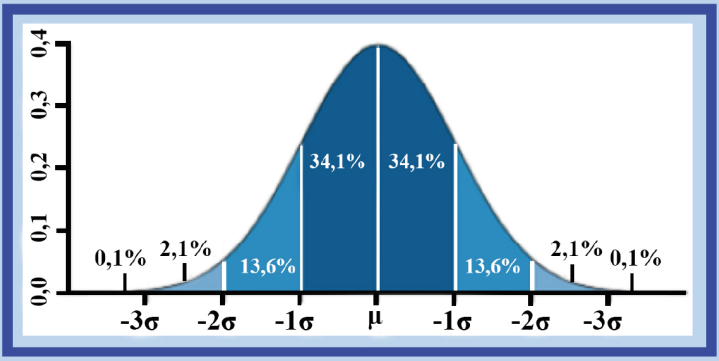
ESTATÍSTICA DESCRITIVA APLICADA À PSICOLOGIA

Com introdução ao SPSS



Edwirde Luiz Silva Camêlo
Dalila Camêlo Aguiar
Ivan Olier
Ramón Gutiérrez Sánchez

30 anos
eduepb

Edwirde Luiz Silva Camêlo
Dalila Camêlo Aguiar
Ivan Olier
Ramón Gutiérrez Sánchez

Estatística descritiva aplicada à psicologia

Com introdução ao SPSS

eduepb
Campina Grande-PB | 2025

Universidade Estadual da Paraíba
Profª. Célia Regina Diniz | *Reitora*
Profª. Ivonildes da Silva Fonseca | *Vice-Reitora*

Editora da Universidade Estadual da Paraíba
Cidoval Morais de Sousa | *Diretor*

Editora indexada no SciELO desde 2012

Editora filiada a ABEU

EDITORA DA UNIVERSIDADE ESTADUAL DA PARAÍBA
Rua Baraúnas, 351 - Bairro Universitário - Campina Grande-PB - CEP 58429-500
Fone: (83) 3315-3381 - http://eduepb.uepb.edu.br - email: eduepb@uepb.edu.br

Editora da Universidade Estadual da Paraíba
Cidoval Morais de Sousa (*Diretor*)

Expediente EDUEPB

Design Gráfico e Editoração
Erick Ferreira Cabral
Jefferson Ricardo Lima A. Nunes
Leonardo Ramos Araujo

Revisão Linguística e Normalização
Antonio de Brito Freire
Elizete Amaral de Medeiros

Assessoria Editorial
Eli Brandão da Silva

Assessoria Técnica
Thaise Cabral Arruda

Divulgação
Danielle Correia Gomes

Comunicação
Efigênio Moura

Depósito legal na Câmara Brasileira do Livro - CDL

E79 Estatística descritiva aplicada à psicologia da saúde [recurso eletrônico] : com aplicação no SPSS / Edwirde Luiz Silva Camêlo … [et al.]. – Campina Grande : EDUEPB, 2025.
188 p. : il. color. ; 15 x 21 cm.

ISBN: 978-65-5221-094-4 (Impresso)
ISBN: 978-65-5221-097-5 (4.000 KB - PDF)
ISBN: 978-65-5221-096-8 (Epub)

1. Estatística Descritiva. 2. Estatística Aplicada à Psicologia. 3. Programa SPSS. 4. Estatística Básica. I. Camêlo, Edwirde Luiz Silva. II. Aguiar, Dalila Camêlo. III. Olier, Ivan. IV. Sánchez, Ramón Gutiérrez. V. Título.

21. ed. CDD 519.53

Ficha catalográfica elaborada por Fernanda Mirelle de Almeida Silva - CRB - 15/483

AGRADECIMENTOS

Gostaríamos de agradecer à EDUEPB, que muito tem apoiado e estimulado a divulgação dos trabalhos dos professores da UEPB. Devemos também agradecer ao trabalho profissional do Prof. Cidoval???, que pacientemente abrilhantou este trabalho realizando as devidas correções gramaticais. Finalmente, nós autores não poderíamos deixar de agradecer aos alunos que nos permitiram a experiência para elaboração deste livro, aos colegas que contribuíram para a realização desta compilação e em especial ao **Eterno, Criador dos céus e da terra** que nos outorgou vida para concluir esta obra.

Lista de Tabelas

Lista de Figuras

Sumário

Prefácio

A estatística básica e sua prática constitui uma introdução à estatística para estudantes de estatística e psicologia I e II e também os tópicos especiais em estatística para o mestrado de Pós-Graduação em Psicologia da Saúde da Universidade Estadual da Paraíba. Neste prefácio, dou uma descrição do livro para permitir aos professores julgarem se ele é adequado aos seus alunos de graduação e pós-graduação.

Em linhas gerais esse livro dar ênfase ao pensamento estatístico aplicado a psícologia; mais dados e conceitos, menos teorias e menos receitas; promover o ensino ativo que é caracterizado como métodos de ensino que fortalecem a formação de vínculos democráticos na relação entre professor, alunos e conteúdo.

O livro é elementar quanto ao nível da matemática exigida e aos processos estatísticos apresentados, o livro tem como objetivo dar ao estudante não só a compreensão das idéias fundamentais da estatística e psicologia como a habilidade necessária para lidar também com data-frame. Os exemplos e exercícios foram aplicados em psicologia com objetivo de proporcionar um fundamento suficiente nas estatísticas descritivas. Com freqüência, faço conclusões que são mais do que um simples número. Os exercicios visa a construção do entendimento de estatística aplicada a psicologia.

Os Capítulos 1 apresenta uma breve introdução ao SPSS. No capítulo 2 apresenta as etapas para um estudo estatístico. No capítulo 3 apresenta um resumo das estatísticas descritivas. No capítulo 4 apresenta as medidas de posição. No capítulo 5 apresenta as medidas de dispersão. No capítulo 6 as medidas de forma. No capítulo 7 as medidas correlacionais. No capítulo 8 e 9 as distribuições para variáveis contínuas e discretas.

Queremos ainda mostrar aos graduandos e pós-graduandos como incorparar os resultados das suas análises e como interpretar os resultados nos artigos. Tentamos simplificar conceitos complexos, e, algumas vezes, bastantes complexos. Entretanto, ao facilitar existe uma perda de acurácia. Nos exercícios propostos após exemplos de cálculos ajudarão a resolver os problemas de intervalo de confiança. Os conjuntos de dados para exemplos e exercícios constam no próprio texto, de onde podem ser lidos para qualquer programa estatístico além do SPSS, como por exemplo, o programa R.

Capítulo 1

Uma breve introdução ao SPSS

Este guia fornece um conjunto de tutoriais para realizar uma análise útil do seu dados em psicologia. Você pode trabalhar com os tutoriais sequencialmente ou pular para os tópicos sobre os quais deseja informações adicionais. Este capítulo apresenta as funções básicas e mostra uma sessão típica. Considere o arquivo INTRODUCAO.sav de dados do IBM SPSS, vamos gerar um resumo estatístico simples e um gráfico.

Os capítulos a seguir incluirão muitos tópicos do programa SPSS. Esperamos poder fornecer a você uma estrutura básica para entender as principais ferramentas para auxiliar no conteúdo programático da disciplina Estatística e Psicologia.

O programa SPSS é um pacote estatístico, composto de diferentes módulos, desenvolvido também na área de psicologia. Está baseado no ambiente Windows, sendo de fácil operação e muito abrangente, pois permite realizar uma grande amplitude de análises

estatísticas e gráficas (análises descritivas de posição, dispersão e forma; teste de hipóteses, intervalos dde confiança, análises multivariadas, módulos gráficos, entre outras).

