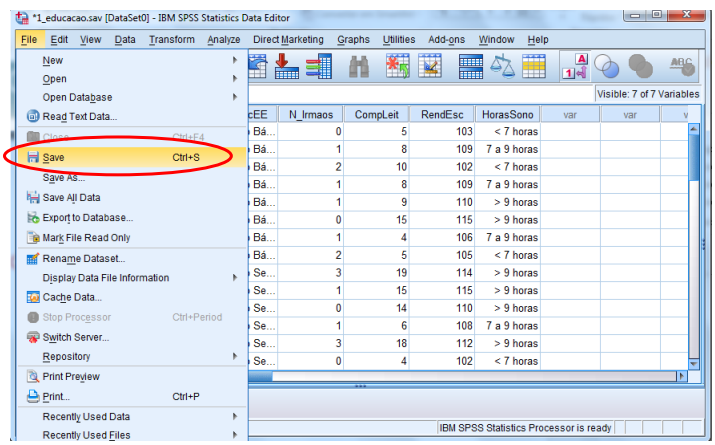


Ou usar a barra de atalho:



Conceção de um Ficheiro de Dados

No final, guarde novamente o ficheiro de dados:



Operações básicas com Ficheiros de Dados

Considere novamente o conjunto de dados referente à avaliação contínua de 27 estudantes.

Suponha agora que se pretendia determinar a média das notas obtidas nos dois testes escritos:

Média = (Teste_1 + Teste_2)/2

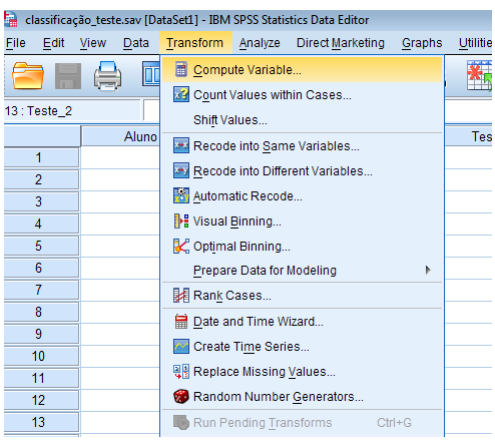
Como proceder?

Nota: Estes dados encontram-se no ficheiro SPSS com o nome 2_avaliação_continua.sav

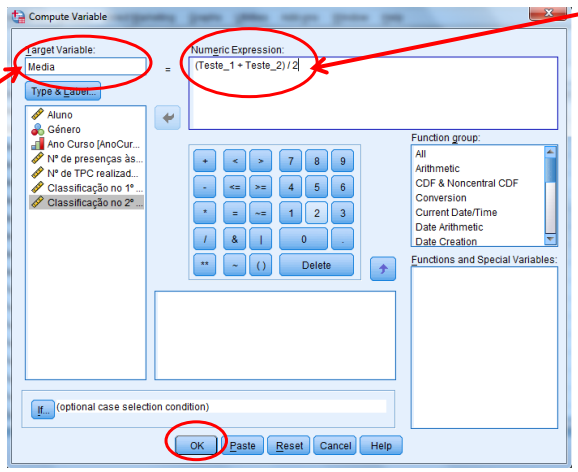
Operações básicas com Ficheiros de Dados

Seleccione os comandos:

Transforme
Compute Variable ...



Operações básicas com Ficheiros de Dados



Nome da nova variável

Expressão matemática para determinar a nova variável

Operações básicas com Ficheiros de Dados

*1_classificação_teste.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

	Aluno	Género	AnoCurso	Presenças	T.P.C	Teste_1	Teste_2	media
1	1	Masculino	1º ano	1	0	2,00	,00	1,00
2	2	Masculino	1º ano	6	4	8,10	5,70	6,90
3	3	Masculino	1º ano	5	0	3,00	9,80	6,40
4	4	Masculino	2º ano	0	0	5,00	3,90	4,45
5	5	Masculino	1º ano	2	0	4,50	12,50	8,50
6	6	Masculino	1º ano	4	4	6,50	11,00	8,75
7	7	Masculino	2º ano	3	0	5,30	8,30	6,80
8	8	Masculino	1º ano	2	2	3,70	5,00	4,35
9	9	Masculino	1º ano	8	0	2,00	3,50	2,75
10	10	Masculino	1º ano	6	2	11,80	11,40	11,60
11	11	Feminino	1º ano	11	8	11,00	19,70	15,35

Operações básicas com Ficheiros de Dados

Suponha agora que a classificação associada à participação é determinada por:

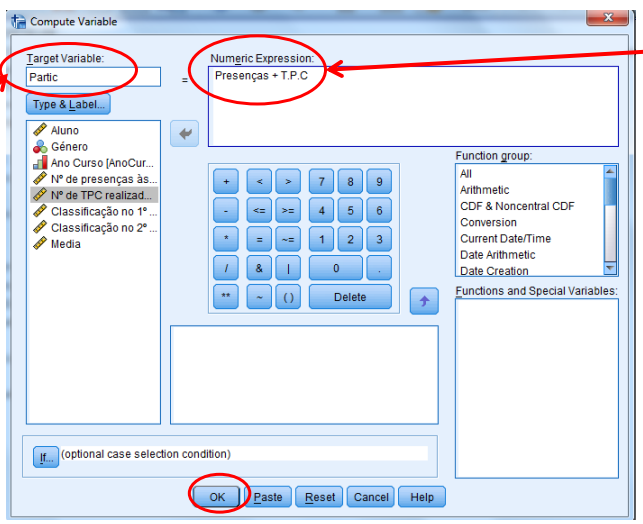
Participação = Presenças + T.P.C.

Como proceder?

Seleccione os comandos:

Transforme
Compute Variable …

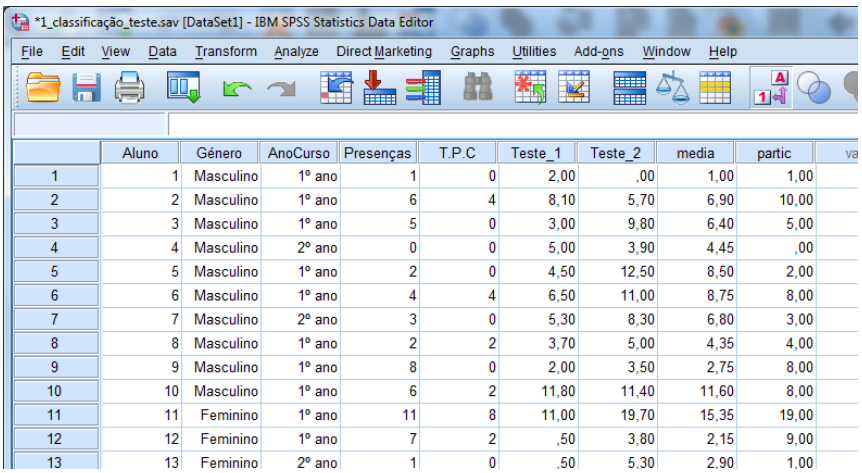
Operações básicas com Ficheiros de Dados



Nome da nova variável

Expressão matemática para determinar a nova variável

Operações básicas com Ficheiros de Dados



	Aluno	Género	AnoCurso	Presenças	T.P.C	Teste_1	Teste_2	media	partic
1	1	Masculino	1º ano	1	0	2,00	,00	1,00	1,00
2	2	Masculino	1º ano	6	4	8,10	5,70	6,90	10,00
3	3	Masculino	1º ano	5	0	3,00	9,80	6,40	5,00
4	4	Masculino	2º ano	0	0	5,00	3,90	4,45	,00
5	5	Masculino	1º ano	2	0	4,50	12,50	8,50	2,00
6	6	Masculino	1º ano	4	4	6,50	11,00	8,75	8,00
7	7	Masculino	2º ano	3	0	5,30	8,30	6,80	3,00
8	8	Masculino	1º ano	2	2	3,70	5,00	4,35	4,00
9	9	Masculino	1º ano	8	0	2,00	3,50	2,75	8,00
10	10	Masculino	1º ano	6	2	11,80	11,40	11,60	8,00
11	11	Feminino	1º ano	11	8	11,00	19,70	15,35	19,00
12	12	Feminino	1º ano	7	2	,50	3,80	2,15	9,00
13	13	Feminino	2º ano	1	0	,50	5,30	2,90	1,00

Operações básicas com Ficheiros de Dados

Suponha agora que a classificação final de cada aluno é dada por:

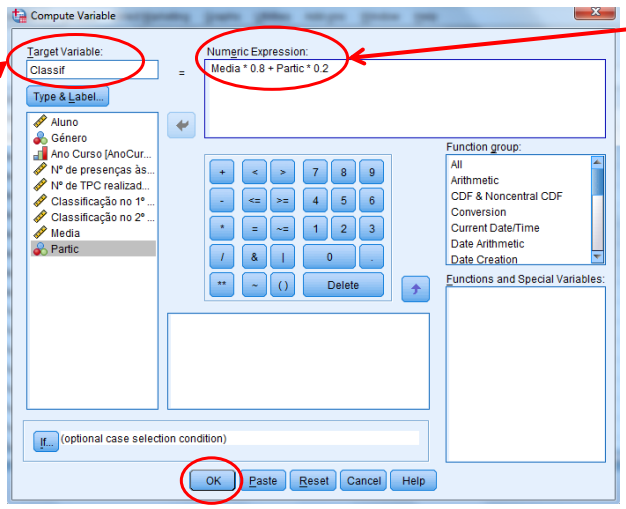
Nota Final = 80% Média dos teste + 20% Participação

Como proceder?

Seleccione os comandos:

Transforme
Compute Variable …

Operações básicas com Ficheiros de Dados



Nome da nova variável

Expressão matemática para determinar a nova variável

Operações básicas com Ficheiros de Dados

A variável nota final de cada aluno, também pode ser codificada da seguinte forma:

Situação Final:

Reprovado se nota final < 9.5

Aprovado se nota final ≥ 9.5

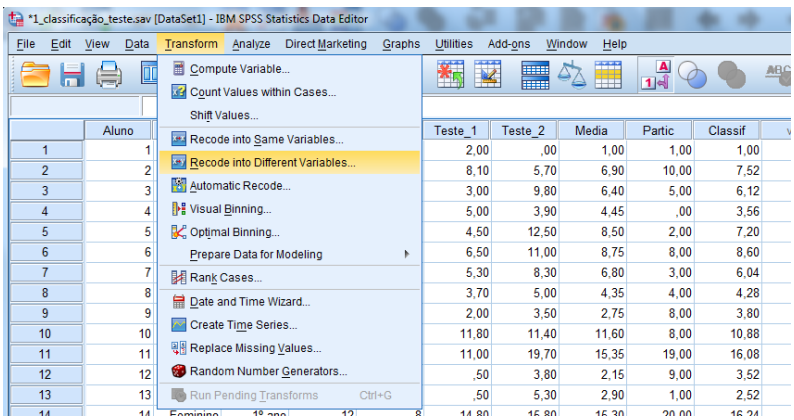
Como proceder?

Operações básicas com Ficheiros de Dados

Seleccione os comandos:

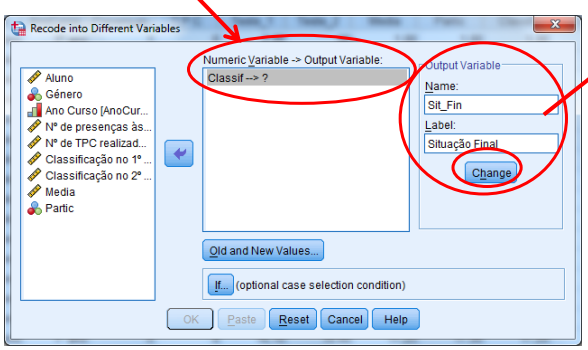
Transforme

Recode into Different Variables



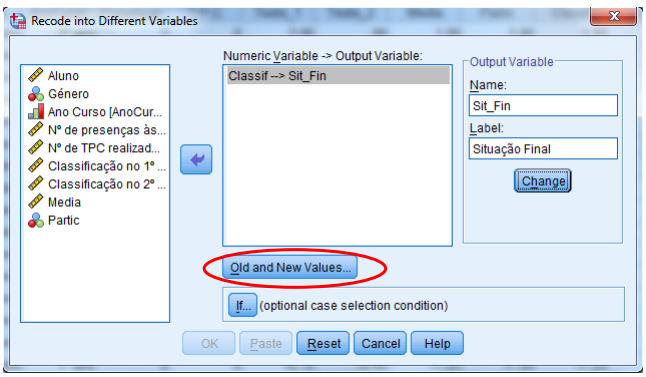
Operações básicas com Ficheiros de Dados

Nome da variável usada para determinar nova variável

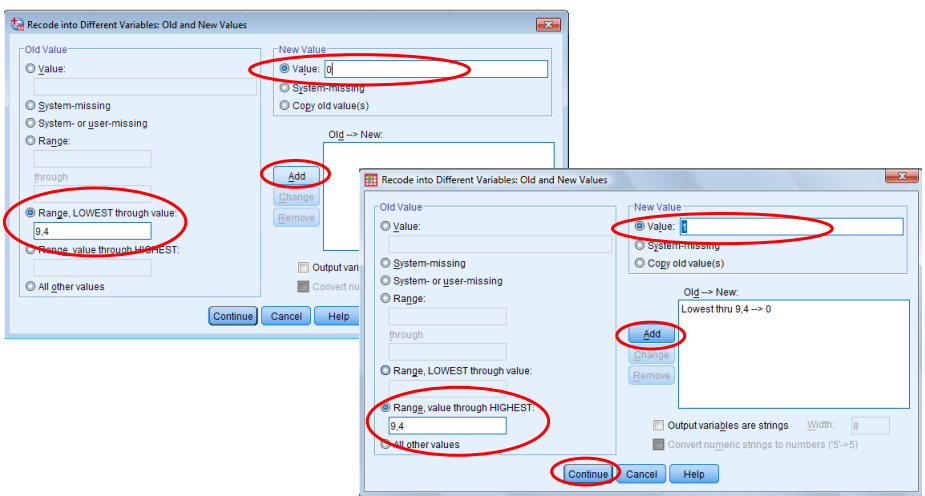


Nome da nova variável

Operações básicas com Ficheiros de Dados

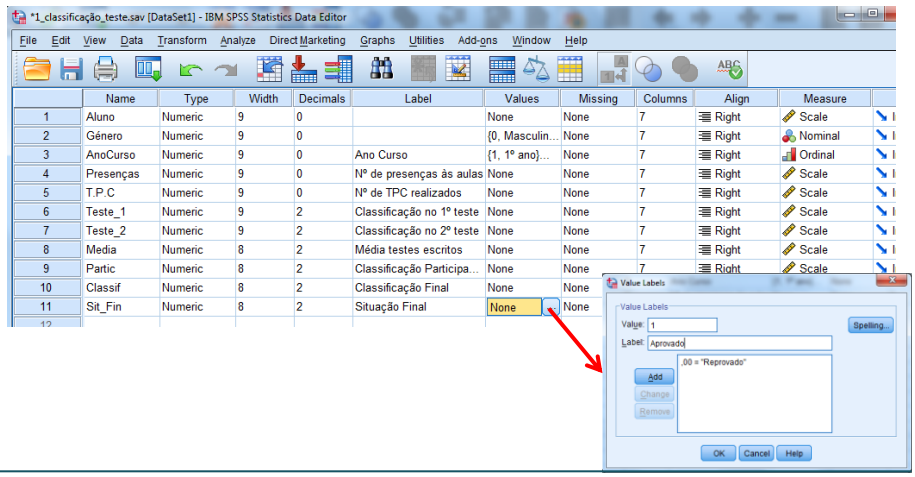


Operações básicas com Ficheiros de Dados



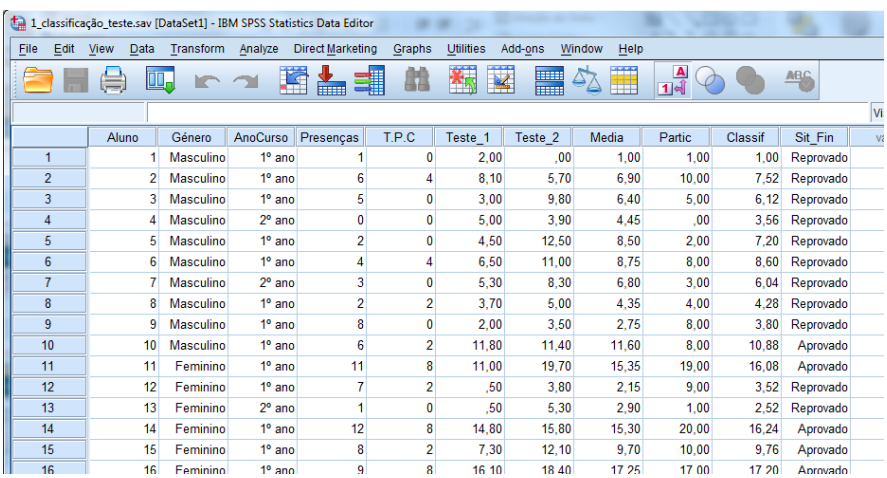
Operações básicas com Ficheiros de Dados

Para facilitar consultas futuras deste ficheiro, deve proceder à codificação da nova variável:



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	Aluno	Numeric	9	0		None	None	7	Right	Scale
2	Género	Numeric	9	0		{0, Masculin...	None	7	Right	Nominal
3	AnoCurso	Numeric	9	0	Ano Curso	{1, 1º ano}...	None	7	Right	Ordinal
4	Presenças	Numeric	9	0	Nº de presenças às aulas	None	None	7	Right	Scale
5	T.P.C	Numeric	9	0	Nº de TPC realizados	None	None	7	Right	Scale
6	Teste_1	Numeric	9	2	Classificação no 1º teste	None	None	7	Right	Scale
7	Teste_2	Numeric	9	2	Classificação no 2º teste	None	None	7	Right	Scale
8	Media	Numeric	8	2	Média testes escritos	None	None	7	Right	Scale
9	Partic	Numeric	8	2	Classificação Participa...	None	None	7	Right	Scale
10	Classif	Numeric	8	2	Classificação Final	None	None			
11	Sit_Fin	Numeric	8	2	Situação Final	None	None			

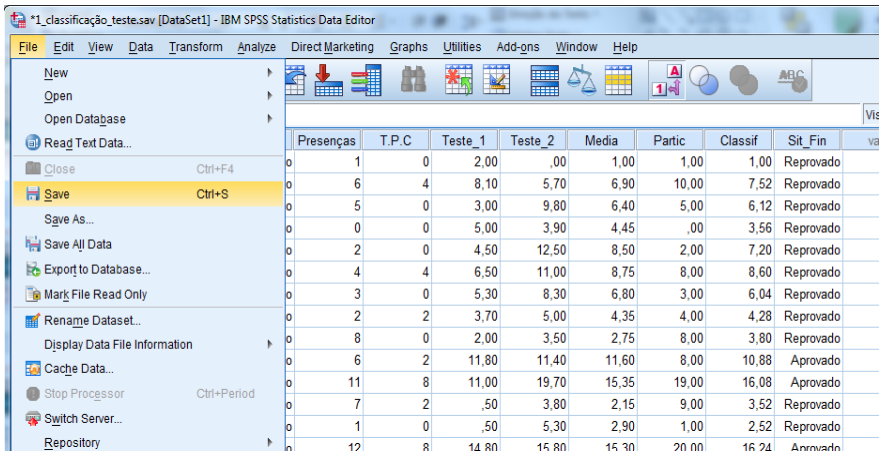
Operações básicas com Ficheiros de Dados



	Aluno	Género	AnoCurso	Presenças	T.P.C	Teste_1	Teste_2	Media	Partic	Classif	Sit_Fin
1	1	Masculino	1º ano	1	0	2,00	,00	1,00	1,00	1,00	Reprovado
2	2	Masculino	1º ano	6	4	8,10	5,70	6,90	10,00	7,52	Reprovado
3	3	Masculino	1º ano	5	0	3,00	9,80	6,40	5,00	6,12	Reprovado
4	4	Masculino	2º ano	0	0	5,00	3,90	4,45	,00	3,56	Reprovado
5	5	Masculino	1º ano	2	0	4,50	12,50	8,50	2,00	7,20	Reprovado
6	6	Masculino	1º ano	4	4	6,50	11,00	8,75	8,00	8,60	Reprovado
7	7	Masculino	2º ano	3	0	5,30	8,30	6,80	3,00	6,04	Reprovado
8	8	Masculino	1º ano	2	2	3,70	5,00	4,35	4,00	4,28	Reprovado
9	9	Masculino	1º ano	8	0	2,00	3,50	2,75	8,00	3,80	Reprovado
10	10	Masculino	1º ano	6	2	11,80	11,40	11,60	8,00	10,88	Aprovado
11	11	Feminino	1º ano	11	8	11,00	19,70	15,35	19,00	16,08	Aprovado
12	12	Feminino	1º ano	7	2	,50	3,80	2,15	9,00	3,52	Reprovado
13	13	Feminino	2º ano	1	0	,50	5,30	2,90	1,00	2,52	Reprovado
14	14	Feminino	1º ano	12	8	14,80	15,80	15,30	20,00	16,24	Aprovado
15	15	Feminino	1º ano	8	2	7,30	12,10	9,70	10,00	9,76	Aprovado
16	16	Feminino	1º ano	9	8	16,10	18,40	17,25	17,00	17,20	Aprovado

Operações básicas com Ficheiros de Dados

No fim, não se esqueça de gravar o ficheiro construído:



	Presenças	T.P.C	Teste_1	Teste_2	Media	Partic	Classif	Sit_Fin
o	1	0	2,00	,00	1,00	1,00	1,00	Reprovado
o	6	4	8,10	5,70	6,90	10,00	7,52	Reprovado
o	5	0	3,00	9,80	6,40	5,00	6,12	Reprovado
o	0	0	5,00	3,90	4,45	,00	3,56	Reprovado
o	2	0	4,50	12,50	8,50	2,00	7,20	Reprovado
o	4	4	6,50	11,00	8,75	8,00	8,60	Reprovado
o	3	0	5,30	8,30	6,80	3,00	6,04	Reprovado
o	2	2	3,70	5,00	4,35	4,00	4,28	Reprovado
o	8	0	2,00	3,50	2,75	8,00	3,80	Reprovado
o	6	2	11,80	11,40	11,60	8,00	10,88	Aprovado
o	11	8	11,00	19,70	15,35	19,00	16,08	Aprovado
o	7	2	,50	3,80	2,15	9,00	3,52	Reprovado
o	1	0	,50	5,30	2,90	1,00	2,52	Reprovado
o	12	8	14,80	15,80	15,30	20,00	16,24	Aprovado

Estatística Descritiva Univariada

A estatística no mundo atual

A **Estatística** é usada no nosso dia a dia, por exemplo em

- Estudos de mercados
- Sondagens
- …

O que é a estatística?
Para que serve a estatística?

