Diapositivos de apoio às disciplinas de

Análise de Dados

Análise Avançada de Dados

Docente: Pedro Sousa (pedro.sousa.mrm@gmail.com)

Enquadramento

Estes diapositivos focam os pontos essenciais da matéria das duas disciplinas acima mencionadas – Análise de Dados (AD) e Análise Avançada de Dados (AAD), obedecendo aos conteúdos programáticos das mesmas. De modo algum dispensam que o aluno prossiga a sua investigação e estudo das matérias leccionadas, consultando para isso títulos constantes na lista bibliográfica cedida, ou outros.

Diapositivos de apoio à disciplina de

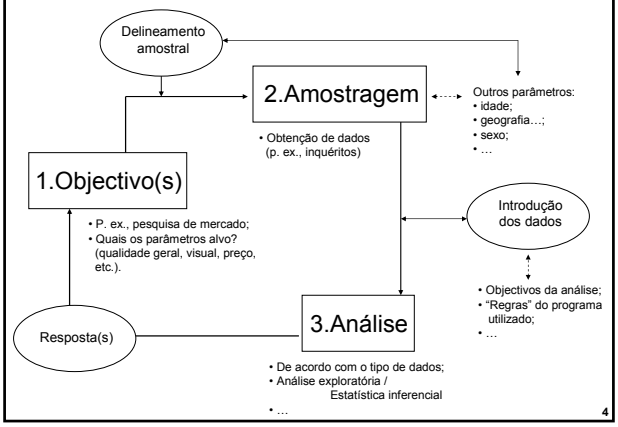
Análise de Dados

Docente: Pedro Sousa (pedro.sousa.mrm@gmail.com)

Tópicos preliminares

- Intro
- Horários das aulas
- Programa, bibliografia e avaliação
- Esquema de trabalho por tema (introdução, exemplos, ficha de trabalho)
- Material de apoio
- Registo no site INUAF (www.inuaf-studia.pt)
- Calendário
- Marcação de provas
- Delegado(s)
- Presenças
- Recomendações

Enquadramento geral...



Objectivos da análise de dados

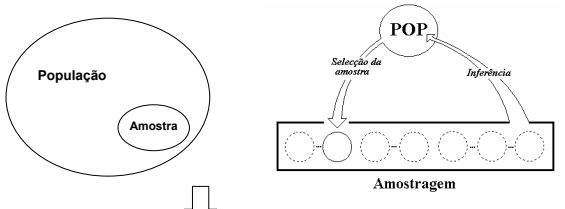
- Identificar e compreender o problema:

⇨ Qual é a População alvo?

⇨ Quais são as perguntas às quais queremos responder?

Objectivos da análise de dados

Intervalo estatístico: "3 MUNDOS"



- **Porquê amostrar?...**
- **Notações; *N*, *n*, fracção amostral**
- **Erro amostral...**
- **Parâmetros, estatísticas e Inferência;**
- **Estatística Descritiva x Inferencial.**

⇨ **A estas notações adicionam-se diferentes tipos de amostragem que dependem essencialmente das características da População sob estudo. <u>Voltaremos a este assunto posteriormente</u>.**

Objectivos

- Identificar e compreender o problema:

Qual é a População alvo? ⟺ Quais são as perguntas às quais queremos responder?

⇩

- *Tipos de fraldas descartáveis;*
- *Intenção de voto para prox. AEINUAF;*
- *Vestuário...*
- *Controle de qualidade de leitores de DVD;*
- *Emissão de gases de combustão automóvel numa dada cidade.*
- *...*

Objectivos

- Quais são as características da População que queremos estudar?

↳ Variáveis aleatórias

↓

• Traduzem a variação que existe na população.
• Obedecem a uma DISTRIBUIÇÃO DE PROBABILIDADES.

→

Objectivos

- A **TEORIA DAS PROBABILIDADES*** permite-nos **INFERIR!**

⇩

• "Qual é a verosimilhança** de observar ou amostrar um dado indivíduo ou um dado conjunto de indivíduos da População (amostra)?".

→ Ex. *informal*...: Lotaria ou totoloto?...

O Programa SPSS

A progressiva utilização de computadores tem disponibilizado um número cada vez maior de métodos estatísticos que, embora já existissem, a sua aplicabilidade era restrita pela complexidade de cálculos.

Por outro lado, o tratamento de grandes quantidades de dados tem também vindo a ser praticável com o auxílio da informática.

Neste contexto, têm sido desenvolvidos vários programas informáticos de estatística, de entre os quais se destacam o ***SPSS***, o ***SAS***, o ***Statistica***, o ***S-PLUS*** e o **R**. Este último tem a grande vantagem de ser gratuito, embora se exija que o utilizador possua conhecimentos médios de programação (mais informações sobre o **R** em http://cran.r-project.org).

De entre estes programas, o que utilizaremos na presente disciplina é o ***SPSS***.

O Programa SPSS

Introdução

Consideremos uma amostra de 6 alunos que prestaram provas psicotécnicas, na qual se registou o *sexo*, *idade* e classificação (*nota*), de 0 a 100, na referida prova. A Tabela abaixo representa esta informação introduzida numa folha de cálculo (não no SPSS).

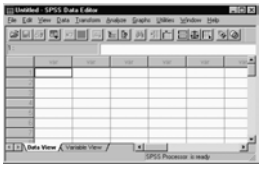
Sexo, *idade* e *nota* são **variáveis**. Os alunos são os **indivíduos**, **casos** ou **observações**.

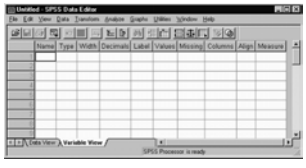
	A	B	C	D
1	Nome	Sexo	Idade	Nota
2	Sara	f	16	87
3	Daniel	m	14	78
4	Pedro	m	17	98
5	Lucia	f	19	67
6	Sandra	f	21	50
7	Cristina	f	12	72

O Programa SPSS

Introdução

Quando abrimos o programa SPSS, ficamos presente o *Data Editor,* o qual possui dois menus: *Data View* (Figura à esquerda) e *Variable View* (Figura à direita). No *Data View*, como o nome indica, são introduzidos e visualizados os dados, tal como no presente exemplo (página anterior). No *Variable View*, são descritas as características das variáveis, tal como o nome, o tipo (numérico ou não), número de casas decimais (*Decimals*), a largura de coluna *(Columns)* e o alinhamento *(Align)*. Os campos restantes serão explicados seguidamente.





O Programa SPSS

Introdução

Quando introduzimos os dados no SPSS tal qual o exemplo da folha de cálculo da página anterior, os dois menus anteriores assumem o aspecto abaixo.

Data View

	nome	sexo	idade	nota
1	Sara	f	16.00	67.00
2	Daniel	m	14.00	78.00
3	Pedro	m	17.00	98.00
4	Lucia	f	19.00	67.00
5	Sandra	f	21.00	50.00
6	Cristina	f	12.00	72.00
7				

Repare-se que o SPSS adoptou duas casas decimais *(Decimals)* para as variáveis *idade* e *nota*. Podemos alterar para 0 casas decimais no campo *Decimals* do *Variable View*. O SPSS também reconheceu que *nome* e *sexo* não são variáveis numéricas, mas sim texto (*String*).

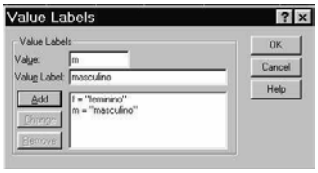
Variable View

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	nome	String	8	0		None	None	8	Left	Nominal
2	sexo	String	8	0		{f, feminino}...	None	8	Left	Nominal
3	idade	Numeric	11	2		None	None	8	Right	Ordinal
4	nota	Numeric	11	2		None	None	8	Right	Ordinal
5										

O Programa SPSS

Introdução

O campo *Values* especifica o significado de siglas, letras ou outro tipo de caracteres que possam ser utilizados em variáveis não numéricas. Por exemplo, neste caso, **"f" = feminino** e **"m"=masculino**. O menu de introdução de *value labels* encontra-se na Figura abaixo.



NOTA IMPORTANTE: a variável "sexo" foi considerada, para este exemplo, como texto (*string*). No entanto, como veremos posteriormente, as variáveis categóricas (ou factores) possuem níveis, os quais assumem valores numéricos. Assim teríamos: *Value*: 1 , *Value label*: masculino; *Value*: 2 , *Value label*: feminino, por exemplo.

O Programa SPSS

Nomes de variáveis - REGRAS

Existem algumas **regras para nomear variaveis** (*Name*, menu *Variable View*), as quais se especificam seguidamente:

- Deve começar por uma letra ou por @ .
- Não pode terminar com ponto.
- Pode conter letras, dígitos ou qualquer dos seguintes caracteres: @, #, _ ou $.
- Não pode conter espaços ou caracteres especiais tais como ! , ? e *, e outros que não os referidos acima .
- Não pode ser uma das palavras-chave do SPSS, ou seja:
  AND, NOT, EQ, BY e ALL .

O Programa SPSS

Tipos de ficheiros SPSS / Gravar ficheiros.

Existem três tipos principais de ficheiros SPSS, nomeadamente, ficheiros de **dados** (***.sav**), de ***output*** (***.spo**) e de ***syntax*** (***.sps**). Os ficheiros de *output* registam todos os procedimentos efectuados e respectivos resultados, em cada sessão de trabalho. O registo dos procedimentos é efectuado através de uma linguagem chamada *syntax*[1]. Esta linguagem, e apenas esta, é utilizada nos ficheiros de *syntax, e* torna-se útil em muitos aspectos. Por exemplo, quando se pretende repetir uma série de procedimentos sem ter que utilizar os menus, ou quando se pretende programar qualquer operação não implementada directamente nos menus do SPSS. Veremos posteriormente casos concretos de aplicação da linguagem *syntax*.

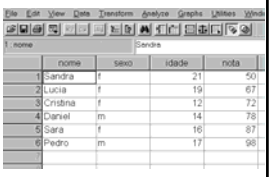
Para gravar ficheiros efectua-se FILE \ SAVE. A Figura ao lado é um exemplo para o nosso ficheiro de dados.

[1] O processo de registar todos os procedimentos no *output (display commands in the log)* é opcional, sendo a sua activação / desactivação efectuada no sub-menu EDIT \ OPTIONS . Este sub-menu será explorado posteriormente.

O Programa SPSS

Ordenação de observações (*sort*)

Para ordenar *(sort)* observações no SPSS utiliza-se o sub-menu DATA \ SORT CASES. Por exemplo, ordenemos crescentemente (*ascending*) os nossos dados pela variável *nota (*Figura abaixo, à esquerda). O resultado no *Data View* encontra-se na Figura da direita.



	nome	sexo	idade	nota
1	Sandra	f	21	50
2	Lucia	f	19	67
3	Cristina	f	12	72
4	Daniel	m	14	78
5	Sara	f	16	87
6	Pedro	m	17	98

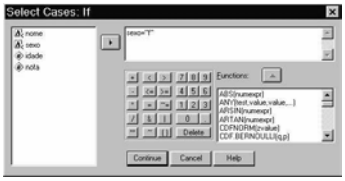
NOTA: Pode utilizar-se mais que uma condição no campo "sort by". Por exemplo, "nota-ascending" e "idade-descending". A primeira condição tem preferência relativamente à segunda. O mesmo se estende a mais condições.

O Programa SPSS

Selecção de observações (*select cases*)

Imaginemos que queremos analisar apenas uma parte do conjunto de dados, por exemplo, achar a média das notas para os indivíduos do sexo feminino. Para isso, utilizamos o menu DATA \ SELECT CASES.

Especifiquemos então que só queremos analisar os indivíduos do sexo feminino, no sub-menu "**If condition is satisfied \ if**". Note-se a variedade de hipóteses de operações de selecção disponíveis neste sub-menu.



O Programa SPSS

Selecção de observações (continuação)

O *Data View* passa então a apresentar o aspecto abaixo, no qual os indivíduos não seleccionados surgem "traçados" (excluídos da análise sub-sequente). Observe-se que o SPSS cria uma variável de *filtragem* (filter_$), de natureza binária (0 ou 1), a qual utiliza na selecção de casos (0 = exclui; 1 = inclui). A elaboração de filtros deste género pode ser muito vantajosa, como veremos posteriormente.

Para voltar a analisar todos os casos novamente, executa-se
DATA \ SELECT CASES\ SELECT **All cases**.

	nome	sexo	idade	nota	filter_$
1	Sara	f	16.00	87.00	1
2	Daniel	m	14.00	78.00	0
3	Pedro	m	17.00	98.00	0
4	Lucia	f	19.00	67.00	1
5	Sandra	f	21.00	50.00	1
6	Cristina	f	12.00	72.00	1
7					

O Programa SPSS

Selecção de observações (continuação):
Filtros mais elaborados – ex.1

Imagine-se que se pretende seleccionar os indivíduos do sexo feminino **e** aqueles com idade igual ou superior a 17 anos. Nesta filtragem, bem como noutras de semelhante natureza, são utilizados os operadores lógicos:

"&" : "e", ou seja, observações que satisfaçam todas as condições explícitas;
"|" : "ou", pelo menos uma das condições.

Podem também ser utilizados todos os operadores matemáticos ou estatísticos disponíveis no SPSS, por exemplo:

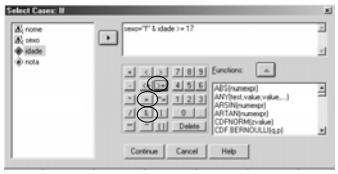
"=" : "igual a"
"~=" : "diferente de"
">" : "maior que"
"<" : "menor que"
"<=" : "menor ou igual a"
">=" : "maior ou igual a"
""** : "elevado a" (exponenciação)

NOTA: Para apagar variáveis ou casos (observações) **NÃO UTILIZAR** a tecla "delete" mas sim, com o **botão direito do rato**, seleccionar **"clear"**.

O Programa SPSS

Selecção de observações (continuação):
Filtros mais elaborados – ex.1

Assim, considerando as condições deste exemplo, temos:



	nome	sexo	idade	nota	filter_$
1	Sandra	f	21	50	Selected
2	Lucia	f	19	67	Selected
3	Cristina	f	12	72	Not Selecte
4	Daniel	m	14	78	Not Selecte
5	Sara	f	16	87	Not Selecte
6	Pedro	m	17	98	Not Selecte

O Programa SPSS

Selecção de observações (continuação):
Filtros mais elaborados – ex.2

Indivíduos masculinos cuja nota se situa entre 70 (inclusive) e 80:

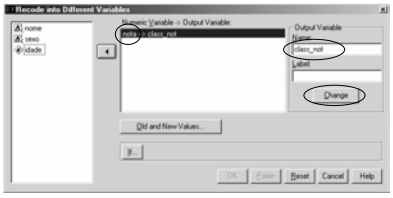
	nome	sexo	idade	nota	filter_$
1	Sandra	f	21	50	Not Selecte
2	Lucia	f	19	67	Not Selecte
3	Cristina	f	12	72	Not Selecte
4	Daniel	m	14	78	Selected
5	Sara	f	16	87	Not Selecte
6	Pedro	m	17	98	Not Selecte

O Programa SPSS

Recodificar variáveis (*recode*)

Considere-se que se pretende classificar as notas em 3 categorias: **1** (50-69), **2** (70-89) e **3** (90-100) numa nova variável, por exemplo "class_not". Para isso utiliza-se o comando **"TRANSFORM \ RECODE \ INTO DIFFERENT VARIABLES".** Se se pretendesse substituir a variável original pela nova, codificada, seria utilizada a opção "**… INTO SAME VARIABLES**":

1. Seleccionar variável original;
2. Nomear *output variable*
3. *Comando* **"change"**
4. Comando **"*old and new values*"**



Passo seguinte

O Programa SPSS

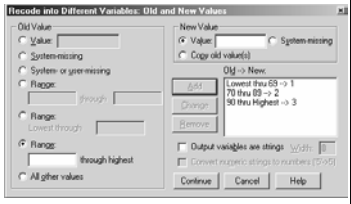
Recodificar variáveis (*recode*)

Especificar valores antigos e novas categorias:

Range: Lowest through **69**
New Value / Value: **1**
Add

Range: **70** *through* **89**
New Value / Value: **2**
Add

Range: **90** *through highest*
New Value / Value: **3**
Add



"Do valor 90 até ao maior dos mesmos, atribui a categoria 3"

"Entre os valores 70 e 89 atribui a categoria 2"

"Do valor original mais baixo até ao 69, atribui a categoria 1"

⇨ ***Continue***
OK

O Programa SPSS

Recodificar variáveis (*recode*)

Uma vez concluída a sequência de operações as alterações no *dataview* são as seguintes:

	nome	sexo	idade	nota	class_not
1	Sandra	f	21	50	1.00
2	Lucia	f	19	67	1.00
3	Cristina	f	12	72	2.00
4	Daniel	m	14	78	2.00
5	Sara	f	16	87	2.00
6	Pedro	m	17	98	3.00
7					

O Programa SPSS

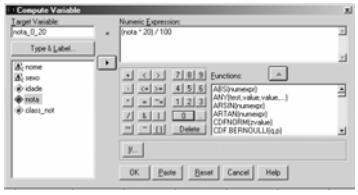
Cálculo de novas variáveis (*compute*)

Considere-se que se pretende converter a variável "nota", cuja amplitude é de 0 a 100, numa nova variável "nota_0_20", de amplitude 0 a 20 valores. A operação matemática consiste em:

nota_0_20 = (nota original x 20) / 100

No *SPSS* utiliza-se o comando **"TRANSFORM \ COMPUTE"**:

1. *Target variable*
   (nome da nova var.)
2. *Numerical expression*
   (expressão de conversão)
3. OK



O Programa SPSS

Cálculo de novas variáveis (*compute*)

O resultado da operação encontra-se abaixo.

	nome	sexo	idade	nota	class_not	nota_0_20
1	Sandra	f	21	50	1	10.00
2	Lucia	f	19	67	1	13.40
3	Cristina	f	12	72	2	14.40
4	Daniel	m	14	78	2	15.60
5	Sara	f	16	87	2	17.40
6	Pedro	m	17	98	3	19.60
7						

NOTA 1: Repare-se que é ainda possível especificar o tipo (*type*) e descrição (*label*) da nova variável, bem como especificar condições com o comando "if".

NOTA 2: O comando "paste", comum a todos os menus de operações em SPSS, faz com que todas as ordens dadas na operação sejam transcritas numa linguagem do SPSS (syntax) e adicionadas a uma janela à parte. Posteriormente estudaremos esta linguagem.

O Programa SPSS

Importação / exportação de ficheiros de dados

Tal como gravámos anteriormente o nosso ficheiro em formato *.SAV (SPSS), também é possível fazê-lo (exportar o ficheiro) em formato **Excel**. Para tal, no sub-menu FILE \ SAVE AS \ SAVE AS TYPE selecciona-se **Excel (*.XLS)**.