1.1 Janela do SPSS

Com o SPSS é possível criar, definir e modificar variáveis quantitativas e qualitativas; realizar cruzamentos de variáveis; gerar diversos tipos de gráficos; verificar a existências de associações o verificar a existências de correlações, etc.

O SPSS mostra a janela de digitação (ou input) de dados SPSS Data Editor, na qual os bancos de dados são gerados e analisados. Na janela do programa SPSS Data Editor as linhas são relativas aos indivíduos (casos), participantes ou respondentes e as colunas relativas as variáveis investigadas que podem ser quantitativa ou qualitativa.

Esta janela gera um dos três tipos de arquivos associados ao SPSS. Esse arquivo tem terminação .sav e armazena todas as informações relativas ao banco de dados, como definição de variáveis e os dados digitados.

A janela acima apresentada possui várias colunas relativas aos principais parâmetros de cada uma das variáveis da planilha. São 11 colunas:

1. Name

   Refere-se ao nome atribuído a variável, composto de até oito caracteres, que será colocado nas colunas da janela de input de dados. Deve-se clicar em qualquer cela dessa coluna para que se possa digitar o nome da variável;

2. Type

   Refere-se ao tipo de variável, ou seja, sua característica de notação (numérica: contínua ou discreta; qualitativa: nominal ou ordinal). Clicando-se na cela desta coluna abre-se a caixa de diálogo. Observe que a caixa de diálogo Variable Type permite a escolha de vários tipos de variáveis. A definição de cada uma pode ser acessada clicando-se no botão direito do mouse colocado em cima da opção. Esta mesma caixa de diálogo também permite a escolha de outras características da variável (width (comprimento) e número de casas decimais).

3. Width

   É o número de caracteres da variável nomeada. Pode ser definido diretamente na cela ou através da caixa de diálogo Variable Type.

4. Decimals

   É o número de casas decimais, a direita da vírgula, que serão apresentadas tanto para um número categorizado como para variáveis métricas.

5. Label

   Define o nome atribuído a uma variável e não possui restrição de número de caracteres (e.g., idade de um idoso estressado, escolaridade de uma pessoa ansiosa, descrição do item de um questionário para análise psicológica, entre outras);

6. Value

   São os valores que os labels podem assumir. Neste parâmetro o pesquisador deve definir todos os possíveis valores que uma

variável pode assumir. Um tipo comum em psicologia é a escala Likert entre outras em psicologia.

7. Missing(ausente)

   Define o tipo de tratamento para os indivíduos ausentes que o pesquisador deseja considerar. O Default do programa geralmente é usado, mas outros valores podem ser definidos como 999 ou 99.

8. Columns(colunas)

   Indicar o tamanho da coluna da variável, que será apresentada na janela de input de dados;

9. Align

   Define o alinhamento dentro de cada célula da planilha de dados (esquerda, centralizado ou direita).

10. Measure

    Aqui informa o tipo de variável (contínua ou discreta). Esta definição é fundamental, pois o SPSS habilitará o uso das variáveis em certos procedimentos a partir do tipo de medida (measure) selecionada.

A maioria dos exemplos fornecidos usa o arquivo de dados no formato .sav (do SPSS). Esses dados é um estudo fictício de 16 pessoas que contém informações básicas sobre idade (em anos) e nível de estresse variando de 1 a 10. O arquivo DadosIdadeEstresse.sav será uma amostra representativa do arquivo de dados original, reduzido considerando apenas 16 casos (indivíduos).

Na Tabela 1.1 mostra alguns dados para inserir no SPSS.

Tabela 1.1: Algumas técnicas estatística apropriada

Idades	Níveis de estresse	Idades	Níveis de estresse
21	4	23	4
45	9	15	6
32	8	33	7
54	6	44	6
21	4	25	4
35	8	31	7
32	9	71	8
51	7	19	5

Passos para inserir esses dados no SPSS:

1. Digitar os dados nas duas primeiras colunas no SPSS

   Antes de executar qualquer análise, precisa-se fornecer os dados ao SPSS. Observa-se a planilha, formada por células, é semelhante ao Excel, que são o encontro das linhas (indivíduos) e colunas (variáveis).

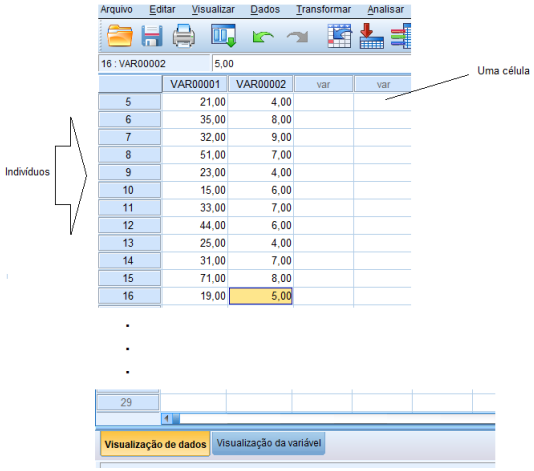


Figura 1.1: Inserindo as duas colunas no SPSS

2. Nomeando as colunas. Click em Visualização das variáveis como se observa na Figura 1.1 e visualizando as linhas.

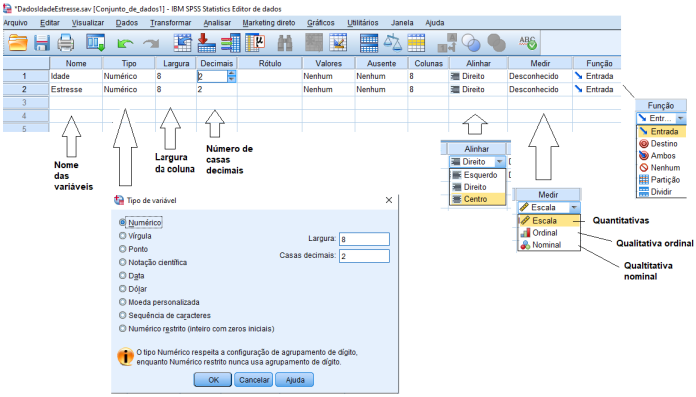


Figura 1.2: Nomeando variáveis

Na tela de visualização de variáveis (variable View), as colunas representam variáveis, e as linhas os indivíduos. Precisa-se fornecer o nome de cada variável. Click em Visualização de variáveis. Em outras palavras, o editor de dados tem dois painéis (views): o de dados e o de variáveis. Nas colunas de dados ficam por exemplo os dois grupos e a quantidade de clínicas em Campina Grande e João Pessoa. No editor de dados entram os dados obviamente dados e o editor de variáveis permite que definamos várias características das variáveis do editor de dados.