A **Estatística** é muitas vezes definida como

- Método matemático de análise de dados
- Ciência que trata os dados

A estatística no mundo atual

Analise o seguinte conjunto de dados, por breves momentos:

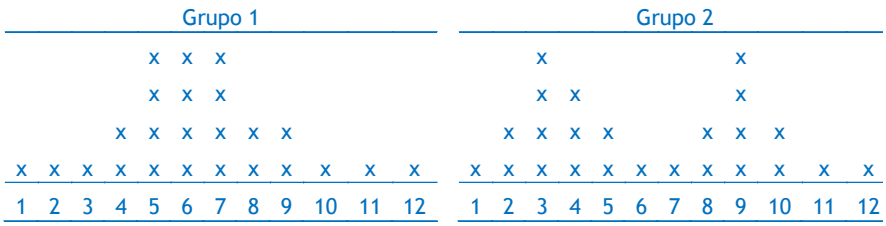
Grupo 1					
6	4	1	12	7	5
3	6	5	8	11	5
2	9	7	9	4	10
8	6	6	7	5	7

Grupo 2					
6	9	2	12	8	4
3	11	1	10	9	3
4	5	4	7	3	9
5	8	10	2	9	3

Quais as principais conclusões?

A estatística no mundo atual

Utilizando algumas "técnicas" de organização dos dados, podem-se obter as seguintes tabelas:



As "diferenças" entre os dois conjuntos de dados são facilmente identificadas

Tratamento dados univariados - Introdução

A forma de organizar, sumarizar e descrever os dados depende basicamente

- da **escala** em que os dados estão medidos
- do **número de variáveis** a analisar

Uma Análise Estatística pode ser:

- **Univariada** - se envolve apenas uma variável
- **Bivariada** - se envolve apenas duas variáveis
- **Multivariada** - se envolve a análise simultânea de mais de 2 variáveis

Dados Qualitativos Nominais: Tabelas de Frequência

A organização dos dados definidos numa escala nominal, baseia-se em **contagens** de elementos pertencentes à mesma caraterística ou modalidade.

Uma forma possível de organizar dados nominais é conseguida através do uso de **tabelas de frequência**.

Como proceder?

1 - listar todas as categorias da variável em estudo

2 - contar os elementos pertencentes a cada modalidade (ou classe).

Dados Qualitativos Nominais: Tabelas de Frequência

Tabela de Frequências para dados Qualitativos Nominais

Frequência absoluta (n_i) número de observações em cada modalidade ou classe

Frequência relativa (f_i) quociente entre a frequência absoluta dessa modalidade (ou classe) e o número total de observações

X	n_i	f_i
x_1	n_1	$f_1 = n_1/N$
x_2	n_2	$f_2 = n_2/N$
...	...	...
x_k	n_k	$f_k = n_k/N$
Total	N	1

Modalidade ou classe

A soma de todas as frequências absolutas é igual a N

A soma de todas as frequências relativas é igual a 1.

Dados Qualitativos Nominais: Tabelas de Frequência - Exemplo

Considere o seguinte conjunto de dados associados à competência leitora dos alunos:

Aluno	Género	Grau Escolaridade Enc. Educ.	Nº irmãos	Compe-tência Leitora	Rendi-mento Escolar	Nº médio horas sono
1	Feminino	E. Básico	0	5	103	<7 horas
2	Feminino	E. Básico	1	8	109	7 a 9 horas
3	Feminino	E. Básico	2	10	102	<7 horas
4	Feminino	E. Básico	1	8	109	7 a 9 horas
5	Feminino	E. Básico	1	9	110	>9 horas
6	Masculino	E. Básico	0	15	115	>9 horas
7	Masculino	E. Básico	1	4	106	7 a 9 horas
8	Masculino	E. Básico	2	5	105	<7 horas
9	Feminino	E. Secundário	3	19	114	>9 horas
10	Feminino	E. Secundário	1	15	115	>9 horas
11	Feminino	E. Secundário	0	14	110	>9 horas
12	Feminino	E. Secundário	1	6	108	7 a 9 horas
13	Feminino	E. Secundário	3	18	112	>9 horas
14	Feminino	E. Secundário	0	4	102	<7 horas
15	Masculino	E. Secundário	3	7	106	<7 horas
16	Masculino	E. Secundário	2	15	109	7 a 9 horas
17	Masculino	E. Secundário	1	10	108	>9 horas
18	Masculino	E. Secundário	0	9	109	>9 horas
19	Masculino	E. Secundário	0	8	108	7 a 9 horas
20	Masculino	E. Secundário	0	11	107	7 a 9 horas
21	Feminino	E. Superior	1	12	109	>9 horas
22	Feminino	E. Superior	1	19	115	>9 horas
23	Feminino	E. Superior	0	14	111	>9 horas
24	Feminino	E. Superior	1	17	115	>9 horas
25	Feminino	E. Superior	2	18	110	>9 horas
26	Feminino	E. Superior	2	12	109	7 a 9 horas
27	Masculino	E. Superior	1	14	108	7 a 9 horas
28	Masculino	E. Superior	1	9	104	<7 horas
29	Masculino	E. Superior	2	5	107	<7 horas
30	Masculino	E. Superior	3	15	102	<7 horas

Nota: Estes dados encontram-se no ficheiro SPSS com o nome 1_educacao.sav

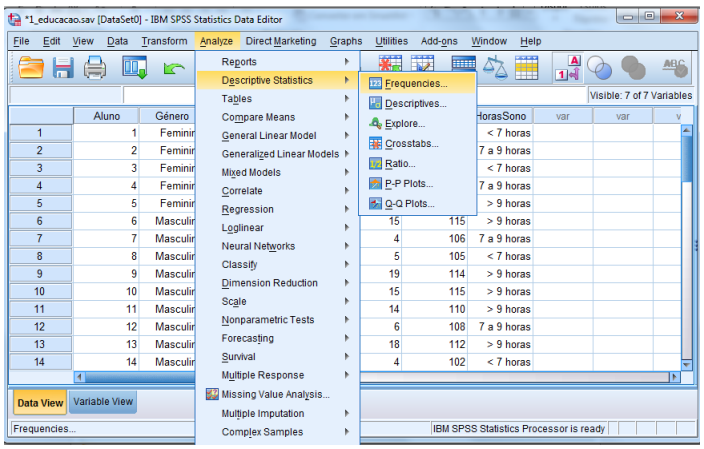
Dados Qualitativos Nominais: Tabelas de Frequência - Exemplo

A **tabela de frequências** para a variável "género" é obtida da seguinte forma:

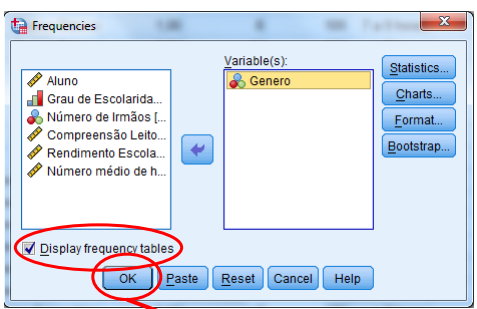
Género	n_i	f_i
Masculino	13	13/30 = 0.43
Feminino	17	17/30 = 0.57
↓	↓	↓
Modalidade (classe)	Frequência absoluta	Frequência relativa

Dados Qualitativos Nominais: Tabelas de Frequência - Exemplo

No SPSS:



Dados Qualitativos Nominais: Tabelas de Frequência - Exemplo



Genero

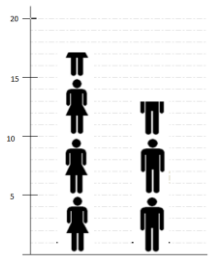
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Masculino	13	43,3	43,3	43,3
	Feminino	17	56,7	56,7	100,0
	Total	30	100,0	100,0	

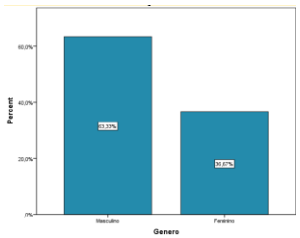
Modalidade (classe) | Frequência absoluta | Percentagem =Frequência relativa x 100

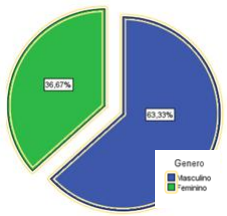
Dados Qualitativos Nominais: Representações Gráficas - Exemplo

As **representações gráficas** mais adequadas para dados expressos numa escala nominal são:

- Pictograma
- Gráfico de barras
- Gráfico circular



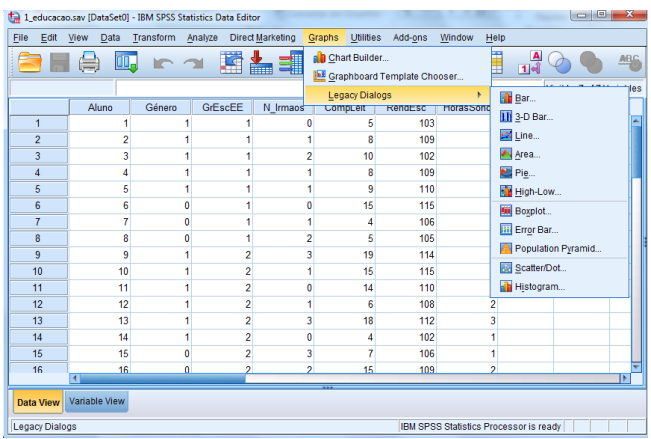




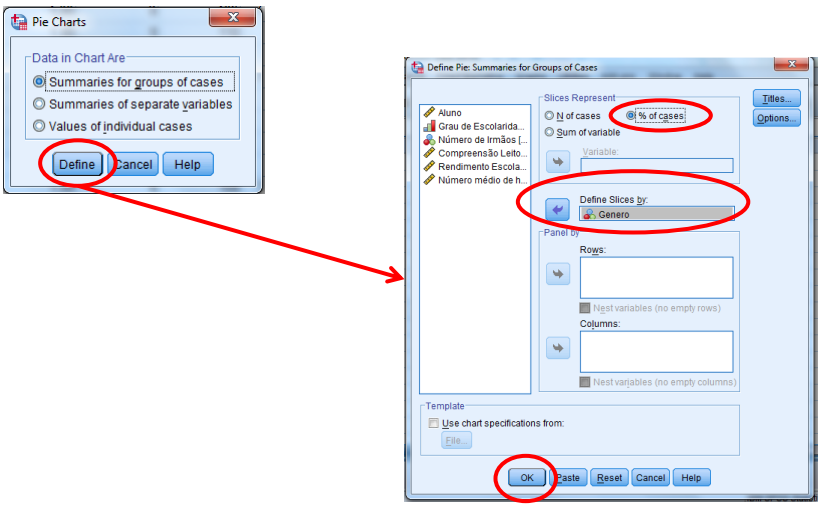
Cada  ou  corresponde a 5 indivíduos

Dados Qualitativos Nominais: Representações Gráficas

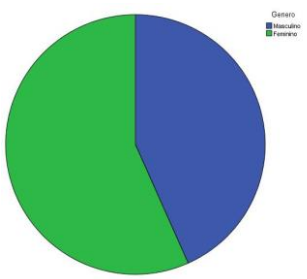
No SPSS:



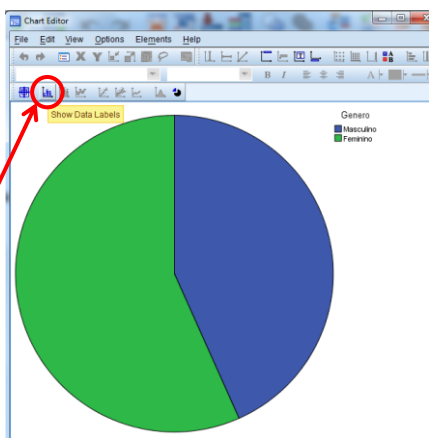
Dados Qualitativos Nominais: Representações Gráficas



Dados Qualitativos Nominais: Representações Gráficas

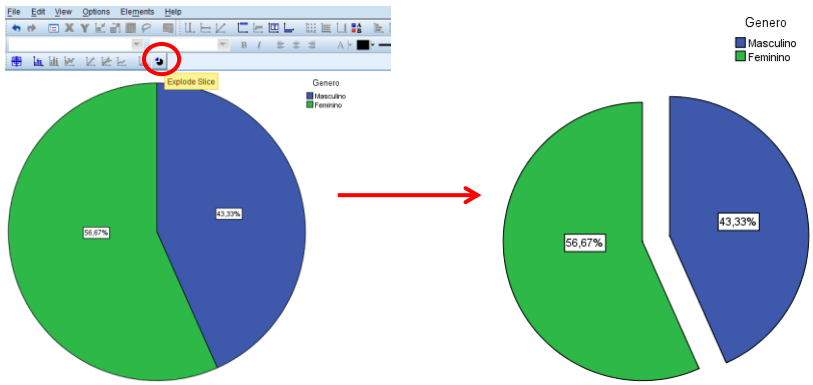


Para alterar as especificações do gráfico clicar duas vezes em cima do gráfico abrindo-se outra janela. Escolher, por exemplo, a opção



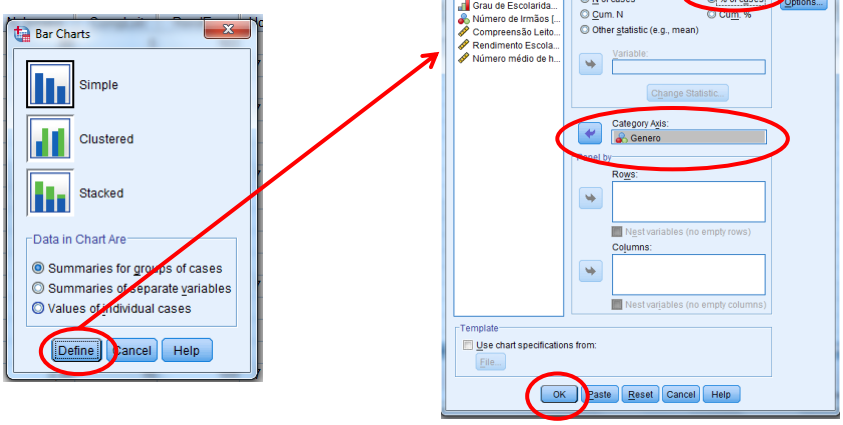
Dados Qualitativos Nominais: Representações Gráficas

Escolhendo a opção

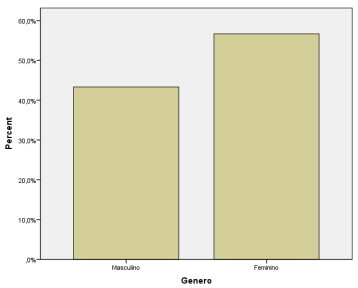


Dados Qualitativos Nominais: Representações Gráficas

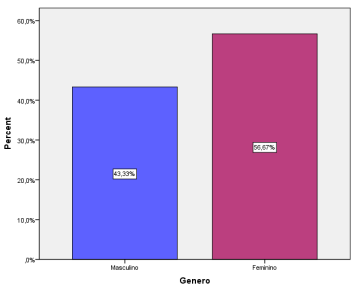
Gráfico Barras no SPSS:
Graphs > Legacy Dialogs > Bar…



Dados Qualitativos Nominais: Representações Gráficas



Clicando duas vezes em cima do gráfico pode-se alterar as definições



Dados Qualitativos Ordinais: Tabelas de Frequência

Dados Qualitativos Ordinais

A organização de dados qualitativos ordinais baseia-se na contagem dos elementos pertencentes às diferentes categorias.

Como é possível estabelecer uma relação de ordem entre as categorias, já é possível determinar as frequências acumuladas (absolutas e relativas)

Como proceder?

1 - listar todas as categorias da variável em estudo e ordená-las

2 - contar os elementos pertencentes a cada modalidade (ou classe).

Dados Qualitativos Ordinais: Tabelas de Frequência

Tabela de Frequências para dados Qualitativos Ordinais

X	n_i	f_i	N_i	F_i
x_1	n_1	$f_1 = n_1/N$	n_1	f_1
x_2	n_2	$f_2 = n_2/N$	n_1+n_2	f_1+f_2
...	...	...	...	...
x_k	n_k	$f_k = n_k/N$	N	1_k
Total	N	1		

Frequência absoluta (n_i) — Frequência relativa (f_i) — Frequência absoluta acumulada (N_i) — Frequência relativa acumulada (F_i)

Dados Qualitativos Ordinais: Tabelas de Frequência - Exemplo

A tabela de frequências para a variável "grau de escolaridade do encarregado de educação" é obtida da seguinte forma:

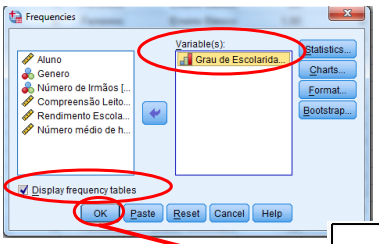
Grau Esc. Enc. Educ.	n_i	f_i	N_i	F_i
Ensino Básico	8	8/30 = 0.27	8	0.27
Ensino Secundário	12	12/30 = 0.40	20	0.67
Ensino Superior	10	10/30 = 0.33	30	1.00

Modalidade (classe) — Frequência absoluta — Frequência relativa — Frequência absoluta acumulada — Frequência relativa acumulada

Dados Qualitativos Ordinais: Tabelas de Frequência - Exemplo

No SPSS:

Analyze > Descriptive Statistics > Frequencies...



		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Ensino Básico	8	26,7	26,7	26,7
	Ensino Secundário	12	40,0	40,0	66,7
	Ensino Superior	10	33,3	33,3	100,0
	Total	30	100,0	100,0	

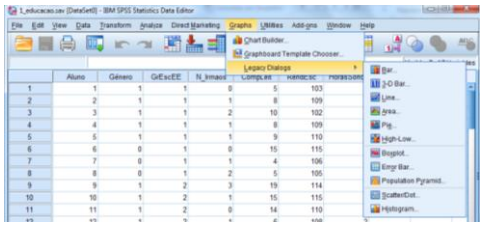
Modalidade (classe) | Frequência absoluta | Percentagem | Percentagem acumulada

Dados Qualitativos Ordinais: Representações Gráficas

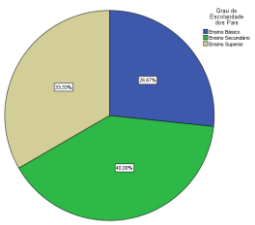
As **representações gráficas** mais adequadas para dados expressos numa escala ordinal são:

- Gráficos circulares
- Gráficos barras
- Gráficos de traços
- Diagramas de extremos e quartis /Caixas de bigodes (*boxplot*) (a ver mais tarde)

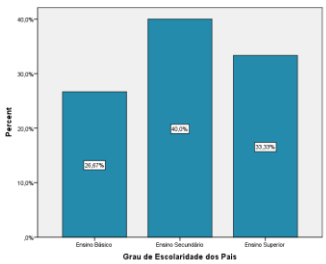
Dados Qualitativos Ordinais: Representações Gráficas - Exemplo



- Gráfico Circular



- Gráfico de barras



Dados Quantitativos Discretos: Tabelas de Frequência - Exemplo

Dados Quantitativos Discretos

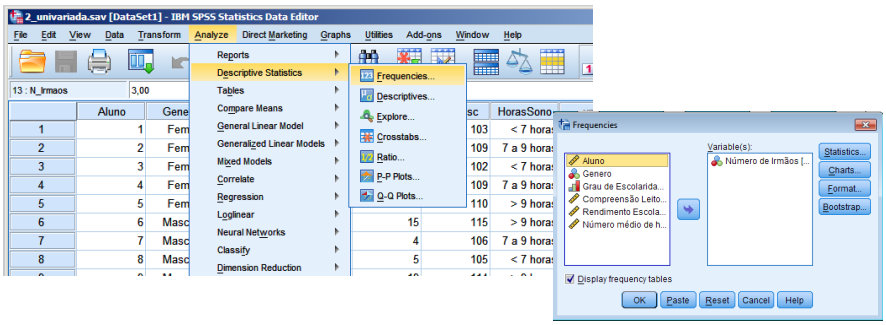
A organização de dados quantitativos discretos é feita de forma semelhante à considerada para dados qualitativos ordinais

A **tabela de frequências** para a variável "nº de irmãos" é obtida da seguinte forma:

Nº de Irmãos	n_i	f_i	N_i	F_i
0	8	8/30 = 0.27	8	0.27
1	12	12/30 = 0.40	20	0.67
2	6	6/30 = 0.20	26	0.87
3	4	4/30 = 0.13	30	1.00
↓	↓	↓	↓	↓
Valores observados para a variável	Frequência absoluta	Frequência relativa	Frequência absoluta acumulada	Frequência relativa acumulada

Dados Quantitativos Discretos: Tabelas de Frequência - Exemplo

No SPSS:



Número de Irmãos

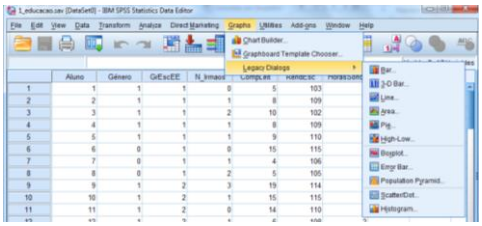
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	,00	8	26,7	26,7	26,7
	1,00	12	40,0	40,0	66,7
	2,00	6	20,0	20,0	86,7
	3,00	4	13,3	13,3	100,0
	Total	30	100,0	100,0	

Dados Quantitativos Discretos: Representações Gráficas - Exemplo

As **representações gráficas** mais adequadas para dados expressos numa escala quantitativa discreta são:

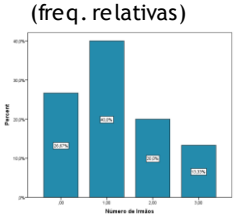
- Gráficos circulares
- Gráficos barras
- Gráficos de traços
- Diagramas de extremos e quartis /Caixas de bigodes (a ver mais tarde)

Dados Quantitativos Discretos: Representações Gráficas - Exemplo

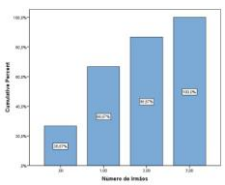


- Gráfico de barras

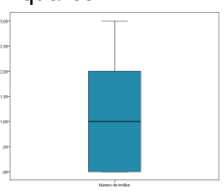
(freq. relativas)



(freq. acumuladas)



- Diagrama de extremos e quartis



Dados Quantitativos (agrup. em classes): Tabelas de Frequência

Dados Quantitativos agrupados em classes

A organização deste tipo de dados requer que os dados sejam agrupados em classes ou intervalos.