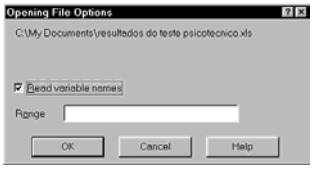
Para importar um ficheiro **Excel**, efectua-se FILE \ OPEN \ DATA \ FILES OF TYPE \ **Excel (*.XLS)**.

Na Figura ao lado encontra-se o exemplo de importação do conjunto de dados que tem vindo a ser trabalhado.

O Programa SPSS

Importação / exportação de ficheiros de dados

Após seleccionarmos ***Open***, surge um novo menu que questiona se queremos que o SPSS leia o nome das variáveis (***read variable names***). Se na primeira linha do ficheiro *Excel* que estamos a importar se encontram os nomes das variáveis, esta opção deve ser seleccionada (Figura abaixo).



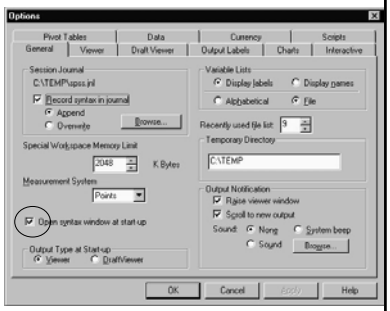
O resultado da importação deste ficheiro para o *Data View* do SPSS é apresentado na Figura do diapositivo 12.

O Programa SPSS

Configuração do *SPSS*

Todas as opções de configuração do SPSS encontram-se no sub-menu EDIT \ OPTIONS, dentro do qual existem 10 outros sub-menus. Abordemos apenas dois deles, nomeadamente, EDIT \ OPTIONS \ GENERAL e EDIT \ OPTIONS \ VIEWER, e dentro destes, duas opções fundamentais.

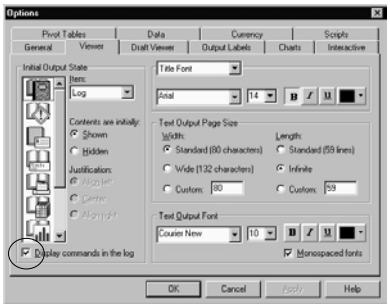
No sub-menu GENERAL é importante assinalar a opção que automatiza a abertura de uma janela de *syntax* no arranque do SPSS:
Open syntax window at start-up (Figura ao lado).



O Programa SPSS

Configuração do *SPSS (continuação)*

No sub-menu Viewer é fundamental assinalar a opção ***Display commands in the log*** que faz com que todos os procedimentos efectuados na sessão de trabalho sejam descritos no ficheiro de *output*, através da linguagem *syntax*.



Tipos de dados

O delineamento e a análise a efectuar depende do tipo de dados!

- Qualitativos ou nominais (categóricos, não-métricos)
  - **São categorias.** Ex., fumador, sexo, profissão. (casos part.: cat. binárias)
  - Ordem subjacente → ordinais: **São categorias ORDENADAS.** Ex., critérios de preferência, grupos etários.
- Quantitativos ou métricos
  - contínuos
    - por intervalos: **Escala contínua SEM ZERO ABSOLUTO.** Ex., temperatura em ºC.
    - por rácios: **Escala contínua COM ZERO ABSOLUTO.** Ex., tempo, altura, peso.
  - discretos: **Números inteiros.** Ex., número de clientes por dia numa dada zona comercial.

Tipos de dados

Adaptado de um exemplo já trabalhado:

Trabalhador	Salário (€)	Categoria	Fumador	Agregado familiar
1	1024	2	1	3
2	567	1	1	4
3	892	1	0	2
4	1232	3	1	1
5	449	1	0	3
6	982	2	1	5
7	1100	2	0	2

- Agregado familiar → Quantitativos, discretos.
- Fumador → Qualitativos, nominais (binários).
- Categoria → Qualitativos, ordinais.
- Salário (€) → Quantitativos, contínuos, por rácios.

Ex. SPSS...

A análise gráfica

- Porquê gráficos?
  (*uma imagem vale mais que mil palavras*)

Mas não só!... ⇨ Caso de conjuntos extensos de dados

(o que está directamente relacionado com a elaboração de tabelas de frequências, como veremos posteriormente)

- Tipos de gráficos **x** tipos de dados!

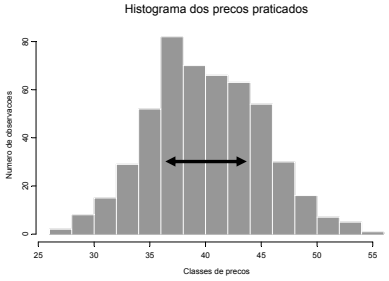
A análise gráfica

- **Pergunta:**
  Com base na informação da tabela abaixo referente a preços (€) de um dado produto em 500 diferentes zonas comerciais da mesma cidade, identifique uma gama de preços mais praticados.
  (sem recorrer ainda ao cálculo de indicadores estatísticos).

43.69	30.12	40.07	37.74	40.52	36.61	37.71	34.57	46.5	37.87	42.75	42.34	40.08	40.83	44.52	30.78	46.9	41.55	43.96	45.16	40.54	42.26
48.08	37.66	38.61	44.5	40.74	37.93	36.83	44.41	48.65	44.32	39.81	46.69	41.96	38.99	36.34	32.39	33.33	43.49	48.26	43.65	44.75	37.13
42.2	48.32	43.03	45.01	39.35	26.26	38.74	41.57	37.87	36.84	41.78	48.41	44	36.39	33.77	44.26	46.14	38.67	29.77	37.39	40.97	49.32
46.46	39.36	31.67	35.18	30.81	36.26	42.74	44.22	38.38	44.62	36.91	40.2	46.27	37.19	41.28	49.64	46.67	40.36	39.33	38.49	37.94	47.79
49.74	37.08	40.01	45.58	40.15	32.22	38.53	43.62	34.99	41.52	44.1	44.67	34.17	37.36	41.9	38.88	38.07	39.49	41.83	46.67	44.81	40.42
44.31	32.44	54.14	41.41	40.9	41.98	38.95	40.42	34.6	32.88	40.08	32.93	44.93	41.92	31.7	42.81	38.55	45.53	30.4	36.75	34.27	41.53
39.95	44.57	35.73	51.48	37.39	46.26	47	37.54	37.65	34.37	33.77	35.08	41.68	35.13	34.97	39.26	38.69	39.8	45.23	36.89	42.14	34.86
37.82	35.58	33.29	50.85	32.78	42.3	37.45	44.58	36.81	39.63	37.64	40	42.81	36.08	46.3	37.06	31.52	53.97	42.51	32.39	27.75	36.71
38.11	37.67	33.15	39.29	49.45	35.11	44.94	34.39	28.07	39.72	39.27	37.56	33.15	40.86	44.61	33.88	34.46	37.74	45.28	35.24	38.04	34.53
43.31	33.73	45.38	45.19	42.99	43.88	39.87	38.36	34.87	36.42	42.14	42.66	37.84	33.85	35.87	43.47	38.83	45.06	37.46	37.79	38.45	42.24
35.98	39.06	37.36	36.05	33.67	41.48	34.79	38.73	44.63	35.61	35.74	40.86	35.64	37.54	42.11	37.76	40.26	35.11	50.07	44.81	44.8	35.12
48.9	46.89	46.27	53.27	44.08	37.82	44.66	38.78	34.49	34.68	44.44	42.6	36.78	41.12	42.71	41.39	41.8	42.79	48.34	29.3	42.56	46.73
45.64	43.88	43.81	38.13	39.84	34.35	52.08	40.44	39.8	37.1	39.07	39.82	37.94	35.63	34.69	36.33	30.39	40.83	43.7	33.01	51.08	32.31
41.49	34.28	39.71	35.07	34.19	51.32	30.42	40.03	47.64	29.42	39.21	47.44	38.46	43.29	37.84	36.54	40.17	50.17	39.47	32.86	39.14	41.7
40.73	43.41	39.85	36.69	34.69	43.28	47.05	35.41	38.21	46.1	47.11	42.55	39.77	46.4	37.55	33.33	45.92	42	30.71	41.86	39.46	31.47
40.17	39.6	48.62	29.73	38.4	38.87	41.03	36.42	32.68	44.61	41.69	33.36	45.69	38.97	37.68	47.61	37.72	47.83	40.3	36.46	37.62	35.13
41.66	42.04	43.35	51.56	29.61	35	44.05	38.22	48.24	42.34	32.28	49.94	37.57	37.99	35.11	41.71	36.95	47.65	37.2	44.93	38.79	43.97
43.22	43.33	41.66	47.84	44.47	52.3	41.59	52.21	34.32	43.73	47.39	42.39	45.59	44.01	36.59	41.54	37.76	45.3	39.06	36.02	42.38	37.83
35.69	37.26	32.95	32.51	37.06	41.62	45.28	37.87	35.91	37.59	39.06	44.35	39.4	35.04	41.78	39.19	42.41	35.79	44.97	42.81	40.71	42.24
42.91	28.45	41.97	30.45	45.95	36.03	36.74	43.27	42.02	40.21	44.39	42.64	44.55	39.99	29.72	34.54	35.63	39.83	40.16	44.71	41.58	38.13

A análise gráfica

- **Resposta:**



A análise gráfica

- **Resposta (**cont.**):**

  Na base da elaboração do **HISTOGRAMA** da página anterior está uma **TABELA DE FREQUÊNCIAS** na qual os dados originais são **agrupados em classes**:

  1. Mínimo observado = 26.26 € ; máximo = 54.14 €.
  2. Limite inferior da menor classe = 25 € ;
     limite superior da maior classe = 55 €.
  3. Estabelecer nº classes com base no intervalo de valores (55 – 25) = 30. No exemplo, 15 classes. Assim, cada classe possui uma **AMPLITUDE** de (30 / 15) = 2 €.
  4. Distribuir as observações pelas classes, obedecendo à regra: ≥ limite inferior da classe e < limite superior da classe*.
  5. Os **CENTROS DE CLASSE*** surgem em consequência dos pontos anteriores.

A análise gráfica

TABELA DE FREQUÊNCIAS:

Classes		Frequências		
Limite inferior	Centro	absolutas	relativas (%)	rel. acumuladas (%)
25	26	1	0.2	0.2
27	28	3	0.6	0.8
29	30	15	3.0	3.8
31	32	20	4.0	7.8
33	34	40	8.0	15.8
35	36	56	11.2	27.0
37	38	86	17.2	44.2
39	40	67	13.4	57.6
41	42	73	14.6	72.2
43	44	64	12.8	85.0
45	46	34	6.8	91.8
47	48	22	4.4	96.2
49	50	9	1.8	98.0
51	52	7	1.4	99.4
53	54	3	0.6	100.0
	total	500		

?:

- **Frequências relativas**
- **Frequências acumuladas**

A análise gráfica

Apliquemos esta metodologia às variáveis do exemplo já trabalhado:

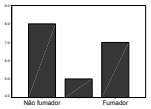
⇩ ⇩ ⇩ ⇩

Trabalhador	Salário (€)	Categoria	Fumador	Agregado familiar
1	1024	2	1	3
2	567	1	1	4
3	892	1	0	2
4	1232	3	1	1
5	449	1	0	3
6	982	2	1	5
7	1100	2	0	2

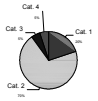
Tipos básicos de gráficos de acordo com a natureza dos dados:

Dados qualitativos (nominais e ordinais) e dados quantitativos discretos

→ **Gráficos de barras (*bar charts*)**



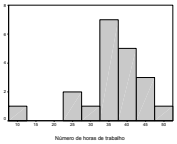
→ **Gráficos "tipo tarte" (*pie charts*)**



Dados quantitativos contínuos

↓

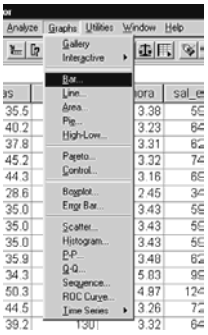
Histogramas (*histograms*)



Elaboração de gráficos Em SPSS

↓

Menu "Graphs" ⇨

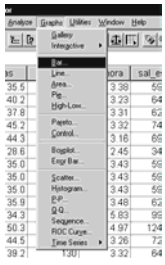


⇨ Cada sub-menu possui várias opções.

Analisemos, como exemplo, as relativas aos gráficos de barras.

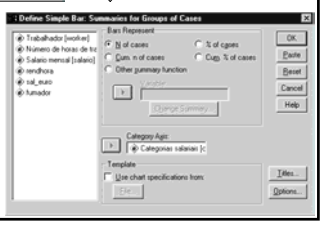
Elaboração de gráficos Em SPSS

Gráficos de barras:
Variável "cat_sal" do exemplo já trabalhado.



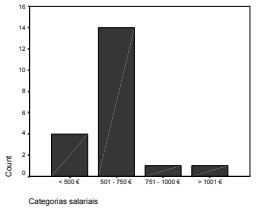
⇨

⇩

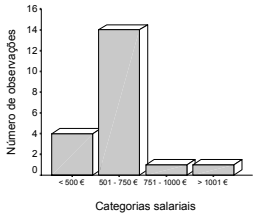


Elaboração de gráficos
Em SPSS

Gráficos de barras
(*output*)



Gráficos de barras
(personalizado no editor de gráficos*)



Tipos de gráficos de barras x opções...

Elaboração de gráficos
Em SPSS:

Gráficos de barras: *Stacked* (empilhado) e *clustered* (agrupado)

Após um debate televisivo no âmbito das eleições presidenciais nos Estados Unidos, recolheu-se, telefonicamente uma amostra *n*=20, tendo sido inquirida a intenção de voto e o nível de escolaridade do indivíduo. Os resultados encontram-se na tabela ao lado.

Pergunta: **Graficamente** será possível discernir alguma relação entre as duas variáveis?

	educacao	voto_usa
1	Não completou ensino básico	Não vota
2	Ensino básico	Republicano
3	Ensino secundário	Democrata
4	Não completou ensino básico	Republicano
5	Não completou ensino básico	Não vota
6	Ensino básico	Republicano
7	Ensino secundário	Democrata
8	Não completou ensino básico	Não vota
9	Não completou ensino básico	Republicano
10	Não completou ensino básico	Democrata
11	Ensino básico	Republicano
12	Ensino secundário	Republicano
13	Ensino secundário	Democrata
14	Bacherlato	Democrata
15	Licenciatura ou superior	Democrata
16	Licenciatura ou superior	Democrata
17	Não completou ensino básico	Republicano
18	Ensino básico	Não vota
19	Ensino secundário	Republicano
20	Ensino secundário	Democrata

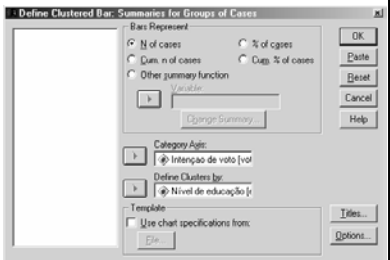
Elaboração de gráficos
Em SPSS:

Gráficos de barras: *Stacked* (empilhado) e *clustered* (agrupado)

Resposta...

**Graphs \
Bar \ *clustered* \
Define...**

OK

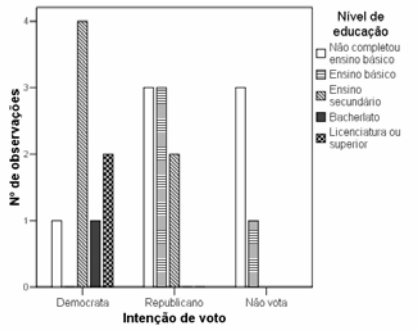


NOTA: O sub-menu de elaboração de gráficos de barras empilhados (stacked) é semelhante ao apresentado ao lado (*clustered*)

Elaboração de gráficos
Em SPSS:

Gráficos de barras: *Stacked* (empilhado) e *clustered* (agrupado)

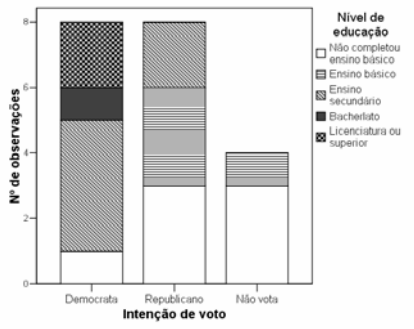
Resultados
(*clustered*) :



Elaboração de gráficos
Em SPSS:

Gráficos de barras: *Stacked* (empilhado) e *clustered* (agrupado)

Resultados
(*stacked*) :

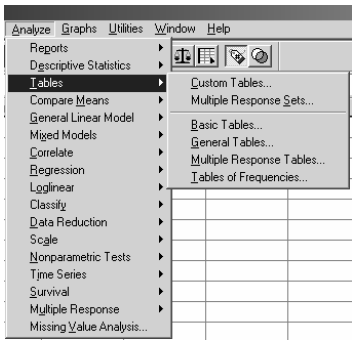


Comentários:

...

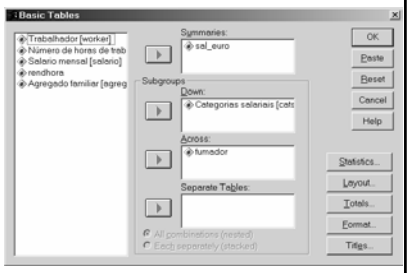
Elaboração de tabelas
Em SPSS

Menu:
Analyze
Tables



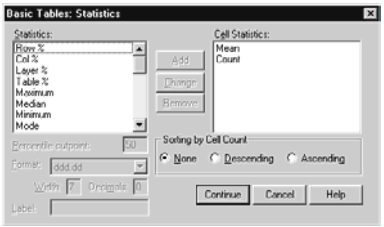
Elaboração de tabelas
Em SPSS: *Basic Tables*

Menu:
Analyze
Tables
Basic Tables



Elaboração de tabelas
Em SPSS: *Basic Tables* (continuação)

Menu:
Analyze
Tables
Basic Tables
Statistics



Elaboração de tabelas
Em SPSS: *Basic Tables* (*output*)

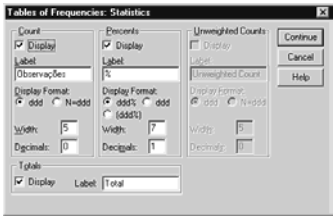
		Não fumador		Ex fumador		Fumador	
		n	Média	n	Média	n	Média
Categorias salariais	< 500 €	3	282.65		.	1	349.16
	501 até 750 €	5	628.49	4	623.50	5	660.41
	751 até 1000 €		.		.	1	997.60
	> 1000 €		.	1	1246.99		.

