Introdução à análise estatística com SPSS

Guião nº3: Operações básicas com o SPSS

Abrir o ficheiro de dados com o SPSS

Abrir o SPSS e depois a opção **open an existing data source** e procurar pelo ficheiro *Ex3_Questionnaire_data.sav*. Alternativa será clicar duas vezes no ficheiro do SPSS com os dados.

Recodificar dados qualitativos

Uma situação que acontece regularmente é ter de recodificar valores ou categorias de uma variável, como por exemplo transformar uma variável qualitativa numa outra variável qualitativa nominal. A variável *health* têm 4 categorias e o objectivo é criar uma nova variável com apenas 2 categorias: 1- *Poor* e 2- Not poor (acceptable, good, very good).

Usando o comando **Transform/Recode/Into Different Variables,** selecionar a variável *Health* e clique na seta de transporte para transferir a variável para a caixa designada por **input variable -> Output variable**. Defina uma nova variável na caixa **Output Variable** (por exemplo, *NewHealth*) e no **label** indicar a etiqueta da nova variável: *New health* e depois clicar em **change**..

Clique na tecla **Old and New Values** e preencha com os correspondentes valores nas caixas **old value** e **new value**, Como a categoria pobre irá se manter, então no quadro **old value**, na opção **value** indicar o valor enquanto no quadro **New Value**, na opção **value** indicar novamente o valor 1. Clicar na opção **add** no quadro **Old→New**. Repetir novamente a operação para os outros códigos: **Old value:** 2**, New value:** 2 e clicar na opção **add**; **Old value:** 3**, New value:** 2 e opção **add**; **Old value:** 4**, New value:** 2 e a opção **add**. No final, o quadro **Old→New** deve apresentar: 1**→1;** 2**→2;** 3**→2;** 4**→2.** Clicar em **Continue** e depois em **Ok.**

A nova variável deve aparecer na última coluna na janela do **data view** dos dados com o nome *NewStatus*. Falta apenas indicar que o valor 1 representa pré-graduado e o valor 2 representa pós graduado na janela do **variable view** na opção **values** para a variável *NewStatus*.

Recodificar dados quantitativos em dados qualitativos

Uma outra situação comum é recodificar uma variável quantitativa numa variável qualitativa. Iremos recodificar a variável altura como alto, médio ou baixo através do comando **Transform/Visual Binning.** Selecionar a variável *Age* e transferir esta variável para o quadro **Variables to Bin** e clique em **continue**. No novo quadro, selecionar novamente *Age* e um histograma irá aparecer. Na caixa **Binned Variable,** escrever um novo nome para a variável a criar como por exemplo *AgeCat***.** No campo **Value** indicar os valores *30.0, 40.0, 50.0, 80.0* e no campo **Label** escreva as seguintes etiquetas: *[20-30];]30;40];]40;50]; >50*. A opção HIGH[1] não é possível apagar e irá aparecer sempre na penúltima célula da coluna **Value**. Ao aceitar a opção **included (<=)** significa que se está a criar os seguintes intervalos: ≤30.0, ≤40.0 (mas excluindo todas as idades inferiores ou iguais a 30 e assim sucessivamente), ≤50.0 e ≤80.0. **Nota:** Em algumas versões do SPSS será necessário substituir o ponto pela vírgula: 30,0). Ao clicar no **Ok**, irá aparecer uma mensagem **"Binning specifications will create 1 variable"**. Ao clica novamente em **Ok**, uma nova variável com o nome *AgeCat* irá ser construída.

Manipulação de dados: Transformação de variáveis

Uma operação frequente é a manipulação de um conjunto de variáveis com o objetivo de se calcular uma nova variável com o resultado dessa transformação. No ficheiro de dados, alguns dos participantes indicaram o seu peso em *Stones* e *Pounds*, outros em quilogramas (*Kilos*). O objectivo é reescrever a variável *Kilos*, com a

[1] Em algumas versões do SPSS é necessário apagar esta categoria após a variável ter sido criada, através da janela das variáveis (**variable view**) e depois na coluna dos **Values** na linha correspondente á nova variável criada (por exemplo: *AgeCat*). Selecionar a categoria a eliminar e depois clicar em **remove**.

informação das unidades britânicas de peso em quilogramas, mas mantendo a informação que a variável *Kilos* já possuí.