3. Estudando as funções: rótulos, valores e ausente. Considere mais uma variável que chamaremos de gênero:

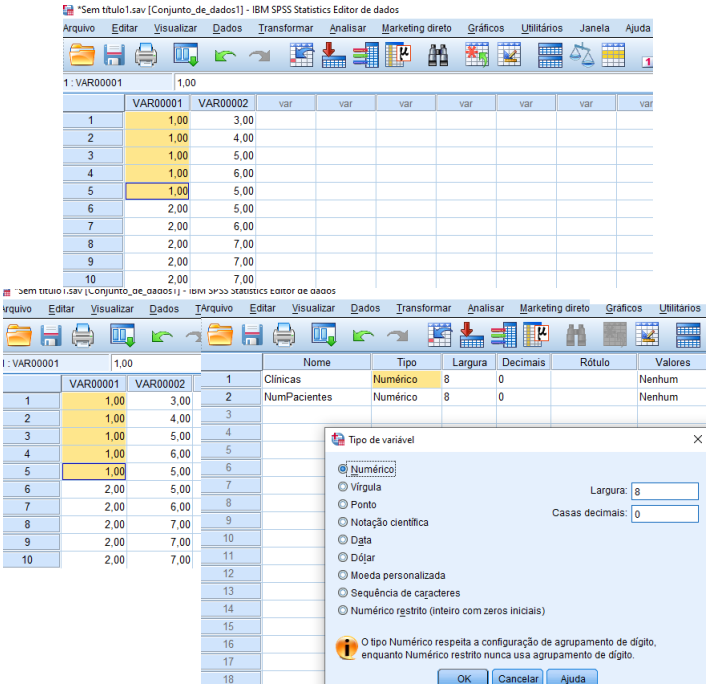


Figura 1.3: Número de decimais, tipos de variáveis, rótulos e valores

4. Valores

   Em Valores click em Nenhum, inserir 1 para Campina Grande 2 para João Pessoa.

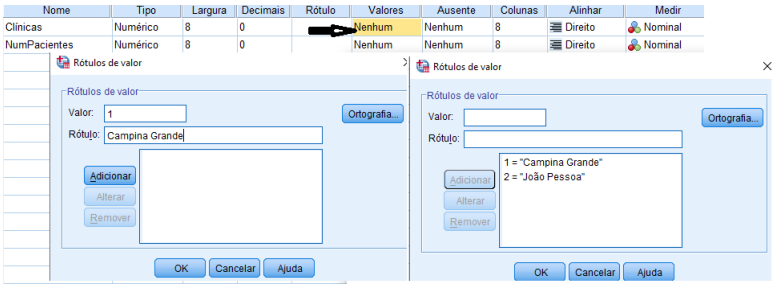


Figura 1.4: Nomes dos valores 1 para CG e 2 para JP

Tudo que temos que fazer é clicar com o mouse: Data View w Variable View. E criar variáveis codificadas. Uma variável codificadora é uma variável que consiste em uma série de números representada em níveis de uma variável de tratamento ou que descreve diferentes números de grupos, no caso, 1 para Campina Grande e 2 para João Pessoa.

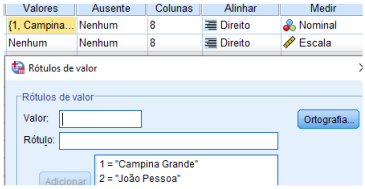


Figura 1.5: Nomes dos valores 1 para CG e 2 para JP

Finalmente, definimos as variáveis de códigos e seus valores no SPSS e os tipos de variáveis utilizadas, quantidade de decimais, alinhamento dos números e nomes dentro das células e a dimesão das colunas.

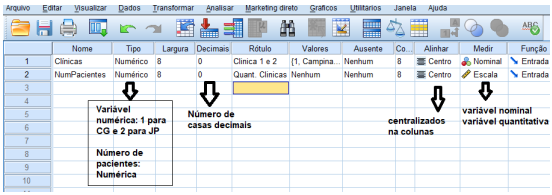


Figura 1.6: Visualizando as variáveis

5. Valores que faltam (missing)

   Muitas vezes ocorre que não temos dados que não podem, por algum causa, ser obtidos, dados que faltam ou são desconhecidos. As vezes, o intrevistado tem vergonha de falar, esquecer-se de responder a algumas questões, etc. Mais adiante veremos como preencher essa lacuna.

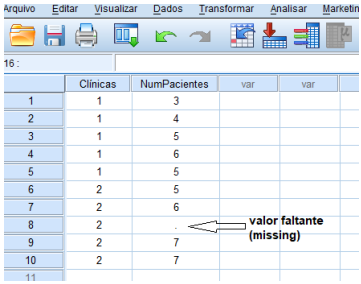


Figura 1.7: Elementos faltante na variável

1.2 Resumo de casos

Vamos iniciar as análise preliminares usando Resumo de Casos, como se observa na Figura 1.8.

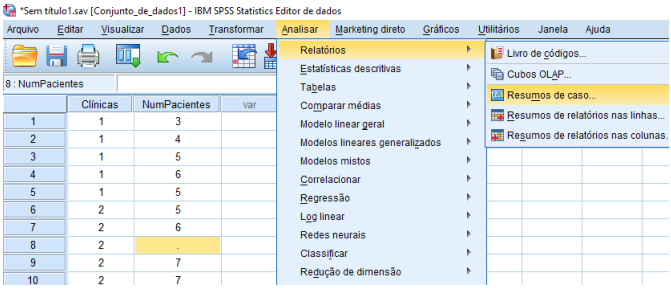


Figura 1.8: Resumos de casos no SPSS

Clicando em Ok, temos:

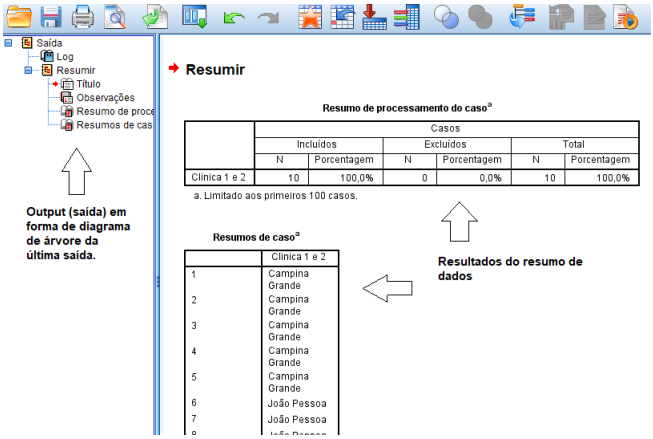


Figura 1.9: Resultado do resumos de casos

Observe que temos 100% dos dados válidos, nenhum caso foi excluído. Na última coluna temos Válido + excluído = total.

Este livro não tem a pretensão de substituir a aprendizagem séria e comprometida do pensamento estatístico. Muito pelo contrário, este manual é apenas uma leitura complementar para uso do SPSS. Portanto, foi concebido como uma ferramenta extra para

servir de apoio ao ensino de estatística, feito com, no mínimo, o uso de um livro texto de estatística voltado para cientistas sociais. Isto significa que o manual não pretende substituir a leitura de um bom livro introdutório ou de análise multivariada de dados, muito menos a realização de exercícios de aplicação de conhecimentos estatísticos, mas sim auxiliar o aluno na realização de procedimentos estatísticos via SPSS. Considera que o desenvolvimento destas competências é parte essencial da formação de pesquisa em estatística e psicologia.

1.3 Menus de comando

Na janela SPSS Data Editor o programa possui uma série de menus de comandos, que possibilitam a manipulação e análise dos dados, bem como os procedimentos do windows para trabalhar com arquivos.

O primeiro desses menus é denominado File e tem como principais funções abrir, salvar e importar, além de outras funções comuns aos diferentes programas da mesma plataforma operacional ou específicos do SPSS.

O segundo menu é o Editar, que possui as seguintes funções: voltar a última ação, copiar, colar, cortar, procurar (find) e opções. O menu opções define-se diferentes parâmetros no SPSS, como formato de apresentação dos dados digitados na SPSS Data Editor, o formato de geração das tabelas na janela de output (saída) de dados, entre outras funções.