O número de classes a considerar (c) é dado pela regra de **Sturges**:

$$c = 1 + 3.322 \log_{10}(N)$$

Nota: O número de classes a considerar deve ser arredondado para o valor inteiro superior.

No caso de se considerar que as classes têm igual amplitude, ela é dada por

$$h = (X_{max} - X_{min}) / c$$

Dados Quantitativos (agrup. em classes): Tabelas de Frequência

Tabela de Frequências para dados Quantitativos (agrup. em classes)

Ponto médio da classe	X	n_i	f_i	N_i	F_i
$x'_1=(l_1+l_2)/2$	$[l_1, l_2[$	n_1	$f_1 = n_1/N$	n_1	f_1
$x'_2=(l_2+l_3)/2$	$[l_2, l_3[$	N_2	$f_2 = n_2/N$	n_1+n_2	f_1+f_2
...	...	...	...	...	...
$x'_k=(l_k+l_{k+1})/2$	$[l_k, l_{k+1}[$	n_k	$f_k = n_k/N$	N	1
	Total	N	1		

Frequência absoluta (n_i) — Frequência relativa (f_i) — Frequência absoluta acumulada (N_i) — Frequência relativa acumulada (F_i)

Dados Quantitativos (agrup. em classes): Tabelas de Frequência - Exemplo

A **tabela de frequências** para a variável "competência leitora" é obtida da seguinte forma:

O número de classes é dados por:
$c = 1 + 3.322 \log_{10}(30) = 5.906 \Rightarrow$ considerar 6 classes

A amplitude de cada classe é dada por:
$h = (19 - 4)/6 = 2.5$ $\Rightarrow$ considerar, por exemplo, h = 2.6

Considerando h = 2.5, as classes seriam:
[4.0, 6.5[
[6.5, 9.0[
[9.0, 11.5[
[11.5, 14.0[
[14.0, 16.5[
[16.5, 19.0[

O valor 19 não pertence a nenhuma classe

Assim, as classes a considerar são:
[4.0, 6.6[
[6.6, 9.2[
[9.2, 11.8[
[11.8, 14.4[
[14.4, 17.0[
[17.0, 19.6[

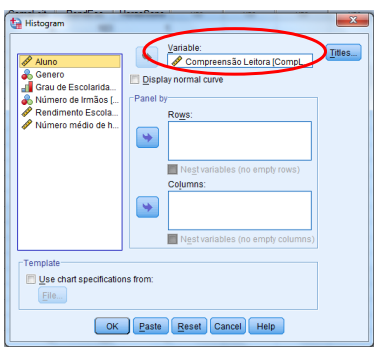
Dados Quantitativos (agrup. em classes): Tabelas de Frequência - Exemplo

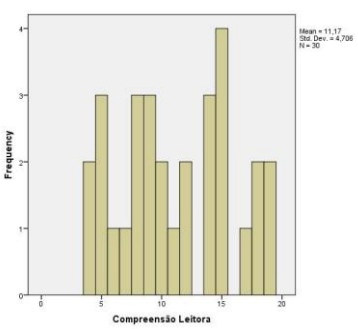
Ponto médio da classe	Classe	n_i	f_i	N_i	F_i
5.3	[4.0,6.6[	6	6/30 =0.20	6	0.20
7.9	[6.6,9.2[	7	7/30 =0.23	13	0.43
10.5	[9.2, 11.8[	3	3/30=0.10	16	0.53
13.1	[11.8, 14.4[	5	5/30=0.17	21	0.70
15.7	[14.4, 17.0[	4	4/30=0.13	25	0.83
18.3	[17.0, 19.6[	5	5/30=0.17	30	1.00
	↓ Classes	↓ Frequência absoluta	↓ Frequência relativa	↓ Frequência absoluta acumulada	↓ Frequência relativa acumulada

Dados Quantitativos (agrup. em classes): Representações Gráficas - Exemplo

Para representar graficamente estes dados usando o SPSS pode-se construir um histograma:

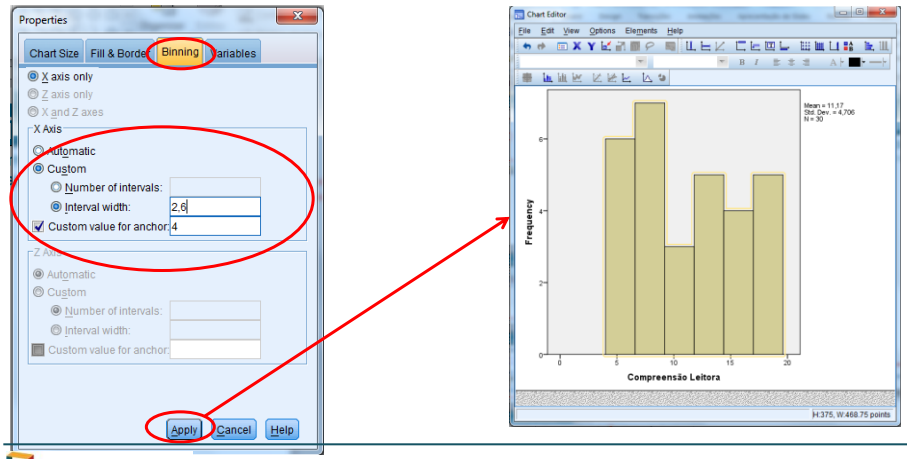
Graph > Legacy Dialogs > Histogram…





Dados Quantitativos (agrup. em classes): Representações Gráficas - Exemplo

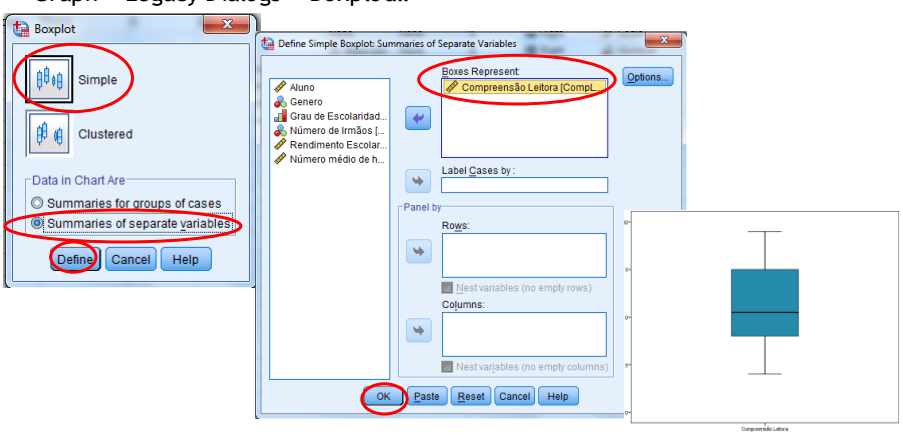
Para definir o número de classes e a amplitude das classes, deve clicar com o rato duas vezes em cima das barras do histograma que aparece no editor de gráficos para obter a seguinte janela:



Dados Quantitativos (agrup. em classes): Representações Gráficas - Exemplo

Para representar graficamente estes dados usando o SPSS pode-se construir um diagrama de extremos e quartis:

Graph > Legacy Dialogs > Boxplot...



Distribuições de Frequência e Representações Gráficas - Exercícios

Considere o novamente o conjunto de dados referente à avaliação contínua de 27 estudantes.
Descreva cada uma das variáveis através de distribuições de frequência e de representações gráficas.

LEGENDA:
Genero: Masculino, Feminino
Ano Curso: 1º ano, 2º ano, 3º ano
Presenças: Nº de presenças às aulas (0-12)
TPC: Nº de TPC realizados (0-8)
Teste_1: Classificação no 1º teste (0-20)
Teste_2: Classificação no 2º teste (0-20)

Nota: Estes dados encontram-se no ficheiro SPSS com o nome "2_avaliação_continua_Exercicio.sav"

	A	B	C	D	E	F	G
1	Aluno	Género	Ano Curso	Presenças	T.P.C.	Teste_1	Teste_2
2	1	Masculino	1	1	0	2,00	0,00
3	2	Masculino	1	6	4	8,10	5,70
4	3	Masculino	1	5	0	3,00	9,80
5	4	Masculino	2	0	0	5,00	3,90
6	5	Masculino	1	2	0	4,50	12,50
7	6	Masculino	1	4	4	6,50	11,00
8	7	Masculino	2	3	0	5,30	8,30
9	8	Masculino	1	2	2	3,70	5,00
10	9	Masculino	1	8	0	2,00	3,50
11	10	Masculino	1	6	2	11,80	11,40
12	11	Feminino	1	11	8	11,00	19,70
13	12	Feminino	1	7	2	0,50	3,80
14	13	Feminino	2	1	0	0,50	5,30
15	14	Feminino	1	12	8	14,80	15,80
16	15	Feminino	1	8	2	7,30	12,10
17	16	Feminino	1	9	8	16,10	18,40
18	17	Feminino	1	11	0	8,80	12,80
19	18	Feminino	1	8	4	6,40	17,40
20	19	Feminino	2	11	6	17,40	14,90
21	20	Feminino	1	12	8	14,50	15,50
22	21	Feminino	1	4	0	0,00	0,80
23	22	Feminino	1	8	6	8,60	7,40
24	23	Feminino	1	12	8	18,80	20,00
25	24	Feminino	2	9	0	15,10	12,00
26	25	Feminino	1	2	0	9,30	3,30
27	26	Feminino	2	2	4	10,10	11,50
28	27	Feminino	1	9	2	5,10	10,50

Estatísticas Descritivas

Os dados também podem ser caraterizados pela obtenção de medidas estatísticas designadas por **Estatísticas Descritivas.**

O cálculo das estatísticas descritivas **depende da escala** em que os dados estão expressos.

As estatísticas descritivas mais frequentes são:

a) Medidas de localização:

Permitem caraterizar a ordem de grandeza dos dados.

b) Medidas de dispersão:

Permitem quantificar a variabilidade dos dados.

Medidas de Localização

As medidas de localização podem ser classificadas em

a) **Medidas de Tendência Central**
As medidas de tendências central indicam o valor central em torno do qual se distribuem os restantes dados em estudo. Na maior parte dos casos esse valor (central) é aquele em torno do qual se agrupam os dados da distribuição.

b) **Medidas de Posição (ou de Tendência não Central)**
As medidas de posição indicam a "posição" de uma determinada observação relativamente às restantes.

Medidas de Tendência Central: Moda

A **moda** (Mo) consiste na observação mais frequente da amostra.

No caso em que os dados estão agrupados em classes fala-se em **classe modal** e representa a classe com frequência absoluta mais elevada.

A moda é a medida de localização menos usada, embora possa ser determinada para **qualquer tipo de dados** estatísticos.

Propriedades da Moda:
- A moda é bastante simples de calcular, no entanto tem o inconveniente de poder não ser única ou mesmo não existir.
- A moda depende da frequência das observações.
- A moda não é afetada por valores extremos.

Medidas de Tendência Central: Moda - Exemplos

Considere as tabelas de frequências obtidas para os dados associados à competência leitora dos alunos:

Genero

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Masculino	13	43,3	43,3	43,3
	Feminino	17	56,7	56,7	100,0
	Total	30	100,0	100,0	

A Moda da variável "Género" corresponde ao género Feminino.

Grau de Escolaridade do Encarregado de Educação

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Ensino Básico	8	26,7	26,7	26,7
	Ensino Secundário	12	40,0	40,0	66,7
	Ensino Superior	10	33,3	33,3	100,0
	Total	30	100,0	100,0	

A Moda da variável "Grau de Escolaridade do Encarregado de Educação" corresponde ao Ensino Secundário.

Medidas de Tendência Central: Moda - Exemplos

Número de Irmãos

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	,00	8	26,7	26,7	26,7
	1,00	12	40,0	40,0	66,7
	2,00	6	20,0	20,0	86,7
	3,00	4	13,3	13,3	100,0
	Total	30	100,0	100,0	

A Moda da variável "Nº de irmãos" corresponde a 1 irmão.

Pto médio da classe	Classe	n_i	F_i	N_i	F_i
5.3	[4.0,6.6[	6	6/30 =0.20	6	0.20
7.9	[6.6,9.2[	7	7/30 =0.23	13	0.43
10.5	[9.2, 11.8[	3	3/30=0.10	16	0.53
13.1	[11.8, 14.4[	5	5/30=0.17	21	0.70
15.7	[14.4, 17.0[	4	4/30=0.13	25	0.83
18.3	[17.0, 19.6[	5	5/30=0.17	30	1.00

A classe modal da variável "Competência Leitora" corresponde à classe [6.6, 9.2[.

Medidas de Tendência Central: Mediana

A mediana (Me) é um valor que divide o **conjunto ordenado** de observações em duas partes iguais.

A mediana só pode ser determinada para dados expressos numa escala **pelo menos ordinal**.

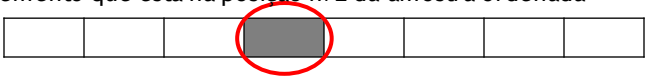
Propriedades da Mediana:
- A mediana é **única**.
- A mediana depende das **posições** ocupadas pelas observações.
- A mediana depende unicamente das **observações centrais**.

A forma de determinar a mediana, depende da escala em que os dados estão expressos:

Medidas de Tendência Central: Mediana

- Se os dados estão expressos numa escala ordinal

Isto é, se a amostra tem dimensão par, a mediana corresponde ao elemento que está na posição n/2 da amostra ordenada



Se a dimensão da amostra é ímpar, a mediana corresponde ao valor central das observações ordenadas;

Medidas de Tendência Central: Mediana

- Se os dados estão expressos numa escala quantitativa (discreta)

Isto é, se a amostra tem dimensão par, a mediana corresponde à média aritmética dos dois valores centrais.

Se a dimensão da amostra é ímpar, a mediana corresponde ao valor central das observações ordenadas;

Medidas de Tendência Central: Mediana - Exemplos

Considere novamente a tabela de frequências obtidas anteriormente para a variável "Grau de Escolaridade do Encarregado de Educação":

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Ensino Básico	8	26,7	26,7	26,7
	Ensino Secundário	12	40,0	40,0	66,7
	Ensino Superior	10	33,3	33,3	100,0
	Total	30	100,0	100,0	

Como se trata de uma variável expressa numa escala ordinal e a dimensão da amostra é par (n=30), a mediana corresponde ao 15º elemento da amostra ordenada.
Logo, Me = "Ens. Secundário"

Medidas de Tendência Central: Mediana - Exemplos

Para a variável "Número de irmãos":

Número de Irmãos

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	,00	8	26,7	26,7	26,7
	1,00	12	40,0	40,0	66,7
	2,00	6	20,0	20,0	86,7
	3,00	4	13,3	13,3	100,0
	Total	30	100,0	100,0	

Como se trata de uma variável expressa numa escala quantitativa (discreta) e a dimensão da amostra é par (n=30), a mediana corresponde à media aritmética dos dois valores centrais da amostra ordenada. Logo,

Assim, Me = "1 irmão"

Medidas de Tendência Central: Média

A Média só pode ser determinada para dados do tipo **quantitativo**

Propriedades da Média:
- A média é única.
- A média depende do valor de cada observação.
- A média é afetada por valores extremos.

A média (aritmética) de uma amostra é obtida dividindo a soma de todos os valores da amostra pelo número de elementos que constitui essa amostra (n), isto é,

Medidas de Tendência Central: Média - Exemplos

Considere as tabelas de frequências obtidas para os dados apresentados na página 140 associados à competência leitora dos alunos:

Número de Irmãos

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	,00	8	26,7	26,7	26,7
	1,00	12	40,0	40,0	66,7
	2,00	6	20,0	20,0	86,7
	3,00	4	13,3	13,3	100,0
	Total	30	100,0	100,0	

A média é dada por,

Medidas de Posição

De uma forma geral as **medidas de posição** são designadas por **quantis**, isto é, são medidas que permitem dividir (uma amostra ordenada) em várias partes iguais.

Os quantis mais frequentemente utilizados são:

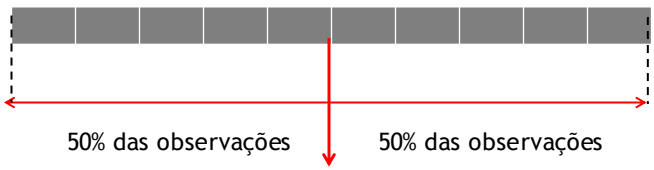
Quartis – (Q_1, Q_2, Q_3)
Dividem a mostra em quatro partes iguais

Decis – (D_1, D_2, …, D_9)
Dividem a amostra em 10 partes iguais

Percentis – (P_1, P_2, …, P_{99})
Dividem a amostra em cem partes iguais

Medidas de Posição

Dizemos que o percentil de ordem p toma o valor *a* ($P_p = a$), quando p% das observações que são inferiores ou iguais a *a* e (100-p)% das observações são superiores ou iguais a *a*.



$Me = Q_2 = P_{50}$

Medidas de Posição - Exemplos

Com base na informação fornecida através da tabela de frequências para a variável "Grau de escolaridade do encarregado de educação", podem-se determinar os quartis.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Ensino Básico	8	26,7	26,7	26,7
	Ensino Secundário	12	40,0	40,0	66,7
	Ensino Superior	10	33,3	33,3	100,0
	Total	30	100,0	100,0	

O primeiro quartil (Q_1) corresponde ao percentil 25
O segundo quartil (Q_2) corresponde ao percentil 50 ou à mediana
O terceiro quartil (Q_3) corresponde ao percentil 75

Medidas de Posição - Exemplos

Como interpretar estes valores?

- Q_1 = "Ensino Básico"
  Para 25% dos inquiridos, o nível de escolaridade do Encarregado de Educação é inferior ou igual ao "ensino básico"
  Para 75% dos inquiridos, o nível de escolaridade do Encarregado de Educação é superior ou igual ao "ensino básico"
- Q_2 = "Ensino Secundário"
  Para 50% dos inquiridos, o nível de escolaridade do Encarregado de Educação é inferior ou igual ao "ensino secundário"
  Para 50% dos inquiridos, o nível de escolaridade do Encarregado de Educação é superior ou igual ao "ensino secundário"
- Q_3 = "Ensino Superior"
  Para 75% dos inquiridos, o nível de escolaridade do Encarregado de Educação é inferior ou igual ao "ensino superior"
  Para 25% dos inquiridos, o nível de escolaridade do Encarregado de Educação é superior ou igual ao "ensino superior"

Medidas de Posição - Exemplos

Número de Irmãos

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	,00	8	26,7	26,7	26,7
	1,00	12	40,0	40,0	66,7
	2,00	6	20,0	20,0	86,7
	3,00	4	13,3	13,3	100,0
	Total	30	100,0	100,0	

1º quartil:
O 1º Quartil corresponde a Zero irmãos

2º quartil:
O 2º Quartil ou mediana corresponde a 1 irmão

3º quartil:
O 3º Quartil corresponde a 2 irmãos

Medidas Localização – Resumo

As medidas de localização permitem caraterizar a ordem de grandeza dos dados.

Medida de localização			Escala		
			Nominal	Ordinal	Quantitativa
Tendência	Central	Moda	✓	✓	✓
		Mediana		✓	✓
		Média			✓
	Não central	Quantis		✓	✓

Medidas de Dispersão

As medidas de dispersão quantificam a dispersão, isto é, a variabilidade dos dados amostrais. Estas medidas, juntamente com as medidas de localização, permitem uma melhor "descrição" das amostras.

Considerem-se os seguintes conjuntos de dados que representam idades de 5 indivíduos (em anos):

Amostra 1	14	15	18	18	25		Mo = 18	Me = 18	
Amostra 2	10	18	18	18	26		Mo = 18	Me = 18	
Amostra 3	18	18	18	18	18		Mo = 18	Me = 18	

Estes dados apresentam a mesma moda, mediana e média, no entanto, os valores que constituem as amostras são diferentes. As medidas de dispersão vão realçar estas diferenças.

Medidas de Dispersão: Amplitude de variação total

A **amplitude de variação total** pode ser calculada para dados expressos numa escala quantitativa.

Esta medida de dispersão indica a diferença entre o valor máximo e mínimo da amostra, isto é, $R = Max(X_i) - Min(X_i)$.

Propriedades da amplitude de variação total:
- A amplitude de variação é muito fácil de calcular.
- A amplitude de variação total tem em conta apenas os valores extremos
- A amplitude de variação total é única.

						AVT
Amostra 1	14	15	18	18	25	11.0
Amostra 2	10	18	18	18	26	16.0
Amostra 3	18	18	18	18	18	0.00

Medidas de Dispersão: Amplitude de Variação Quartílica

As amplitudes de variação **interquartílica** e **semi-interquartílica** podem ser calculadas para dados expressos numa escala quantitativa.

A **amplitude de variação interquartílica** é dada pela diferença entre o terceiro e o primeiro quartil, isto é,

$$AIQ = Q_3 - Q_1,$$

A **amplitude de variação semi-interquartílica** representa a "semi-distância" entre o 1º e o 3º quartil, isto é,

$$ASIQ = (Q_3 - Q_1)/2$$

Nota: No caso de dados ordinais, os valores que estas diferenças apresentam não têm qualquer significado matemático. Logo não tem qualquer sentido calcular estas medidas de dispersão!

Medidas de Dispersão: Amplitude de Variação Quartílica - exemplo

						AVT	AIQ	ASIQ
Amostra 1	14	15	18	18	25	11.0	3.00	1.50
Amostra 2	10	18	18	18	26	16.0	0.00	0.00
Amostra 3	18	18	18	18	18	0.00	0.00	0.00

Propriedades da amplitude de variação quartílica:
- A amplitude de variação interquartílica e semi-interquartílica são **únicas**.
- A amplitude de variação interquartílica e semi-interquartílica **não dependem dos valores extremos**.
- A amplitude de variação interquartílica e semi-interquartílica **dependem das posições dos quartis**.