Através do comando **transform/compute variable**, aparece um quadro que permite fazer várias manipulações. Na caixa **target variable**, escreva a variável *Kilos*. Na caixa **numeric expression**, vai-se introduzir a expressão para a conversão: *(Stones*14+Pounds)*0.453*. Para não haver enganos na escrita das variáveis Stones e Pounds, o utilizador deve utilizar a lista de variáveis que estão no painel esquerdo, fazendo a sua seleção e depois clicando na tecla do cursor (que está situado entre a lista das variáveis e a caixa do **Numeric Expression**), transferir o nome da variável para a caixa do **Numeric Expression.** Mas não clique em ok ainda! Existe um pequeno problema, como se está a reescrever a variável *Kilos*, se uma pessoa indicou o seu peso em quilogramas e não o indicou nas unidades britânicas, o resultado da transformação irá ser um dado omisso na variável *Kilos*, indicado no SPSS indicado como um ponto "." e o valor do peso em quilogramas que lá estaria anteriormente irá ser perdido. Para se evitar perda de informação, através do clique na tecla **if** (colocado na parte inferior do menu **compute variable**) e depois selecionar a opção **include if cases satisfies condition**, cria-se aquilo que se designa por um filtro[2]. A condição deste filtro será *Stones > 0*, que significa que o programa só irá calcular a conversão proposta se a variável *Stones* for um não-zero, ou seja, excluídos todos valores omissos ou iguais a 0 desta variável. Clicar em **Continue** e depois em **Ok**. Irá aparecer uma mensagem com a seguinte pergunta "**change existing variable**?". Ao clicar no **OK**, a variável *Kilos* irá ser preenchida com os valores em falta.

Questão 1: Repita o procedimento para a altura. Faça a conversão de pés (*feet*) e polegadas (*inches*) para metros (*Meters*), sabendo que existe 12 polegadas para cada pé e que cada polegada é 0.0254 metros. **Dica**: Modifique a expressão para a conversão anterior e altere a condição do filtro para: *Feet > 0*.

[2] Um filtro corresponde á situação que se vai selecionar apenas parte da informação e não toda a informação disponível (ver seção Manipulação de dados: Seleção de dados através de um filtro).

Manipulação de dados: Funções pré-definidas no SPSS

Outras transformações utilizam funções pré-definidas existentes no menu **compute variable** através das opções presentes no quadro **Function group** (categorias) e **functions and special variables** (lista de funções) como por exemplo gerar ou trabalhar com variáveis aleatórias de uma determinada distribuição de probabilidade, transformações logarítmicas ou exponenciais, valores absolutos de uma variável, cálculo médias/medianas/máximo entre duas ou mais variáveis ou ainda operações com datas.

Como exercício vai-se gerar duas varáveis aleatórias, como por exemplo a pontuação num teste psicotécnico na primeira vez (designada por pontuação_1) e pontuação num teste psicotécnico na segunda vez (designada por pontuação_2) com distribuição normal de média 90 e desvio-padrão de 15 unidades para a primeira variável e 95 de média e 10 de desvio-padrão para a segunda variável. Vai-se começar por construir a primeira variável. Através do comando **transform/compute variable**, na caixa **target variable**, escreva a variável *Pontuação_1*. Na caixa **numeric expression**, vai-se introduzir a expressão para gerar a distribuição Normal através da procura da função *rv.normal* no quadro: **Function group** (***Random numbers***) e **functions and special variables** (*Rv.Normal*) e clicar na tecla com seta para cima, de forma a função apareça na caixa **numeric expression**. Esta função tem 2 parâmetros, o primeiro representa a média e o segundo o desvio-padrão. Após inserção dos valores a expressão deverá ter esta forma: *RV.NORMAL(90,15)*. Clicar no **Ok** e repetir o mesmo para a segunda variável alterando o nome da variável para *Pontuação_2* e os valores correspondentes da média (95) e do desvio-padrão (10). Verifique que na janela do **data view**, existem duas novas variáveis.