Elaboração de tabelas
Em SPSS: Tabelas de frequências (*tables of frequencies*)

Menu:
Analyze
Tables
Tables of frequencies

Elaboração de tabelas
Em SPSS: Tabelas de frequências (*tables of frequencies*)
(continuação)

Menu:
Analyze
Tables
Tables of frequencies
Statistics



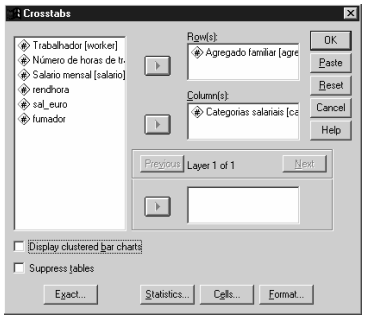
Elaboração de tabelas
Em SPSS: Tabelas de frequências (*tables of frequencies*)
(*output*)

	Observações	%
Não fumador	8	40.0%
Ex fumador	5	25.0%
Fumador	7	35.0%
Total	20	100.0%

Elaboração de tabelas
Em SPSS: Tabelas cruzadas (*crosstabs*)

Menu:
Analyze
Descriptive Stats
Crosstabs

Weight cases...



Elaboração de tabelas
Em SPSS: Tabelas cruzadas (*crosstabs*)

(*output*)

Agregado familiar * Categorias salariais Crosstabulation

Count

		Categorias salariais				
		< 500 €	501 até 750 €	751 até 1000 €	> 1000 €	Total
Agregado familiar	1		7			7
	2	1	2		1	4
	3	2	4			6
	4	1	1	1		3
Total		4	14	1	1	20

Análise exploratória

Indicadores de localização

Média (*mean*) $\bar{x} = \frac{1}{n}\sum_i x_i f_i$ (dados agrupados) $\bar{x} = \frac{1}{n}\sum_i x_i$ (dados não agrupados)

Mediana (*median*) $M = \frac{N+1}{2}$ (para *N* ímpar) $M = \frac{x_{(N/2)} + x_{(N/2)+1}}{2}$ (para *N* par)

Moda (*mode*): valor mais observado.

Quantis: Indicadores que têm a ver com a ordem das observações. São tanto mais úteis quanto maior fôr o conjunto de dados. Consistem em quartis (*quartiles*) quando se referem ao 1º, 2º e 3º quartos (25%, 50% e 75%) do número total de observações ordenadas crescentemente, ou em percentis *(percentiles)*, quando se referem a qualquer outra localização referente aos valores ordenados. Por exemplo, o quantil$_{10}$ (percentil$_{10}$) refere-se à observação correspondente a ordem 10% do número total de observações ordenadas.

NOTA SOBRE: Observações atípicas: observações aberrantes relativamente ao conjunto de dados. A sua existência tanto pode ser devida a erros de introdução de dados, como pode traduzir um dado fenómeno real. Este facto sugere que sempre que existam observações atípicas se proceda a um estudo pormenorizado sobre a natureza das mesmas.

Análise exploratória

Indicadores de dispersão

Amplitude total (*range*): $A_{pt} = x_n - x_i$ Amplitude inter-quartil: $A_{iq} = q_{75} - q_{25}$

Variância (*variance*): $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Desvio padrão (*standard deviation*): $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Coeficiente de variação (*coefficient of variation*):
Consiste num indicador de dispersão relativa e é dado por: $cv = \frac{s}{\bar{x}}$

Outros indicadores

Assimetria (*skewness*): Tem a ver com a simetricidade da distribuição dos valores observados. Por exemplo, uma curva Normal possui assimetria 0. Valores de assimetria negativa indicam uma cauda da distribuição mais alongada para a esquerda. Valores positivos indicam o contrário.

Curtose (*kurtosis*): Tem a ver com o "peso das caudas" da distribuição e é comparado à curva Normal. Valores superiores a zero indicam caudas menos achatadas do que a Normal, e por isso de menor importância. Valores inferiores a zero indicam caudas mais achatadas do que a curva Normal, ou seja, de maior importância.

Análise exploratória

Representação gráfica de indicadores

Diagramas de caule-e-folhas (*stem-and-leaf plots*): Consistem em diagramas nos quais é representada uma categorização da variável em estudo, indicando os dígitos dominantes das observações (caules), os restantes dígitos das mesmas (folhas), a frequência do número de observações por categoria e ainda a presença de *outliers* ou de valores extremos.

Histogramas (*histograms*): São gráficos que categorizam a variável em estudo e nos quais a área de cada barra é proporcional à frequência de observações dentro da respectiva gama de valores.

Diagramas de extremos e quartis ou caixas de bigodes (*boxplots*): Neste diagramas são simultaneamente representados indicadores de localização e de dispersão. Dentro de algumas variantes, os mais comuns apresentam os valores mínimos e máximo (limites dos bigodes), mediana (2º quartil), e 1º e 3º quartis (limites das caixas).

A Figura seguinte descreve um exemplo de caixa de bigodes.

Análise exploratória

Caixas de bigodes (*boxplots*) - continuação:

Na figura ao lado apresenta-se um *boxplot* hipotético no qual:

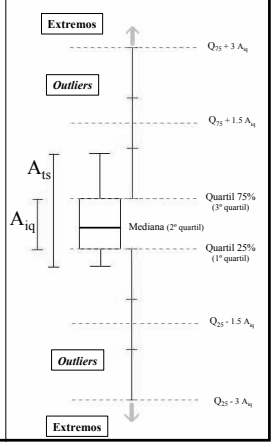
A_{iq} : Amplitude inter-quartil (limites da caixa).
A_{ts} : Amplitude total com excepção de *outliers* e extremos (limites dos bigodes).

Repare-se que existem dois tipos de barreiras (a tracejado) que se estendem a partir dos limites da caixa, com base na Amplitude inter-quartil.

As **barreiras interiores** são definidas como:
$Q_{75} + 1.5\ A_{iq}$
$Q_{25} - 1.5\ A_{iq}$

As **barreiras exteriores** são definidas como:
$Q_{75} + 3\ A_{iq}$
$Q_{25} - 3\ A_{iq}$

Assim, ***outliers*** são observações que se situam entre as barreiras interiores e as exteriores e **extremos** são observações que se situam além das barreiras exteriores.



Análise exploratória

Procedimentos em SPSS

EXEMPLO: A seguinte amostra traduz as quantidades (kg) de cana de açúcar cortada num dia por 15 trabalhadores num país da América latina:

140, 141, 145, 132, 112, 140, 141, 122, 111, 137, 132, 160, 128, 144, 129

Indicadores estatísticos

Analyze
Descriptive Statistics
Frequencies
Statistics Mean Median…kurtosis (indicadores)

Análise exploratória

Procedimentos em SPSS

Indicadores estatísticos (***output***):

Statistics

Cana de açucar cortada (kg)

N	Valid	15
	Missing	0
Mean		134.27
Median		137.00
Mode		132[a]
Std. Deviation		12.820
Variance		164.352
Minimum		111
Maximum		160
Sum		2014
Percentiles	25	128.00
	50	137.00
	75	141.00

a. Multiple modes exist. The smallest value is shown

Análise exploratória

Procedimentos em SPSS

<u>Representação gráfica de indicadores</u>:

- diagramas de caule-e-folhas.
- histogramas.
- caixas de bigodes.

Analyze
Descriptive Statistics
Explore
Plots stem-and-leaf (caule-e-folhas)
boxplots (caixas de bigodes)
histogram

(***output***)

Análise exploratória

Procedimentos em SPSS

Representação gráfica de indicadores (***output***):

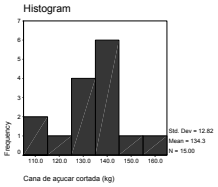
- diagramas de caule-e-folhas
- histogramas
- caixas de bigodes

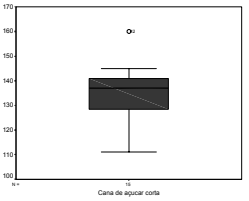
```
Cana de açucar cortada (kg) Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     2.00       11 .  12
     3.00       12 .  289
     3.00       13 .  227
     6.00       14 .  001145
     1.00 Extremes    (>=160)

 Stem width:        10
 Each leaf:       1 case(s)
```





Análise exploratória

Noção de relação entre duas variáveis métricas

Consideremos o seguinte exemplo: à medida que a idade de uma criança aumenta, a sua estatura aumenta também. Diz-se que há uma relação entre as variáveis "idade" e "estatura". Poderá dizer-se ainda que uma variável ("estatura") é função da outra ("idade"). Neste caso, a "estatura" que é função da "idade", isto é, que depende da "idade", denomina-se variável **dependente**. Por seu lado, a "idade", denomina-se variável **independente**.

Representação gráfica da relação entre duas variáveis métricas

Um dos primeiros passos no estudo da relação entre duas variáveis é elaborar uma **nuvem de pontos** (***scatterplot***), isto é, um gráfico que traduza a variação da variável dependente em função da variação da variável independente. Seguindo o nosso exemplo anterior, consideremos uma amostra de 8 crianças e adolescentes dos quais se registou a idade e a estatura.

(**exemplo**)

Análise exploratória

Elaboração de uma nuvem de pontos (*scatterplot*)

Exemplo: consideremos uma amostra de 8 crianças e adolescentes dos quais se registou a idade e a estatura (cm):

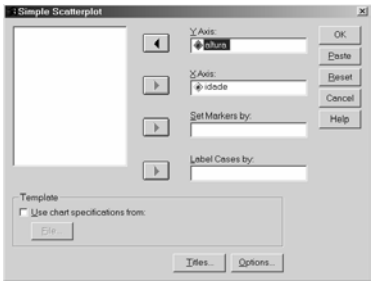
	criancas	idade	estatura
1	1	14	155.00
2	2	6	120.00
3	3	11	145.00
4	4	7	123.00
5	5	13	161.00
6	6	12	150.00
7	7	2	91.00
8	8	4	103.00

⇨ Nuvem de pontos (SPSS):

Análise exploratória

Nuvem de pontos (SPSS):

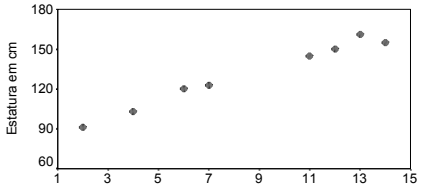
Graphs
Scatter…
Simple (define)
x axis
y axis



(***output***)

Análise exploratória

Nuvem de pontos (*output*):



Através da análise da nuvem de pontos podemos observar que, como esperaríamos, a "estatura" tende claramente a aumentar com a "idade", para os indivíduos amostrados. Neste caso, dizemos que parece existir uma **relação positiva** entre as duas variáveis. Noutros casos, a observação de uma dada nuvem de pontos pode sugerir que a relação entre duas variáveis é **nula**, ou que ela existe mas é **negativa**. Por outro lado, pode acontecer também que, embora seja evidente uma relação entre duas variáveis a mesma **não parece ser linear**. Exemplos…

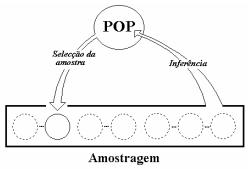
Estatística Inferencial

***A análise exploratória permite APENAS conhecer a(s) amostra(s)*:**

Respondemos a questões do género:

- *Qual o formato da distribuição dos indivíduos da amostra?*
- *A maioria dos indivíduos amostrados são fumadores ou não-fumadores?*
- *Dos indivíduos amostrados, quem vai mais às compras, homens ou mulheres? Quais são os tipos de produtos comprados por uns e outros?*
- …

A Estatística Inferencial permite-nos ir mais além do universo amostral e tirar conclusões acerca da população da qual a amostra foi seleccionada.



Estatística Inferencial

Por exemplo, **com base numa amostra *p*** de *n* indivíduos extraída de uma população **P** de dimensão *N (n << N)* na qual se descrevem as variáveis *sexo*, dimensão do agregado familiar e rendimento mensal do agregado, podemos então responder a questões do género:

- *Será que a proporção de homens na população **P** é 0.43?*
- *A média do número de indivíduos dos agregados familiares da população **P** é superior a 3?*
- *Será 890 € mensais o rendimento médio das famílias pertencentes à população **P**?*

Estatística Inferencial

Este tipo de suposições tem por base apenas uma população. Mas a Inferência pode estender-se a duas ou mais populações. Ambos os casos serão estudados posteriormente mas considerando duas populações $\mathbf{P}_1$ e $\mathbf{P}_2$ e duas amostras $\boldsymbol{p}_1$ e $\boldsymbol{p}_2$ extraídas respectivamente de $\mathbf{P}_1$ e $\mathbf{P}_2$ com as mesmas variáveis do exemplo acima, podemos então responder a questões que comparam efectivamente as pops:

- *O rendimento médio dos agregados é diferente* nas pops $\boldsymbol{P}_1$ e $\boldsymbol{P}_2$?*
- *A proporção de homens em $\boldsymbol{P}_2$ é superior a $\boldsymbol{P}_1$?*

* (des)igualdade matemática x estatística...

Estatística Inferencial

Os conceitos referidos abaixo são fundamentais em testes de hipóteses. Desenvolvamos cada um deles…

- Parâmetros (populacionais) e estatísticas (amostrais);
- Hipóteses nula (H_0) e alternativa (H_1);
- Nível de significância (α) de um teste;
- *p-value* (significância) do teste
- Testes unilaterais e bilaterais;
- Decisão (resposta formal) do teste;
- Tipos de erros (α e β);
- Pressupostos dos testes.

Estatística Inferencial

Os dois grandes grupos de testes: Paramétricos e não-paramétricos.

De uma forma breve (iremos abordar este assunto posteriormente) a diferença entre testes paramétricos e não-paramétricos tem a ver com PRESSUPOSTOS.

A aplicabilidade dos **testes paramétricos**, mais convencionais, dependem de pressupostos tais como a **Normalidade das observações***, entre outros. Muitas vezes este facto constitui uma limitação à análise quando os pressupostos não se verificam.

Os testes **não-paramétricos** não exigem este tipo de pressupostos**. Por este motivo a aplicabilidade dos mesmos é facilitada, embora de um modo geral estes métodos sejam ligeiramente menos robustos*** que os paramétricos.

Utilizaremos posteriormente testes paramétricos e não-paramétricos em paralelo, de acordo com o tipo de análise a efectuar.

Estatística Inferencial

Testes de hipóteses sobre parâmetros

Breve introdução teórica

Média : Comparação de $\bar{x}$ a μ_0 Estatística do teste: $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$

Teste bilateral Hipóteses: H_0: $\mu = \mu_0$; H_1: $\mu \neq \mu_0$

Decisão: Rejeitamos H_0 a um nível de significância α, se $|t| \geq t_{(\alpha/2)(n-1)}$

Teste unilateral Hipóteses:

H_0: $\mu = \mu_0$; H_1: $\mu > \mu_0$ Decisão: Rejeitamos H_0 a um nível de significância α, se $t > t_{\alpha(n-1)}$

H_0: $\mu = \mu_0$; H_1: $\mu < \mu_0$ Decisão: Rejeitamos H_0 a um nível de significância α, se $t < -t_{\alpha(n-1)}$

Estatística Inferencial

Testes de hipóteses sobre parâmetros

Breve introdução teórica

Proporção : Comparação de $\hat{p}$ a p_0

Hipóteses: H_0: $p = p_0$; H_1: $p \neq p_0$

Estatística do teste: $u = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$

Decisão: Rejeitamos H_0 a um nível de significância α, se $|u| \geq z_{(\alpha/2)}$

Estatística Inferencial

Testes de hipóteses sobre parâmetros

Breve introdução teórica

Variância : Comparação de $\hat{s}^2$ a σ_0^2

Hipóteses: H_0: $\sigma^2 = \sigma^2_0$; H_1: $\sigma^2 \neq \sigma^2_0$

Estatística do teste: $$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

Decisão: Rejeitamos H_0 a um nível de significância α, se $|\chi^2| \geq \chi^2_{(\alpha/2)(n-1)}$

Estatística Inferencial

Intervalos de confiança* na estimação de parâmetros

Breve introdução teórica

Consideremos três casos: média, proporção e variância.

Média

$$IC = \left[\bar{x} \pm t_{(\alpha/2)(n-1)} \frac{s}{\sqrt{n}} \right]$$

Proporção

$$IC = \left[\hat{p} \pm z_{(\alpha/2)} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

Variância

$$IC = \left[\frac{(n-1)s^2}{\chi^2_{(\alpha/2)(n-1)}}, \frac{(n-1)s^2}{\chi^2_{1-(\alpha/2)(n-1)}} \right]$$

Notações:

IC = Intervalo de confiança

$t_{(\alpha/2)(n-1)}$ = valor crítico de ordem (α/2) de uma lei de Student com (n-1) graus de liberdade.

α = nível de significância

$z_{(\alpha/2)}$ = valor crítico de ordem (α/2) de uma lei Normal reduzida.

$\chi^2_{1-(\alpha/2)(n-1)}$ e $\chi^2_{1-(\alpha/2)(n-1)}$ = valores críticos de ordens 1-(α/2) e (α/2), respectivamente, de uma lei de Qui-quadrado com (n-1) graus de liberdade.

Estatística Inferencial

Comparação (paramétrica) entre duas populações

Breve introdução teórica

Médias :

Comparação de médias de duas amostras independentes

Hipóteses: H_0: $\mu_1 = \mu_0$; H_1: $\mu_1 \neq \mu_0$

Estatística do teste: $$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

sendo: $$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}$$

Decisão: Rejeitamos H_0 a um nível de significância α, se $|t| \geq t_{(\alpha/2)(n_1+n_2-2)}$

Amostras independentes / amostras emparelhadas

Considerando duas amostras (1 e 2), independência implica que cada observação da amostra 1 não possui nenhum tipo de relação com nenhuma observação da amostra 2.

Emparelhamento implica que cada observação da amostra 1 possui uma determinada relação com uma observação da amostra 2, de tal modo que se pode considerar que os dados ocorrem aos pares.

Exemplo: Níveis de colesterol em 10 pacientes antes (amostra 1) e depois (amostra 2) de um programa de dieta e exercício. Note-se que os indivíduos das amostras 1 e 2 são os mesmos. Por esta razão, existe um emparelhamento.

Estatística Inferencial

Comparação (paramétrica) entre duas populações

Breve introdução teórica

Médias :

Comparação de médias de duas amostras emparelhadas

A idéia fundamental no tratamento de amostra emparelhadas é substituir o conjunto das duas amostras por uma única amostra, a amostra das diferenças dos valores de cada par. A partir daqui podemos recorrer aos métodos de inferência sobre parâmetros de localização, entre os quais, ao teste *t* , já referido, aplicado neste caso para inferir sobre a média populacional das diferenças entre as duas amostras.

Assim, tendo $\bar{x}_d$ = média da amostra das diferenças, queremos comparar $\bar{x}_d$ a μ_0.