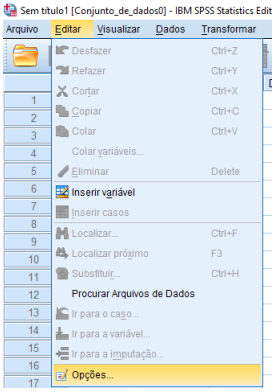


Figura 1.10: Menu de opções no SPSS

A Figura 1.11 apresenta a janela de diálogo do menu editar em opções. Este menu de opções é importante para adaptar o padrão de apresentação do programa às características ou preferências de trabalho do pesquisador. Idioma que o pesquisador deseja na saída do SPSS; o tipo de fonte das tabelas, configurações de gráficos, $\cdots$.

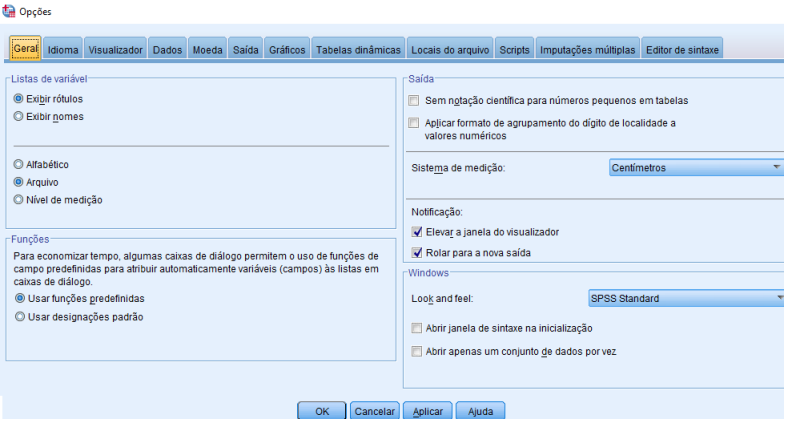


Figura 1.11: Geral no Menu de opções no SPSS

O próximo menu da barra de comandos é denominada Data. Esse menu possibilita manipular o arquivo de dados de diferentes maneiras. A Figura 1.3 apresenta a tela do SPSS com este menu aberto.

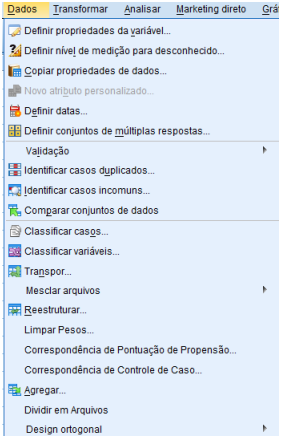


Figura 1.12: Menu de dados no SPSS

Vamos estudar alguns desses. Como pode ser observado na Figura 1.13 o menu Data possui vários comandos. Alguns serão

apresentados em tópicos a seguir. Para isso considere a seguinte amostra:

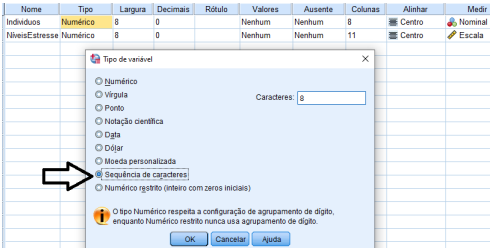
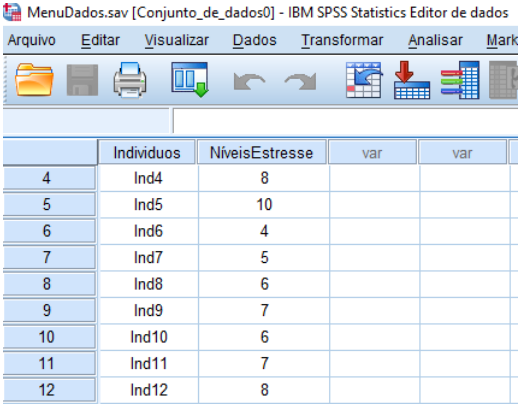


Figura 1.13: Tipos de variáveis no SPSS

Observe que em indivíduos o tipo de evariável considerada foi: Sequência de caracteres.



	Individuos	NíveisEstresse	var	var
4	Ind4	8		
5	Ind5	10		
6	Ind6	4		
7	Ind7	5		
8	Ind8	6		
9	Ind9	7		
10	Ind10	6		
11	Ind11	7		
12	Ind12	8		

Figura 1.14: Inserindo as variáveis indivíduos e níveis de estresse

O comando select cases possibilita a seleção de um grupo de casos em um mesmo arquivo ou a criação de um outro arquivo a partir de um grupo de casos inicial.

- Selecionando indivíduos

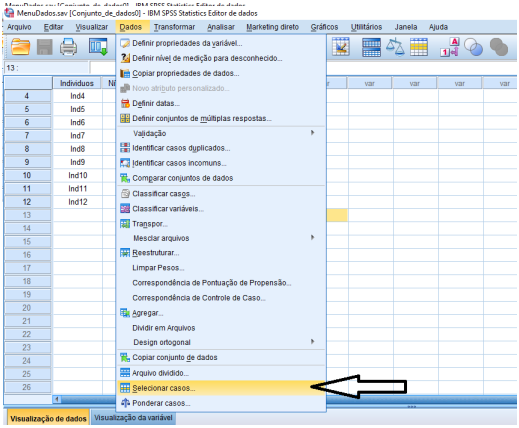


Figura 1.15: Seleção de casos

Seleciona casos: se a condição for cumprida. Continuar e Ok

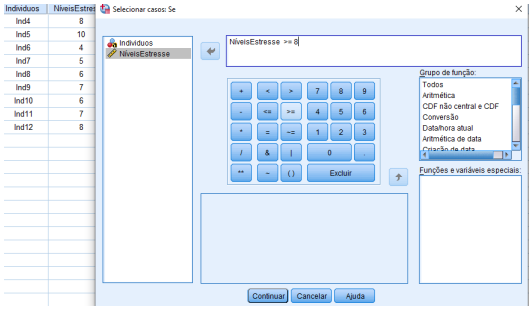


Figura 1.16: Seleção de casos para nível de estresse maior ou igual a 8

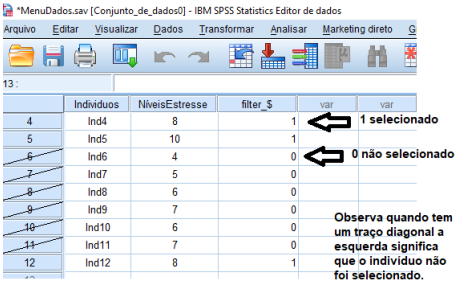


Figura 1.17: Os casos considerados

- Selecionando indivíduos aleatórios. Dados - Selecionar casos. Amostra aleatória de casos, como se observa na Figura 1.18. Clicar em continuar e ok.