No próximo exercício consiste no cálculo do máximo destas duas variáveis como pontuação final. Através do comando **transform/compute variable**, na caixa **target variable**, escreva a variável *Pontuação_F*. Na caixa **numeric expression**, vai-se introduzir a expressão para cálculo do máximo através da procura da função *max* no quadro: **Function group** (*Search*) e **functions and special variables** (*Max*) e clicar na tecla com seta para cima, de forma que a função apareça na caixa **numeric expression**. Esta função pode ter n parâmetros, consoante o número de variáveis que se pretende utilizar (neste caso serão apenas duas variáveis). No quadro onde estão listadas todas as

variáveis do ficheiro, selecionar a variável *Pontuação_1* e clicar na tecla com a seta do lado direito de forma a colocar esta variável no quadro **numeric expression** dentro dos parêntesis da função *MAX()* e repetir o mesmo para a variável *Pontuação_2* de forma a ficar desta forma: MAX(*Pontuação_1, Pontuação_2*). Clicar no **Ok**. Verifique que na janela do **data view** o resultado apresentado para esta variável.

O exercício seguinte consiste em calcular o módulo para a expressão *Pontuação_1-Pontuação_2* (ou seja, todos valores negativos da diferença serão transformados em positivos). Através do comando **transform/compute variable**, na caixa **target variable**, escreva a variável *Pont_Dif_Mod*. Na caixa **numeric expression**, vai-se introduzir a expressão para cálculo do módulo através da procura da função *abs* no quadro: **Function group** (*Arithmetic*) e **functions and special variables** (*Abs*) e clicar na tecla com seta para cima, de forma a função apareça na caixa **numeric expression**. Esta função só pode ter 1 parâmetro, ou seja, só permite uma variável ou o correspondente a uma expressão. No quadro das variáveis do ficheiro, selecionar a variável *Pontuação_1* e clicar na tecla com a seta do lado direito de forma a colocar esta variável no quadro **numeric expression** dentro dos parêntesis da função *abs()*, repetir o mesmo para a variável *Pontuação_2*, de forma que a expressão final fique: *ABS(Pontuação_1-Pontuação_2)*. Clicar no **Ok**. Verifique que na variável *Pont_Dif_Mod* não tem valores negativos.

Na janela dos dados (**data editor**) estão apresentados todos os resultados destas transformações (procurar nas últimas colunas) e na janela das variáveis (**variable view**) está apresentada a descrição destas variáveis (procurar nas últimas linhas).

Manipulação de dados: Transformações para se obter a simetria

Uma outra situação muito usual é utilizar transformações de dados para se obter simetria e desta forma uma melhor aproximação a uma distribuição de probabilidade teórica (maior parte será a distribuição Normal). A desvantagem é que a unidade física associada á variável original também é transformada. Por exemplo, se a unidade física for kgs e transformação for logarítmica então a unidade do peso também é transformada em log(kg).

Para identificar qual a transformação a utilizar é necessário conhecer o histograma da variável quantitativa. Através do comando **Analyse/Descriptive Statistics/explore**,

selecionar a variável *Age*. De seguida, clicar em **plots** e ativar a opção **histogram** e desactivar a opção **Stem-and-leaf**. Clicar no **Continue** e depois no **ok**.

Questão 2: Analise o histograma e indique se o considera simétrico? **Dica:** Na análise de dados reais, é muito pouco provável encontrar simetrias perfeitas.

Faça as seguintes transformações através do menu: **transform/compute variable** (os nomes das variáveis estão definidas na parte esquerda da equação e expressão matemática correspondente encontra-se na parte direita da equação):

- Age2=Lg10(Age+1)
- Age3=1/(Age+1)

Questão nº3: O que observa dos histogramas? Indique os que considera simétricos.