Medidas de Dispersão: Variância e desvio-padrão

A **variância** mede a dispersão das observações em torno da média, isto é, é uma medida de dispersão relativamente à média.

O **desvio-padrão** é a raiz quadrada da variância.

Este tipo de medidas de dispersão só pode ser calculado para dados do tipo quantitativo.

Propriedades da variância:
- A variância mede a dispersão dos valores em relação à média.
- A variância depende do valor de cada observação.
- A variância é única.
- A variância é sempre um valor não negativo.

Medidas de Dispersão: Variância e desvio-padrão

A **variância** de uma amostra pode ser determinada através de:

$$\frac{(\quad)}{\quad} \quad \frac{\quad}{\quad}$$

O **desvio-padrão** é a raiz quadrada da variância.

						AVT	AIQ	ASIQ	S'	S'^2
Amostra 1	14	15	18	18	25	11.0	3.00	1.50	2.07	4.30
Amostra 2	10	18	18	18	26	16.0	0.00	0.00	2.38	5.66
Amostra 3	18	18	18	18	18	0.00	0.00	0.00	0.00	0.00

Medidas de Dispersão: Coeficiente de Variação - Exemplo

O Coeficiente de Variação é uma outra medida de dispersão que pode ser determinada pelo quociente entre o desvio-padrão e a média, isto é,

$$\frac{\quad}{\bar{\quad}}$$

						AVT	AIQ	ASIQ	S'	S'^2	CV
Amostra 1	14	15	18	18	25	11.0	3.00	1.50	2.07	4.30	0.115
Amostra 2	10	18	18	18	26	16.0	0.00	0.00	2.38	5.66	0.132
Amostra 3	18	18	18	18	18	0.00	0.00	0.00	0.00	0.00	0.000

Medidas de Dispersão

Considerando novamente os dados que representam as idades de 5 indivíduos:
Embora as medidas de localização não evidenciassem diferenças,

Amostra 1	14	15	18	18	25	Mo = 18	Me = 18	
Amostra 2	10	18	18	18	26	Mo = 18	Me = 18	
Amostra 3	18	18	18	18	18	Mo = 18	Me = 18	

as medidas de dispersão já vão permitir detetar diferenças entre as amostras:

						AVT	AIQ	ASIQ	S'	S'^2	CV
Amostra 1	14	15	18	18	25	11.0	3.00	1.50	2.07	4.30	0.115
Amostra 2	10	18	18	18	26	16.0	0.00	0.00	2.38	5.66	0.132
Amostra 3	18	18	18	18	18	0.00	0.00	0.00	0.00	0.00	0.000

Medidas de Dispersão - Resumo

As **medidas de Dispersão** permitem quantificar a variabilidade dos dados.

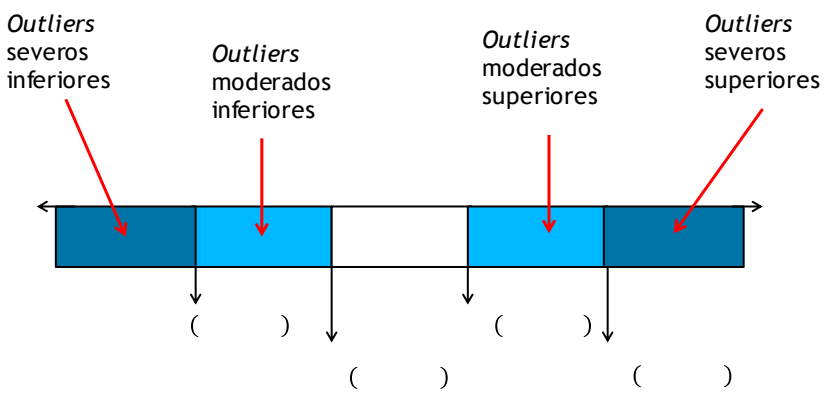
Medida de dispersão	Escala Nominal	Ordinal	Quantitativa
Amplitude de variação total (Max-Min)			✓
Amplitude de variação semi-interquartílica			✓
Variância			✓
Desvio-padrão			✓
Coeficiente de variação			✓

Medidas de Dispersão - *outliers*

Além destas medidas também é usual fazer referência à existência de observações que se afastam bastante das restantes. Estas observações são designadas por ***outliers*** (moderados ou severos) e podem ser detetados da seguinte forma:

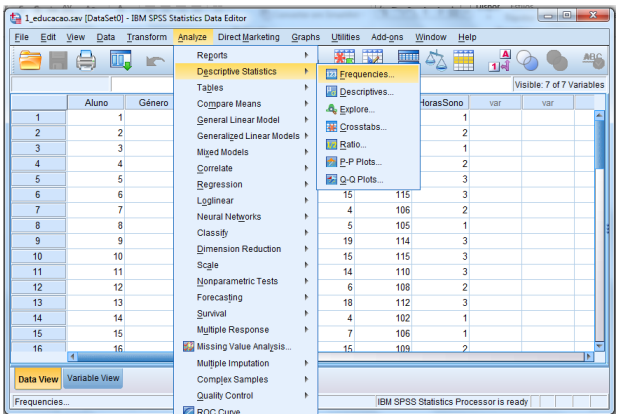
- Moderado inferior:
  ( ) ( )
- Moderado superior:
  ( ) ( )
- Severo inferior:
  ( )
- Severo superior:
  ( )

Medidas de Dispersão - *outliers*

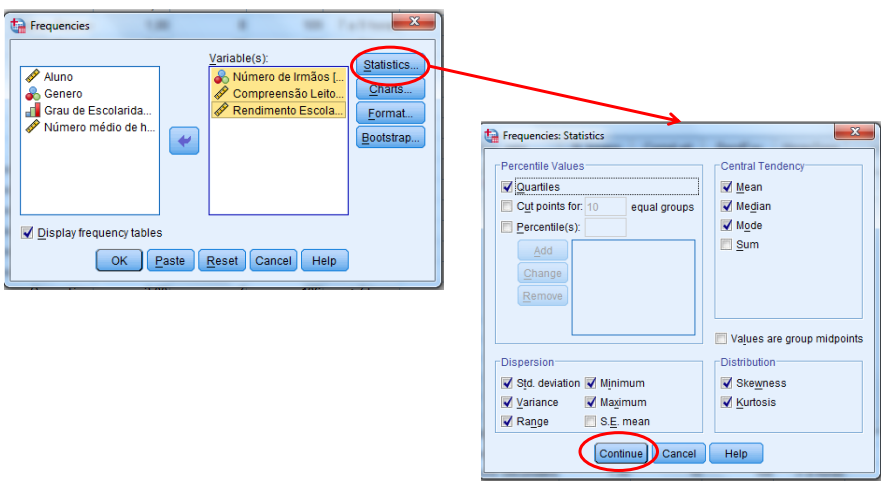


Estatística descritiva univariada

No SPSS as estatísticas descritivas univariadas para as variáveis quantitativas apresentadas no ficheiro 1_educacao.sav podem ser obtidas através dos comandos:



Estatística descritiva univariada



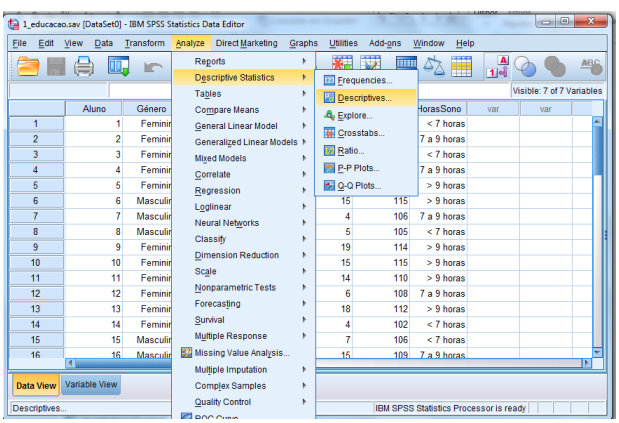
Estatística descritiva univariada

Statistics

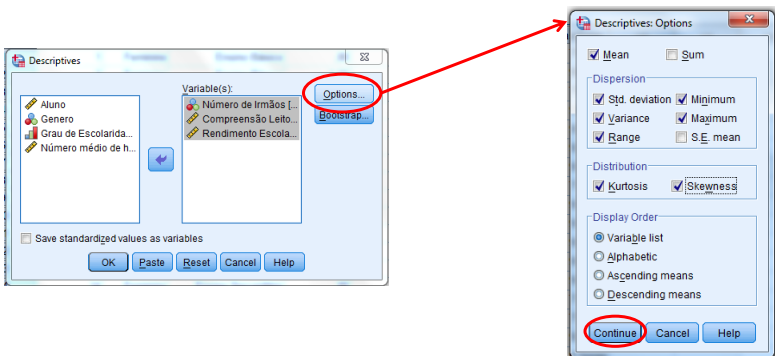
		Número de Irmãos	Compreensão Leitora	Rendimento Escolar do Aluno
N	Valid	30	30	30
	Missing	0	0	0
Mean		1,2000	11,17	108,57
Median		1,0000	10,50	109,00
Mode		1,00	15	109
Std. Deviation		,99655	4,706	3,866
Variance		,993	22,144	14,944
Skewness		,466	,111	,054
Std. Error of Skewness		,427	,427	,427
Kurtosis		-,711	-1,165	-,461
Std. Error of Kurtosis		,833	,833	,833
Range		3,00	15	13
Minimum		,00	4	102
Maximum		3,00	19	115
Percentiles	25	,0000	7,75	106,00
	50	1,0000	10,50	109,00
	75	2,0000	15,00	110,25

Estatística descritiva univariada

Ou alternativamente, através de



Estatística descritiva univariada



Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Número de Irmãos	30	3,00	,00	3,00	1,2000	,99655	,993	,466	,427	-,711	,833
Compreensão Leitora	30	15	4	19	11,17	4,706	22,144	,111	,427	-1,165	,833
Rendimento Escolar do Aluno	30	13	102	115	108,57	3,866	14,944	,054	,427	-,461	,833
Valid N (listwise)	30										

Estatística descritiva univariada - Exercício

Considere o novamente o conjunto de dados referente à avaliação contínua de 27 estudantes.
Obtenha as medidas de localização, de dispersão, de assimetria e de achatamento mais adequadas para os diferentes tipos de dados.

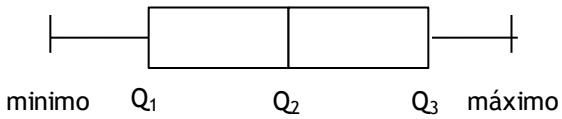
LEGENDA:
Genero: Masculino, Feminino
Ano Curso: 1º ano, 2º ano
Presenças: Nº de presenças às aulas (0-12)
TPC: Nº de TPC realizados (0-8)
Teste_1: Classificação no 1º teste (0-20)
Teste_2: Classificação no 2º teste (0-20)

Nota: Estes dados encontram-se no ficheiro SPSS com o nome 2_avaliação_continua.sav

	A	B	C	D	E	F	G
1	Aluno	Género	Ano Curso	Presenças	T.P.C.	Teste_1	Teste_2
2	1	Masculino	1	1	0	2,00	0,00
3	2	Masculino	1	6	4	8,10	5,70
4	3	Masculino	1	5	0	3,00	9,80
5	4	Masculino	2	0	0	5,00	3,90
6	5	Masculino	1	2	0	4,50	12,50
7	6	Masculino	1	4	4	6,50	11,00
8	7	Masculino	2	3	0	5,30	8,30
9	8	Masculino	1	2	2	3,70	5,00
10	9	Masculino	1	8	0	2,00	3,50
11	10	Masculino	1	6	2	11,80	11,40
12	11	Feminino	1	11	8	11,00	19,70
13	12	Feminino	1	7	2	0,50	3,80
14	13	Feminino	2	1	0	0,50	5,30
15	14	Feminino	1	12	8	14,80	15,80
16	15	Feminino	1	8	2	7,30	12,10
17	16	Feminino	1	9	8	16,10	18,40
18	17	Feminino	1	11	0	8,80	12,80
19	18	Feminino	1	8	4	6,40	17,40
20	19	Feminino	2	11	6	17,40	14,90
21	20	Feminino	1	12	8	14,50	15,50
22	21	Feminino	1	4	0	0,00	0,80
23	22	Feminino	1	8	6	8,60	7,40
24	23	Feminino	1	12	8	18,80	20,00
25	24	Feminino	2	9	0	15,10	12,00
26	25	Feminino	1	2	0	9,30	3,30
27	26	Feminino	2	2	4	10,10	11,50
28	27	Feminino	1	9	2	5,10	10,50

Diagrama de extremos e quartis

Para representar graficamente dados expressos numa escala pelo menos ordinal, é por vezes utilizado o **diagrama de extremos e quartis**, também designado por **caixa de bigodes** ou ***boxplot***.



Este tipo de representação gráfica permite obter informação sobre:

- Medidas de tendência central: Mediana
- Medidas de tendência não central: 1º e 3º quartil
- Medidas de dispersão: amplitude e distância interquartilica

Diagrama de extremos e quartis

Este diagrama também dá indicação sobre a possível existência de *outliers*:

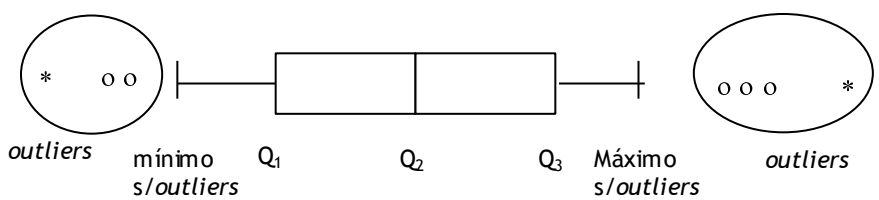
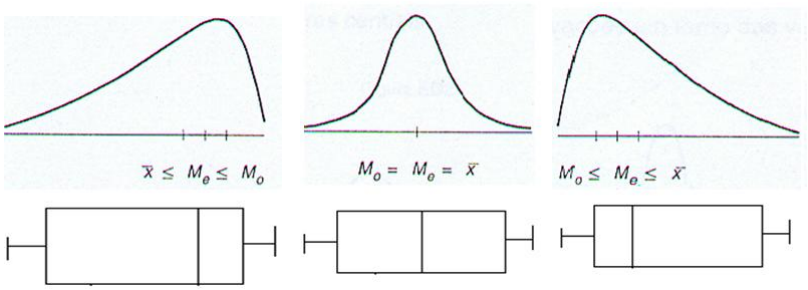


Diagrama de extremos e quartis

E sobre a assimetria da distribuição, tendo em conta a posição relativa da mediana e o comprimento dos "bigodes"



assimétrica negativa | simétrica | assimétrica positiva

Diagrama de extremos e quartis

No SPSS

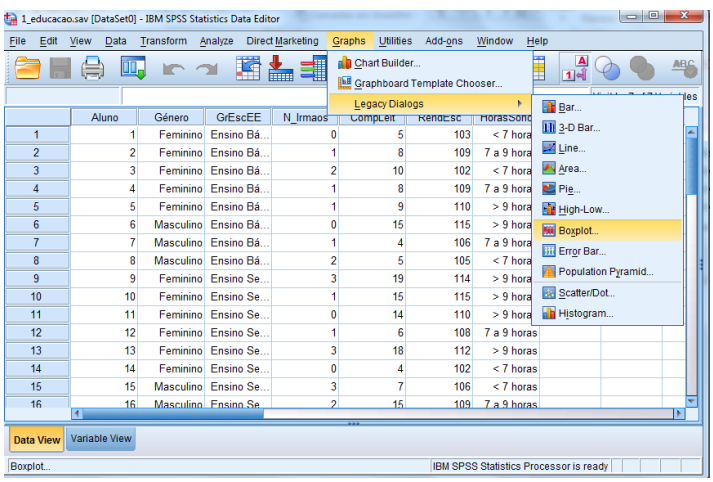


Diagrama de extremos e quartis

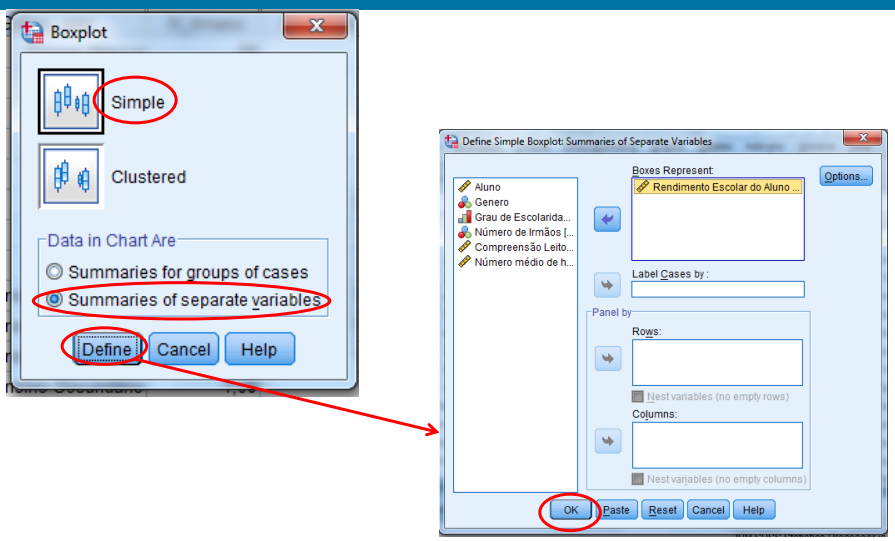
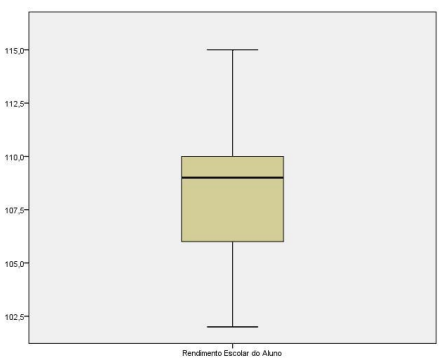


Diagrama de extremos e quartis



assimetria negativa

Diagrama de extremos e quartis- Exercício

Considere o novamente o conjunto de dados referente à avaliação contínua de 27 estudantes.
Obtenha obtenha o diagrama de extremos e quartis para as variáveis teste_1 e teste_2.

	A	B	C	D	E	F	G
1	Aluno	Género	Ano Curso	Presenças	T.P.C.	Teste_1	Teste_2
2	1	Masculino	1	1	0	2,00	0,00
3	2	Masculino	1	6	4	8,10	5,70
4	3	Masculino	1	5	0	3,00	9,80
5	4	Masculino	2	0	0	5,00	3,90
6	5	Masculino	1	2	0	4,50	12,50
7	6	Masculino	1	4	4	6,50	11,00
8	7	Masculino	2	3	0	5,30	8,30
9	8	Masculino	1	2	2	3,70	5,00
10	9	Masculino	1	8	0	2,00	3,50
11	10	Masculino	1	6	2	11,80	11,40
12	11	Feminino	1	11	8	11,00	19,70
13	12	Feminino	1	7	2	0,50	3,80
14	13	Feminino	2	1	0	0,50	5,30
15	14	Feminino	1	12	8	14,80	15,80
16	15	Feminino	1	8	2	7,30	12,10
17	16	Feminino	1	9	8	16,10	18,40
18	17	Feminino	1	11	0	8,80	12,80
19	18	Feminino	1	8	4	6,40	17,40
20	19	Feminino	2	11	6	17,40	14,90
21	20	Feminino	1	12	8	14,50	15,50
22	21	Feminino	1	4	0	0,00	0,80
23	22	Feminino	1	8	6	8,60	7,40
24	23	Feminino	1	12	8	18,80	20,00
25	24	Feminino	2	9	0	15,10	12,00
26	25	Feminino	1	2	0	9,30	3,30
27	26	Feminino	2	2	4	10,10	11,50
28	27	Feminino	1	9	2	5,10	10,50

LEGENDA:
Genero: Masculino, Feminino
Ano Curso: 1º ano, 2º ano, 3º ano
Presenças: Nº de presenças às aulas (0-12)
TPC: Nº de TPC realizados (0-8)
Teste_1: Classificação no 1º teste (0-20)
Teste_2: Classificação no 2º teste (0-20)

Nota: Estes dados encontram-se no ficheiro SPSS com o nome 2_avaliação_continua_exercicio.sav

Diagrama de extremos e quartis- Exercício

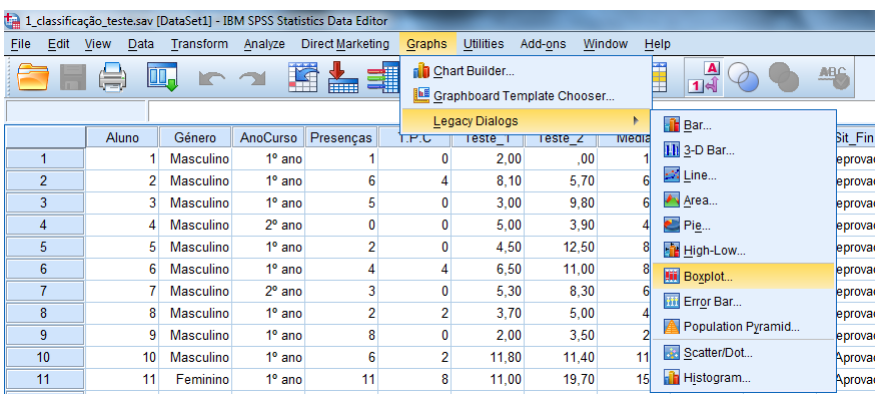


Diagrama de extremos e quartis- Exercício

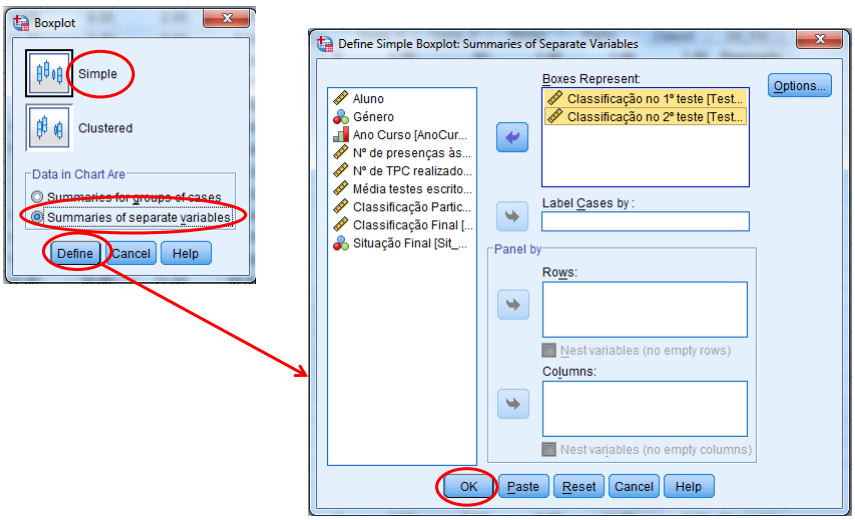
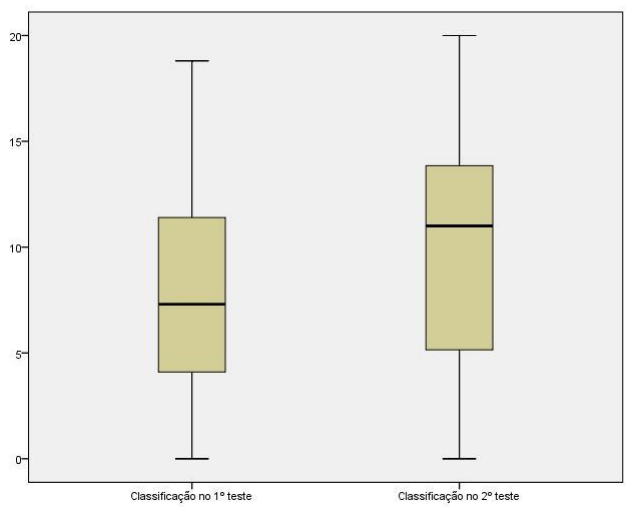