Re-escrevamos então as hipóteses nula e alternativa do seguinte modo:
H_0: $\mu_d = \mu_0$; H_1: $\mu_d \neq \mu_0$, sendo $\mu_d = \mu_1 - \mu_2$

Estatística do teste:

$$t = \frac{\bar{x}_d - \mu_0}{s_d / \sqrt{n}}$$

Decisão: Rejeitamos H_0 a um nível de significância α, se $|t| \geq t_{(\alpha/2)(n_1+n_2-2)}$

Estatística Inferencial

Comparação (paramétrica) entre duas populações

Breve introdução teórica

Variâncias

Hipóteses: $H_0 : \sigma_1^2 = \sigma_2^2$; $H_1 : \sigma_1^2 \neq \sigma_2^2$

Estatística do teste: $F = \frac{s_1^2}{s_2^2}$ isto é, o quociente das variâncias amostrais. Na prática, pode inverter-se o numerador / denominador de forma a que o quociente seja >1 (variância do numerador maior). Isto porque se simplifica a decisão do teste:

Decisão: Rejeitamos H_0 a um nível de significância α, se $F \geq F_{(\alpha/2)(n_1-1)(n_2-1)}$

Nota: Se o quociente entre as duas variâncias amostrais fosse <1 (variância do numerador menor que a do denominador), a região crítica do teste seria $F_{1-(\alpha/2)(n_1-1)(n_2-1)}$

Estatística Inferencial

Procedimentos em SPSS

Duas advertências sobre testes de hipóteses em SPSS:

1. REGRAS DE DECISÃO DO TESTE: *p-values*

Até aqui falámos no cálculo de estatísticas e na sua comparação com valores críticos tabelados, de forma a suportar a decisão do nosso teste. O SPSS, tal como a maioria dos outros programas de estatística, não nos fala de valores tabelados, mas sim dá-nos directamente o *p-value*, ou seja, a probabilidade exacta de, sendo H_0 verdadeira, obtermos um valor de teste igual ao que calculámos. Assim, num teste bilateral, se o nosso nível de significância (α) fôr de 0.05, rejeitamos H_0 se obtivermos um *p-value* inferior a 0.05. Por outro lado, num teste unilateral, e considerando o mesmo valor de **α** , rejeitamos H_0 se o **quociente (*p-value* / 2)** fôr inferior a 0.05.

No caso onde pode haver bilateralidade ou unilateralidade do teste estatístico (caso de comparação entre valores amostrais a valores padrão, ou caso de comparação entre duas amostras), as significâncias calculadas pelo SPSS são sempre bilaterais - ***Sig. (2-tailed)*** ou ***Asymp. Sig. (2-tailed)***. Se o teste elaborado pelo utilizador possuir carácter unilateral ter-se-á que proceder à transformação acima referida.

Consideremos o exemplo seguinte:

Estatística Inferencial

Procedimentos em SPSS

Duas advertências prévias sobre testes de hipóteses em SPSS:

1. REGRAS DE DECISÃO DO TESTE: *p-values* (continuação)

EXEMPLO* (*):** Num teste *t* de comparação entre duas médias amostrais, o valor de *t* calculado foi 2.718. O valor de α considerado foi 0.05, e o tamanho da amostra $n = 12$. A significância bilateral do teste (p-value*** ou ***sig. (2-tailed)***) calculada pelo SPSS é 0.02. Assim, num teste bilateral, rejeitamos H_0 visto o valor da significância ser inferior a α. Se o teste fôr unilateral, a respectiva significância seria 0.02/2 = 0.01. Também neste caso se rejeita H_0 uma vez que o valor da significância unilateral do teste é inferior a α.

2. Estatísticas e testes que o SPSS não possui directamente

Na realidade o SPSS não permite calcular directamente intervalos de confiança para proporções e para variâncias, testes paramétricos sobre proporções e testes a variâncias de uma ou duas amostras. No entanto, existe uma alternativa não-paramétrica para testes a proporções (teste binomial), e também a possibilidade de calcularmos as estatísticas e significâncias acima referidas através de funções em *syntax*, como veremos posteriormente.

(*) – Este exemplo foi elaborado com base numa tabela de valores críticos da distribuição *t*.

Estatística Inferencial

Procedimentos em SPSS

Particularidade na introdução de dados:

No caso de comparação entre **duas amostras independentes**, os valores das duas amostras terão que ser introduzidos numa mesma variável (variável de teste). Uma segunda variável especificará a pertença de cada valor à primeira e à segunda amostra (ex. ao lado).

	teste	amostras
1	14	1
2	10	1
3	10	1
4	10	1
5	10	1
6	13	2
7	10	2
8	14	2
9	14	2
10	12	2
11	14	2

Estatística Inferencial

Procedimentos em SPSS

Particularidade na introdução de dados:

No caso de comparação entre **duas amostras emparelhadas**, as amostras são introduzidas em variáveis separadas (ex. ao lado).

	amost1	amost2
1	38.8	26.8
2	51.3	34.8
3	43.8	31.8
4	64.9	56.6
5	29.8	29.0
6	43.4	37.2
7	44.8	36.3
8	33.9	25.2
9	62.7	42.2
10	40.1	29.3
11	55.3	44.1
12	47.4	46.1

Estatística Inferencial

Procedimentos em SPSS

Intervalos de confiança para médias	Analyze Descriptive statistics Explore Statistics Descriptives **confidence interval for mean**
Teste à média de uma população	Analyze Compare means One-sample t test **nome da variável / valor a testar**
Teste à diferença entre as médias de duas populações com base em amostras independentes	Analyze Compare means Independent samples t test test variable **variável de estudo** grouping variable **variável de agrupamento** group 1 (**código 1**) group 2 **(código 2)**
Teste à diferença entre as médias de duas populações com base em amostras emparelhadas	Analyze Compare means Paired sample t test **variável 1** **variável 2**

Estatística Inferencial

Procedimentos em SPSS

Cálculo de estatísticas de teste e de *p-values* em *SYNTAX*

Como referido anteriormente, quando o SPSS não dispõe de caixas de diálogo para um determinado teste, pode recorrer-se à linguagem de syntax para calcular a estatística do teste e o *p-value*, de forma a decidirmos sobre a rejeição ou não da nossa H_0.

Exemplo: comparação de $\hat{p}$ = 0.66 a p_0 = 0.5, com base numa amostra *n* = 20.

Na janela de syntax:

`matrix.`	inicio do comando "matrix"
`compute u=(0.66-0.5)/sqrt((0.5*(1-0.5)/20)).`	calculo de u de acordo com a exp. da pag 3
`compute p=1-cdfnorm(u).`	calculo do p-value (p)
`print u.`	imprimir o valor de u no output
`print p.`	imprimir o valor de p no output
`end matrix.`	fim do comando "matrix".

Resultados (janela de output)

```
Run MATRIX procedure:
U
   1.431083506
P
  10 ** -2   X
   7.620314198
------ END MATRIX -----
```

COMENTÁRIOS: "**sqrt**" = raiz quadrada; "**cdfnorm**" = função do SPSS que calcula a função distribuição acumulada da Normal reduzida até ao valor que lhe damos, neste caso, u=1.431. Como o que queremos não é o acumulado mas a área "restante", isto é, a área de decisão, calculamos 1-cdfnorm(u) para o cálculo do *p-value*. Neste caso, p=0.076. A decisão do teste bilateral a α=0.05 seria não rejeitar H_0, pois o *p-value* não é inferior a 0.05.

Estatística Inferencial

Procedimentos em SPSS

Cálculo de estatísticas de teste e de *p-values* em *SYNTAX*

Outras funções importantes no cálculo de *p-values* em syntax:

CHICDF(m,v) Função distribuição de qui-quadrado. Com base na estatística por nós calculada (m) e dos graus de liberdade (v) a função dá-nos o *p-value*.
Exemplos de aplicação: estimação de intervalos de confiança para variâncias, teste de comparação de uma variância a um valor padrão.

FCDF(m,v1,v2) Função distribuição de *F*. Do mesmo modo: "m"=estatística do teste; "v1"=graus de liberdade do numerador ; "v2"=graus de liberdade do denominador.
Exemplos de aplicação: teste de comparação de duas variâncias.

TCDF(m,v) Função distribuição de *t* - Student. "m"=estatística do teste ; "v"=graus de liberdade.
Exemplos de aplicação: Cálculo imediato de significâncias do valor testado.

Estatística Inferencial N-P

Comparação entre estatística paramétrica e não-paramétrica

Já abordámos previamente o problema da comparação de valores de uma determinada variável em duas populações diferentes, com o objectivo de decidir se é ou não admissível a hipótese de a generalidade dos valores das 2 populações serem globalmente análogos. O ponto de partida para esse estudo são duas amostras (uma de cada população), que podem ser extraídas de forma independente ou através de um delineamento que garanta o seu emparelhamento. A comparação das populações fazia-se mediante a comparação dos indicadores de localização média populacional, concretamente, através de testes à diferença das médias das duas populações.

No entanto, a validade dos testes estudados anteriormente depende principalmente da hipótese de normalidade da distribuição dos valores (ou das diferenças, no caso de amostras emparelhadas) populacionais. Embora os resultados relativos a médias sejam relativamente robustos a desvios à normalidade, têm vindo a ser cada vez mais utilizados os métodos não-paramétricos para um conjunto de testes cuja validade não depende da normalidade dos indivíduos da população.

Seguidamente serão estudados dois desses testes que são utilizados na comparação da localização de populações sem a hipótese de normalidade: **o teste de Wilcoxon-Mann-Whitney** (*Rank sum test*) válido no caso de amostras independentes e **o teste de Wilcoxon** (*Signed rank test*), válido no caso de amostras emparelhadas. Ambos os testes partilham a particularidade de substituirem os valores observados pela ordem *(rank)* dos mesmos.

Estatística Inferencial N-P

O teste de Mann-Whitney* *(rank sum test)*

Consideremos duas amostras independentes ordenadas por ordem crescente (ou decrescente), sendo n_1 = número de observações da amostra 1 e n_2 = número de observações da amostra 2. Suponhamos que, para efeitos de generalização, $n_1 \neq n_2$. À menor das observações do conjunto das duas amostras atribui-se a ordem 1, à segunda menor observação atribui-se a ordem 2, e assim consecutivamente até os $N = n_1 + n_2$ indivíduos se encontrarem ordenados. Note-se que, na ordenação não se perde a correspondência entre o indivíduo e a sua ordem respectiva.

É calculada então a estatística *U* de WMW: $U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$

Sendo R_1 a soma das ordens da amostra 1. No caso do teste ser bilateral é necessário também calcular a estatística *U'*. O maior valor entre *U* e *U'* é então comparado com o valor crítico $U_{(2),(n_1,n_2)}$. *U'* é então dado por:

$$U' = n_1 n_2 - U$$

* Este teste foi primeiramente elaborado por Wilcoxon e posteriormente desenvolvido por Mann e Whitney. No entanto, na bibliografia e programas de estatística encontra-se referido, na maioria das vezes, apenas por Mann-Whitney.

Estatística Inferencial N-P

O teste de Mann-Whitney *(rank sum test)*

Exemplo*: Queremos testar, a um nível de significância α=0.05, a hipótese bilateral de não existir diferenças entre a altura de estudantes do sexo masculinos e feminino.

H_0 : Alunos e alunas são da mesma altura
H_1 : Alunos e alunas não são da mesma altura.

Altura do alunos	Altura das alunas	Ordens das Alturas do alunos	Ordens da Alturas das alunas
170	163	4	1
178	165	7	2
180	168	8	3
183	173	9	5
185	175	10	6
188		11	
193		12	
n_1=7	n_2=5	R_1=61	R_2=17

U = 7*5+((7*(7+1))/2)-61
U = 2

U' = 7*5-2
U' = 33

*Adaptado de Zar (1984).

Estatística Inferencial N-P

O teste de Mann-Whitney *(rank sum test)*

Exemplo (continuação): Consultemos então uma tabela de valores críticos da distribuição U de WMW, para um nível de significância de 0.05, $n_1 = 5$, $n_2 = 7$:

TABLE B.10 (cont.) CRITICAL VALUES OF THE MANN-WHITNEY *U* DISTRIBUTIC

	a(2):	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
	a(1):	0.10	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
n_1	n_2								
4	32		91	98	104	110	114	117	120	122
	33		94	101	107	113	117	120	124	126
	34		96	104	110	116	120	124	127	130
	35		99	107	113	120	124	127	131	133
	36		102	110	116	123	127	131	135	137
	37		105	113	119	126	131	134	138	141
	38		107	116	122	130	134	138	142	144
	39		110	118	125	133	137	141	145	148
4	40		113	121	129	136	141	145	149	152
5	5		20	21	23	24	25	--	--	--
	6		23	25	27	28	29	30	--	--
	7		25	29	30	32	34	35	--	--
	8		30	32	34	36	38	39	40	--
	9		33	36	38	40	42	43	44	45
	10		37	39	42	44	46	47	49	50
	11		40	43	46	48	50	52	53	54

(Continua)

Decisão:

Uma vez que o valor crítico do teste é $U_{0.05(2),(5,7)}$= 30 (tabela), rejeitamos H_0 pois U'=33 > 30.

Concluímos que, a um nível de significância α=0.05, os alunos e alunas não são da mesma altura.

Estatística Inferencial N-P

O teste de Mann-Whitney *(rank sum test)*

Casos particulares:

1. A existência de "empates" (*tied ranks*)

No exemplo anterior não encontramos, no conjunto das duas amostras, observações iguais. No entanto, pode acontecer, e neste caso estamos em presença de "empates" *(tied ranks*). Nesse caso, as ordens de observações de igual valor são a média das ordens que teriam sido atribuídas a estas observações, caso não existisse empate. Como exemplo consideremos o exercício anterior. Se o 4º valor mais elevado das alunas fosse 170 e não 173, teríamos um empate com o valor mais baixo dos alunos. Concretamente, entre as ordens 4 e 5. Neste caso atribuir-se-ia a ordem (4+5)/2 = 4.5 a cada uma das observações. Uma vez efectuada esta alteração, o procedimento segue como referido anteriormente.

2. O teste unilateral

Nos testes unilaterais há que decidir se calcular *U* ou *U'*, de acordo com o tipo de hipótese a testar (quadro ao lado). O valor crítico tabelado será $U\alpha_{(1),(n_1,n_2)}$.

	H_0: grupo 1 ≥ grupo 2 H_1: grupo 1 < grupo 2	H_0: grupo 1 ≤ grupo 2 H_1: grupo 1 > grupo 2
Ordenação crescente	*U*	*U'*
Ordenação decrescente	*U'*	*U*

Estatística Inferencial N-P

O teste de Mann-Withney *(rank sum test)*

Procedimento em SPSS

Introdução dos dados — À semelhança do que já vimos para testes *t* a duas amostras independentes, uma variável contém os valores das duas amostra, havendo uma segunda variável que especifica os códigos de pertença de cada valor à variável 1 ou 2.

Menus

Analyze
Nonparametric Tests
2 Independent Samples
Test variable / grouping variable
Test Type **Mann-Whitney U**
OK

Estatística Inferencial N-P

O teste de Wilcoxon *(signed-rank test)*

Do mesmo modo que o teste de Wilcoxon-Mann-Whitney constitui uma alternativa não-paramétrica ao teste *t* para duas amostras independentes, o teste de Wilcoxon apresenta-se como alternativa não-paramétrica ao teste *t* para duas amostras emparelhadas.

O teste consiste nos seguintes procedimentos sequenciais:
1. Cálculo das diferenças entre os pares de observações (tal como o teste *t* para amostras emparelhadas).
2. Ordenação crescente dos valores absolutos das diferenças.
3. Atribuição do sinal de cada diferença (- ou +) aos valores ordenados anteriormente.
4. Cálculo de T_+ (soma das ordens de sinal positivo) e de T_- (soma das ordens de sinal negativo).

Decisão do teste: Rejeitamos H_0 se T_+ ou T_- fôr menor que o valor crítico $T_{\alpha(2),n}$.

NOTA: No caso da existência de empates, o procedimento é igual ao que foi referido para o teste de WMW, ou seja, atribui-se a média das ordens que as duas observações possuiriam se não ocorresse empate.

Estatística Inferencial N-P

O teste de Wilcoxon *(signed-rank test)*

Exemplo: Queremos testar, a um nível de significância α=0.05, a hipótese bilateral de não existirem diferenças entre o grau de queimadura solar com um dado protector (ombro esquerdo dos indivíduos) e sem protector solar (ombro direito dos indivíduos).

H_0 : Não há diferenças entre os graus de queimadura solar (o protector não faz efeito).
H_1 : Há diferenças entre os graus de queimadura solar (o protector faz efeito).

Indivíduo	Ombro sem protector	Ombro com protector	Diferenças	Ordem das diferenças absolutas	Ordens afectadas de sinal
1	142	138	4	4.5	4.5
2	140	136	4	4.5	4.5
3	144	147	-3	3	-3
4	144	139	5	7	7
5	142	143	-1	1	-1
6	146	141	5	7	7
7	149	143	6	9.5	9.5
8	150	145	5	7	7
9	142	136	6	9.5	9.5
10	148	146	2	2	2

n = 10
T_+= 4.5+4.5+7+7+9.5+7+9.5+2 = **51**
T_-= 3+1 = **4**
$T_{0.05\,(2),10}$= **8** (tabela)

Estatística Inferencial N-P

O teste de Wilcoxon *(signed-rank test)*

Exemplo (continuação): Consultemos então uma tabela de valores críticos da distribuição *T* de Wilcoxon, para um nível de significância de 0.05 e n = 10:

TABLE B.11 CRITICAL VALUES OF THE WILCOXON *T* DISTRIBUTION

	$\alpha(2) = 0.50$	0.20	0.10	0.05	0.02	0.01
	$\alpha(1) = 0.25$	0.10	0.05	0.025	0.01	0.005
n						
4	2	0				
5	4	2	0			
6	6	3	2	0		
7	9	5	3	2	0	
8	12	8	5	3	1	0
9	16	10	8	5	3	1
10	20	14	10	8	5	3
11	24	17	13	10	7	5
					9	7

(Continua)

Decisão:

Uma vez que o valor crítico do teste é $U_{0.05(2)(10)}$= 8 (tabela), rejeitamos H_0 pois T_-= 4 < 8.

Concluímos que, a um nível de significância α=0.05, o protector solar faz efeito.

Estatística Inferencial N-P

O teste de Wilcoxon *(signed-rank test)*

Procedimento em SPSS

Introdução dos dados — À semelhança do que já vimos para testes *t* a duas amostras emparelhadas, cada variável contém os valores de cada amostra.