Uma alternativa aos histogramas é inspeccionar visualmente os *QQ plots*. Estas medidas gráficas permitem analisar o comportamento da distribuição amostral (representada por círculos) e a distribuição teórica (representada pela linha) e se o comportamento da distribuição amostral for próximo da distribuição teórica, então pode-se indicar que amostra provém de uma determinada distribuição teórica. Através do comando **Analyse/Descriptive Statistics/QQ plot** selecionar as variáveis *age*, *age2* e *age3* para o quadro **variables**. Verifique se no quadro **test distribution** a opção *Normal* se encontra ativada. Depois clicar no **Ok**.

Questão nº4: Qual dos resultados mais se aproxima da distribuição Normal teórica? **Dica:** Observe as escalas utilizadas no gráfico designado por **Detrended Normal QQ plot**, que indica a diferença entre o valor amostral e correspondente valor teórico e portanto quanto menor for o erro melhor será a aproximação. Qual dos três gráficos apresenta um erro menor?

Manipulação de dados: Operação com datas

Operações com datas são também muito frequentes numa análise de dados. As operações relevantes que se irá apresentar são o cálculo da diferença entre as duas datas e entre uma data (por exemplo, a data de nascimento) e a data atual.

Através do comando **transform/ data and time wizard**, vai se explorar a opção 4 (**calculate with dates and times**). Para o cálculo da diferença entre as duas datas, vai-se escolher a opção 4 e depois a opção **next**. No menu seguinte, irá se escolher a opção "**calculate the number of time units between two dates**" e clicar em **next**. No menu seguinte para o campo **Date1**, escolhe-se a variável "*questionnaire ending*" e para o campo **minus Date 2**: "*questionnaire beggining*". Para calcular a diferença têm-se que especificar a unidade (**Unit**) que se pretende e pode ser em Anos (**Years**), Meses (**Months**) indo até á unidade mais pequena que são segundos (**seconds**). O resultado ainda pode ser apresentado truncado (**truncate to integer**) ou arredondado (**round to integer**) para o inteiro mais próximo. Escolher **hours** na opção **Unit** e **round to integer**. Clicar em **next** e na caixa de texto **result variable**, escrever o nome da variável resultante (*duração_horas*) e clicar em **Finish**. Uma nova variável será criada na última coluna da janela **data view** do ficheiro de dados.

O próximo exercício é adicionar uma determinada quantidade a uma data (por exemplo, adicionar o tempo de um tratamento variável a uma data inicial). Para o cálculo da diferença entre as duas datas, vai-se escolher novamente a opção 4 e depois a opção **next**. No menu seguinte, irá se escolher a opção "**add or subtract a duration from a data**" e clicar em **next**. No menu seguinte para o campo **Date**, escolhe-se a variável "*questionnaire beggining*" e para o campo **duration variable**: "*duração_horas*". Também é possível em vez de uma duração variável introduzir uma duração constante (por exemplo, adicionar 5 horas). Na opção Units, deve-se escolher a opção hours e no quadro operation, verificar que a operação selecionada é a adição (addition) e clicar em next. Na caixa de texto **result variable**, escrever o nome da variável resultante (*data_final*) e clicar em **Finish**.

Questão nº5: Verifique os resultados obtidos com a variável "*questionnaire ending*".

O último exercício é calcular o tempo entre a data do inicio do questionário e o tempo atual. Vai-se escolher novamente a opção 4 e depois a opção **next**. No menu seguinte, irá se escolher a opção "**calculate the number of time units between two dates**" e clicar em **next**. No menu seguinte para o campo **Date1**, escolhe-se a variável "*current date and time*" e para o campo **minus Date 2**: "*questionnaire beggining*". Para calcular a diferença têm-se que especificar a unidade (**Unit: Months**) e opção **round to integer**. Clicar em **next** e na caixa de texto **result variable**, escrever o nome da variável resultante (*duração_meses*) e clicar em **Finish**. Uma nova variável será criada na última coluna da janela **data view** do ficheiro de dados.