Diagrama de extremos e quartis- Exercício



Estatísticas Descritivas: Exercícios

Exercício:

Realizou-se um inquérito a 34 funcionários seleccionados aleatoriamente de uma dada empresa, tendo-se analisado as seguintes variáveis:

- **Antig**: Antiguidade do funcionário na empresa (menos de 5 anos, 5 a 10 anos, mais de 10 anos)
- **Conf_Emp**: Nível de Confiança dos funcionários relativamente à empresa (medido numa escala de 1 a 10)
- **Cred_Emp**: Nível de credibilidade da empresa transmitido ao público pelos seus funcionários (medido numa escala de 1 a 10)

Os valores obtidos encontram-se na tabela seguinte:

Nota: Estes dados encontram-se no ficheiro SPSS com o nome 3_funcionarios.sav

Estatísticas Descritivas: Exercícios

Ind	Antig	Com_Emp	Cred_Emp	Ind	Antig	Com_Emp	Cred_Emp
1	Mais de 10 anos	2	4	18	Mais de 10 anos	7	10
2	5 a 10 anos	6	9	19	5 a 10 anos	5	1
3	5 a 10 anos	1	5	20	Mais de 10 anos	3	2
4	Mais de 10 anos	9	10	21	Menos de 5 anos	9	5
5	5 a 10 anos	2	4	22	Menos de 5 anos	10	7
6	Mais de 10 anos	9	4	23	Mais de 10 anos	4	8
7	Mais de 10 anos	9	7	24	Menos de 5 anos	4	2
8	5 a 10 anos	9	6	25	5 a 10 anos	10	8
9	5 a 10 anos	8	7	26	Mais de 10 anos	7	9
10	Menos de 5 anos	5	5	27	5 a 10 anos	6	10
11	5 a 10 anos	4	10	28	5 a 10 anos	6	10
12	Mais de 10 anos	4	2	29	5 a 10 anos	4	2
13	Mais de 10 anos	1	2	30	Menos de 5 anos	2	3
14	5 a 10 anos	5	10	31	Mais de 10 anos	3	2
15	Mais de 10 anos	3	7	32	Menos de 5 anos	5	8
16	Mais de 10 anos	6	4	33	Menos de 5 anos	4	2
17	Mais de 10 anos	6	6	34	Mais de 10 anos	4	6

Estatísticas Descritivas: Exercícios

a) Para cada uma das variáveis, indique a sua escala de medida e as medidas de localização e de dispersão mais adequadas para as caraterizar.

b) Construa uma tabela de frequências para a variável "Antig".

c) Represente graficamente os valores associados à variável "Antig".

d) Determine os percentis de ordem 10 e 50 para as variáveis "Antig" e "Conf_Emp" e interprete esses valores.

e) Estude a existência de possíveis *outliers* para a variável "Cred_Emp" e construa o diagrama de extremos e quartis.

f) Comente a seguinte afirmação "*O nível de confiança dos funcionários relativamente à empresa apresenta uma maior variabilidade de valores do que o nível de credibilidade da empresa transmitido ao público por parte dos seus funcionários*".

g) O que pode concluir sobre a assimetria e achatamento das variáveis "Cred_Emp" e "Conf_Emp"?

Estatística Descritiva Bivariada

Estatística Descritiva Bivariada

A **Estatística Bivariada** considera o estudo simultâneo de duas variáveis estatísticas.

Cada indivíduo (objeto) é avaliado segundo duas modalidades, uma pertencendo à primeira variável e a outra à segunda variável, isto é,

Indivíduo	Variável X	Variável Y
1	X_1	Y_1
2	X_2	Y_2
...	...	...
n	X_n	Y_n

Cruzamento de variáveis Tabelas de Contingência

O resumo de dados bivariados depende do nível de mensuração das variáveis.

Os dados bivariados podem ser representados através de tabelas de dupla entrada designadas por **tabelas de contingência**.

X\Y	Y_1	Y_2	...	Y_j	total
X_1	O_{11}	O_{12}	...	O_{1j}	L_1
X_2	O_{21}	O_{22}	...	O_{2j}	L_2
...	...	...	...	...	...
X_i	O_{i1}	O_{i2}	...	O_{ij}	L_i
total	C_1	C_2	...	C_j	n

Distribuição marginal de X

Distribuição marginal de Y

Tabelas de Contingência

Na construção de tabelas de contingência para os dados do tipo quantitativo pode-se considerar os dados agrupados ou não agrupados.

No caso de os dados não estarem agrupados em classes, considera-se cada uma das variáveis em separado e recorre-se às diferentes técnicas de agrupamento de dados já estudados.

O número de classes a considerar para cada uma das variáveis pode não ser o mesmo e a amplitude das classes de uma variável não é necessariamente igual à amplitude das classes considerada para a outra variável.

Tabelas de Contingência - Exemplo

Considere o seguinte conjunto de dados já apresentado anteriormente:

Aluno	Género	Grau Escolaridade Enc. Educ.	Nº irmãos	Competência Leitora	Rendimento Escolar	Nº médio horas sono
1	Feminino	E. Básico	0	5	103	<7 horas
2	Feminino	E. Básico	1	8	109	7 a 9 horas
3	Feminino	E. Básico	2	10	102	<7 horas
4	Feminino	E. Básico	1	8	109	7 a 9 horas
5	Feminino	E. Básico	1	9	110	>9 horas
6	Masculino	E. Básico	0	15	115	>9 horas
7	Masculino	E. Básico	1	4	106	7 a 9 horas
8	Masculino	E. Básico	2	5	105	<7 horas
9	Feminino	E. Secundário	3	19	114	>9 horas
10	Feminino	E. Secundário	1	15	115	>9 horas
11	Feminino	E. Secundário	0	14	110	>9 horas
12	Feminino	E. Secundário	1	6	108	7 a 9 horas
13	Feminino	E. Secundário	3	18	112	>9 horas
14	Feminino	E. Secundário	0	4	102	<7 horas
15	Masculino	E. Secundário	3	7	106	<7 horas
16	Masculino	E. Secundário	2	15	109	7 a 9 horas
17	Masculino	E. Secundário	1	10	108	>9 horas
18	Masculino	E. Secundário	0	9	109	>9 horas
19	Masculino	E. Secundário	0	8	108	7 a 9 horas
20	Masculino	E. Secundário	0	11	107	7 a 9 horas
21	Feminino	E. Superior	1	12	109	>9 horas
22	Feminino	E. Superior	1	19	115	>9 horas
23	Feminino	E. Superior	0	14	111	>9 horas
24	Feminino	E. Superior	1	17	115	>9 horas
25	Feminino	E. Superior	2	18	110	>9 horas
26	Feminino	E. Superior	2	12	109	7 a 9 horas
27	Masculino	E. Superior	1	14	108	7 a 9 horas
28	Masculino	E. Superior	1	9	104	<7 horas
29	Masculino	E. Superior	2	5	107	<7 horas
30	Masculino	E. Superior	3	15	102	<7 horas

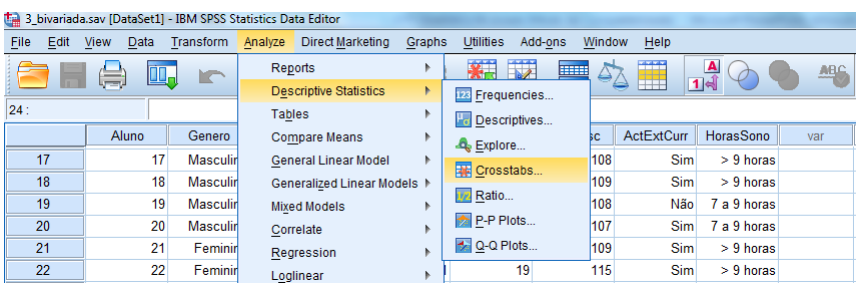
NOTA: Estes dados encontram-se no ficheiro de dados do SPSS com o nome 1_educacao.sav

Tabelas de Contingência - Exemplo

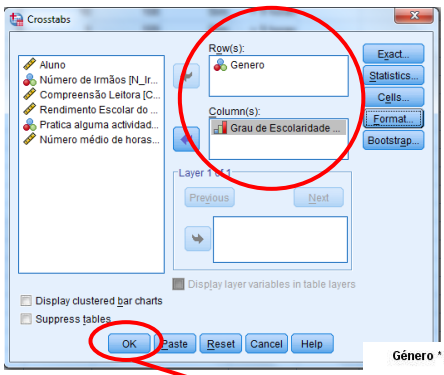
Comp leitora \ Ensino	Ensino Básico	Ensino Secundário	Ensino Superior	total
[4.0;6.6[	3	2	1	6
[6.6;9.2[	3	3	1	7
[9.2;11.8[	1	2	0	3
[11.8;14.4[	0	1	4	5
[14.4;17.0[	1	2	1	4
[17.0;19.6[	0	2	3	5
total	8	12	10	30

Tabelas de Contingência - Exemplo

No SPSS as tabelas de contingência podem ser obtidas através dos comandos:



Tabelas de Contingência - Exemplo



Género * Grau de Escolaridade do Encarregado de Educação Crosstabulation

Count

		Grau de Escolaridade do Encarregado de Educação			
		Ensino Básico	Ensino Secundário	Ensino Superior	Total
Género	Masculino	3	6	4	13
	Feminino	5	6	6	17
Total		8	12	10	30

Associação Estatística

A **"associação estatística"** entre duas variáveis pode ser estudada considerando:

- a forma de ligação de duas variáveis
  linear/não linear,
- a sua intensidade
  forte, média ou fraca
- o seu sentido
  positivo ou negativo

Associação Estatística

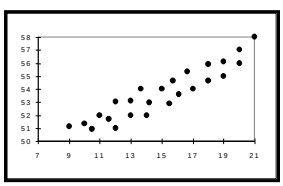
A "associação estatística" diz-se

- **positiva**
  se as variáveis variam no mesmo sentido, isto é, se para valores elevados de uma variável se observam valores elevados da outra e, simultaneamente, para valores reduzidos das duas variáveis é verificada a mesma associação.

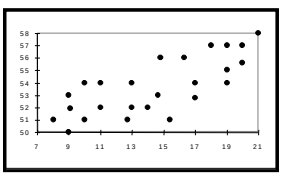
- **negativa**
  se as variáveis variarem em sentidos opostos, isto é, a valores elevados de uma variável estão associados valores baixos da outra variável e vice-versa.

Associação Estatística
Diagramas de Dispersão

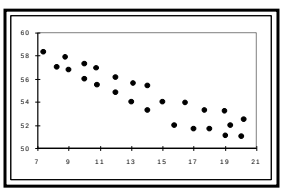
A existência (ou não) de associação estatística entre duas variáveis pode ser analisada graficamente através de **diagramas de dispersão:**



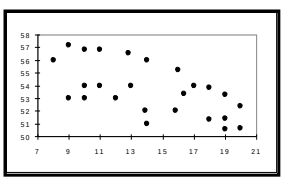
Associação linear positiva forte



Associação linear positiva fraca

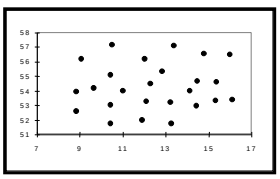


Associação linear negativa forte

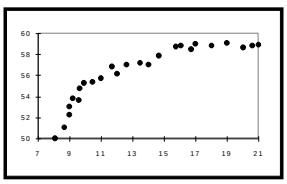


Associação linear negativa fraca

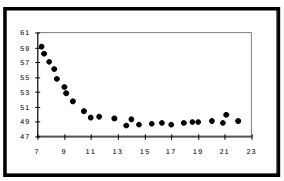
Associação Estatística
Diagramas de Dispersão



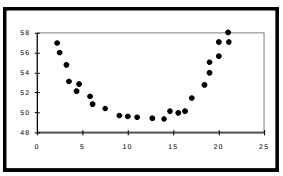
Não existe qualquer tipo de associação entre as variáveis



Existe associação entre as variáveis mas não do tipo linear



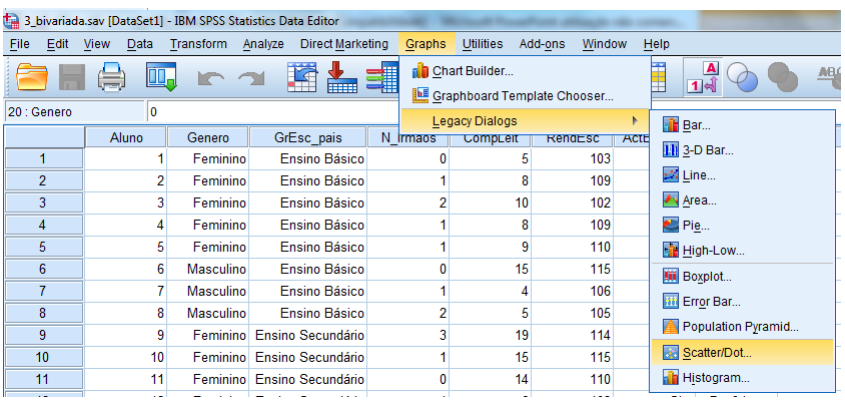
Existe associação entre as variáveis mas não do tipo linear



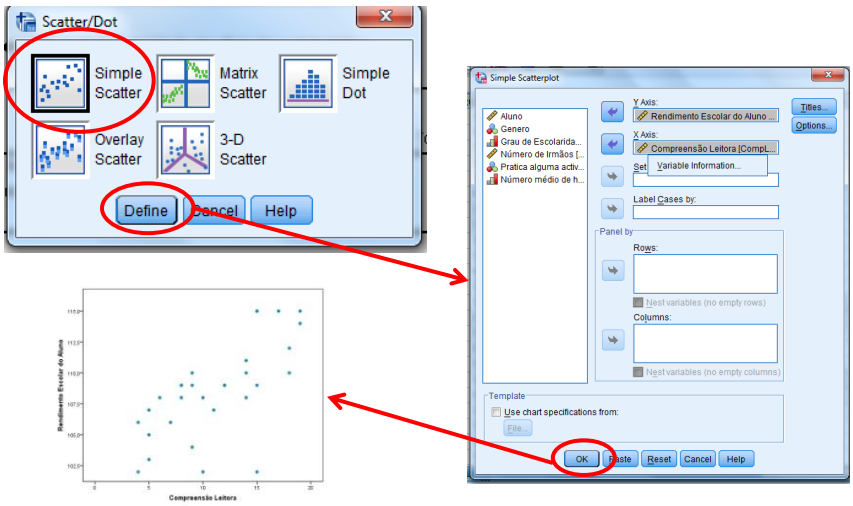
Existe associação entre as variáveis mas não do tipo linear

Diagramas de Dispersão - Exemplo

No SPSS, os diagramas de dispersão podem ser obtidos através dos comandos:



Diagramas de Dispersão - Exemplo



Exercícios

Considere o seguinte ficheiro de dados SPSS 2_avaliação_continua_exercício, referente à avaliação contínua de 27 estudantes a uma cadeira de estatística:

1_classificação_teste.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

	Aluno	Género	AnoCurso	Presenças	T.P.C	Teste_1	Teste_2	Media	Partic	Classif	Sit_Fin
1	1	Masculino	1º ano	1	0	2,00	,00	1,00	1,00	1,00	Reprovado
2	2	Masculino	1º ano	6	4	8,10	5,70	6,90	10,00	7,52	Reprovado
3	3	Masculino	1º ano	5	0	3,00	9,80	6,40	5,00	6,12	Reprovado
4	4	Masculino	2º ano	0	0	5,00	3,90	4,45	,00	3,56	Reprovado
5	5	Masculino	1º ano	2	0	4,50	12,50	8,50	2,00	7,20	Reprovado
6	6	Masculino	1º ano	4	4	6,50	11,00	8,75	8,00	8,60	Reprovado
7	7	Masculino	2º ano	3	0	5,30	8,30	6,80	3,00	6,04	Reprovado
8	8	Masculino	1º ano	2	2	3,70	5,00	4,35	4,00	4,28	Reprovado
9	9	Masculino	1º ano	8	0	2,00	3,50	2,75	8,00	3,80	Reprovado
10	10	Masculino	1º ano	6	2	11,80	11,40	11,60	8,00	10,88	Aprovado
11	11	Feminino	1º ano	11	8	11,00	19,70	15,35	19,00	16,08	Aprovado
12	12	Feminino	1º ano	7	2	,50	3,80	2,15	9,00	3,52	Reprovado
13	13	Feminino	2º ano	1	0	,50	5,30	2,90	1,00	2,52	Reprovado
14	14	Feminino	1º ano	12	8	14,80	15,80	15,30	20,00	16,24	Aprovado

Exercícios

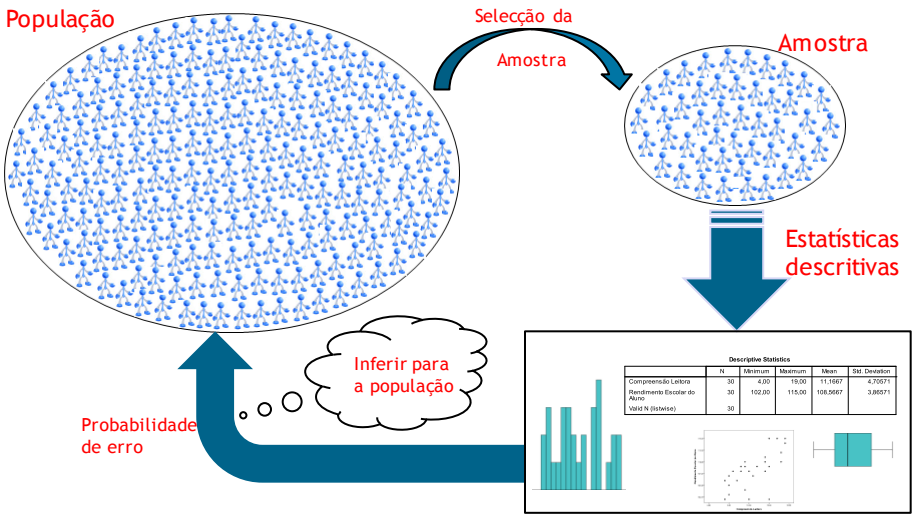
a) Estude a associação estatística existente entre a média nos testes escritos e a classificação obtida na participação.

b) Estude a associação estatística existente entre a classificação no 1º teste escrito e a nota final à cadeira. Comente os resultados obtidos.

c) Estude a associação estatística existente entre o ano do curso e a classificação final à cadeira. Comente os resultados obtidos.

Inferência Estatística

Introdução à Inferência Estatística



Introdução à Inferência Estatística

A **inferência estatística** tem por objetivo estimar parâmetros populacionais a partir do estudo de uma amostra.

Os **parâmetros** (representados por letras gregas) são caraterísticas populacionais que têm um valor exato, embora sejam normalmente desconhecidos. Os parâmetros são estimados por **estatísticas**.

As **estatísticas** (representadas por legras latinas) são caraterísticas amostrais que podem ser calculadas, embora possam ser diferentes de amostra para amostra.

Introdução à Inferência Estatística

Caraterística	Parâmetro (Populacional)	Estatística (amostral)
Dimensão	N	n
Valor médio	μ	$\hat{\mu}$ ou $\bar{x}$
Desvio padrão	s	$\hat{\sigma}$ ou s'
Proporção	p	$\hat{\pi}$ ou p
Coeficiente de Correlação	r	$\hat{\rho}$ ou r

Métodos de Estimação

Os "mecanismos" mais usuais para estimar esses parâmetros são:

Estimação Pontual:
Obter um valor numérico único (a partir da amostra) para estimar o correspondente parâmetro populacional.

Estimação Intervalar:
Obter um intervalo de valores que contenha o(s) parâmetro(s) desejado(s) com uma probabilidade especificada.

Testes de hipóteses:
Avaliar, através de técnicas estatísticas apropriadas, se uma determinada hipótese ou conjetura que se faz sobre os valores possíveis do(s) parâmetro(s) tem ou não razão de existir.

Estimação Pontual

A **estimação pontual** consiste em determinar um valor (único) para estimar o verdadeiro valor do parâmetro populacional desconhecido.

Este método tem bastantes desvantagens, uma vez que **não existe nenhum grau de certeza relativamente à qualidade da estimativa obtida.**

Introdução à Inferência Estatística

Exemplo:
Suponha uma população consistindo nas idades de 5 crianças:

6, 8, 10, 12, 14.