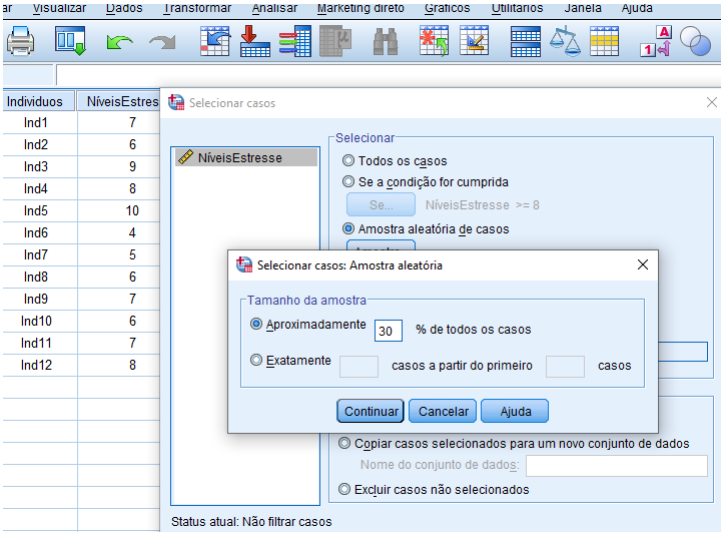


Figura 1.18: Seleção de apenas 30% dos casos

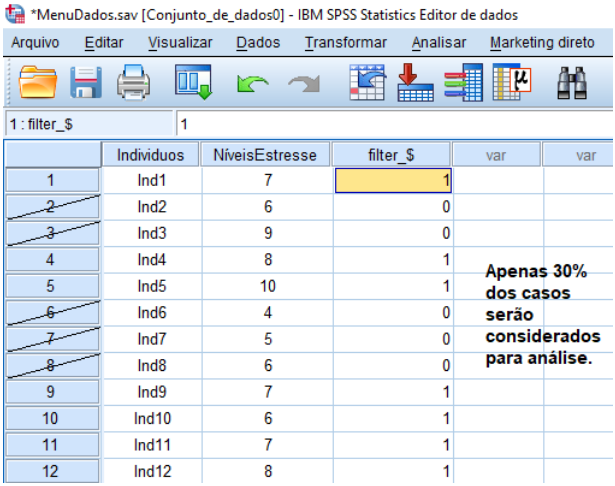


Figura 1.19: Seleção de apenas 30% dos casos

- Seleção considerando intervalo de casos. No caso tivemos uma mostra aleatória de 5 indivíduos e selecionou apenas 3 indivíduos aleatoriamente dos cinco.

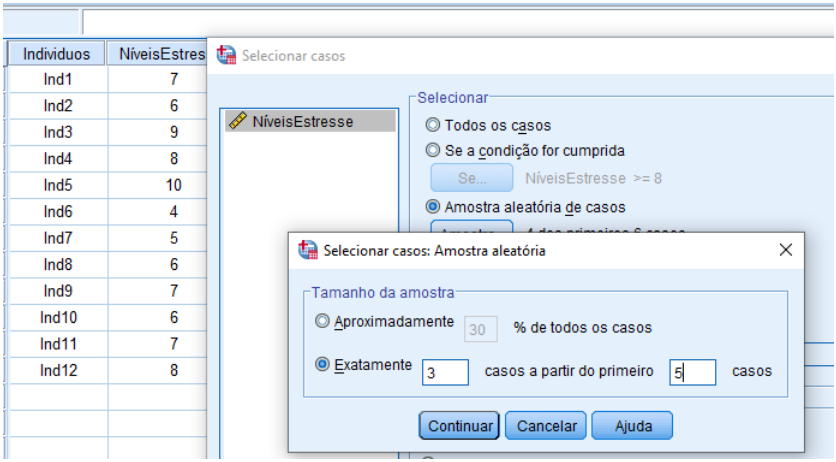


Figura 1.20: Seleção de apenas 3 de 5 indivíduos

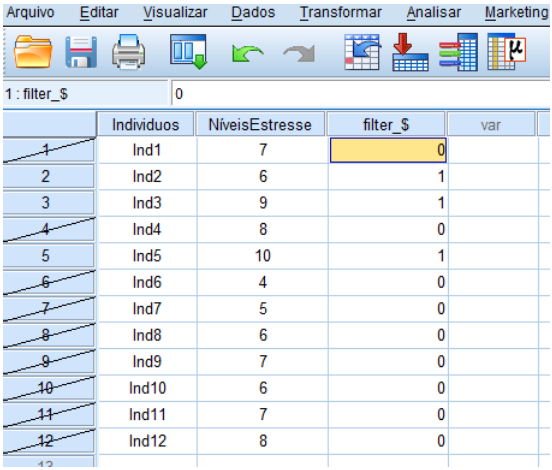


Figura 1.21: Seleção de apenas 3 de considerando 5 indivíduos

1.4 Operações usando o SPSS

Operadores	Funções
x + y	Soma
x - y	Subtração
x * y	Multiplicação
x / y	Divisão
x x^y	Eleva a distintas potências
x = y	Igualdada
x <> y	Não igualdade
x > y	Maior que
x ≥ y	Maior ou igual que
x < y	Menor que
x ≤ y	Menor ou igual que
ABS(x)	Valor absoluto
AVG (numvar)	Calcula a média aritmética de uma variável
CELL (var,n)	Selecione uma linha de uma coluna específica, ou seja, uma célula.
CHISQUARE (n ,df)	Calcula a probabilidade cumulativa de uma distribuição qui-quadrado com df graus de liberdade para um determinado ponto.
COMPRESS (var,logical)	Selecione as linhas que atendem a uma condição.
COUNT (start,end,step)	Crie um vetor de números sequencialmente, começando em estrela, até o valor final e pulando por passos.
CV (numvar)	Calcule o coeficiente de variação.
DIFF (numvar)	Calcula a diferença entre dois valores consecutivos da variável.
DROP(var,n)	Selecione todas as linhas de uma variável, exceto o primeiro n.
DROPLAST (var,n)	Selecione todas as linhas de uma variável, exceto o último n.
FACT(x)	Calcule o fatorial.
FIRST (n)	Retorna um 1 para as primeiras n linhas e 0 para o restante.
FIRSTROWS (var,n)	Selecione as primeiras n linhas e substitua as outras linhas pelos códigos de valor ausentes
GEOMEAN (numvar)	Calcule a média geométrica
INVCHISQUARE (n ,df)	Calcula o valor crítico para uma distribuição qui-quadrado com df graus de liberdade dada uma probabilidade.
INVNORMAL (n ,mean,sdev)	Calcula o valor crítico para uma distribuição normal dada uma probabilidade.

Tabela 1.2: Operadores e suas funções no SPSS

Segue a continuação de operadores e funções no SPSS.

Operadores	Funções
KURTOSIS (numvar)	Calcule o coeficiente de curtose.
LAST (n)	Gera 1 para as últimas n linhas e 0 para as outras linhas
MAX (numvar)	Valor máximo de uma variável.
MIN (numvar)	Valor mínimo de uma variável.
MODE (numvar)	Moda de uma variável.
NORMAL (n ,mean,sdev)	Calcula a probabilidade cumulativa de uma distribuição normal dado um ponto.
PERCENTILE (numvar,n)	Calcule os percentis.
Q25 (numvar)	Calcule o primeiro quartil.
Q75 (numvar)	Calcule o terceiro quartil.
RANDOM (n)	Ele gera n que atribui aleatoriamente às linhas de um arquivo, ao restante das linhas atribui zeros.
INVSNEDECOR (n ,df1,df2)	Calcula o valor crítico para uma distribuição Snedecor F dada uma probabilidade.
INVSTUDENT (n ,df)	Calcula o valor crítico para uma distribuição t de Student dada uma probabilidade.
IQR (numvar)	Calcule o intervalo interquartil.
JOIN (var,var)	Junte duas colunas uma no final da outra
JOIN3 (var,var,var)	Junte três colunas uma após a outra.
JOIN4 (var,var,var,var)	Junte quatro colunas uma após a outra.
Operadores	Funções
RANDOM (n)	Ele gera n que atribui aleatoriamente às linhas de um arquivo, ao restante das linhas atribui zeros.
RANGE (numvar)	Calcule o intervalo.
REP (var,n)	Repetir cada valor de var n vezes.
REPLACE (numvar ,old,new)	Substitue todas as ocorrência valor antigo (old) da variável con outro número diferente new (novo).
RNORMAL (n,mean,sdev)	Gera números aleatórios de uma distribução normal.