Menus

Analyze
Nonparametric Tests
2 Related Samples **Variable 1** -- **Variable 2**
Test Type **Wilcoxon**
OK

Lista de alguns *sites* de ajuda *online* em Estatística e Análise de dados:

- http://davidmlane.com/hyperstat/index.html
- http://math.about.com/cs/statistics/
- http://www.meandeviation.com/tutorials/stats/
- http://biology.nebrwesleyan.edu/empiricist/sources/tips/stats1.html
- http://www.math.ist.utl.pt/stat/

Diapositivos de apoio à disciplina de

Análise Avançada de Dados

Docente: Pedro Sousa (pedro.sousa.mrm@gmail.com)

Estatística? Para quê?… Relembremos os conceitos básicos

A Estatística é a *ferramenta* que nos permite conhecer melhor as características de um dado conjunto de indivíduos, ao qual chamamos **população**, e também tirar conclusões válidas acerca dessa população. Uma vez que a análise de todos os indivíduos da população é, na maioria das vezes, impraticável (ver exemplos abaixo), pode-se estudar apenas uma pequena parte dessa população à qual chamamos **amostra**. O objectivo é **inferir** certos factos da população, a partir dos resultados observados na amostra. Este processo denomina-se **inferência estatística**. Ao processo de obtenção de amostras chamamos **amostragem**.

Neste contexto, temos dois grandes tipos de aplicações da estatística: a **Estatística Descritiva**, quando apenas se pretende explorar e conhecer melhor um conjunto de dados, e a **Estatística Inferencial**, quando se pretende inferir, ou seja, tirar conclusões acerca da população, com base nas características da amostra.

Estatística? Para quê?… Relembremos os conceitos básicos

Exemplo 1: Pretende-se tirar conclusões acerca das alturas de 20 000 alunos que frequentam uma dada Universidade (**população**), com base numa **amostra** de 100 desses alunos.

Exemplo 2: Pretende-se saber se a aplicação de determinada vacina aos habitantes de um país (**população**) é eficiente ou não, com base nos resultados da aplicação da vacina a 200 desses habitantes (**amostra**).

Deste modo, a estatística é hoje, e cada vez mais, fundamental em Medicina, Farmacologia, Ciências da comunicação, Engenharias, Biologia, Pescas, Ecologia, Psicologia e em todas as muitas outras áreas científicas.

Medidas de associação Estatística:

I. Covariância e Coeficiente de Correlação Linear (Pearson)

Considerando que é sugerida a existência de uma relação linear (positiva ou negativa) entre duas variáveis métricas *X* e *Y*, torna-se necessário quantificar essa relação. Neste contexto, definamos o conceito de covariância. Do mesmo modo que a variância de uma dada variável consiste na soma média de quadrados dos desvios das observações relativamente à sua média, a covariância consiste no produto médio dos desvios das observações relativos às respectivas médias:

$$Cov(X,Y) = \sum (x_i - \bar{x})(y_i - \bar{y}) / (n-1)$$

A covariância é uma medida do grau de relação entre duas variáveis, mas não é um índice, pois depende das unidades das variáveis. Se padronizarmos a cov(X,Y) pelo produto dos desvios padrão amostrais obtemos o coeficiente de correlação linear de Pearson *(r)*:

Covariância e Coeficiente de Correlação Linear (Pearson)

$$r = \frac{\mathrm{cov}(X,Y)}{s_x s_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$$r = \frac{n\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}$$

O coeficiente de Pearson varia de -1 (correlação negativa perfeita), a 1 (correlação positiva perfeita), significando um valor de 0 ausência de correlação linear.

Para verificar a significância de um dado valor de *r*, recorre-se à respectiva tabela de valores críticos. Em alternativa, calcula-se a estatística *t* com base no número de pares de observações e no valor de *r* . Esta estatística distribui-se como um *t* de *Student* com (n-2) graus de liberdade, sob H_0.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Em *SPSS* a significância de *r* é tabelada e basta-nos comparar este valor com α (regra universal).

Covariância e Coeficiente de Correlação Linear (Pearson)

Etapas de análise em SPSS

1. Descrever graficamente a relação entre as duas variáveis em estudo (nuvem de pontos):

   Graphs
   Scatter…
   Simple Define **y axis** **x axis** (label cases by)
   OK

2. Se é sugerida a existência de uma relação linear positiva ou negativa entre as variáveis, calcular o coeficiente de Pearson e a respectiva significância:

   Analyze
   Correlate
   Bivariate
   variables **correlation coefficient**
   OK

Covariância e Coeficiente de Correlação Linear (Pearson)

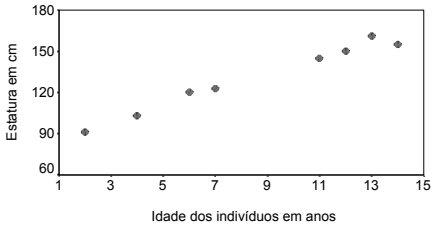
Exemplo: amostra de 8 crianças e adolescentes dos quais se registou a idade e a estatura (cm) - DADOS:

	criancas	idade	estatura
1	1	14	155.00
2	2	6	120.00
3	3	11	145.00
4	4	7	123.00
5	5	13	161.00
6	6	12	150.00
7	7	2	91.00
8	8	4	103.00

⇨ Nuvem de pontos:

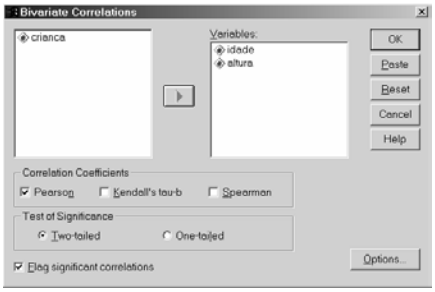
Covariância e Coeficiente de Correlação Linear (Pearson)

Exemplo: amostra de 8 crianças e adolescentes dos quais se registou a idade e a estatura (cm) – NUVEM DE PONTOS:



Covariância e Coeficiente de Correlação Linear (Pearson)

Exemplo: amostra de 8 crianças e adolescentes dos quais se registou a idade e a estatura (cm) – SPSS *DIALOG BOX*:



Covariância e Coeficiente de Correlação Linear (Pearson)

Exemplo: amostra de 8 crianças e adolescentes dos quais se registou a idade e a estatura (cm) - RESULTADOS:

Correlations

		IDADE	ALTURA
IDADE	Pearson Correlation	1	.990**
	Sig. (2-tailed)	.	.000
	N	8	8
ALTURA	Pearson Correlation	.990**	1
	Sig. (2-tailed)	.000	.
	N	8	8

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

Correlação não-paramétrica: Coeficiente de Spearman

Se os nossos dados se afastam bastante da normalidade, o coeficiente de Pearson não é, na generalidade, aplicável. A alternativa mais conhecida não-paramétrica ao coeficiente de Pearson é o coeficiente ordinal de Spearman, o qual utiliza as ordens das observações de cada variável. O teste pode ser resumido nos seguintes procedimentos:

1. Atribuição de ordens a cada uma das variáveis, separadamente (ao contrário do efectuado para o teste não-paramétrico de Wilcoxon para duas amostras emparelhadas).
2. Calcular o vector de diferenças entre os pares de ordens.

Correlação não-paramétrica: Coeficiente de Spearman

3. Calcular a estatística do teste, dada por:

$$r_s = 1 - \frac{6\sum d_i^2}{n^3 - n}$$

Sendo: "r_s" o coeficiente amostral de Spearman, "d_i^2" o quadrado das diferenças das ordens e "n" o número de pares de observações.

4. Decisão do teste: comparação de r_s com o respectivo valor crítico tabelado (α,n). Rejeita-se H_0 se o valor calculado fôr superior ao tabelado. Em SPSS, utiliza-se a regra universal de decisão (*).

Correlação não-paramétrica: Coeficiente de Spearman

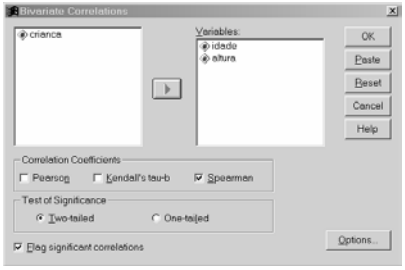
Exemplo 1:
Considere-se uma dada amostra bi-variada ($\mathbf{X}_1$, $\mathbf{X}_2$) n = 7.
Adopte-se α = 0.05.

		Ordens			
X_1	X_2	O_{X1}	O_{X2}	Diferença (O_{x1} - O_{x2})	Diferença^2
4.1	12.0	2	2	0	0
4.0	13.1	1	3	-2	4
6.0	11.0	3	1	2	4
6.3	16.2	4	4	0	0
9.8	19.7	5	5	0	0
12.1	20.8	7	6	1	1
11.3	21.4	6	7	-1	1

O somatório das diferenças elevadas ao quadrado é, neste caso, 10. Aplicando a fórmula da página anterior, $\boldsymbol{r_s}$ **= 0.821**. Uma vez que o valor crítico tabelado é **0.786** (teste bilateral), rejeita-se H_0 pois r_s é superior ao valor crítico, ou seja, as duas variáveis são correlacionadas. **Em SPSS, utiliza-se a regra universal de decisão** (*).

Correlação não-paramétrica: Coeficiente de Spearman

Exemplo 2: amostra de 8 crianças e adolescentes dos quais se registou a idade e a estatura (cm) – SPSS *DIALOG BOX*:



Correlação não-paramétrica: Coeficiente de Spearman

Exemplo 2: amostra de 8 crianças e adolescentes dos quais se registou a idade e a estatura (cm) - RESULTADOS:

Correlations

			IDADE	ALTURA
Spearman's rho	IDADE	Correlation Coefficient	1.000	.976**
		Sig. (2-tailed)	.	.000
		N	8	8
	ALTURA	Correlation Coefficient	.976**	1.000
		Sig. (2-tailed)	.000	.
		N	8	8

**. Correlation is significant at the .01 level (2-tailed).

Medidas de associação Estatística:

II. Análise de Tabelas de Contingência: Teste de Qui-quadrado

A secção anterior abordou métodos de quantificação do grau de associação entre duas variáveis métricas, utilizando métodos paramétricos e não-paramétricos.

No caso das duas (ou mais) variáveis para as quais se pretende estudar a relação serem de natureza **categórica** (nominais ou ordinais) o procedimento é, consequentemente, diferente, o qual se abordará na presente secção.

Análise de Tabelas de Contingência: Teste de Qui-quadrado

Considere-se, como exemplo, que se fez um questionário a 15 pessoas no qual se registou o sexo (feminino = 0, masculino = 1) e se o indivíduo fuma (1) ou não (0).

Os resultados (já numa folha de dados do SPSS) foram os apresentados ao lado (códigos e respectivos *labels*).

O **objectivo** é verificar se existe uma relação significativa entre sexo e fumador, isto é, se as mulheres fumam mais que os homens, ou vice-versa.

Códigos		*Labels*	
sexo	fumador	sexo	fumador
1	0	masc.	não
0	0	fem.	não
0	0	fem.	não
0	1	fem.	sim
1	0	masc.	não
1	0	masc.	não
0	1	fem.	sim
1	0	masc.	não
1	0	masc.	não
0	1	fem.	sim
0	1	fem.	sim
0	1	fem.	sim
1	1	masc.	sim
1	0	masc.	não
0	1	fem.	sim

Análise de Tabelas de Contingência: Teste de Qui-quadrado

O cruzamento das duas variáveis, isto é, a contagem do número de observações que correspondem a cada par de categorias, resulta numa **TABELA DE CONTINGÊNCIA**. Neste exemplo, esta tabela assume a forma:

SEXO * FUMADOR Crosstabulation

Count

		FUMADOR		
		não	sim	Total
SEXO	fem.	2	6	8
	masc.	6	1	7
Total		8	7	15

Análise de Tabelas de Contingência: Teste de Qui-quadrado

Neste ponto, urgem as seguintes observações:

- Já estimámos a correlação entre variáveis métricas (contínuas), quer através do coeficiente linear de Pearson ou do coef. não-paramétrico de Spearman. Neste caso, as variáveis não são métricas mas sim **categóricas (nominais)**.
- Na tabela de contingência anterior, além das contagens existem os totais marginais e o número total de observações. Os totais marginais das linhas (variável sexo) dão-nos o número total de indivíduos do sexo masculino e feminino. Os totais marginais das colunas (variável fumador) dão-nos o número total de indivíduos que fumam e que não fumam.

Análise de Tabelas de Contingência: Teste de Qui-quadrado

Mais observações:

- O número total de observações corresponde, assim, tanto à soma das observações no "interior" da tabela, como à soma dos totais marginais das linhas ou das colunas.
- Tratando-se de variáveis categóricas (nominais), o número de categorias pode ser 2, ou mais. Assim, atribui-se a nomenclatura **tabela de contingência 2 × 2** (como neste exemplo), **2 × 3**, **4 × 4**, etc., de acordo com o número de categorias que cada variável possui (ver exemplo seguinte).

Análise de Tabelas de Contingência: Teste de Qui-quadrado

Exemplo de uma tabela de contingência 2 × 4:

	Cor do cabelo				
Sexo	Negro	Castanho	Louro	Ruivo	Total
Masculino	32	43	16	9	100
	(29.0)	(36.0)	(26.67)	(8.33)	
Feminino	55	65	64	16	200
	(58.0)	(72.0)	(53.33)	(16.67)	
Total	87	108	80	25	300

Pergunta-se: "o que são os números entre parêntesis?"…
Para responder, voltemos ao primeiro exemplo:

Análise de Tabelas de Contingência

			FUMADOR não		FUMADOR sim		Total
SEXO	fem.	número de observações	2		6		8
		% do total de "sexo"	25	*1	75	*2	100
		% do total de "fumador"	25		85.7		53.3 *7
	masc.	número de observações	6		1		7
		% do total de "sexo"	85.7	*3	14.3	*4	100
		% do total de "fumador"	75		14.3		46.7 *8
Total		número de observações	8		7		15
		% do total de "sexo"	53.3	*5	46.7	*6	100
		% do total de "fumador"	100		100		100

Legenda:

*1: (2 / 8)*100 = 25% (% do número de indivíduos do sexo fem. que não fumam)
*2: (6 / 8)*100 = 75% (% do número de indivíduos do sexo fem. que fumam)
*3: (6 / 7)*100 = 85.7% (% do número de indivíduos do sexo masc. que não fumam)
*4: (1 / 7)*100 = 14.3% (% do número de indivíduos do sexo masc. que fumam)
*5: (8 / 15)*100 = 53.3% (% do número total de indivíduos que não fumam)
*6: (7 / 15)*100 = 46.7% (% do número total de indivíduos que fumam)
*7: (8 / 15)*100 = 53.3% (% do número total de indivíduos do sexo feminino)
*8: (7 / 15)*100 = 46.7% (% do número total de indivíduos do sexo masculino)

Análise de Tabelas de Contingência: Teste de Qui-quadrado

O princípio de análise tem a ver com um raciocínio do tipo:

Havendo um total de 53.3% de indivíduos do sexo feminino, **ESPERA-SE** que, se não existir relação entre as duas variáveis, 53.3% do número total de **não fumadores** serão mulheres (53.3% de 8 = **4.3**) e 46.7 % serão homens (46.7% de 8 = **3.7**) . Do mesmo modo, 53.3% do número total de **fumadores** serão mulheres (53.3% de 7 = **3.7**) e 46.7 % serão homens (46.7% de 7 = **3.3**).

Estes são os **VALORES ESPERADOS** do cruzamento das categorias das duas variáveis, atrás assinalados entre parêntesis. Neste exemplo, assinalem-se na tabela seguinte.

O teste estatístico que permite verificar se existe uma relação significativa entre as duas variáveis compara os valores observados com os esperados (ver formulação seguinte).

Análise de Tabelas de Contingência: Teste de Qui-quadrado

Tabela de contingência com valores observados *(count)* e esperados *(expected count*):

SEXO * FUMADOR Crosstabulation

			FUMADOR		
			não	sim	Total
SEXO	fem.	Count	2	6	8
		Expected Count	4.3	3.7	8.0
	masc.	Count	6	1	7
		Expected Count	3.7	3.3	7.0
Total		Count	8	7	15
		Expected Count	8.0	7.0	15.0

Análise de Tabelas de Contingência: Teste de Qui-quadrado

O valor do teste (Qui-quadrado de Pearson) é-nos dado por:

$$\chi^2 = \sum_{i=1}^{l}\sum_{j=1}^{c}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

Onde O_{ij} são os valores observados e E_{ij} os observados.

No caso do nosso exemplo este valor seria obtido do seguinte modo:

$$\chi^2 = \frac{(2-4.3)^2}{4.3} + \frac{(6-3.7)^2}{3.7} + \frac{(6-3.7)^2}{3.7} + \frac{(1-3.3)^2}{3.3} \approx 5.6$$

Análise de Tabelas de Contingência: Teste de Qui-quadrado

O grau de associação entre as variáveis é-nos dado pelo coeficiente de associação (ou de contingência), calculado com base no valor de qui-quadrado anterior:

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}}$$

Este valor, situa-se entre 0 e 1 (máxima associação).

A **significância** do valor de qui-quadrado (*chi-square*) e do coeficiente de associação *(contingency coefficient)* é-nos dado pelo SPSS. A regra de decisão estatística* é posteriormente aplicada. As hipóteses previamente estabelecidas são:

H_0: As variáveis são independentes (não há relação entre as mesmas).

H_1: Há relação significativa entre as variáveis.

Análise de Tabelas de Contingência: Teste de Qui-quadrado

Procedimento em SPSS

A introdução dos dados:

Há duas situações distintas: ou os dados são introduzidos *em bruto*, isto é, segundo o exemplo do *slide* 114, ou num formato mais parecido com a tabela de contingência, que já resume a informação anterior. Vejamos esta última alternativa, usando os mesmo dados do *slide* 114:

Códigos

sexo	fumador	observ
0	0	2
0	1	6
1	0	6
1	1	1

Ou

Labels

sexo	fumador	observ
fem.	não fuma	2
fem.	fuma	6
masc.	não fuma	6
masc.	fuma	1

Análise de Tabelas de Contingência: Teste de Qui-quadrado

Procedimento em SPSS

A introdução dos dados: dados ponderados

Neste caso os dados estão ponderados pela variável "observ", isto é, lendo as linhas da tabela, o número de indivíduos do sexo feminino (0) que não fumam (0) é 2; o número de indivíduos do sexo feminino (0) que fumam (1) é 6; etc.

Por este motivo, neste caso, há que "informar" o SPSS que existe esta ponderação:

Data

Weight cases

Weight cases by **nome da variável** OK

Deste modo, este procedimento não precisa de ser efectuado se se utilizar uma tabela de dados do primeiro tipo (*slide* 114).