A estimação pontual de μ e σ^2, vai depender da amostra que se extrai:

Amostra	Estimativa pontual de μ	Estimativa pontual de σ^2
(6, 10, 14)	10.00	16.00
(14,12,14)	13.33	1.33
(8, 6, 14)	9.33	17.33
(6, 8, 10)	8.00	4.00

Amostra	Estimativa pontual de μ	Estimativa pontual de σ^2
(10, 8, 14)	10.67	9.33
(10, 8, 10)	9.33	1.33
(14, 6, 12)	10.67	17.33
(10, 14, 12)	12.00	4.00

Amostras diferentes produzem estatísticas diferentes!

Estimação Intervalar

A **estimação intervalar**, permite obter um determinado intervalo de valores que contém o verdadeiro valor do parâmetro populacional, com um certo grau de certeza.

Exemplo:

A classificação média num teste de matemática situa-se entre os 8 e 11 valores, com uma probabilidade de 95%.

Em cada 100 intervalos obtidos, 95 desses intervalos conterão o verdadeiro valor do parâmetro (apenas 5 desses intervalos não contêm o valor do parâmetro).

Estimação Intervalar

Resultados Oficiais		RTP1		SIC		TVI	
50,59		49,0	54,0	50,4	54,6	50,0	54,8
20,72		20,0	23,0	17,7	21,5	18,4	22,4
14,34		11,0	14,0	12,5	16,3	11,0	16,0
8,59		8,0	10,0	6,4	8,6	7,0	10,5
5,31		4,0	6,0	4,1	6,3	3,4	6,4
0,44		0,0	1,0	0,5	1,1	0,3	1,5
Sondagens		Melhor		Pior		Media	

Estimação Intervalar

Intervalo de Confiança a 95% =]49,0 ; 54,0[

Temos 95% de confiança de que a percentagem de votos no candidato Cavaco Silva esteja entre 49% e 54%.

Estimação Intervalar

Margem de Erro:

- é metade da amplitude do intervalo de confiança
- é a medida da precisão

Quanto menor a margem de erro maior a precisão da estimativa.

Num intervalo de confiança,

- a **margem de erro** indica a precisão da estimativa
- o **nível de confiança** indica-nos a confiança que temos em que o intervalo contenha o valor do parâmetro

Quanto maior for o intervalo, maior é o grau de confiança, mas menor a precisão da estimativa.

Intervalo de Confiança para a Proporção

Intervalo de confiança a 95% para a proporção

- quando se conhece a dimensão da população

$$IC_{95\%}: \left] \hat{p} - 1{,}96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}\left(1-\frac{n}{N}\right)}\ ;\ \hat{p} + 1{,}96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}\left(1-\frac{n}{N}\right)} \right[$$

onde *p* é a estimativa da proporção

proporção = percentagem / 100

- assumindo que a dimensão da população (N) é grande.

$$IC_{95\%} = \left] \hat{p} - 1{,}96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ;\ \hat{p} + 1{,}96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right[$$

Intervalo de Confiança para a Proporção

Exercício 1: Estudo sobre sites de notícias em Portugal

População: indivíduos registados nos vários sites analisados.

Estime os resultados para a população, sabendo que obtivemos com uma amostra de 300 indivíduos o seguinte resultado:

a) 45% dizem que recomendariam o site a amigos.

b) 20% dizem que consideram o site mal estruturado.

c) 20% dizem que consideram o site mal estruturado, neste caso a população considerada foi de 5000 indivíduos.

Intervalo de Confiança para a Proporção

Exercício 2: (Frequência de Estatística – Janeiro 2015)

Calcula e interpreta o intervalo de confiança a 95% para a proporção de alunos de Jornalismo que escolheram o curso pela vocação.

Crosstab

			Vocação, gosto pelas matérias		
			não escolheu	Escolheu	Total
Curso	Publicidade e Marketing	Count	21	82	103
		Adjusted Residual	1,0	-1,0	
	Publicidade e Marketing (regime pós-laboral)	Count	5	45	50
		Adjusted Residual	-1,4	1,4	
	Relações Públicas e Comunicação Empresarial	Count	30	74	104
		Adjusted Residual	3,6	-3,6	
	Relações Públicas e Comunicação Empresarial (regime pós-laboral)	Count	10	35	45
		Adjusted Residual	1,0	-1,0	
	Audiovisual e Multimédia	Count	17	121	138
		Adjusted Residual	-1,7	1,7	
	Jornalismo	Count	10	97	107
		Adjusted Residual	-2,4	2,4	
Total		Count	93	454	547

Testes de Hipóteses

Testes de hipóteses: Introdução

Teste de Hipótese

Objetivo:

Avaliar se uma determinada hipótese ou conjetura que se faz sobre um parâmetro ou população tem ou não razão de existir.

Nota:

Os testes de hipóteses só devem ser aplicados a amostras aleatórias.

Existem diferenças entre os resultados de uma amostra e os resultados da população de onde a amostra foi retirada, ou seja, existem diferenças entre as estimativas amostrais e os valores previamente fixados correspondentes a parâmetros populacionais ?

Testes de hipóteses: Generalidades

Para realizar um teste de hipóteses devem-se considerar os seguintes passos:

1º – Identificação do Teste

2º - Definição das hipóteses estatísticas

3º – Determinação da significância do teste (obtida através do *software*)

4º - Decisão

Se *sig.* $\leq \alpha \Rightarrow$ Rejeita-se H_0

Se *sig.* $> \alpha \Rightarrow$ Não se rejeita H_0

5º – Conclusão

Testes de hipóteses: Generalidades

1º) Identificação do Teste:

Este primeiro passo consiste na identificação do teste a utilizar.

Ao identificar o teste mais adequado para analisar a questão em estudo deve-se ter em conta os seguintes aspectos:

- quais as variáveis envolvidas no estudo,
- as escalas em que essas variáveis estão definidas,
- qual o parâmetro a avaliar (média, desvio-padrão, coeficiente de correlação, ….)
- os pressupostos (condições de aplicação) subjacentes ao teste que se pretende utilizar.

Testes de hipóteses: Generalidades

2º) Definição das hipóteses estatísticas:

Num teste de hipóteses há sempre duas hipóteses estatísticas em confronto:

Hipótese nula (H_0)	vs	**Hipótese alternativa** (H_1)
↓		↓
provisoriamente aceite como verdadeira e que é submetida a uma comprovação experimental		hipótese complementar à H_0.
↓		↓
contém sempre uma condição de igualdade ($=, \leq$ ou $\geq$).		contém sempre condição de não igualdade ($\neq$) ou uma desigualdade ($<$ ou $>$).

Testes de hipóteses: Generalidades

Exemplo:

Pretende-se analisar se a média das notas de acesso dos alunos de PM não difere da média das notas de acesso dos alunos de RPCE.

H_0:

vs

H_1:

Testes de hipóteses: Generalidades

3º) Determinação da significância do teste:

A significância do teste será obtida por recurso ao *software* estatístico SPSS.

4º) Decisão:

Quando se realiza um teste de hipóteses, pode-se tomar uma de duas decisões:

Decisão do teste { rejeita-se a hipótese nula ou não se rejeita a hipótese nula

Em ambos os casos corre-se o risco de errar.

Uma das caraterísticas dos testes de hipóteses é minimizar esse risco.

Testes de hipóteses: Generalidades

Ao tomar a decisão de rejeitar ou não rejeitar a hipótese nula, podem-se cometer os seguintes erros:

		Situação real	
		H_0 Verdadeira	H_0 Falsa
Decisão	Rejeitar H_0	**Erro tipo I**	Decisão correta
	Não rejeitar H_0	Decisão correta	**Erro tipo II**

Testes de hipóteses: Generalidades

Estes erros podem ser quantificados em termos de probabilidades:

- α = P(cometer erro tipo I) = P(rejeitar H_0 quando H_0 verdadeira)

Nível de significância

- **1-α** = P(não cometer erro tipo I) =
  = P(não rejeitar H_0 quando H_0 verdadeira)

Grau de Confiança

- β = P(cometer erro tipo II) = P(não rejeitar H_0 quando H_0 falsa)
- **1-β** = P(não cometer erro tipo II) = P(rejeitar H_0 quando H_0 falsa)

Potência do teste

Testes de hipóteses: Generalidades

Escolher a margem de erro associada a um teste de hipótese

Escolher o erro tipo I ou nível de significância

Os níveis de significância usuais são $\alpha = 0.01$; $\alpha = 0.05$ ou $\alpha = 0.10$

		Situação real	
		H_0 Verdadeira	H_0 Falsa
Decisão	Rejeitar H_0	**Erro tipo I (α)**	Decisão correta ($1-\beta$)
	Não rejeitar H_0	Decisão correta ($1-\alpha$)	**Erro tipo II (β)**

Testes de hipóteses: Generalidades

- A Estatística de teste é uma fórmula matemática que compara os dados amostrais com a suposição feita sobre a população (sob a validade de H_0). **Nota:** A Estatística de Teste é obtida através do *software* SPSS.

- A decisão do Teste consiste em comparar a significância do teste obtida através do *software* SPSS com o nível de significância pré-definido. Assim, se **sig. do teste > $\alpha \Rightarrow$ não rejeito H_0** **sig. do teste ≤ $\alpha \Rightarrow$ rejeito H_0**

5º) Conclusão:

- se a **hipótese nula não é rejeitada**, diz-se que os dados sobre os quais o teste foi realizado **não apresentam evidências suficientes para levar à rejeição desta hipótese**;
- se a **hipótese nula é rejeitada**, diz-se que os dados em estudo **não são compatíveis com a hipótese nula.**

Testes de Hipóteses

- Testes de independência
  - Teste de independência do Qui-Quadrado
- Testes ao coeficiente de Correlação
  - Teste ao coeficiente de Correlação de Pearson
  - Teste ao Coeficiente de Correlação de Spearman
- Teste de Comparação de Médias
  - Teste para comparação de médias a partir de 2 amostras independentes
  - Teste para comparação de médias a partir de mais de 2 amostras independentes.

Testes de Independência do Qui-Quadrado

Teste de Independência do Qui-Quadrado

❖ **Condições de aplicabilidade:**

- As variáveis podem estar expressas em qualquer escala, desde que categorizadas;
- A dimensão da amostra deve ser superior a 30 elementos;
- não pode haver mais de 20% das células com frequência esperada inferior a 5;
- cada célula tem de ter frequência esperada igual ou superior a 1.

❖ **Hipóteses:**

H_0: As variáveis X e Y são independentes

vs

H_1: As variáveis X e Y não são independentes

Teste de Independência do Qui-Quadrado

Exemplo:

Consideremos as variáveis

Acreditar na astrologia

Género

Género * Acredita na astrologia? Crosstabulation

Count

		Acredita na astrologia?				
		Nada	Pouco	Algo	Muito	Total
Género	Feminino	15	48	79	9	151
	Masculino	25	21	9	1	56
Total		40	69	88	10	207

Teste de Independência do Qui-Quadrado

1º) Identificação do teste a utilizar

Pretende-se verificar se existe independência entre duas variáveis "género" e "acreditar na astrologia", definidas numa escala qualitativa. O teste a utilizar (caso se verifiquem os pressupostos de aplicação) será o teste de independência do Qui-Quadrado.

2º) Definição das hipóteses

H_0: As variáveis "Género" e "Acreditar na astrologia" são independentes

vs

H_1: As variáveis "Género" e "Acreditar na astrologia" não são independentes

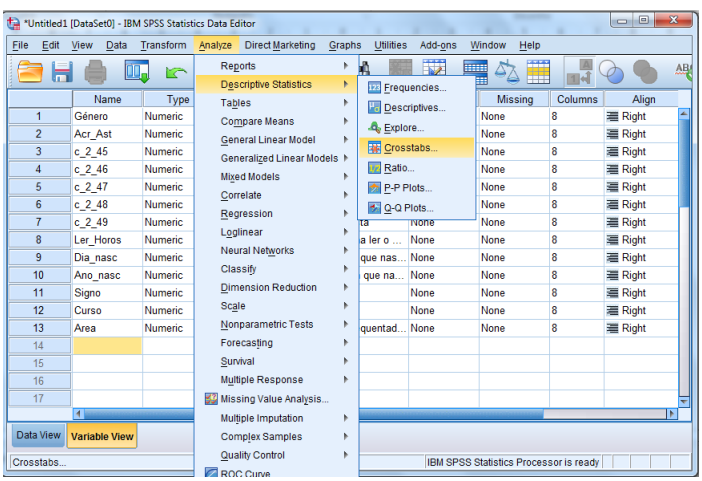
3ª) Determinação da significância do teste

Analyze

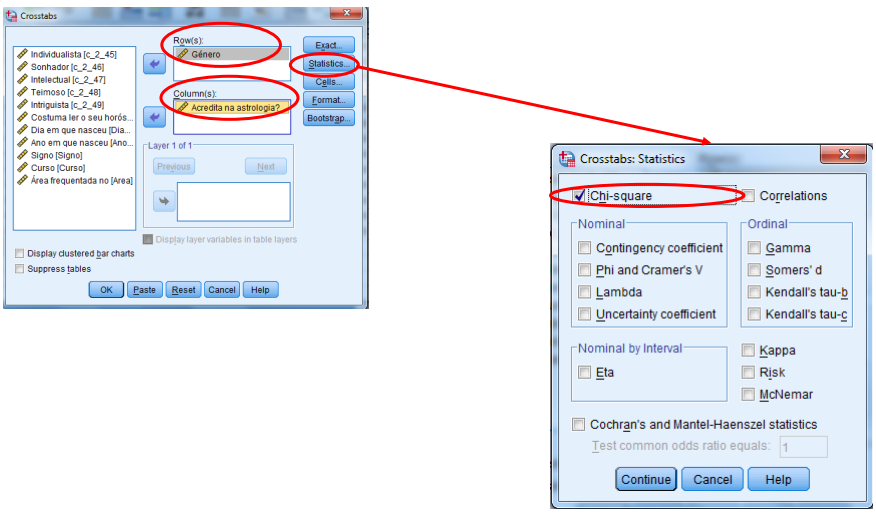
Descriptive Statistics

Crosstabs…

Teste de Independência do Qui-Quadrado



Teste de Independência do Qui-Quadrado



Teste de Independência do Qui-Quadrado

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	39,966[a]	3	,000
Likelihood Ratio	39,371	3	,000
Linear-by-Linear Association	36,334	1	,000
N of Valid Cases	207		

a. 1 cells (12,5%) have expected count less than 5. The minimum expected count is 2,71.

Teste de Independência do Qui-Quadrado

4ª) Decisão do teste

Como Sig < 0.001 < 0,05 $\Rightarrow$ Rejeita-se H_0

As variáveis "Género" e "Acreditar na astrologia" não são independentes.

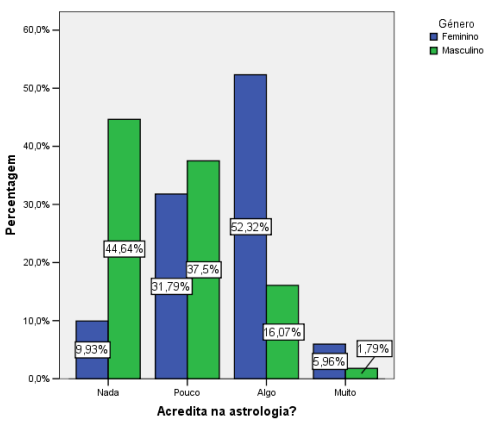
Nota: o fato de se ter rejeitado a hipótese nula não significa que as variáveis "género" e "acreditar na astrologia" sejam dependentes. Apenas se sabe que existe qualquer tipo de associação entre elas.

Podemos tentar explicar essa associação através da análise:

- gráfica
- de resíduos

Teste de Independência do Qui-Quadrado

Análise gráfica



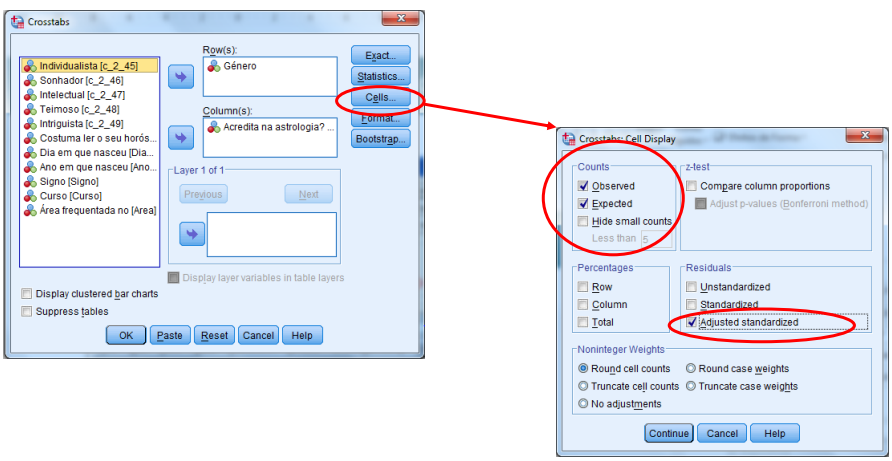
Teste de Independência do Qui-Quadrado

Análise de resíduos:

- Se H_0 for verdadeira, ou seja, se houver independência, os resíduos estariam entre **-2 e 2.**
- A existência de não independência é explicada pelos resíduos que se afastam muito de -2 e 2.

Teste de Independência do Qui-Quadrado

Análise de resíduos



Teste de Independência do Qui-Quadrado

4º) Decisão

Análise de resíduos

Género * Acredita na astrologia? Crosstabulation

			Acredita na astrologia?				
			Nada	Pouco	Algo	Muito	Total
Género	Feminino	Count	15	48	79	9	151
		Expected Count	29,2	50,3	64,2	7,3	151,0
		Adjusted Residual	-5,6	-,8	4,7	1,2	
	Masculino	Count	25	21	9	1	56
		Expected Count	10,8	18,7	23,8	2,7	56,0
		Adjusted Residual	5,6	,8	-4,7	-1,2	
Total		Count	40	69	88	10	207
		Expected Count	40,0	69,0	88,0	10,0	207,0

Teste de Independência do Qui-Quadrado

5º) Conclusão

Há uma maior tendência para os indivíduos do género feminino serem mais crentes na astrologia. O inverso é verificado relativamente aos indivíduos do género masculino que se mostram menos crentes.

Teste de Independência do Qui-Quadrado

Exemplos:

Considere a base de dados 4_Consumo.sav

a) Verifique se o fato de os alunos considerarem ser importante estar na moda é independente do género.

b) O fato de o aluno considerar que é divertido fazer compras é independente do género?

c) Será que existe independência entre as variáveis "Moda" e "Trabalhador"?

d) As variáveis "Orçamento mensal" e "Fazer compras é divertido" são independentes?

Testes aos coeficientes de correlação

Associação Estatística

A "associação estatística" entre duas variáveis pode ser estudada considerando:

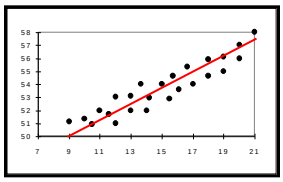
- a forma de ligação de duas variáveis
  linear/não linear,
- a sua intensidade
  forte, média ou fraca
- o seu sentido
  positivo ou negativo

Associação Estatística

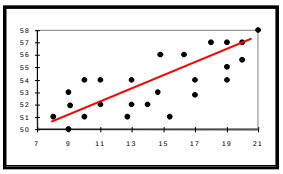
A "associação estatística" diz-se

- positiva
  se as variáveis variam no mesmo sentido, isto é, se para valores elevados de uma variável se observam valores elevados da outra e, simultaneamente, para valores reduzidos das duas variáveis é verificada a mesma associação.
- negativa
  se as variáveis variarem em sentidos opostos, isto é, a valores elevados de uma variável estão associados valores baixos da outra variável e vice-versa.

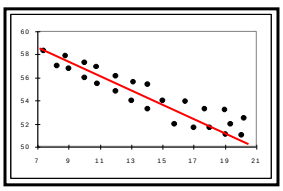
Associação Estatística



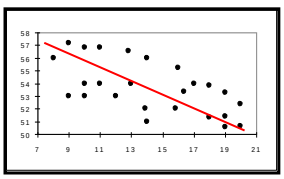
Associação linear positiva forte



Associação linear positiva fraca

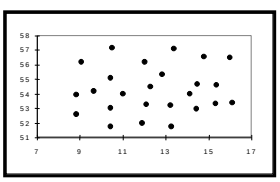


Associação linear negativa forte fraca

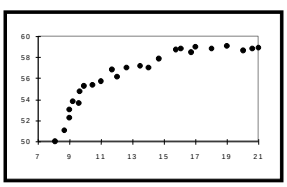


Associação linear negativa

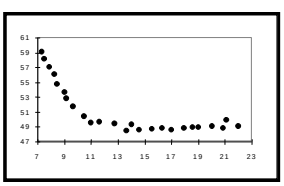
Associação Estatística



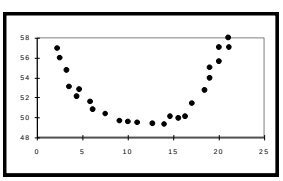
Não existe qualquer tipo de associação entre as variáveis



Existe associação entre as variáveis mas não do tipo linear



Existe associação entre as variáveis mas não do tipo linear



Existe associação entre as variáveis mas não do tipo linear

Associação Estatística

O grau de associação linear entre duas variáveis pode ser estudado através dos coeficientes de associação/correlação de

- Pearson (ρ)

Caso de variáveis quantitativas

- Spearman (ρ_S)

Caso de variáveis pelo menos ordinais

Coeficiente de associação de Pearson

O coeficiente de correlação de Pearson

- mede o grau de associação linear entre duas variáveis expressas numa escala quantitativa
- não depende das unidades de medida das variáveis
- os seus valores variam sempre entre –1 e 1

O sinal do coeficiente de correlação de Pearson indica a direcção da associação linear:

- se o sinal for **positivo**, existe uma tendência para as duas variáveis variarem no mesmo sentido;
- se o sinal for **negativo**, existe uma tendência para as duas variáveis variarem em sentido contrário

Coeficiente de associação de Pearson

De uma forma geral, pode-se considerar que:

- Se $r_{XY} = 1$ ou $r_{XY} = -1$ existe correlação linear perfeita
- Se $r_{XY} = 0$, não existe qualquer tipo de correlação linear entre as duas variáveis em estudo.
  (embora possa existir correlação de outro tipo que não o linear)
- Se $0 < | r_{XY} | < 0.3$, existe correlação linear baixa
- Se $0.3 \leq | r_{XY} | < 0.7$, existe correlação linear média
- Se $0.7 \leq | r_{XY} | < 1$, existe correlação linear forte

Teste de Pearson

1º) Identificação do teste

Pretende-se avaliar se existe associação estatística entre duas variáveis definidas numa escala quantitativa. O teste a realizar é o teste ao coeficiente de correlação de Pearson.