Continuação dos operadores do SPSS

ROUND (x)	Aproxima ao inteiro mais próximo.
RUNIFORM (n ,lower,upper)	Gera números aleatórios de uma distribução uniforme.
SD (numvar)	Calcula o desvio padrão.
SELECT (var,logical)	Selecione as linhas que atendem a uma condição.
SERROR (numvar)	Calcule o erro padrão.
SKEWNESS (numvar)	Calcule o coeficiente de assimetria.
SNEDECOR (n ,df1,df2)	Calcula a probabilidade cumulativa de uma distribuição normal de Snedecor F dado um ponto.
SQRT (x)	Calcule a raiz quadrada.
STANDARDIZE (numvar)	Digite os valores de uma variável.
STRIPBLANKS (charvar)	Remova os espaços em branco duplos de uma variável de caractere.
STUDENT (n ,df)	Calcula a probabilidade cumulativa da distribuição t de um aluno dado um ponto.
SUM (numvar)	Calcula a soma de todos os valores da variável.
VARIANCE (numvar)	Calcule a variação.

Tabela 1.3: Operadores e suas funções no SPSS

Capítulo 2

Etapas para um estudo estatístico

As diferentes etapas que um estudo estatístico são as seguintes:

Formulação de uma hipóteses sobre a população

Objetivo do estudo

Plano de estudo

Plano de investigação

Planejamento Decide-se que dados vamos coletar:

- Que indivíduos pertencem ao estudo (amostra)
- Que dados coletamos dos mesmos (variáveis)

Amostra Especificamos como coletar os dados:

- Amostra aleatória simples

- Amostra aleatória estratificada
- etc.

Análise descritiva

Análise análise inferencial

- Testes de hipóteses, modelos de regressão, $\cdots$.
- Nívis de confiança, valor p, $\cdots$.

Formulação de novas hipóteses. Conclusões e generalização de resultados

Em psicologia, muitos dos experimentos e investigações são estudos de séries de dados nos quais nenhuma técnica de amostragem é usada, temos apenas uma hipótese que queremos verificar, alguns grupos de pessoas nas quais se pretende testar.

2.1 Tipos de variáveis e transformação

Inicialmente o SPSS considera três tipos de variáveis, a saber, variável quantitativa, variável qualitativa: nominal e ordinal, como se observa na figura abaixo.

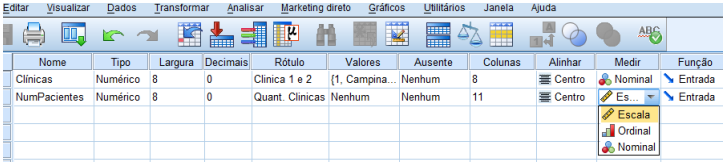


Figura 2.1: Três tipos de variáveis no SPSS

Quando se deseja medir ou atribuir nomes as modalidades de uma variável. Quando medimos a idade, atribuímos números que

são valores numéricos das diferentes modalidades de idades. Também fazemos quando mensuramos as diferentes modalidades de crenças religiosas (protestante, católicos, etc). Segundo a escala de medidas se distingue em variáveis:

- **Quantitativa**: quando, entre um conjunto de observações, qualquer observação individual corresponde a um número que representa uma quantidade ou contagem. Uma variável quantitativa toma valores numéricos com os quais tem sentido efetuar operações aritméticas, como adicionar ou tomar uma média amostral da idade de idosos estressados. Variáveis quantitativas são divididas em:

  Discreta - admite apenas números inteiros (conjunto dos números Naturais). Exemplo: quantidade de pessoas com transtorno de ansiedade numa escola privada.

  Contínua - admite números fracionários (conjunto dos números Reais). Exemplo: peso(kg) de pessoas com transtorno alimentar restritivo evitativo.

  As variáveis quantitativas contínuas são aquelas que podem tomar qualquer valor numa continuidade, de modo que não tem valores consecutivos, já que entre dois valores qualquer seguem existindo infinitos valores possíveis. As variáveis contínuas podem se agrupar em categorias, mas de umna forma arbitrária. Por exemplo, a variável dos pacientes se pode dividir em categorias como pequeno, normal ou alto, e os limites de cada uma dessas categorias de forma arbitrária.

  A diferença entre essas quatro tipos de variáveis é importante por vários motivos:

  - O cálculo de alguas medidas de posição ou de dispersão não tem sentido com variáveis qualitativas, por exemplo no caso da variável sexo.
  - Para a aplicação correta de técnicas de análise estatística: assm a maioria dos testes não paramétrico requer que a variável seja ao menos ordinal, e muitos métodos de análise multivariada exige que as variáveis sejam quantitativas (por exemplo, análise fatorial ou análise discriminante).

- **Qualitativa**: quando, entre um conjunto de observações, qualquer observação individual corresponde a uma palavra ou a um código que representa uma classe ou categoria. Indica a qual de diversos grupos ou categorias um indivíduo pertence. Variáveis qualitativas são divididas em:

  Nominal - quando nomeia em categorias ou espécies mutuamente exclusivas mas com idênticas propriedades. Exemplo: nacionalidade, status social, religião com jovens com estresse.

  Ordinal - dispõe as informações segundo dada ordem, posição hierárquica ou sequência classificatória. Exemplo: classes sociais, grau de instrução, nível de estresse categorizado de 1 a 5.

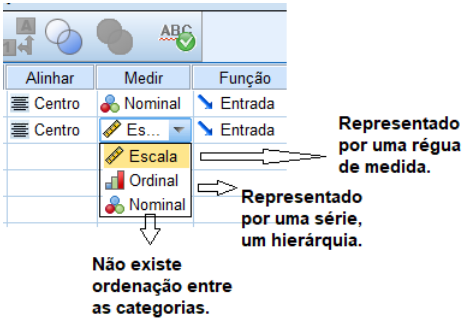


Figura 2.2: Tipos de variáveis no SPSS

2.1.1 Transformação de variáveis

É importante realizar a transformação de dados. Converter valores simbílicos em valores numéricos. Como visto anteriormente os valores simbólicos são nominais ou ordinais, diferentes técnicas podem ser empregadas.

- Transformação Categórico-Numérico

  Quando a variável é do tipo nominal e assume apenas dois valores, se os valores denotam a presença ou ausência de uma característica ou se apresentam uma relação de ordem, um dígito binário é suficiente, ou seja, o valor 0 indica ausência e o valor 1, a presença da característica. O menor valor ordinal pode assumir o valor 0 e o outro o valor 1. Para mais de dois valores, a técnica utilizada na transformação depende da variável ser nominal ou ordinal. Dependendo de valores nominais, a sequência binária para representar cada valor pode ficar muito longo. Segundo Sternberg dependendo da combinação de intimidade, paixão e compromisso produz diferentes dimensões num relacionamento. Vamos categorizar de 1

a 6 essas dimensões.

Amor	Código
Gostar	1
Amor companheiro	2
Amor vazio	3
Encantamento	4
Amor romântico	5
Amor completo	6

Tabela 2.1: Diferentes dimensões num relacionamento

Seria necessário utilizar vetores com n elementos. Uma alternativa seria utilizar uma representação dos possíveis valores nominais por conjunto de pseudoatributos. Os valores pseudoatributo podem ser do tipo binário, inteiro ou real. Na Tabela representa uma conversão de valor ordinal para inteiro.