Manipulação de dados: Seleção de dados através de um filtro

Uma outra situação muito recorrente é ter de seleccionar casos específicos de um conjunto grande de dados. Vamos supor que queremos analisar apenas o sexo masculino. Através do comando **Data/selected cases**, impor a condição através dos comandos **if condition is satisfied** e depois clicar na tecla **if**. A condição será *sex=1*. Deve-se seleccionar a variável *sex* do quadro da esquerda e depois transferi-la para o quadro da direita através da tecla com a seta para a direita e depois escrever o resto da expressão. O valor 1 representa o valor numérico atribuído ao sexo masculino. Depois clicar em **continue** e no **ok**. O próximo passo é verificar se o filtro criado responde ao que se pretende. Na janela **data view**, na coluna numérica cinzenta se tiver um traço diagonal, significa que essa linha não irá fazer parte dos resultados futuros enquanto este filtro estiver ativo. Neste caso, todos os elementos do sexo feminino serão "removidos" da análise. Para desativar o filtro é necessário voltar ao menu através do comando **Data/selected cases** e escolher a opção **all cases.**

Questão nº6: Com o filtro ativo, preencha a tabela seguinte para a variável peso dos fumadores e não fumadores do sexo masculino. **Dica**: Para cruzar uma variável quantitativa (**Kilos**) por uma qualitativa (**Smoker**) deve utilizar o comando **Analyse/descriptive statistics/explore** e clicar na tecla **statistics** e depois ativar a opção **Percentiles** (ver guião nº2).

Tabela 1- Resultados da estatística descritiva para variável peso (em kgs) de fumadores (n = _____) e não fumadores (n = 166) do sexo masculino

Peso (kg)	Média	D.P.	1º Quartil	Mediana	3º Quartil
Fumadores				70.9	76.1
Não fumadores	70.6				

Quando existem *outliers* e ou extremos é necessário analisar a influência destes valores, dado que estes potencialmente afetam os valores obtidos para as médias e o desvios-padrão. Se os resultados sem os *outliers* e ou extremos forem muito diferentes, então deve remover estes valores da análise, no caso de querer apresentar resultados para as médias e desvios-padrão.

Questão nº7: Com as caixas de bigodes obtidas para a tabela anterior, verificar se existem *outliers* e ou extremos e no caso de afirmativo, indicar os valores que correspondem a estes *outliers*/extremos. **Dica:** Os números apresentados ao lado dos *outliers* (representados pelo círculos) e dos extremos (representados por asteriscos) indicam a linha da base de dados correspondente a esses valores (na janela **data view**). Por exemplo, o valor correspondente ao outlier 266 corresponde ao valor de 92 kilos da linha 266 da barra azulada.

Uma das formas para remoção de *outliers* e ou extremos e através da construção de um filtro. Através do comando **Data/selected cases/if condition is satisfied** e depois no **if**. A condição *sex=1* terá que ser mantida e portanto teremos que adicionar uma condição que elimine os *outliers*/extremos existentes. Experimente o seguinte comando:

Sex = 1 & ((Kilos<91 & Smoker=1) | (Kilos<95 & Smoker=2))

Questão nº8: O operador *&* representa uma intercepção enquanto o operador | representa uma reunião. Qual o resultado previsto? **Dica:** Observe se os valores dos outliers/extremos encontrados anteriormente são superiores a 91 kgs no caso de ser fumador e superiores a 95 kgs no caso de não ser um não fumador.

Calcule as estatísticas descritivas sem os valores dos outliers, utilizando o comando **Analyse/descriptive statistics/explore**. Não é necessário alterar as variáveis anteriormente colocadas nos quadros.

Questão nº9: Preencha a seguinte tabela e indique qual deve ser a decisão final: remove-se ou não os outliers/extremos? **Nota:** Após a remoção dos *outliers*/estremos das caixas de bigodes, poderão aparecer novos valores considerados *outliers*/extremos, mas estes não são considerados "verdadeiros" *outliers*/extremos, mas sim resultantes da eliminação dos primeiros *outliers*/extremos e portanto não devem ser removidos.

Tabela 2 – Resultados da estatística descritiva para variável peso (em kgs) de fumadores (n = _____) e não fumadores (n = 165) do sexo masculino sem os outliers

Peso (kg)	Média	D.P.	1º Quartil	Mediana	3º Quartil
Fumadores		7.5			
Não fumadores			64.1		