2º) Definição das hipóteses

H_0: Não existe correlação do tipo linear entre as variáveis X e Y
vs
H_1: Existe correlação do tipo linear entre as variáveis X e Y

ou então,

H_0: $\rho = 0$
vs
H_1: $\rho \neq 0$

Teste de Pearson

Exemplo:

Existirá alguma associação entre a cilindrada e as rotações de um automóvel?

1º) Identificação do teste

Pretende-se avaliar se existe associação estatística entre duas variáveis definidas numa escala quantitativa. O teste a realizar é o teste ao coeficiente de correlação de Pearson.

2º) Definição das hipóteses

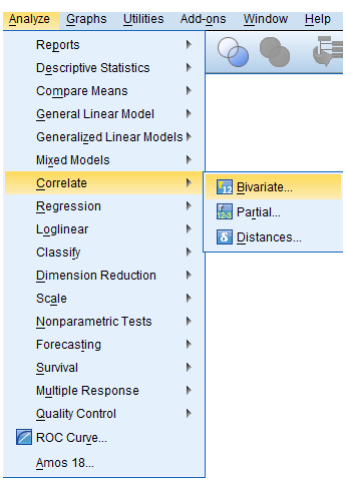
H_0:

vs

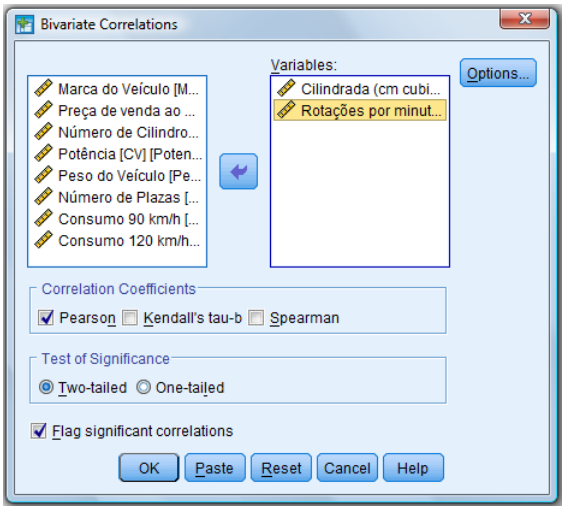
H_1:

Teste de Pearson

3º) Determinação da significância do teste



Teste de Pearson



Teste de Pearson

Correlations

		Cilindrada (cm cúbicos)	Rotações por minuto
Cilindrada (cm cúbicos)	Pearson Correlation	1	-,442**
	Sig. (2-tailed)		,000
	N	125	125
Rotações por minuto	Pearson Correlation	-,442**	1
	Sig. (2-tailed)	,000	
	N	125	125

**. Correlation is significant at the 0.01 level (2-tailed).

4º) Decisão do teste

Como sig < 0.001 (inferior a 0.05) rejeita-se a hipótese nula,

5º) Conclusão:

Existe associação (do tipo linear) entre a cilindrada e as rotações de um automóvel.

O valor do coeficiente de correlação de Pearson é -0,442, logo a associação é negativa média. Ou seja, os automóveis com mais rotações têm menos cilindrada

Teste de Pearson - Exercícios

Exercícios:

Considere os dados fornecidos no ficheiro *5_forest_fire.sav* referentes ao índice Meteorológico de Perigo de Incêndio (FWI - *Fire Weather Index*) registados no parque de Montesinho.

a) Será que se pode admitir que quando o índice de humidade dos combustíveis compactos (DMC) aumenta há tendência para o valor do índice de seca do sistema (DC) aumentar?

b) Verifique se existe alguma relação estatística do tipo linear entre a humidade relativa e a área de floresta ardida?

c) Nos dias de maior humidade relativa há tendência para uma diminuição da temperatura?

Coeficiente de associação de Spearman

O coeficiente de correlação de Spearman é um caso particular do coeficiente de correlação de Pearson, aplicado a variáveis expressas numa escala pelo menos ordinal.

As propriedades do coeficiente de correlação de Spearman são idênticas às do coeficiente de correlação de Pearson.

- O valor absoluto do coeficiente de correlação de Spearman mede o grau de associação linear de duas variáveis expressas numa escala pelo menos ordinal.
- O coeficiente de correlação de Spearman está sempre entre -1 e 1.
- Se o coeficiente de correlação de Spearman tomar o valor zero, não existe qualquer tipo de associação linear entre as variáveis em estudo.
- O sinal do coeficiente de correlação de Spearman dá a direcção da associação linear entre as variáveis e estudo,
  - Se for positivo, as variáveis evoluem no mesmo sentido;
  - se for negativo, as variáveis evoluem em sentidos opostos

Teste de Spearman

O grau de associação entre duas variáveis definidas numa escala pelo menos ordinal pode ser estudado através do teste de associação de Spearman:

Definição das hipóteses

H_0: As variáveis X e Y não estão correlacionadas
vs
H_1: As variáveis X e Y estão correlacionadas

ou então,

H_0: $\rho_S = 0$
vs
H_1: $\rho_S \neq 0$

Teste de Spearman

Exemplo:

Será que as pessoas mais comunicadoras têm tendência para ser mais eloquentes?

1º) Identificação do teste

Pretende-se avaliar se existe associação estatística entre duas variáveis definidas numa escala qualitativa ordinal. O teste a realizar é o teste ao coeficiente de correlação de Spearman.

2º) Definição das hipóteses

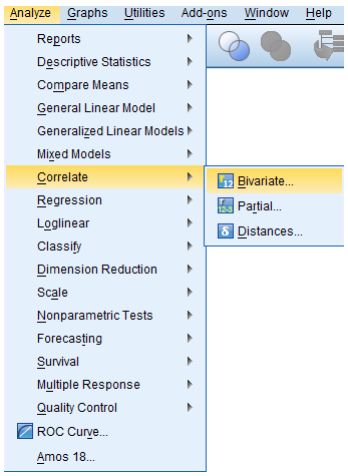
H_0:

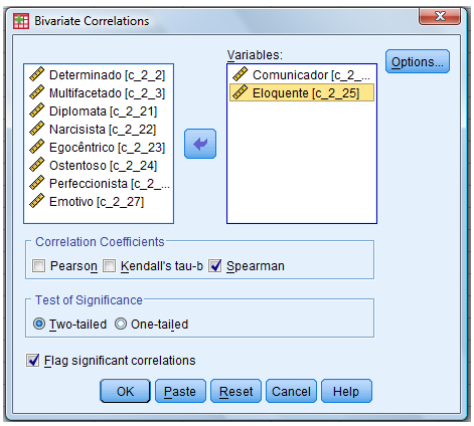
vs

H_1:

Teste de Spearman

3º) Determinação da significância do teste





Teste de Spearman

Correlations

			Comunicador	Eloquente
Spearman's rho	Comunicador	Correlation Coefficient	1,000	,418**
		Sig. (2-tailed)	.	,000
		N	209	206
	Eloquente	Correlation Coefficient	,418**	1,000
		Sig. (2-tailed)	,000	.
		N	206	206

**. Correlation is significant at the 0.01 level (2-tailed).

4º) Decisão do teste

Como sig < 0.001 (inferior a 0.05) rejeita-se a hipótese nula,

5º) Conclusão:

Rejeita-se a hipótese nula, logo existe associação do tipo linear entre as variáveis "Comunicador" e "Eloquente".

O valor do coeficiente ordinal de Spearman é 0,418, logo a associação é positiva média. Ou seja, as pessoas mais comunicadoras são mais eloquentes.

Teste de Spearman - Exercícios

Exercícios:

Considere a base de dados *4_consumo.sav*

a) Verifique se as pessoas que consideram importante estar na moda e ser chique têm tendência a considerarem que fazer compras lhes dá imenso prazer?

b) Será que se pode admitir que quanto mais velhas são os inquiridos menos satisfeitos estão com a sua situação financeira actual?

Testes para comparação de valores médios

Testes para comparação de valores médios

Quando se pretende comparar o valor médio de uma mesma variável (quantitativa) em dois ou mais grupos independentes, podemos recorrer ao teste:

- **Teste *t-Student*** para amostras independentes

  comparar o valor médio de uma mesma variável (quantitativa) em dois grupos independentes

- **ANOVA** *one-way* – Análise de Variância

  comparar o valor médio de uma mesma variável (quantitativa) em mais de 2 grupos independentes

Teste de *t-Student* para amostras independentes

Teste *t-Student*

Para que serve?

Comparar o valor médio de uma mesma variável em dois grupos independentes, como por exemplo: Homens vs Mulheres; Curso diurno vs Curso pós-laboral; Grupo controlo vs Grupo experimental…

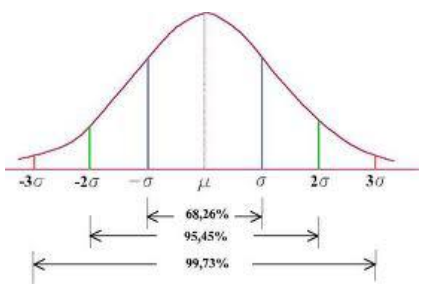
Quando se aplica (Pressupostos)?

i. variável dependente é quantitativa;
ii. amostras independentes;
iii. variável dependente com distribuição normal em ambos os grupos em estudo;

Teste *t-Student - normalidade*

Validação do pressuposto da n*ormalidade*

O teste para comparação de valores médios para duas amostras independentes (teste *t-student*) é um teste paramétrico, e como todos os testes paramétricos exigem que a distribuição amostral seja conhecida. Neste caso em particular, exige-se a distribuição **Normal ou Gaussiana.**



Teste *t-Student - normalidade*

Para verificar se uma variável provém de uma população com distribuição Normal usa-se um destes testes:

Kolmogorov-Smirnov ou ***Shapiro-Wilk***

Hipóteses:

H_0: A variável em estudo provém de uma população com distribuição Normal

vs

H_1: A variável em estudo não provém de uma população com distribuição Normal

Teste *t-Student - normalidade*

Para realizar o teste *t-Student* é necessário verificar sempre se a variável (quantitativa) provém de uma população com distribuição Normal em cada um dos grupos? **Não!**

NOTA: Nos testes *t-Student* para duas amostras se a dimensão da amostra for:

- **Superior ou igual a 30**, assume-se que a distribuição assintótica é Normal e aplica-se o teste sem a verificação formal do pressuposto da normalidade.
- **Inferior a 30** é necessário verificar se a variável segue uma distribuição Normal através do teste de *Shapiro-Wilk*.

Teste *t-Student*

Hipóteses

H_0: Não há diferenças significativas entre os valores médios nos dois grupos

vs

H_1: Há diferenças significativas entre os valores médios nos dois grupos

ou H_0: $\mu_1=\mu_2$ vs H_1: $\mu_1\neq\mu_2$

Testes *t-Student*

Exemplo 1:

Teste a seguinte hipótese:

A "Valorização do estudo" difere entre as raparigas e os rapazes que frequentam a ESCS.

Os dados encontram-se no ficheiro 6_novos_alunos_13_15.sav

Hipóteses:

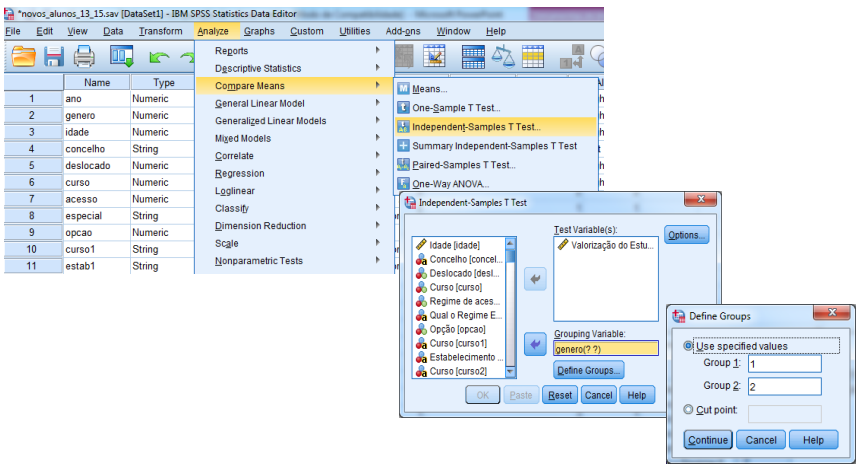
H_0: O valor médio da valorização do estudo não difere entre as raparigas e os rapazes.

vs

H_1: O valor médio da valorização do estudo difere entre as raparigas e os rapazes.

Testes *t-Student*

No SPSS



Teste *t-Student*

Outputs **da análise:**

Group Statistics

	Sexo	N	Mean	Std. Deviation	Std. Error Mean
Valorização do Estudo	Fem	184	8,6150	1,24279	,09162
	Masc	72	7,9144	1,41227	,16644

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Valorização do Estudo	Equal variances assumed	1,061	,304	3,900	254	,000	,70068	,17966	,34688	1,05449
	Equal variances not assumed			3,688	116,401	,000	,70068	,18999	,32440	1,07697

Teste *t-Student*

Resolução:

Validação de Pressupostos:

- v. dependente→ Valorização do Estudo→ Quantitativa
- v. independente → Género → Qualitativa nominal com 2 grupos: feminino e masculino
- Amostras independentes
- n_{Fem}=182; n_{Mas}= 71 (ver output Teste *t-Student*) → assume-se que a distribuição assintótica é Normal

Group Statistics

	Sexo	N	Mean	Std. Deviation	Std. Error Mean
Valorização do Estudo	Fem	184	8,6150	1,24279	,09162
	Masc	72	7,9144	1,41227	,16644

Como se verificam os pressupostos de aplicabilidade do teste t-Student, vamos analisá-lo.

Teste *t-Student - teste de Levene*

Para se analisar o teste t é necessário saber se a variável dependente (valorização do estudo) apresenta variâncias homogéneas nos grupos (feminino e masculino). Para isso utiliza-se o **teste de Levene.**

Hipóteses:

H_o: As variâncias populacionais são homogéneas em ambos os grupos em estudo.

vs

H_1: As variâncias populacionais não são homogéneas em ambos os grupos em estudo.

Nota: No SPSS o teste de *Levene* é feito conjuntamente com o teste *t-Student*.

Teste *t-Student* - *teste de Levene*

Teste de Levene

Hipóteses:

H_o: As variâncias populacionais da valorização do estudo são homogéneas entre o género masculino e feminino.

vs

H_1: As variâncias populacionais da valorização do estudo não são homogéneas entre o género masculino e feminino.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Valorização do Estudo	Equal variances assumed	1,061	,304	3,900	254	,000	,70068	,17966	,34688	1,05449
	Equal variances not assumed			3,688	116,401	,000	,70068	,18999	,32440	1,07697

Testes *t-Student* - *teste de Levene*

Análise do teste de Levene:

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Valorização do Estudo	Equal variances assumed	1,061	,304	3,900	254	,000	,70068	,17966	,34688	1,05449
	Equal variances not assumed			3,688	116,401	,000	,70068	,18999	,32440	1,07697

Decisão:

Como a sig.= 0.304 >0.05, não se rejeitar H_0.

Conclusão:

As variâncias populacionais da valorização do estudo são homogéneas em ambos os géneros.

Nota: Como existe homogeneidade de variâncias iremos analisar a 1ª linha do teste *t-Student*.

Testes *t-Student*

Análise do teste t-Student:

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Valorização do Estudo	Equal variances assumed	1,061	,304	3,900	254	,000	,70068	,17966	,34688	1,05449
	Equal variances not assumed			3,688	116,401	,000	,70068	,18999	,32440	1,07697

Decisão:

Como a sig.< 0.001< 0.05, rejeitar-se H_0.

Conclusão:

A valorização média do estudo difere consoante o género.

Testes *t-Student*

Concluiu-se que:

A "média da valorização do estudo" difere significativamente entre rapazes e raparigas que frequentam a ESCS.

Como encontrar essas diferenças?

Group Statistics

	Sexo	N	Mean	Std. Deviation	Std. Error Mean
Valorização do Estudo	Fem	184	8,6150	1,24279	,09162
	Masc	72	7,9144	1,41227	,16644

As raparigas valorizam mais o estudo (Média feminina=8.62, DP= 1.243) do que os rapazes (Média masculina= 7.91, DP=1.412).

Testes *t-Student*

Exercício:

Com base nos outputs apresentados, teste a seguinte hipótese:

O nº médio de "anos de escolaridade" difere consoante o género.

Group Statistics

	Género	N	Mean	Std. Deviation	Std. Error Mean
Anos de escolaridade	Masculino	564	9,15	4,760	,200
	Feminino	689	8,52	5,492	,209

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Anos de escolaridade	Equal variances assumed	23,160	,000	2,148	1251	,032	,631	,294	,055	1,208
	Equal variances not assumed			2,178	1246,904	,030	,631	,290	,063	1,200

ANOVA *one-way*

ANOVA *one-way*

Para que serve?

A análise de variância – ANOVA - é uma técnica paramétrica que permite verificar se existem diferenças significativas entre as médias de 3 ou mais populações.

Objetivo:

A ANOVA é usada quando se quer perceber se as diferenças amostrais observadas são causadas por diferenças significativas entre as populações ou se são consequência da variabilidade amostral, partindo do pressuposto que a variabilidade decorrente da análise de amostras só pode ser responsável por pequenos desvios.

Nota:

Esta técnica estatística, a ANOVA, é usada para averiguar se os valores médios são estatisticamente diferentes e não para ver se as variâncias são diferentes!

ANOVA *one-way*

… mas se precisamos comparar *k* grupos porque não usamos o teste t-Student várias vezes?

Por exemplo, se é necessário comparar 3 grupos, porque não usamos 3 testes t-Student (grupo 1 *vs* grupo 2, grupo 1 *vs* grupo 3 e grupo 2 *vs* grupo 3)?

Para manter a probabilidade de erro de tipo I no seu nível nominal a (ou inferior) é necessário usar um método que considere todas as comparações em simultâneo - **ANOVA**

ANOVA *one-way*

Quando se aplica? (Pressupostos)

i) variável dependente quantitativa;

ii) amostras independentes;

iii) variável dependente com distribuição normal em todos os grupos em estudo (Teste de *Kolmogorov-Smirnov* ou *Shapiro-Wilk*);

iv) variável dependente com variâncias homogéneas nos grupos em estudo (teste de *Levene*)

Hipóteses

H_0: Não existem diferenças significativas entre as médias das k populações

vs

H_1: Existe pelo menos um par de médias significativamente diferentes

ou

H_0: $\mu_1=\mu_2=...=\mu_k$ vs H_1: $\exists i,j$: $\mu_i \neq \mu_j$ $(i \neq j;\ i.j=1.....k)$

ANOVA *one-way*

Quando **não se rejeita H_0**

Não existem diferenças significativas entre os valores médios dos grupos em comparação

Quando **se rejeita H_0**

Existem diferenças significativas entre os valores médios de pelo menos dois dos grupos em comparação

"descobrir entre que grupos ocorrem estas diferenças"

Para identificar quais as médias que diferem entre si é preciso fazer testes *Post-Hoc*: O **teste de *Scheffé*** é dos mais utilizados, pois é dos mais potentes!

Comparação Múltipla de Médias : Teste de Scheffé

Teste de Scheffé

Pressupostos:

Os mesmos da ANOVA *one-way*

Hipóteses:

H_0: $\mu_i=\mu_j$ vs. H_1: $\mu_i\neq\mu_j$ para todos os pares *i*,*j* de médias possíveis

É necessário analisar a significância associada a cada par de médias correspondente aos grupos *i* e *j* .

Decisão :

Rejeita-se H_0 se sig. $\leq$ 0,05 concluindo assim que existem diferenças significativas entre os grupos i e j.

ANOVA *one-way*

Exemplo 1:
Considere o ficheiro de dados 1_educação.sav.
Utilize a metodologia mais adequada para averiguar se a **"Competência Leitora"** difere consoante o grau de escolaridade do encarregado de educação.

Resolução:

1)Validação de Pressupostos:

- v. dependente→ Competência Leitora→ Quantitativa
- v. independente → Grau de Escolaridade do encarregado de educação → Qualitativa ordinal (define 3 grupos independentes)
- Amostras independentes
- Normalidade (verificar usando o **Teste de Kolmogorv-Smirnov** ou o **Teste de Shapiro-Wilks**)
- Homogeneidade de variâncias (verificar usando **o Teste Levene**)

ANOVA *one-way*

Validação de Pressupostos- continuação

1. Normalidade:

Hipóteses:

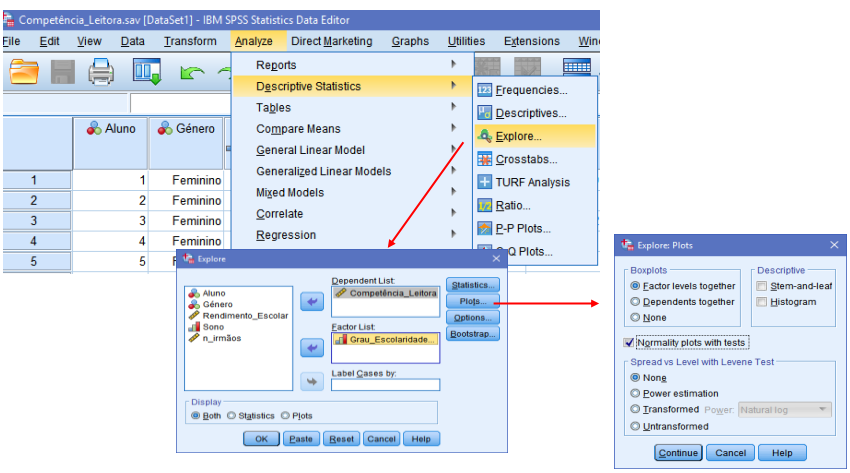
H_0: A variável "Competência Leitora" provém de uma população com distribuição Normal nos diferentes grupos de escolaridade em estudo.

vs

H_1: Existe pelo menos um grupo de escolaridade para o qual a variável "Competência Leitora" não provém de uma população com distribuição normal.