- Transformação Numérico - Categórico

  Alguns estratégias podem ser utilizadas nesta conversão. São elas:

  a) Larguras iguais: divide o intervalo original de valores em subintervalos com mesma largura.

  b) Frequência iguais: Ao realizar uma distribuição de frequência se atribui o mesmo números de cados a cada subintervalos. Estas estratégias pode gerar intervalos de tamanho muito diferentes.

  c) Gráficos de dispersão tais como: *Boxplots*, histogramas e *QQ-plots*. Estes gráficos ajudam a obter uma inspeção visual dos dados visando a conversão.

Faixa etária	Código
$[18-20[$	A
$[20-22[$	B
$[22-24[$	C
$[24-26[$	D
$[26-28[$	E

Tabela 2.2: Diferentes faixa etária

- Transformação de Variáveis numéricas

  Quando os limites de valores das variáveis são muito diferentes, isso leva a uma grande variação de valores, ou ainda quando várias variáveis estão em escalas diferentes, logo, o valor numérico de uma variável precisa ser transformada. Isso é feito para que evite que uma variável predomine sobre outra. No entanto, poderá existir situações em que esta variação deve ser conservada por ser importante para a estimação de um bom modelo. A principal transformação utilizada em estatística é a normalização dos dados, que consiste em subtrair de sua média e dividi-los pelo desvio padrão da respectiva variável. Se as medidas de localização e de escala forem a média (μ) e o desvio padrão (σ), respectivamente, os valores de uma variável são transformados para um novo conjunto de valores com média 0 e desvio padrão 1, que é obtido pela seguinte Equação 3.1 nos valores originais da variável.

$$Z_i = \frac{X_i - \mu_i}{\sigma_i} \tag{2.1}$$

  Para detectar os *outliers*, observações que fogem da dimen-

são esperada. Para detectá-los, pode-se calcular o escore padronizado Z_i e considerar *outliers* as observações cujos escores, em valor absoluto sejam maiores que 3. Vamos padronizar a variável quantitativa denominada Idade.

Definição 2.1.1. *Valores discrepantes são observações que estão fora do padrão de uma distribuição. Procure sempre localizar valores discrepantes e explicá-los.*

Padronizar as variáveis resultar deixar a média amostral ou populacional igual a 0 e um desvio padrão amostral ou populacional igual a 1. Já normalizar tem como objetivo colocar as variáveis dentro do intervalo $[0, 1]$, quando têm valores negativo.

Exemplo 2.1.1. *AAAAA*

2.2 Criando variáveis no SPSS

No SPSS podemos criar variável métrica (representada por uma régua), nominal (representadas por três círculos com cores diferentes) e ordinal (representada por um gráfico de barra indicando uma sárie ou uma hierarquia).

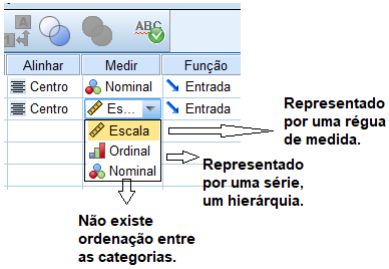


Figura 2.3: Tipos de variáveis no SPSS

2.3 Transformações de variáveis

Um pesquisador em psicologia as vezes precisa normalizar as variáveis de estudo com objetivo de miniminar as distorções em algumas análises.

Abaixo estão as transformações de potência para os diagramas de potência de dispersão por nível. Para transformar os dados, devemos selecionar uma potência para a transformação. Podemos escolher uma das seguintes alternativas:

- Log natural. Transformação de logaritmo natural. Esta é no SPSS uma transformação por default (por defeito).

  Escala original tipo contagens, x. A transformação pode ser $\ln x$.

- *raizquadrada*. Para cada valor de x, calcula-se a raiz quadrada ou $\sqrt{x}$.

- $1/raizquadrada$. Para cada valor de x, calcula-se o inverso da raiz, $\frac{1}{\sqrt{x}}$.

- Porporções, p. A escala transformada será: $logit(p) = \frac{1}{2}\ln\left(\frac{p}{1-p}\right)$.

- Correlações, r. A escala transformada será: $Z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$. Essa escala transformada se denomina Z de Fisher.

Transformações para diminuir os valores de X.

- $x^{-1} = \frac{1}{x}$
- $x^{1/3} = \sqrt[3]{1}$.
- $x^{1/4} = \sqrt[4]{1}$.

Transformações para aumentar os valores de x: x^2, x^3 $\cdots$.

As padronizações de variáveis deve ser utilizadas com cautela, pois existe alguma relação real refletida nas escalas das variáveis originai. Outras padronização podem ser:

- O método Z scores normalizar cada variável (x) de forma a apresentar média 0 e desvio padrão 1. É calculada de usando a fórmula: $Z_i = \frac{x-\bar{x}}{s}$.

  Sendo $\bar{x}$ a média de x, e s o desvio padrão amostral. Se utilizarmos dados populacionais seria μ (média populacional) e σ (desvio padrão populacional).

- Método range $[-1;1]$. A variável padronizada tenha amplitude 1. Amplitude (A) é igual ao valor máximo subtraído do valor mínimo: $p = \frac{x}{amplitude}$.

- Método range $[0;1]$. A variável padronizada apresenta variação de 0 a 1: $p = \frac{x-\min}{amplitude}$.

- Método de máxima amplitude: confere à variável o valor máximo de 1: $p = \frac{x}{\max}$.

- Método de média 1. Transforma a variável de maneira que apresente média 1: $p = \frac{x}{\bar{x}}$.

- Método do desvio padrão 1. Transforma a variável de maneira que apresente desvio padrão 1: $p = \frac{x}{s}$.

Exemplo 2.3.1. *Vamos transformar os valores da variável Óbitos por Residência em 2020, de forma que tenha aproximadamente (por causa do tamanha da amostra) média 0 e desvio padrão 1. Os passos no SPSS serão:*

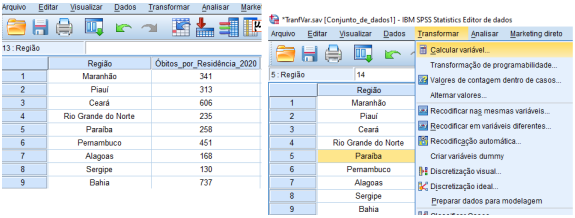


Figura 2.4: Transformação de variáveis

No SPSS seria:

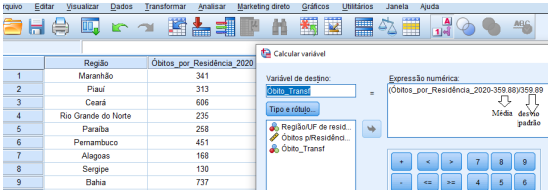
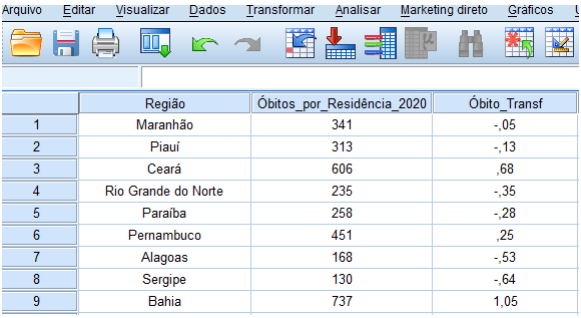


Figura 2.5: Transformação de variáveis passo a passo

A terceira coluna informa os valores já transformado da variável óbito.