Análise de Tabelas de Contingência: Teste de Qui-quadrado

Exemplo em SPSS
Análise de qui-quadrado

Procedamos agora à análise de contingência:

Analyze

Descriptive statistics

Crosstabs

Row(s) **nome da variável *linha***

Column(s) **nome da variável *coluna***

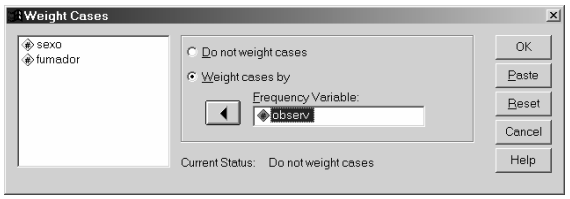
Statistics **chi-square** **contingency coefficient** CONTINUE

Cells **observed** **expected** CONTINUE

OK

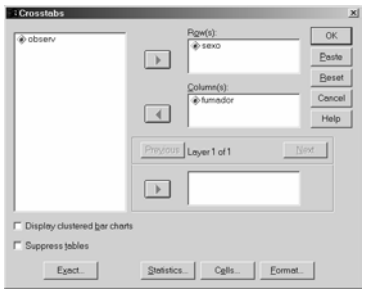
Análise de Tabelas de Contingência: Teste de Qui-quadrado

Procedimento em SPSS
Ponderação de observações *(weight cases)*:
Menu: Data \ Weight cases



Análise de Tabelas de Contingência: Teste de Qui-quadrado

Procedimento em SPSS
Análise de contingência:
Menu: Analyze \ Descriptive Statistics \ Crosstabs

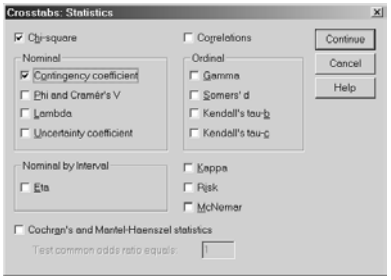


Análise de Tabelas de Contingência: Teste de Qui-quadrado

Procedimento em SPSS

Análise de contingência:

Menu: Analyze \ Descriptive Statistics \ Crosstabs \ Statistics

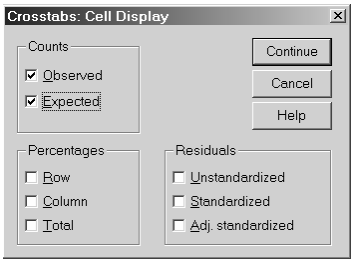


Análise de Tabelas de Contingência: Teste de Qui-quadrado

Procedimento em SPSS

Análise de contingência:

Menu: Analyze \ Descriptive Statistics \ Crosstabs \ Cells



Análise de Tabelas de Contingência: Teste de Qui-quadrado

Procedimento em SPSS

Análise de contingência:

Resultados: tabela de contingência: *slide* 79; tabela de valores críticos e significâncias:

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	5.529[b]	1	.019		
Continuity Correction[a]	3.359	1	.067		
Likelihood Ratio	5.989	1	.014		
Fisher's Exact Test				.041	.032
Linear-by-Linear Association	5.161	1	.023		
N of Valid Cases	15				

Análise de Tabelas de Contingência: Teste de Qui-quadrado

Procedimento em SPSS

Análise de contingência:

Resultados: tabela de valores críticos e significâncias (discussão).

Estão realçados o valor de qui-quadrado (5.529) e a respectiva significância (0.019): para um alfa de 0.05, rejeitamos H_0, isto é, as variáveis são correlacionadas.

Observando a tabela de contingência verifica-se que o valor observado de mulheres que fumam (6) é superior ao esperado (3.7) e que o valor observado de homens que fumam (1) é inferior ao esperado (3.3).

O teste concluiu que estas diferenças são significativas (α=0.05), ou seja, pode-se afirmar que, a um nível de significância α=0.05, as mulheres fumam mais que os homens, na população amostrada.

Análise de Tabelas de Contingência: Teste de Qui-quadrado

Procedimento em SPSS

Análise de contingência:

Resultados: tabela do coeficiente de contingência:

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Contingency Coefficient	.519	.019
N of Valid Cases		15	

Estão realçados o valor do coeficiente de contingência (0.519) e a respectiva significância (0.019): para um alfa de 0.05, este valor de C é significativo.

ANÁLISE DE VARIÂNCIA

(ANOVA - *Analysis Of Variance)*

Comparações entre k > 2 amostras

INTRODUÇÃO

Verificámos anteriormente que, quando queremos comparar as médias de duas amostras referentes a uma dada variável, utilizamos o teste *t*. No entanto, frequentemente queremos comparar mais que duas amostras. Poderá ser tentador utilizar testes *t* a todos os possíveis pares de amostras, mas este procedimento para k > 2 amostras é inválido: o que sucede é que, enquanto entre duas amostras a probabilidade de cometermos um erro de tipo I (rejeitar H_0 quando verdadeira) é α (por exemplo 5%), se aplicarmos este procedimento a mais que duas amostras, é demonstrado (Zar, 1984), que este erro aumenta proporcionalmente em relação ao número de amostras que consideramos. Por exemplo, se estabelecemos α=5% e são comparadas 3 médias, temos 13% de probabilidade de concluir erradamente uma diferença entre as duas médias mais extremas (erro tipo I). Para 10 médias esta probabilidade aumenta para 63% e considerando 20 médias temos 93% de probabilidade de cometer um erro tipo I.

ANÁLISE DE VARIÂNCIA
(ANOVA - *Analysis Of Variance)*
Comparações entre k > 2 amostras

INTRODUÇÃO

Neste contexto, a Análise de Variância (ANOVA) permite verificar qual é o efeito de uma ou mais variáveis independentes (factores) numa variável dependente. A questão central da ANOVA consiste em saber se as populações têm ou não, estatisticamente, médias iguais. Formulemos então as hipóteses nula e alternativa:

H_0: $\mu_1 = \mu_2 = ... = \mu_k$

H_1: As médias das populações não são iguais.

Um aspecto importante da hipótese alternativa é que a mesma contempla não só o caso em que existe desigualdade estatística entre todas as *k* médias, mas também casos em que apenas uma média difira do conjunto das restantes *k*-1.

ANÁLISE DE VARIÂNCIA
(ANOVA - *Analysis Of Variance)*
Comparações entre k > 2 amostras

INTRODUÇÃO

Na análise de variância os factores podem ser fixos (determinados *a priori*) ou aleatórios (escolhidos aleatoriamente). Consoante a aleatoriedade ou não dos factores temos 3 tipos de ANOVA: Tipo I (factores fixos), Tipo II (factores aleatórios) e Tipo III (factores fixos e aleatórios). Considerando o número de factores, temos ANOVA a 1 factor (*one-way ANOVA*) a 2 factores (*Two-way ANOVA*), etc.

A idéia do teste estatístico (*F*) consiste em comparar a variação entre os *k* grupos (amostras) com a variação dentro dos grupos. Se a variação entre os grupos fôr grande comparativamente à variação dentro dos grupos, podemos dizer que há diferenças entre as médias que queremos comparar.

ANÁLISE DE VARIÂNCIA
(ANOVA - *Analysis Of Variance)*
Comparações entre k > 2 amostras

INTRODUÇÃO

Mais concretamente, considerando uma ANOVA a 1 factor fixo (Tipo I), o teste estatístico consiste no cálculo do seguinte valor de *F*:

$$F_{(k-1,n-k)} = \frac{\sum_{i=1}^{k} n_i \left(\bar{x}_i - \bar{\bar{x}}\right)^2 \Big/ k-1}{\sum_{i=1}^{k}\sum_{j=1}^{n} \left(x_{ij} - \bar{x}_i\right)^2 \Big/ n-k}$$

k = número de grupos
n = número total de observações
n_i = número de observações do grupo *i*
x_{ij} = observação *j* do grupo *i*
$\bar{x}_i$ = média das observações do grupo *i*
$\bar{\bar{x}}$ = média de todas as *n* observações

ANÁLISE DE VARIÂNCIA
(ANOVA - *Analysis Of Variance)*
Comparações entre k > 2 amostras

INTRODUÇÃO

É frequente elaborar-se o cálculo da estatística F atrás referida, por etapas, num quadro de ANOVA. Para uma ANOVA a um factor fixo, o quadro genérico é representado abaixo.

Fonte de variação	Soma de quadrados	Graus de liberdade	Quadrados Médios	F
Entre grupos	SQE	k - 1	QME=SQE/(k-1)	QME/QMR
Dentro grupos (resíduo, erro)	SQR	n - k	QMR=SQR/(n-k)	
Total	SQT	n - 1		

Como pode ser observado na fórmula da página anterior, SQE e SQR são dados pelas expressões abaixo:

$$SQE = \sum_{i=1}^{k} n_i \left(\bar{x}_i - \bar{\bar{x}}\right)^2 \qquad SQR = \sum_{i=1}^{k} \left[\sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_i\right)^2 \right]$$

Comparações múltiplas posteriores à ANOVA: O teste de Tukey

Supondo que o valor da estatística F é significativo (α), podemos querer saber entre que grupos (amostras) é que existem diferenças. Isto porque, basta duas de entre as k amostras serem diferentes estatisticamente (embora todas as outras iguais) para rejeitarmos H_0, isto é, para o valor de F calculado ser significativo. O teste mais conhecido para verificar estas diferenças é o de TUKEY:

Consideremos que, para k grupos, ocorrem k (k -1)/2 combinações de pares de grupos. Por exemplo, para 3 grupos (A, B e C) temos 3(3-1)/2 = **3** pares de grupos, nomeadamente, A-B, A-C e B-C. O procedimento do teste de Tukey é representado na tabela abaixo:

Comparação	Diferenças	SE	q	q (α,n,k)	Decisão do teste
A vs. B	$\overline{X}_{A,B} = \overline{x}_A - \overline{x}_B$	$SE_{A,B} = \sqrt{QMR/n_{A,B}}$	$q_{A,B} = \overline{X}_{A,B} / SE_{A,B}$	Tabela (q_t)	Rejeito H_0 se q>q_t
A vs. C	$\overline{X}_{A,C} = \overline{x}_A - \overline{x}_C$	$SE_{A,C} = \sqrt{QMR/n_{A,C}}$	$q_{A,C} = \overline{X}_{A,C} / SE_{A,C}$	Tabela (q_t)	Rejeito H_0 se q>q_t
B vs. C	$\overline{X}_{B,C} = \overline{x}_B - \overline{x}_C$	$SE_{B,C} = \sqrt{QMR/n_{B,C}}$	$q_{B,C} = \overline{X}_{B,C} / SE_{B,C}$	Tabela (q_t)	Rejeito H_0 se q>q_t

Comparações múltiplas posteriores à ANOVA: O teste de Tukey

No quadro anterior, $n_{A,B}$ é o número de observações comuns aos grupos A e B. No caso deste número ser diferente, utiliza-se uma ponderação [ver Zar (1984)]. QMR é o valor do quadrado médio dos resíduos (variação dentro dos grupos) da ANOVA. Considerando 3 grupos (A, B, C) os resultados possíveis seriam os seguintes:

$\mu_A \neq \mu_B \neq \mu_C$	Todos os grupos diferentes
$\mu_A = \mu_B \neq \mu_C$	Grupos A e B sem ≠ entre si mas ≠ de C
$\mu_A \neq \mu_B = \mu_C$	Grupos B e C sem ≠ entre si mas ≠ de A
$\mu_A = \mu_C \neq \mu_B$	Grupos A e C sem ≠ entre si mas ≠ de B
$\mu_A = \mu_C = \mu_B$	Todos os grupos sem ≠ significativas

Em SPSS, os resultados do teste de Tukey são apresentados numa tabela relacionando cada par de grupos, tal como será analisado posteriormente.

ANÁLISE DE VARIÂNCIA
(ANOVA - *Analysis Of Variance)*

Pressupostos da ANOVA

Tal como para o teste *t* para duas amostras, temos que assumir que as observações de cada amostra se distribuem normalmente**.** e que possuem variâncias iguais (homoescedasticidade). Se ocorrerem desvios muito grandes a estes pressupostos, podemos transformar os dados, ou utilizar testes não-paramétricos, que, como já vimos anteriormente, não dependem deste tipo de pressupostos.

ANÁLISE DE VARIÂNCIA
(ANOVA - *Analysis Of Variance)*

Procedimentos em SPSS

1. Introdução dos dados:

Tal como para testes de comparação entre duas amostras independentes, uma variável contém os valores das amostras (grupos) e a outra a indicação 1…*k* correspondente aos *k* grupos.

2: Testes à Normalidade das observações em cada grupo

Analyze
Descriptive Statistics
Explore
Dependent list (variável com os valores)
Factor list (variável com a indicação dos grupos)
Plots **Normality plots with tests** *Continue* *OK*

ANÁLISE DE VARIÂNCIA
(ANOVA - *Analysis Of Variance)*

Procedimentos em SPSS

3. Anova a 1 factor e teste de Tukey de comparações posteriores (*Post-Hoc*):

Analyze
Compare Means
One-Way Anova
Dependent list (variável com os valores)
Factor list (variável com a indicação dos grupos)
Post-Hoc
Equal variances assumed **Tukey**
Significance level (valor de alfa) *Continue*
Options
Homogeneity of variance
Means plot
Continue *OK*

ANÁLISE DE VARIÂNCIA
(ANOVA - *Analysis Of Variance)*

Exemplo em SPSS

1. Dados:

Grupo I	Grupo II	Grupo III
1.53	3.15	8.18
1.61	3.96	5.64
3.75	3.59	7.36
2.89	1.89	8.82
	4.32	

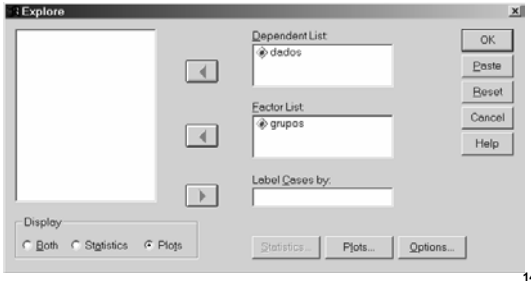
Em SPSS:

	dados	grupos
1	1.53	1
2	1.61	1
3	3.75	1
4	2.89	1
5	3.15	2
6	3.96	2
7	3.59	2
8	1.89	2
9	4.32	2
10	8.18	3
11	5.64	3
12	7.36	3
13	8.82	3

ANÁLISE DE VARIÂNCIA
(ANOVA - *Analysis Of Variance)*

Exemplo em SPSS

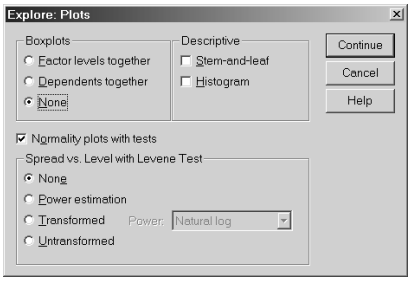
2. Testes à Normalidade : sub-menu *Explore*



ANÁLISE DE VARIÂNCIA
(ANOVA - *Analysis Of Variance)*

Exemplo em SPSS

2. Testes à Normalidade: sub-menu *Explore / Plots*



ANÁLISE DE VARIÂNCIA
(ANOVA - *Analysis Of Variance)*

Exemplo em SPSS

2. Testes à Normalidade: Resultados

Tests of Normality

	GRUPOS	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
DADOS	1	.282	4	.	.880	4	.338
	2	.203	5	.200*	.927	5	.575
	3	.209	4	.	.948	4	.707

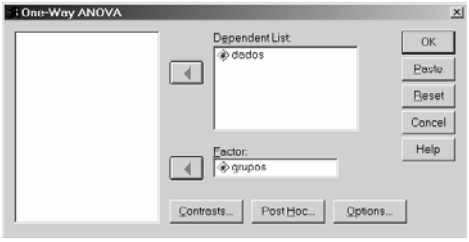
*. This is a lower bound of the true significance.
a. Lilliefors Significance Correction

Relembrando quais são as hipóteses…
Decidimos que HÁ NORMALIDADE em todos os grupos
(regra univ. de decisão estatística em SPSS…)

ANÁLISE DE VARIÂNCIA
(ANOVA - *Analysis Of Variance)*

Exemplo em SPSS

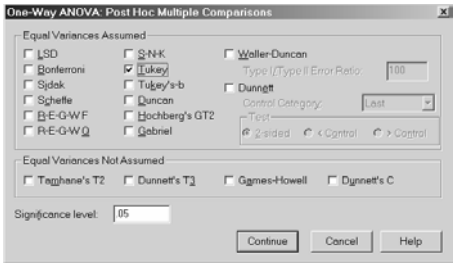
3. ANOVA a 1 factor e teste de Tukey: sub-menu *One-way ANOVA*



ANÁLISE DE VARIÂNCIA
(ANOVA - *Analysis Of Variance)*

Exemplo em SPSS

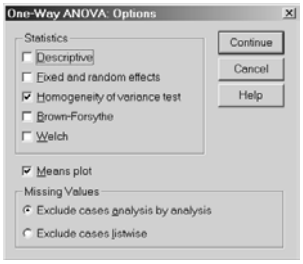
3. ANOVA a 1 factor e teste de Tukey:
sub-menu *One-way ANOVA* / *Post Hoc*



ANÁLISE DE VARIÂNCIA
(ANOVA - *Analysis Of Variance)*

Exemplo em SPSS

3. ANOVA a 1 factor e teste de Tukey:
sub-menus *One-way ANOVA / Options*



ANÁLISE DE VARIÂNCIA
(ANOVA - *Analysis Of Variance)*

Exemplo em SPSS

3. ANOVA a 1 factor e teste de Tukey - RESULTADOS:
Homogeneidade de variâncias e quadro da ANOVA

Test of Homogeneity of Variances

DADOS

Levene Statistic	df1	df2	Sig.
.343	2	10	.718

ANOVA

DADOS

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	58.890	2	29.445	23.264	.000
Within Groups	12.657	10	1.266		
Total	71.546	12			

ANÁLISE DE VARIÂNCIA
(ANOVA - *Analysis Of Variance)*

Exemplo em SPSS

3. ANOVA a 1 factor e teste de Tukey - RESULTADOS:
Quadro de Tukey

Multiple Comparisons

Dependent Variable: DADOS
Tukey HSD

(I) GRUPOS	(J) GRUPOS	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-.9370	.75468	.457	-3.0058	1.1318
	3	-5.0550*	.79551	.000	-7.2357	-2.8743
2	1	.9370	.75468	.457	-1.1318	3.0058
	3	-4.1180*	.75468	.001	-6.1868	-2.0492
3	1	5.0550*	.79551	.000	2.8743	7.2357
	2	4.1180*	.75468	.001	2.0492	6.1868

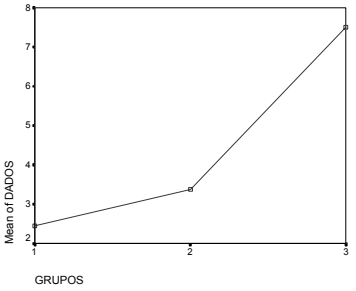
*. The mean difference is significant at the .05 level.