ANOVA *one-way*

No SPSS:



ANOVA *one-way*

Outputs parciais:

Tests of Normality

	Grau de Escolaridade do Encarregado de Educação	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Compreensão Leitora	Ensino Básico	,176	8	,200*	,908	8	,338
	Ensino Secundário	,126	12	,200*	,958	12	,759
	Ensino Superior	,162	10	,200*	,951	10	,683

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Decisão:

Em todos os graus de escolaridade, a significância do teste de Shapiro-Wilk para a variável Competência leitora é sempre superior a 0,05 → Não Rejeitar H_0.

Conclusão:

A "Competência Leitora" segue distribuição normal em todos os grupos graus de escolaridade em estudo (grau de escolaridade dos EE).

ANOVA *one-way*

2. Homogeneidade de variâncias:

Para verificar se a variável dependente apresenta variâncias homogéneas nos grupos em estudo utiliza-se o **teste de *Levene*.**

Hipóteses:

H_0: As variâncias populacionais da variável "Competência Leitora" são homogéneas/não diferem nos diferentes "Graus de escolaridade" .

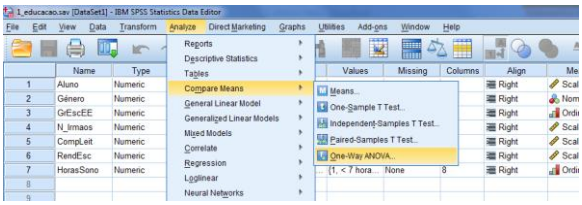
vs

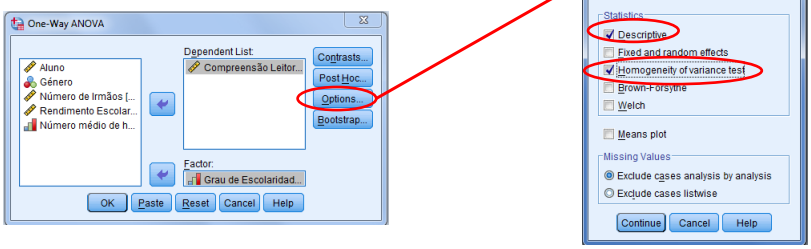
H_1: Existe pelo menos um "Grau de Escolaridade" para o qual a variável "Competência Leitora" apresenta variância populacional diferente dos restantes.

Nota: No SPSS o teste de Levene pode ser feito no mesmo conjunto de comandos da ANOVA

ANOVA *one-way*

No SPSS:





ANOVA *one-way*

Outputs parciais:

Test of Homogeneity of Variances

Competência_Leitora

Levene Statistic	df1	df2	Sig.
1,027	2	27	,372

Decisão:

Como a sig.=0.372 >0.05 → Não Rejeitar H_0.

Conclusão:

A variável "Competência Leitora" apresenta variâncias homogéneas nos diferentes Graus de Escolaridade.

ANOVA *one-way*

Teste ANOVA - Hipóteses:

H_0: A média da "Competência Leitora" não difere entre os "Graus de Escolaridade" dos EE.

vs

H_1: Existe pelo menos um par de " Graus de Escolaridade" em que a média da "Competência Leitora" difere significativamente.

Outputs parciais:

Descriptives

Competência_Leitora

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Ensino Básico	8	8,0000	3,54562	1,25357	5,0358	10,9642	4,00	15,00
Ensino Secundário	12	11,3333	4,83046	1,39443	8,2642	14,4025	4,00	19,00
Ensino Superior	10	13,5000	4,24918	1,34371	10,4603	16,5397	5,00	19,00
Total	30	11,1667	4,70571	,85914	9,4095	12,9238	4,00	19,00

ANOVA

Competência_Leitora

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	135,000	2	67,500	3,593	,041
Within Groups	507,167	27	18,784		
Total	642,167	29			

ANOVA *one-way*

ANOVA

Competência_Leitora

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	135,000	2	67,500	3,593	,041
Within Groups	507,167	27	18,784		
Total	642,167	29			

Decisão:

sig.= 0.041<0.05 → Rejeitar H_0.

Conclusão:

A média da "Competência Leitora" difere significativamente entre pelo menos dois dos graus de escolaridade dos pais.

Entre que grupos** (graus de escolaridade dos EE) **ocorrem as diferenças?

↓

Teste de Scheffé

ANOVA *one-way*

Teste de Scheffé – 3 hipóteses a testar:

H_o: A média da "Competência Leitora não difere em crianças com EE com instrução básica e secundária.
vs
H_1: A média da "Competência Leitora" é diferente em crianças com EE com instrução básica e secundária.

e

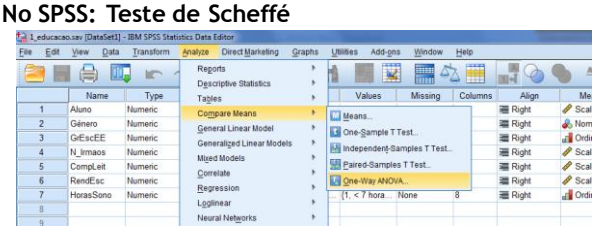
H_o: A média da "Competência Leitora" não difere em crianças com EE com instrução básica e superior.
vs
H_1: A média da "Competência Leitora" é diferente em crianças com EE com instrução básica e superior.

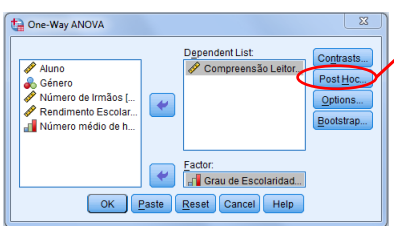
e

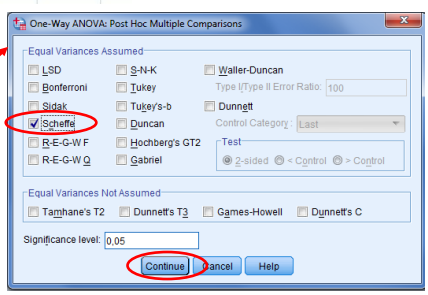
H_o: A média da "Competência Leitora" não difere em crianças com EE com instrução secundária e superior.
vs
H_1: A média da "Competência Leitora" é diferente em crianças com EE com instrução secundária e superior.

ANOVA *one-way*

No SPSS: Teste de Scheffé







ANOVA *one-way*

Outputs parciais:

Multiple Comparisons

Dependent Variable: Compreensão Leitora

Scheffe

(I) Grau de Escolaridade do Encarregado de Educação	(J) Grau de Escolaridade do Encarregado de Educação	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Ensino Básico	Ensino Secundário	-3,333	1,978	,259	-8,46	1,79
	Ensino Superior	-5,500*	2,056	,042	-10,82	-,18
Ensino Secundário	Ensino Básico	3,333	1,978	,259	-1,79	8,46
	Ensino Superior	-2,167	1,856	,514	-6,97	2,64
Ensino Superior	Ensino Básico	5,500*	2,056	,042	,18	10,82
	Ensino Secundário	2,167	1,856	,514	-2,64	6,97

*. The mean difference is significant at the 0.05 level.

Decisão:

Apenas na comparação entre o grupo com o ensino primário e com o ensino superior se observa uma sig. inferior a 0.05 (0.042). Pelo que apenas neste caso se rejeita a H_0

Conclusão:

A "Competência Leitora" média difere significativamente entre crianças com EE com o ensino primário das que têm EE com o ensino superior. A compreensão média das crianças com EE com o ensino superior é significativamente superior (Média $_{Ens\_Sup}$=13.5; DP_{Ens_Sup}=4.249) à média das crianças com EE com o ensino básico ($Média_{Ens\_Bas}$=8.0; DP_{Ens_Bas}=3.546)

ANOVA *one-way*

Exemplo 2:

Considere o ficheiro de dados 6_novos_alunos_13_15.sav.
Utilize a uma metodologia mais adequada para averiguar se: O **Prazer de estudar** é idêntico entre os alunos que frequentam as diferentes licenciaturas da ESCS"

Resolução:

1) Validação de Pressupostos:

- v. dependente→ Prazer de estudar → Quantitativa
- v. independente → Curso → Qualitativa nominal (define 4 grupos independentes)
- Amostras independentes
- Normalidade (**Teste de *Kolmogorv-Smirnov*** ou **Teste de *Shapiro-Wilk***)
- Homogeneidade de variâncias **(Teste *Levene*)**

ANOVA *one-way*

1. Normalidade:

Para verificar se uma variável provém de uma população com distribuição Normal usa-se o teste de:

Kolmogorov-Smirnov ou ***Shapiro-Wilk***

Hipóteses:

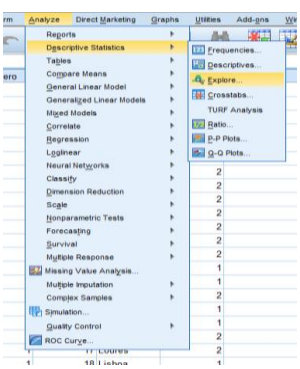
H_0: A variável "Prazer de estudar" provém de uma população com distribuição Normal nos diferentes cursos em estudo.

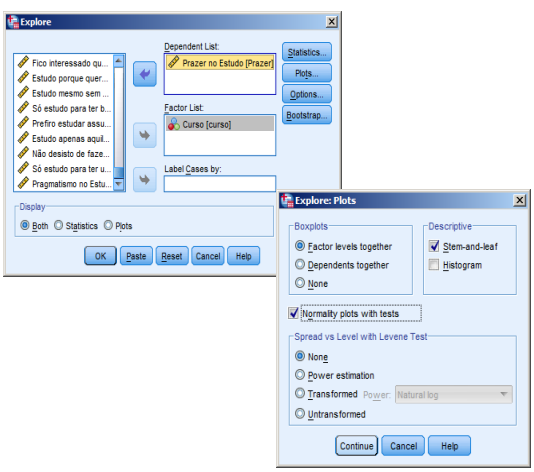
vs

H_1: Existe pelo menos um curso para o qual a variável "Prazer de Estudar" não provém de uma população com distribuição normal.

ANOVA *one-way*

No SPSS:





ANOVA *one-way*

Outputs parciais:

Tests of Normality

	Curso	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Prazer no Estudo	Publicidade e Marketing	,088	56	,200*	,926	56	,002
	Relações Públicas e Comunicação Empresarial	,121	50	,063	,923	50	,003
	Audiovisual e Multimédia	,111	55	,088	,884	55	,000
	Jornalismo	,139	50	,017	,917	50	,002

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Decisão:

O "Prazer de estudar" segue distribuição normal em todos os cursos (sig>0.05) excepto no curso de jornalismo (sig.= 0.017<0.05).

Conclusão:

A variável prazer de estudar **não segue distribuição normal em todos os cursos** pelo que se deverá recorrer a uma alternativa não paramétrica (teste de *Kruskal-Wallis*).

Testes de Hipóteses não Paramétricos

Teste de Kruskal-Wallis

Teste Kruskal-Wallis

Para que serve?

- *Comparar as distribuições populacionais de uma mesma variável dependente em dois ou mais grupos independentes.*
- *Alternativa não paramétrica ao teste t-Student para duas amostras independentes e à ANOVA one-way* (a utilizar quando falham os pressupostos de aplicação)

Quando se aplica (Pressupostos)?

i) variável dependente é pelo menos ordinal;

ii) amostras independentes.

Hipóteses:

H_0: Não existem diferenças significativas entre as distribuições populacionais nos diferentes grupos em estudo

vs

H_1: Existe pelo menos um grupo cuja distribuição populacional difere dos restantes.

Comparações Múltiplas de Médias de ordens

Ao analisar o teste de Kruskal-Wallis, quando **se rejeita H_0** é necessário identificar qual ou quais populações que diferem significativamente entre si, para isso recorre-se ao **Teste de Dunn.**

Pressupostos

Os mesmos do teste de Kruskal-Wallis;

Este procedimento só se deve realizar quando se rejeita H_0 do teste de *K-W* .

Hipóteses

O teste de Dunn realiza todas as comparações duas a duas para identificar entre que grupos existem as diferenças detetadas pelo teste de KrusKal-Wallis. Assim, vamos ter tantos conjuntos de hipóteses quantos os necessários para comparar os grupos 2 a 2:

H_0:Não existem diferenças significativas na distribuição da variável dependente no grupo i e j.

vs

H_1: Existem diferenças significativas na distribuição da variável dependente no grupo i e j. ($i \neq j$; $i,j=1,\ldots,k$, sendo k o número de grupos existentes)

Teste Kruskal-Wallis

Exemplo 1:

Considere o ficheiro de dados 6_novos_alunos_13_15.sav.
Utilize a metodologia mais adequada para averiguar se: O **"Prazer de estudar"** é idêntico entre os alunos que frequentam as diferentes licenciaturas da ESCS.

Resolução:

Validação de Pressupostos:

- v. dependente→ Prazer de estudar → Quantitativa
- v. independente → Curso → Qualitativa nominal (define 4 grupos independentes)
- As amostras são independentes
- Normalidade - verificar usando o **Teste de Shapiro-Wilk** ou o **Teste de Kolmogorv-Smirnov,** se a dimensão do grupo for menos ou igual a 30 ou superior a 30 respetivamante
- Homogeneidade de variâncias - verificar usando o **Teste de Levene**

Teste Kruskal-Wallis

Validação do pressuposto da Normalidade:

Hipóteses:

H_0: A variável "Prazer de estudar" provém de uma população com distribuição Normal no curso Y.

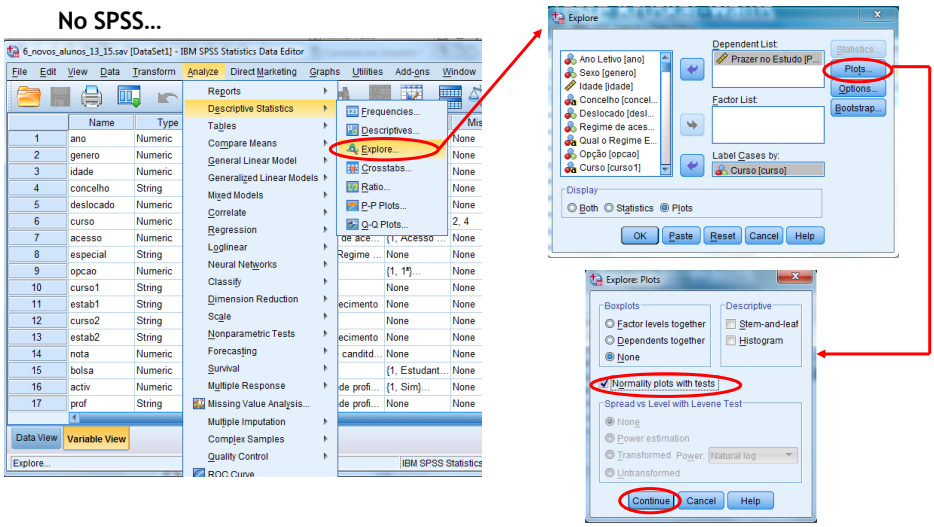
vs

H_1: A variável "Prazer de estudar" não provém de uma população com distribuição Normal no curso Y.

Y={PM, RPCE, AM, JOR}

Teste Kruskal-Wallis

No SPSS…



Teste Kruskal-Wallis

Tests of Normality

Curso		Kolmogorov-Smirnov[a] Statistic	df	Sig.	Shapiro-Wilk Statistic	df	Sig.
Prazer no Estudo	Publicidade e Marketing	,088	56	,200*	,926	56	,002
	Relações Públicas e Comunicação Empresarial	,121	50	,063	,923	50	,003
	Audiovisual e Multimédia	,111	55	,088	,884	55	,000
	Jornalismo	,139	50	,017	,917	50	,002

*. This is a lower bound of the true significance.
a. Lilliefors Significance Correction

Decisão:

O "Prazer de estudar" segue distribuição normal em todos os cursos (sig>0.05) exceto no curso de jornalismo (sig.= 0.017<0.05).

Conclusão:

A variável prazer de estudar **não segue distribuição normal em todos os cursos** pelo que se deverá recorrer a uma alternativa não paramétrica (teste de Kruskal-Wallis).

Teste Kruskal-Wallis

Teste de Kruskal-Wallis:

Hipóteses:

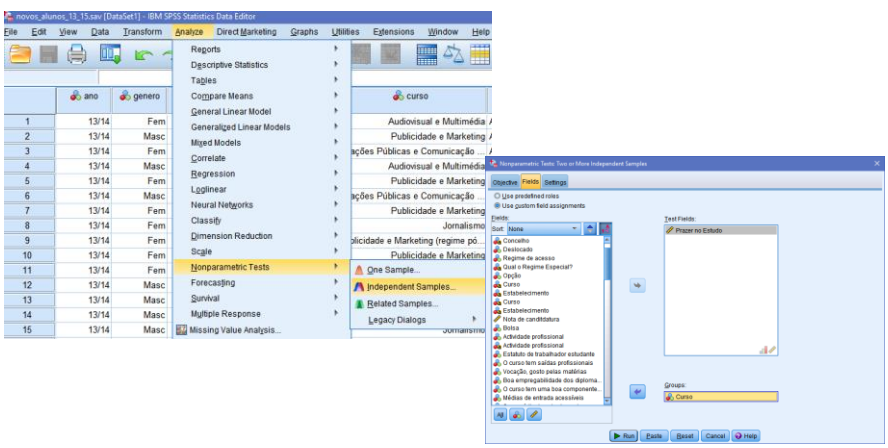
H_0: Não existem diferenças significativas entre as distribuições populacionais do "Prazer de estudar" nos diferentes "cursos" ministrados na ESCS.

vs

H_1: Existe pelo menos um "curso" cuja distribuição populacional do "Prazer de estudar" difere dos restantes.

Teste Kruskal-Wallis

No SPSS:



Teste Kruskal-Wallis

Output:

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Prazer no Estudo is the same across categories of Curso.	Independent-Samples Kruskal-Wallis Test	,064	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Decisão:

Como a significância do teste é 0.064 que é superior a 0.05, então não se rejeita H_0.

Conclusão:

Não existem diferenças estatisticamente significativas na distribuição dos valores do "Prazer de Estudar" nos diferentes cursos.

Teste Kruskal-Wallis

Exercício:

Considere o ficheiro de dados 7_valores_humanos_2016.sav.

Averigúe se os indivíduos das diferentes regiões identificam-se de igual forma com a afirmação **"Importante sentir-se bem"**

Resolução:

Validação de Pressupostos:

- v. dependente → Importante sentir-se bem → Qualitativa ordinal
- v. independente → Região → Qualitativa (define 5 grupos independentes)
- Amostras independentes

Hipóteses:

H_0: Não existem diferenças significativas entre as distribuições populacionais de "Importante sentir-se bem" nas diferentes "Regiões".

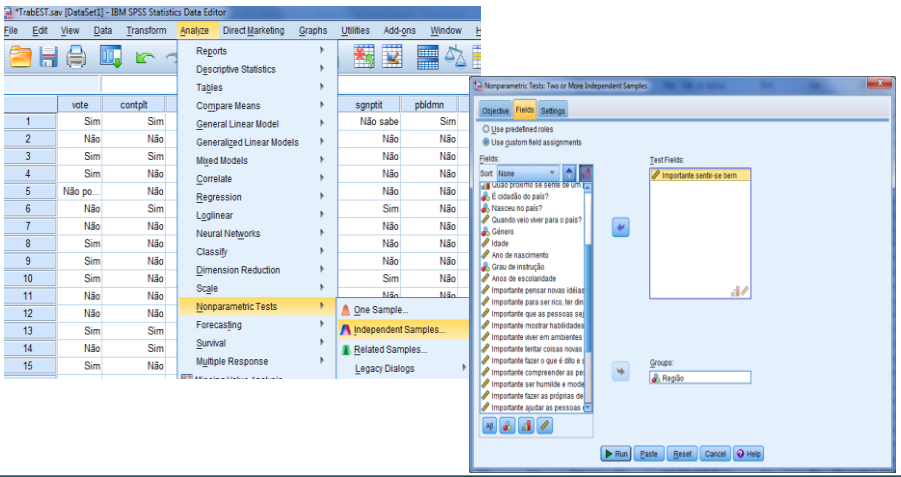
vs

H_1: Existe pelo menos uma "Região" cuja distribuição populacional de "Importante sentir-se bem" difere dos restantes.

Teste Kruskal-Wallis

Resolução:

No SPSS:



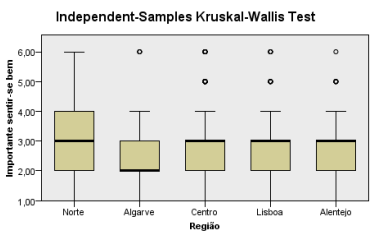
Teste Kruskal-Wallis

Output:

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Importante sentir-se bem is the same across categories of Região.	Independent-Samples Kruskal-Wallis Test	,000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.



Decisão:

sig.< 0.001<0.05 → Rejeitar H_0.

Conclusão:

Existem diferenças estatisticamente significativas na distribuição dos valores da "Importante sair-se bem" em pelo menos um par de regiões.

Para identificar onde estão essas diferenças → ***Teste de Dunn***
(basta dar um duplo clique na tabela anterior no SPSS e escolher a opção "*Pairwise Comparisons*").

Teste Kruskal-Wallis

Teste de Dunn

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
Algarve-Lisboa	-22,881	46,418	-,493	,622	1,000
Algarve-Alentejo	-85,657	53,133	-1,612	,107	1,000
Algarve-Centro	-105,786	46,447	-2,278	,023	,228
Algarve-Norte	125,289	44,866	2,793	,005	,052
Lisboa-Alentejo	-62,776	38,368	-1,636	,102	1,000
Lisboa-Centro	82,905	28,396	2,920	,004	,035
Lisboa-Norte	102,408	25,728	3,980	,000	,001
Alentejo-Centro	20,129	38,403	,524	,600	1,000
Alentejo-Norte	39,632	36,475	1,087	,277	1,000
Centro-Norte	19,503	25,780	,757	,449	1,000

Decisão:

Apenas há diferenças entre as regiões:

- Algarve e Centro → sig. =0.023 < 0.05
- Algarve e Norte → sig. =0.005 < 0.05
- Lisboa e Centro → sig. =0.004 < 0.05
- Lisboa e Norte → sig. <0.001 < 0.05

Conclusão:

As regiões Norte e Centro registam diferenças estatisticamente significativas das regiões de Lisboa e Algarve quanto à distribuição dos valores de ser "Importante sair-se bem".