ANÁLISE DE VARIÂNCIA
(ANOVA - *Analysis Of Variance)*

Exemplo em SPSS

3. ANOVA a 1 factor e teste de Tukey – RESULTADOS:

Quadro de média dos grupos *(means plot)*



ANÁLISE DE VARIÂNCIA
(ANOVA - *Analysis Of Variance)*

Exemplo em SPSS

3. ANOVA a 1 factor e teste de Tukey

CONCLUSÕES:

1. **Há homogeneidade ($\alpha = 0.05$) entre as variâncias dos gupos (quadro *homogeneity of variances - Sig. = 0.718*);**
2. **Há diferenças significativas ($\alpha = 0.05$) entre grupos (quadro da ANOVA – *Sig. < 0.000*);**
3. **Os grupos 1 e 2 não diferem entre si ($\alpha = 0.05$) mas diferem do grupo 3 (quadro de Tukey);**
4. **O grupo 3 difere por excesso (valores maiores) comparativamente aos grupos 1 e 2 (*means plot*).**

Dados NÃO NORMAIS: Quais os passos a seguir?

Como já foi referido, um dos pressupostos na aplicação de testes paramétricos tais como a ANOVA, é que as observações se distribuam segundo uma lei Normal.

Quando a Análise de Variância é inviabilizada pela quebra deste pressuposto temos duas alternativas:

1. Proceder a transformações nos dados com o objectivo de atingir a Normalidade;

2. Proceder a uma Análise de Variância Não-Paramétrica: teste de Kruskal-Wallis.

Dados NÃO NORMAIS: Transformação

Exemplo em SPSS: dados

Consideremos o conjunto de dados ao lado, referente a uma experiência com 3 diferentes níveis de um dado factor.

Analisemos se os indivíduos se distribuem Normalmente em cada grupo (α = 0.05).

Grupo1	Grupo2	Grupo3
9.58	5.26	1.08
9.97	7.46	1.88
7.10	6.30	9.78
15.96	7.54	45.00
7.46	12.43	7.10
7.61	4.53	9.97
10.70	7.24	23.40
7.69	3.63	6.62
8.08	10.38	1.30
7.32	12.68	
6.23	9.78	
5.37		

Dados NÃO NORMAIS: Transformação

Exemplo em SPSS: Normalidade?

Tests of Normality

		Kolmogorov-Smirnov			Shapiro-Wilk		
	GRUPOS	Statistic	df	Sig.	Statistic	df	Sig.
DADOS	1	.239	12	.057	.828	12	.020
	2	.187	11	.200	.944	11	.563
	3	.329	9	.006	.754	9	.006

Podemos observar que não se verifica a Normalidade nos grupos 1 e 3, embora se verifique no grupo 2 (α = 0.05).

Dados NÃO NORMAIS: Transformação

Exemplo em SPSS: transformação – Ln (dados)

Transformemos então os dados: uma das transformações que mais se utiliza com vista a obter a Normalidade é a **logaritmização em base natural**.

Transform
- Compute
  - Target variable (*nome da nova var. logaritmizada*)
  - Numeric expression : LN(*nome da var. original*)
- OK

Façamos novamente o teste de Shapiro-Wilk à Normalidade, agora dos dados logaritmizados:

Dados NÃO NORMAIS:
Transformação

Exemplo em SPSS: Normalidade dos dados transformados?

Tests of Normality

		Kolmogorov-Smirnov			Shapiro-Wilk		
	GRUPOS	Statistic	df	Sig.	Statistic	df	Sig.
LN_DADOS	1	.196	12	.200	.930	12	.382
	2	.120	11	.200	.958	11	.748
	3	.189	9	.200	.937	9	.555

Pode verificar-se que a transformação (logaritmização) conferiu Normalidade à distribuição das observações, uma vez que não se rejeita H_0 para nenhum dos grupos, considerando α=0.05. Havendo Normalidade, podemos proceder à Análise de Variância que nos propusemos.

Dados NÃO NORMAIS:
Transformação

Advertências sobre a transformação de dados

1. A logaritmização não é o único processo de transformação que se pode experimentar visando atingir a Normalidade. Na realidade, a escolha do tipo de transformação a efectuar depende fortemente do tipo de dados que se possui e dos objectivos da análise (Zar, 1984).
2. Após a transformação dos dados, todos os parâmetros estimados têm de sofrer a transformação contrária à efectuada de forma a ser possível compará-los nas unidades originais.
3. Nem sempre a transformação resolve o problema. Em muitos casos a não-normalidade mantêm-se e a solução passa pela utilização de testes não-paramétricos.

Dados NÃO NORMAIS:
ANOVA Não-Paramétrica

O teste de Kruskal-Wallis (*k* > 2 amostras independentes)

Se um determinado conjunto de dados consiste em mais que duas amostras independentes (grupos) sobre um dado factor, o teste não-paramétrico de Kruskal-Wallis (K-W) pode ser aplicado para testar as diferenças entre as médias dos *k* grupos. O teste de K-W pode ser aplicado em todas as situações em que é possível utilizar a ANOVA paramétrica a 1 factor, e também em situações em que a ANOVA paramétrica não é aplicável tais como quando as *k* amostras são provenientes de populações não-Normais, ou com variâncias não homogéneas (heteroescedasticidade). O teste de K-W pode ser considerado uma extensão do teste de Mann-Whitney (M-W, já referido) para *k* > 2 amostras. Para *k* = 2 amostras, o teste de M-W é o indicado.

Dados NÃO NORMAIS: ANOVA Não-Paramétrica

O teste de Kruskal-Wallis (k > 2 amostras independentes)

O teste de K-W pode ser descrito nas seguintes etapas:

1. Atribuição das ordens das observações em cada grupo, relativamente ao total das n observações.

2. Cálculo da estatística H (fórmula abaixo) na qual R_i consiste na soma das ordens do grupo i, n = número total de observações e n_i = número de observações do grupo i.

3. Decisão do teste: Para pequenas amostras (~n < 15) em que $k \leq 5$ os valores críticos do teste são dados por uma tabela própria. Para amostras maiores e/ou k > 5, H pode ser considerada aproximada por um χ^2 com k - 1 graus de liberdade.

Em SPSS utiliza-se a regra generalizada de decisão*

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

Dados NÃO NORMAIS: ANOVA Não-Paramétrica

O teste de Kruskal-Wallis : procedimento em SPSS

Introdução dos dados

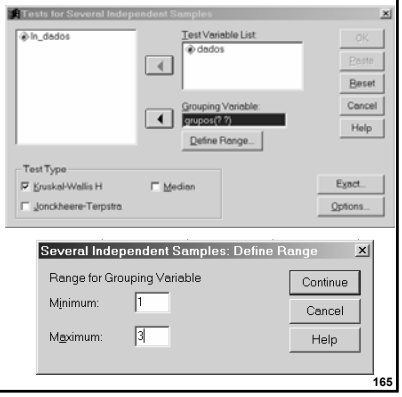
Tal como para a comparação paramétrica entre k amostras independentes (ANOVA) uma variável contém os valores das amostras (grupos) e a outra a indicação 1…k correspondente aos k grupos.

Estatística de K-W e respectiva significância

Analyze
Nonparametric Tests
k independent samples
test variable list (variável com os valores)
grouping variable (var. com a indicação dos grupos)
define range **1** ***k***
test type **Kruskal-Wallis H**
OK

Dados NÃO NORMAIS: ANOVA Não-Paramétrica

O teste de Kruskal-Wallis : procedimento em SPSS - MENUS



Dados NÃO NORMAIS: ANOVA Não-Paramétrica

O teste de Kruskal-Wallis :

Resultados, utilizando o conjunto de dados atrás referido no tema *transformação de dados*.

Ranks

	GRUPOS	N	Mean Rank
DADOS	1	12	17.96
	2	11	15.91
	3	9	15.28
	Total	32	

Neste caso não se rejeita H_o visto que a significância do teste (0.784) é superior a α (e.g., 0.05).

Test Statistics[a,b]

	DADOS
Chi-Square	.487
df	2
Asymp. Sig.	.784

a. Kruskal Wallis Test

b. Grouping Variable: GRUPOS

Assim, não há diferenças entre os grupos.

Comparação N-P com *k* > 2 amostras emparelhadas: o teste de Friedman

Introdução

Vimos anteriormente como, através do teste não-paramétrico de Wilcoxon, se podem comparar duas amostras emparelhadas. O teste de Friedman consiste numa generalização do teste de Wilcoxon para *k* > 2 amostras emparelhadas. Mais, tal como o teste de Kruskal-Wallis é uma alternativa não-paramétrica à ANOVA a 1 factor, o teste de Friedman consiste numa alternativa não-paramétrica para a ANOVA a 2 factores.

Deste modo, temos dois factores na experiência; um factor ***a*** que consiste na variável de teste, e um factor ***b*** que define o emparelhamento das amostras, ou seja, os indivíduos, grupos ou blocos. Consideremos o exemplo da tabela abaixo, no qual se consideram as classificações finais de 3 estudantes seleccionados aleatoriamente (***b*** = 3), a 3 disciplinas diferentes (***a*** = 3). Pretende-se saber se o desempenho da população de alunos da qual esta amostra foi extraída é igual em todas as disciplinas ($\mathbf{H_o}$), ou se é diferente, pelo menos, em uma delas ($\mathbf{H_1}$).

Comparação N-P com *k* > 2 amostras emparelhadas: o teste de Friedman

Introdução

	Disciplina		
Estudante	I. Aplicada 2	Inglês	Economia
A	18	8	12
B	12	10	15
C	15	14	13

Note-se que as classificações obtidas às disciplinas (variáveis - factor ***a***), estão emparelhadas pelo factor ***b*** (estudante), por exemplo, as classificações "18", "8" e "12" estão relacionadas (emparelhadas) entre si uma vez que foram obtidas pelo mesmo estudante.

Comparação N-P com *k* > 2 amostras emparelhadas: o teste de Friedman

Introdução

O teste de Friedman pode ser descrito nas seguintes etapas:

1. Atribuição das ordens (crescentes) das observações em cada bloco, ou seja, em cada linha da tabela, relativamente ao total das ***a*** observações nesse bloco (tabela de ordens das observações).

Tabela de ordens das observações

	Disciplina		
Estudante	I. Aplicada 2	Inglês	Economia
A	3	1	2
B	2	1	3
C	3	2	1
R_i	8	4	6

Comparação N-P com *k* > 2 amostras emparelhadas: o teste de Friedman

Introdução

2. Cálculo da estatística de Friedman χ_r^2 (fórmula abaixo) na qual R_i consiste na soma das ordens do nível *i* da variável de teste, ***a*** = número de níveis da variável de teste e *b* = número de indivíduos.

$$\chi_r^2 = \frac{12}{b\,a\,(a+1)}\sum_{i=1}^{a} R_i^2 - 3b(a+1)$$

3. Decisão do teste: Para pequenas amostras ($a \leq 5$ e $b \leq 8$) os valores críticos do teste são dados por uma tabela própria de Friedman. Para amostras maiores a estatística de Friedman pode ser considerada aproximada por um χ^2 com (***a*** - 1) graus de liberdade. Rejeita-se H_0 se o valor calculado da estatística de Friedman superar o tabelado (α).

EM SPSS, UTILIZA-SE A REGRA UNIVERSAL DE DECISÃO*

Comparação N-P com *k* > 2 amostras emparelhadas: o teste de Friedman

Procedimento em SPSS

Introdução dos dados — Tal como para a comparação não-paramétrica entre 2 amostras emparelhadas (teste de Wilcoxon) cada amostra é introduzida numa variável separada.

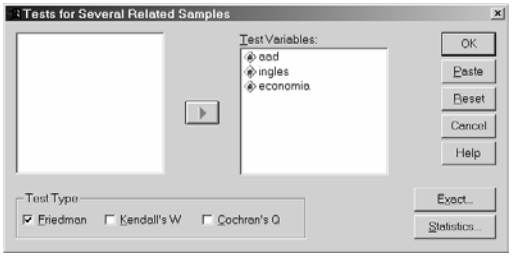
Estatística de Friedman e respectiva significância

Analyze
Nonparametric Tests
k related samples
test variables **vars. correspondentes às amostras**
test type **Friedman**

OK

Comparação N-P com *k* > 2 amostras emparelhadas: o teste de Friedman

Exemplo em SPSS:
Menu "Analyze \ Nonparametric tests \ k related samples"



Comparação N-P com *k* > 2 amostras emparelhadas: o teste de Friedman

Exemplo em SPSS:
RESULTADOS

Neste caso não se rejeita H_o visto que a significância do teste (*Asymp. Sig.* = 0.264) é superior a α (e.g., 0.05).

Assim, conclui-se (α = 0.05) que não há diferenças entre os grupos.

Nota: NESTE EXEMPLO, a pequena dimensão da amostra (*n* = 3) confere elevado grau de incerteza na rejeição de H_o. Por este motivo, apesar de visualmente se verificar que as classificações a AAD são relativamente superiores, estatisticamente não se pode fazer esta afirmação (= não se rejeita H_o).

Ranks

	Mean Rank
AAD	2.67
INGLES	1.33
ECONOMIA	2.00

Test Statistics[a]

N	3
Chi-Square	2.667
df	2
Asymp. Sig.	.264

a. Friedman Test

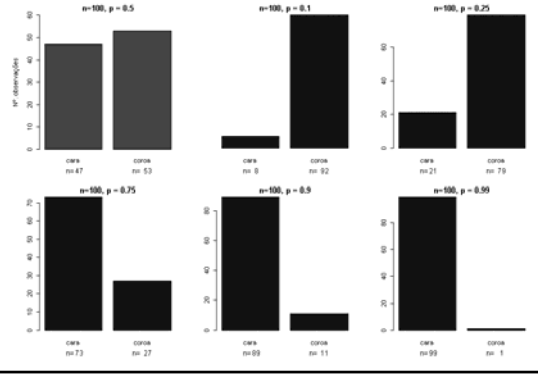
Testes baseados na Distribuição Binomial

0. Breve referência à distribuição Binomial

A distribuição Binomial descreve a probabilidade de se obter exactamente *k* sucessos, com probabilidade de sucesso *p*, e de insucesso (1-*p*), numa série de eventos independentes *n*.

Vejamos o exemplo seguinte, no qual seis moedas são atiradas ao ar, uma delas não viciada, ou seja, com probabilidade de sucesso *p* (cara) = probabilidade de insucesso q=(1-*p)* (coroa) = 0.5. As restantes moedas são viciadas de 5 modos diferentes…

Testes baseados na Distribuição Binomial



Testes baseados na Distribuição Binomial

I. O teste Binomial

Assim, se os dados consistem de *n* eventos independentes, cada resultado será ou "classe 1" ou "classe 2", mas nunca ambos. O número de observações nas classe 1 e 2 serão "O1" e "O2", respectivamente.

⇨ Neste ponto, cabe relembrar que os dados poderão ser agrupados ou não, sendo que, no primeiro caso haverá que efectuar uma ponderação prévia (***weight cases***) a todas as análises a efectuar.

Vejamos o exemplo seguinte:

Testes baseados na Distribuição Binomial

I. O teste Binomial: exemplo

Os dados ao lado correspondem a uma amostra de n=16 indivíduos adultos:
"Classe 1" = 0 = "s/ curso superior"
"Classe 2" = 1 = "c/ curso superior".

Questão:

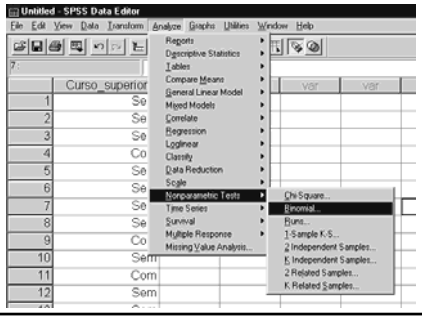
A um nível de sig. de 0.01 será 0.4 a proporção de indivíduos sem curso superior?

H_0: p = 0.4
H_1: p ≠ 0.4
(teste bilateral)

	Curso_superior
1	Sem
2	Sem
3	Sem
4	Com
5	Sem
6	Sem
7	Sem
8	Sem
9	Com
10	Sem
11	Com
12	Sem
13	Sem
14	Sem
15	Com
16	Sem

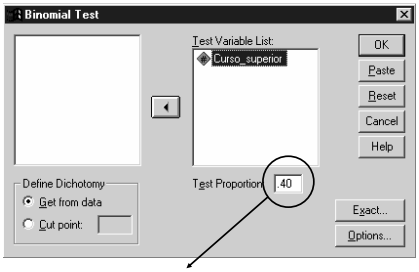
Testes baseados na Distribuição Binomial

I. O teste Binomial: exemplo – procedimento em SPSS: (1)



Testes baseados na Distribuição Binomial

I. O teste Binomial: exemplo – procedimento em SPSS: (2)



Proporção a testar ("classe 1")

Testes baseados na Distribuição Binomial

I. O teste Binomial: exemplo – procedimento em SPSS: <u>Resultado</u>

Binomial Test

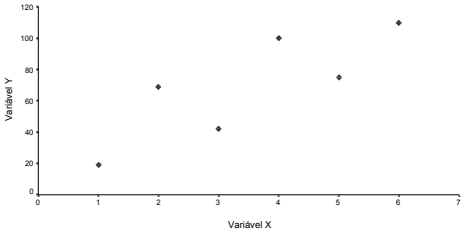
		Category	N	Observed Prop.	Test Prop.	Exact Sig. (1-tailed)
curso_sup	Group 1	sem curso	12	,75	,4	,005
	Group 2	com curso	4	,25		
	Total		16	1,0		

Uma vez que a sig. é menor que α (0.01), rejeita-se H_0, ou seja, a proporção de indivíduos sem curso superior é diferente de 0.4. Pela tabela verifica-se que esta diferença é por excesso (porquê?…)

Regressão Linear

Introdução

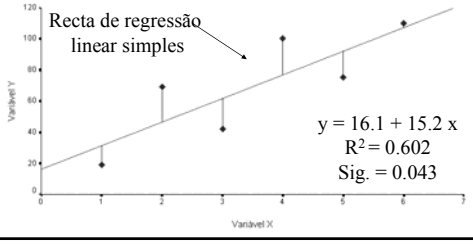
Como vimos anteriormente, um dos primeiros passos no estudo da relação entre duas variáveis métricas é elaborar uma nuvem de pontos, isto é, um gráfico que traduza a variação da variável dependente (Y) em função da variação da variável independente (X). Consideremos a nuvem de pontos representada na Figura abaixo que sugere existir uma tendência linear positiva. Se os coeficientes de correlação quantificam a relação, a **Análise de Regressão** é utilizada para modelar essa relação, isto é, construir uma equação que traduza a variação de y dado x.



Regressão Linear

Introdução

Ao modelo atrás referido chama-se recta de regressão, representada na Figura abaixo. O critério de ajuste da recta de regressão consiste na minimização dos quadrados dos desvios da recta relativamente aos pontos observados. Por outras palavras, pretende-se obter a recta que torne mínimos os quadrados das distâncias verticais dos pontos observados à recta ajustada - erros (ver Figura abaixo). Na Regressão é testada, entre outros parâmetros, a significância do modelo estimado, e a qualidade do ajuste, dada pelo coeficiente de determinação - R^2.

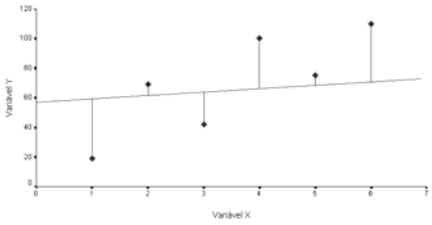


Regressão Linear

Introdução

Na Figura abaixo é representado um exemplo de um modelo cuja qualidade do ajuste é menor relativamente ao anterior, ou seja, os desvios verticais são maiores no modelo da Figura anterior. O modelo de Regressão linear simples (uma variável independente) possui dois parâmetros que definem a recta de regressão; a ordenada na origem (β_0 ou *a*) e o declive da recta (β_1 , *b* ou *m*). Utilizando a nomenclatura β_0 e β_1, a equação da recta de regressão de y dado x é dada por:

$$y = \beta_0 + \beta_1 * x$$



NOTA:
A Regressão é uma técnica paramétrica que exige normalidade dos erros, entre outros pressupostos (Snedecor & Cochran, 1989).

Regressão Linear

Introdução

No cálculo dos parâmetros da regressão são utilizadas as seguintes fórmulas:

$$\beta_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \qquad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

Para testar a significância da regressão procede-se a uma análise de variância (*):

Fonte de variação	Soma de quadrados (SQ)	gls	Quadrados médios	*F*
Regressão	$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	1	SQ regressão / gls regressão	SQ regressão / gls regressão ÷ SQ erro / gls erro
Erro	$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	*n* - 2	SQ erro / gls erro	
Total	$\sum_{i=1}^{n} (y_i - \bar{y})^2$	*n* - 1		

A estatística calculada *F* distribui-se, sob H_0, como um $F_{(\alpha,1,n-2)}$.

Regressão Linear

Introdução

O coeficiente de determinação (R^2) é dado pela expressão abaixo. Note-se que, no caso da regressão linear simples, R^2 é o quadrado do coeficiente de correlação linear de Pearson entre as duas variáveis.

$$R^2 = \frac{\text{SQ regressão}}{\text{SQ total}} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

R^2 varia entre 0 e 1 (ajuste perfeito, isto é, a recta é ajustada sobre todos os pontos observados.

Procedimentos em SPSS

Analyze
  Regression
    Linear
      Dependent **variável dependente**
      Independent(s) **variável independente**
      Statistics **estimates** **model fit** *Continue*
      Options **include constant in equation** *Continue*
OK

Regressão Linear

Exemplo em SPSS

Dados

Consideremos o exemplo já trabalhado: amostra de 8 crianças e adolescentes dos quais se registou a idade e a estatura (cm). Como objectivo, pretendemos criar um modelo que explique a variação da estatura (variável dependente - **Y**) em função da idade (variável independente – **X**).

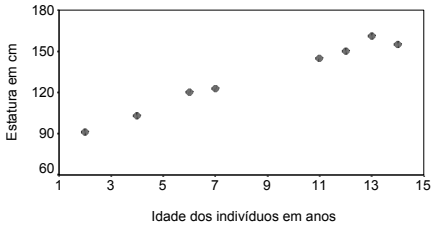
	criancas	idade	estatura
1	1	14	155.00
2	2	6	120.00
3	3	11	145.00
4	4	7	123.00
5	5	13	161.00
6	6	12	150.00
7	7	2	91.00
8	8	4	103.00

Regressão Linear

Exemplo em SPSS

Nuvem de pontos

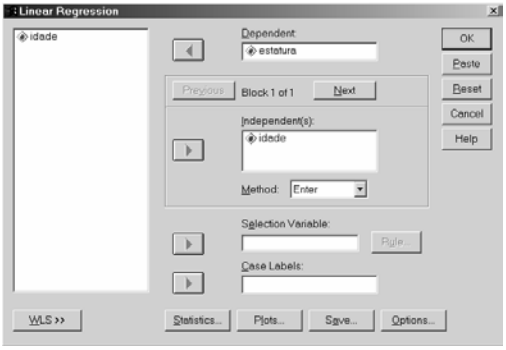
O primeiro passo é sempre analisar a nuvem de pontos. Se a mesma não sugerir uma relação linear, o modelo de regressão linear não é, obviamente, adequado.



Regressão Linear

Exemplo em SPSS

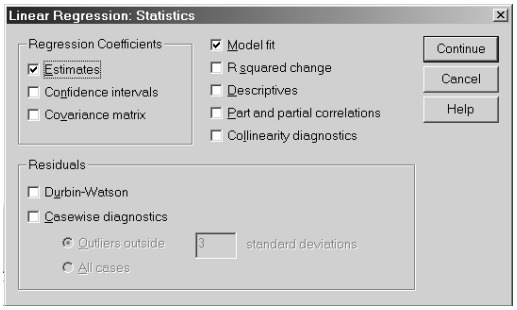
Menu: Analyze \ Regression \ Linear



Regressão Linear

Exemplo em SPSS

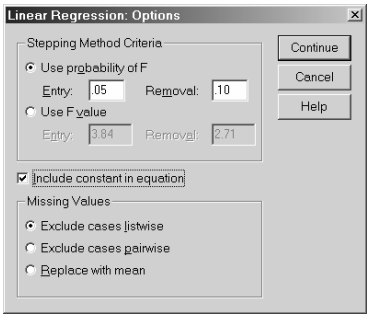
Menu: Analyze \ Regression \ Linear \ Statistics



Regressão Linear

Exemplo em SPSS

Menu: Analyze \ Regression \ Linear \ Options



Regressão Linear

Exemplo em SPSS

Resultados:

Resumo do modelo: Coeficiente de determinação

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.990[a]	.979	.976	3.976

a. Predictors: (Constant), IDADE

O valor do coeficiente de determinação *(R-Square* = 0.979*)* é muito elevado - perto do valor máximo possível: 1. Deste modo existe uma boa qualidade de ajuste do modelo aos dados observados.

Regressão Linear

Exemplo em SPSS

Resultados: Quadro de ANOVA

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4507.139	1	4507.139	285.077	.000[a]
	Residual	94.861	6	15.810		
	Total	4602.000	7			

a. Predictors: (Constant), IDADE

b. Dependent Variable: ESTATURA

A regressão é significativa para um nível de significância de, p. ex., 0.05, pois *Sig.* < 0.000. Por outras palavras (relembrando…), a probabilidade de se cometer um erro tipo I (α), ou seja, rejeitar H_o quando é verdadeira, é muito pequena.

Regressão Linear

Exemplo em SPSS

Resultados: Coeficientes

Já sabendo que a regressão é significativa, quais são os parâmetros do modelo que permitem definir uma equação que traduza a variação de Y dado X?

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	82.040	3.223		25.458	.000
	IDADE	5.676	.336	.990	16.884	.000

a. Dependent Variable: ESTATURA

B_0 = 82.04 ; B_1 = 5.68. Então a equação toma a forma:

Estatura = 82.0 + 5.7 * Idade

Regressão Linear

Exemplo em SPSS

Resultados: Comentários gerais.

É importante relembrar que a utilização do modelo apenas é correcta no intervalo de valores de X com o qual foi elaborado: neste caso, 2 a 14 anos.

Com base nesta equação estimemos a estatura de uma criança com 8 anos de idade:

Estatura com 8 anos = 82.0 + 5.7 * 8

Estatura com 8 anos = 127.6 cm

Em SPSS a linguagem **syntax** permite efectuar estes e outros cálculos. Façamos de seguida uma breve introdução a esta linguagem.

A linguagem syntax

Todas as acções efectuadas por menus em SPSS podem ser opcionalmente* traduzidas em linguagem *syntax* no ficheiro *output*. Por exemplo, a syntax dum histograma de uma variável "var1" é:

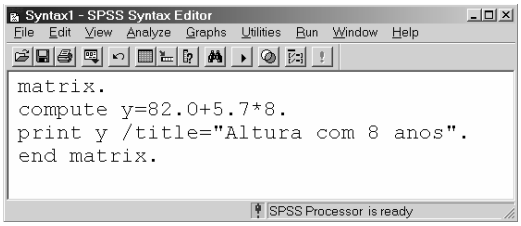
```
GRAPH
  /HISTOGRAM=var1 .
```

Esta e outras "ordens" podem ser dadas / reproduzidas numa janela de syntax. Também a grande maioria das operações matemáticas e de cálculo podem ser efectuadas em syntax. Por exemplo, para se efectuar o cálculo anterior que nos dá a resposta do modelo a uma idade de 8 anos:

A linguagem syntax

Janela de syntax:

exemplo de aplicação do modelo linear estimado anteriormente



A linguagem syntax

Janela de syntax:

O que são os comandos atrás?

`matrix.`
PROCEDIMENTOS
`end matrix.`
	As expressões `matrix` e `end matrix` são, como regra, colocadas ao início e ao fim do procedimento geral a efectuar.

`compute y=`FÓRMULA .	"`compute y=`" significa "calcula o valor de y igual a..."
`print y /title"`TÍTULO".	"`print y`" é o comando usado para visualizar / imprimir o valor calculado no ficheiro de *output* neste caso com a opção de título "`/title"`TÍTULO"

Uma regra fundamental é colocar um ponto "." no final de cada linha ("ordem") definida.

A linguagem syntax

Janela de syntax:

Depois de escrito o pequeno programa que pede o cálculo de y dado x=8, faz-se *correr* o mesmo pelo menu **RUN \ ALL**. O resultado do nosso exemplo é:

```
matrix.
compute y=82.0+5.7*8.
print y /title="Altura com 8 anos".
end matrix.

Run MATRIX procedure:

Altura com 8 anos
   127.6000000

------ END MATRIX -----
```

Nota final: A potencialidade e utilidade da linguagem syntax estende-se muito para além deste pequeno exemplo: o SPSS possui manuais de ajuda através do menu HELP \ SYNTAX GUIDE \ BASE.

Introdução à Amostragem

Enquadramento...

Como nota prévia é de referir que a amostragem consiste numa grande *sub-secção* da estatística, bastante bem documentada na literatura. Este capítulo apresenta apenas uma (muito) breve introdução a toda a problemática da amostragem que, por si só, mereceria um curso completo de um ou mais semestres.

Para futuras abordagens, ficam as referências:

Cochran, W., 1977. *Sampling Techniques*, 3rd. ed., Wiley. 428 pp.
(uma referência importante da amostragem)

Malhotra, K. and Birks, D., 2000. *Marketing Research, An Applied Approach*. 3rd. ed., Finantial Times – Prentice-Hall. 731 pp.
(uma referência mais *ligeira*, aplicada ao marketing)

Introdução à Amostragem

Os passos indispensáveis na amostragem

Os conceitos "População", "amostra" e "inferência" já foram previamente abordados. As etapas fundamentais num processo de amostragem consistem na definição dos seguintes pontos:

1. Objectivos do estudo;
2. População a amostrar;
3. Dados a recolher;
4. Grau de precisão;
5. Métodos de medida;
6. Unidades amostrais;
7. Selecção da amostra.

Introdução à Amostragem

O conceito de aleatoriedade.

A amostragem probabilistica obedece a determinadas propriedades matemáticas* que permitem poder-se desenvolver *a posteriori* toda uma teoria de amostragem.

O conceito de aleatório tem a ver com conhecer-se *a priori* a probabilidade de ser seleccionada uma certa unidade amostral ou amostra. No caso mais elementar, quando esta probabilidade é igual para todas as unidades amostrais / amostras, estamos em presença da **amostragem aleatória simples** (voltaremos a este assunto posteriormente).

Introdução à Amostragem

? O que é uma boa amostra ?

- *Representa o melhor possível as características da População.*
- Formalmente, **maximiza a PRECISÃO*** e **minimiza o ENVIESAMENTO***.

Introdução à Amostragem

ENVIESAMENTO *(bias)*

1. Porque é que existe enviesamento nos exs. abaixo?
2. Sugestões para melhorar o problema...

- Referendo à população portuguesa sobre autoestrada Braga – Vigo ou conclusão da Via do Infante (Algarve); população inquirida: Distrito do Porto.
- Um governo pretende saber o número de desempregados do país. Para isso, extrai de uma listagem geral uma amostra de 860 indivíduos com idades superiores a 12 anos.
- Percentagem da população portuguesa que é católica e praticante: amostra de 240 indivíduos da Beira Interior.

Introdução à Amostragem

PRECISÃO / RIGOR*

1. Porque é que existe imprecisão nos exs. abaixo?
2. Sugestões para melhorar o problema...

- A análise de uma amostra dos estudantes do ensino superior algarvio conclui que o tempo médio de deslocação à instituição de ensino é de 55 minutos, e que os mesmos demoram entre 2 minutos e 1h 52' na deslocação.

Introdução à Amostragem

Métodos de selecção da amostra

Os diferentes métodos de selecção da(s) amostra(s) estão associados ao modo como se trata as características da população alvo. Por exemplo, se a população é finita ou não*1, ou se existe reposição*2 ou não (particularmente importante no caso de populações finitas).

Tendo em conta este tipo de condicionantes, referem-se abaixo alguns tipos de amostragem.

1. Aleatória simples (a.a.s.)
2. Aleatória estratificada
3. Por conglomerados
4. Sistemática

Exemplos…

Introdução à Amostragem

A a.a. simples

Consiste em seleccionar *n* indivíduos (amostra) de entre *N* (população) de tal modo que todas as combinações possíveis de *n* possuam a mesma probabilidade de serem seleccionadas. Na prática o processo aleatório pode ser simulado de diferentes maneiras, a mais simples das quais através de uma tabela de números aleatórios*. Existem também algoritmos de geração de números aleatórios implementados em muitos programas de computador**.

Ex: a extracção do totoloto consiste num processo de a.a.s. sem reposição onde uma amostra $n = 6$ é seleccionada de uma população finita $N = 49$ e no qual o processo aleatório é simulado mecanicamente.

Façamos as contas à probabilidade de ficarmos milionários:
totoloto vs. lotaria (revisitado)…

Introdução à Amostragem

As características a observar

Num processo de amostragem decide-se qual ou quais as características a medir / observar nos indivíduos seleccionados: o peso, sexo, preferência por um dado produto, o número da bola no caso do totoloto, etc. No entanto, de modo geral, o interesse centra-se nas seguintes quatro características da população:

1.Média (p.ex., rendimento médio anual das pme's do Algarve)
2.Total (p.ex., número total de peixes num manancial)
3.Razão (p.ex., entre rendimentos líquidos e brutos das famílias de uma região)
4.Proporção (p.ex., proporção de fumadores masculinos de uma Universidade)

Introdução à Amostragem

Amostragem aleatória simples: os estimadores

No caso da a.a.s., os estimadores amostrais utilizados na inferência destas quatro características da população são:

Média populacional	$\bar{Y}$	$\hat{\bar{Y}} = \bar{y} = \textit{média amostral}$
Total populacional	Y	$\hat{Y} = N\,\bar{y} = N\dfrac{\sum_{i=1}^{n} y_i}{n}$
Razão populacional	R	$\hat{R} = \dfrac{\bar{y}}{\bar{x}} = \dfrac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$
Proporção populacional	P	$\hat{P} = \dfrac{a}{n}$

Introdução à Amostragem

O tamanho da amostra: definições

Os métodos de determinação do tamanho da amostra variam de acordo com as características do processo de amostragem. Analisemos dois casos simples: adicionemos as seguintes definições às atrás referidas:

- *t* — desvio da curva Normal (abcissa) correspondente à confiança probabililística: p. ex., para um nível de confiança de 95%, *t* = 1.96; para 99%, *t* = 2.58.
- *p* — estimativa da proporção de indivíduos numa classe *C* (por ex., fumadores).
- *q* — estimativa da proporção de indivíduos na classe complementar *C'* (p.ex., não fumadores): *q* = 1 - *p*.

Introdução à Amostragem

O tamanho da amostra: definições - continuação

- *d* — erro estabelecido *a priori*.
- *s* — estimativa do desvio padrão da população.
- n_0 — primeira aproximação ao tamanho amostral (fórmulas abaixo).
- *n* — tamanho amostral após análise da fracção amostral: se a mesma fôr superior a 10%, corrigir n_0 através das fórmulas abaixo (correcção para população finita - *cpf*); se *f* < 10%, considera-se $n = n_0$.

Introdução à Amostragem

O tamanho da amostra:

Fórmulas

Proporções	Dados contínuos
$n_0 = \frac{t^2\, p\, q}{d^2}$	$n_0 = \frac{t^2\, s^2}{d^2}$

Se $f\,(n_0 / N) > 0.1$ (10%), corrigir com:

$$n = \frac{n_0}{1 + \frac{(n_0 - 1)}{N}}$$

Caso contrário tomar $n = n_0$

Introdução à Amostragem

Estimação do tamanho da amostra - 2 exemplos:

1. Pretende-se estimar n associado à estimativa $p = 0.35$, sendo P = proporção de indivíduos fumadores do INUAF, adoptando um erro de 0.05, um nível de confiança de 95% ($t = 1.96$), e $N = 2000$.

$$n_0 = \frac{1.96^2\, 0.35(1 - 0.35)}{0.05^2} \cong 350 \quad \text{, sendo} \quad f = \frac{350}{2000} \cong 0.17$$

Uma vez que a fracção amostral f supera 0.1 procede-se à correcção para pop. finita:

$$n = \frac{350}{1 + \frac{(350 - 1)}{2000}} = 298$$

Introdução à Amostragem

Estimação do tamanho da amostra - 2 exemplos:

2. Qual o tamanho da amostra quando se pretende estimar o rendimento médio mensal por agregado familiar no concelho de Loulé? Considere 700 € como estimativa do desvio padrão da população, 100 € como o erro admissível, 99% como nível de confiança ($t = 2.58$) e 20 000 habitantes como a estimativa da população do concelho.

$$n_0 = \frac{2.58^2\, 700^2}{100^2} = 326 \quad \text{, sendo} \quad f = \frac{326}{20000} \cong 0.02$$

Uma vez que a fracção amostral f é menor que 0.1, $n = n_0 = 326$

Introdução à Amostragem

Tamanho da amostra:
Notas finais:

1. Em muitas situações pretende-se a estimação de mais que um parâmetro. Por exemplo, um questionário com múltiplas perguntas. Neste caso há que estimar o tamanho amostral separadamente para cada um dos parâmetros mais importantes e **seleccionar o maior** destes valores.

2. Existem muitas outras abordagens na estimação de *n*, não só dependentes do tipo de amostragem efectuado mas também resultantes da incorporação de outros parâmetros na estimação como por exemplo o custo da amostragem (ver, para este efeito, Cochran, 1977).