	Região	Óbitos_por_Residência_2020	Óbito_Transf
1	Maranhão	341	-,05
2	Piauí	313	-,13
3	Ceará	606	,68
4	Rio Grande do Norte	235	-,35
5	Paraíba	258	-,28
6	Pernambuco	451	,25
7	Alagoas	168	-,53
8	Sergipe	130	-,64
9	Bahia	737	1,05

Figura 2.6: Variável óbito transformada

De forma análoga pode-se realizar as demais transformações.

2.4 Exercícios

1. Preencher o quadro abaixo

Qual tipo de variável?	Numérico	Sequência	Data	Hora
		n/d		
				
				

Figura 2.7: Tipos de variáveis

2. Que tipo de variável pode ser considerado : uma região, um código postal ou denominação religiosa?

3. Que tipo de variável pode ser considerado como níveis de satisfação de um atendimento psicológico. Os níveis dessa variável variam de muito insatisfeito a muito satisfeito?

4. Qual a diferença entre variável nominal e variável nominal ordinal?

5. Que tipo variável tem essas características?

    - A nota em um exame: reprovação, aprovação, notável, excelente.
    - Posição alcançada em um evento esportivo: 1^o, 2^o, 3^o, $\cdots$.
    - Medalhas de um evento esportivo: ouro, prata, bronze.
    - Temperatura dentro da sala de aula na disciplina psicologia e estatística I.

Capítulo 3

Análise descritivo

Primeiramente vamos estudar as variáveis categóricas aplicadas em psicologia da saúde.

Definição 3.0.1. *Uma variável categórica indica a qual de diversos grupos ou categorias um indivíduos pertence.*

Uma vez que os dados tenham sido coletados, tabulados e preparados para análise, a primeira tarefa a ser enfrentada (independentemente do tipo de análise que você pretende realizar) para cobrir os objetivos de um estudo) é formar uma ideia o mais exata possível sobre as características de cada variável.

Isso poderia ser feito fazendo uma lista de todos os valores, uma vez que, independentemente natureza de uma variável, uma listagem de todos os seus valores contém todas as informações disponíveis. No entanto, uma listagem de dados é de pouca (ou talvez nenhuma) utilidade. O que realmente é útil poder organizar e resumir esses dados de alguma forma que esclareça a situação para facilitar compreensão. No entanto, como um resumo não é uma

descrição detalhada dos dados, você só poderá capturar um aspecto parcial deles. Portanto, já desde o início, é importante conhecer as ferramentas estatísticas que permitem resumir os dados e conseguir escolher aqueles que darão uma resposta adequada às questões que você tem interesse em estudar.

Nesse sentido, se uma variável é categórica ou quantitativa, obter uma descrição completa geralmente requer atenção a três características claramente diferenciadas: a medida de posição, a medida de dispersão e a medida de forma da distribuição. Este capítulo apresenta as ferramentas estatísticas comumente usadas para descrever variáveis não métricas.

Uma das primeiras ações a serem feitas com dados estatísticos é sua descrição, por meio de estatísticas descritivas, como freqüências, médias, medianas, variância, entre outras. Muitas vezes também é objetivo da análise de dados a realização de cruzamento de diferentes informações, para se ter uma noção da relação existente entre variáveis da pesquisa psicológica. Todos os procedimentos referentes a essa descrição de dados, e conseqüentemente todo o conteúdo desse segundo capítulo da apostila, estão concentrados no menu Analyze $\rightarrow$ Descriptives Statistics do SPSS.

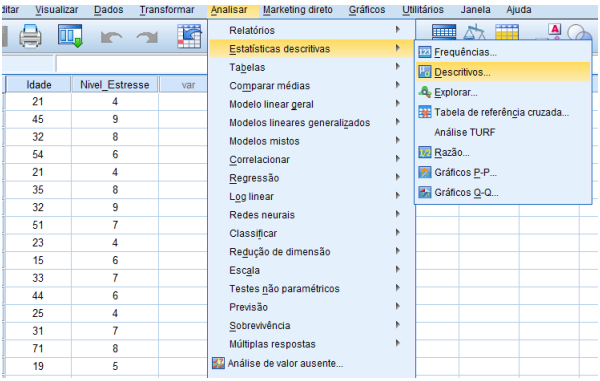


Figura 3.1: Opção descriptive statistic

Existem outras opções para encontrar estatísticas descritivas, mas a opção Explore é a mais flexível, como se observa na Figura 3.2. Essa opção permite que se acesse diversas estatísticas descritivas e é, desta forma, uma opção para se utilizar. As várias opções na janeta de diálogo são:

- Lista de variáveis (quantitativas ou qualitativas)
- Caixa para variáveis dependentes (dependent list)
- Caixa para variáveis de agrupamento (factor list)
- Opção de apresentação (display - embaixo à esquerda)
- Várias opções de botões (estatísticas, plots, diagramas, options)

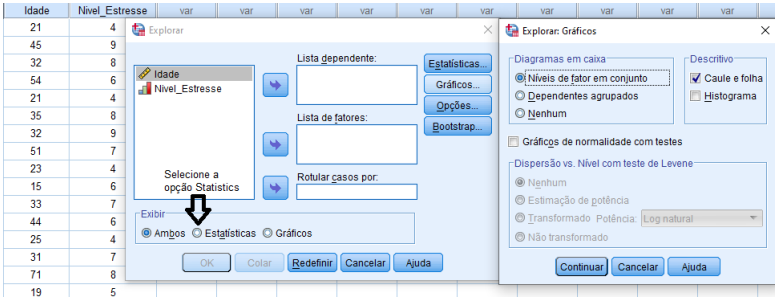


Figura 3.2: Opção Explore

3.1 Tabelas de frequência

A frequência pode descrever a quantidade de registros em cada categoria ou item contabilizado, podendo ser apresentado pelo número de ocorrência registradas pelos percentual que expressa esse número no conjunto de dados coletados. Além da frequência e sua proporção, pode-se incluir a frequência e o percentual acumulado como se observa na 3.1.

Uma tabela ou distribuição de frequência é uma maneira particular de ordenar dados com base nos valores específicos que uma variável categórica assume e no número de vezes que é repita cada valor. O objetivo de uma tabela de frequência é organizar as informações e, sobretudo, tudo, resumir. A Tabela abaixo mostra as frequências obtidas pela classificação de uma amostra de n = 16 sujeitos na variável X = estressados.

Considere três variáveis categóricas: X_1 = 'estressado', X_2 = 'ex-estressado' e X_3 = 'não estressado'(o subscrito i refere-se a cada um dos diferentes valores que a variável assume; portanto i = 1, 2, 3). A primeira coluna da tabela coleta os três valores da variável. A segunda coluna mostra as frequências absolutas (n_i), ou seja, o

número de vezes que cada valor é repetido. A terceira coluna contém as frequências relativas (Pi), que são obtidas dividindo-se as frequências absolutas correspondentes entre o número total de casos:

$$P_i = \frac{n_i}{n} \tag{3.1}$$

Essas frequências indicam a proporção de vezes que cada valor é repetido (também são chamadas de proporções). As frequências relativas (proporções) são amplamente utilizadas para fazer comparações entre grupos. Multiplicando por 100 as frequências relativas definidas, obtêm-se as frequências percentuais (% i) que aparecem na última coluna da tabela: $i\%100P_i$.

Tabela 3.1: Frequências dos estudos

Drogas	Freq. Abs. n_i	Freq. Rel. P_i	Freq. Perc. % i	Freq. Abs. Acum. na_i	Freq. Rel. Acum Pa_i
Estimulantes	5	5/500 = 0,01	1	5	5/500 = 0,01
Depressores	20	20/500 =0,04	4	25	30/500 = 0,06
Opiáceos	270	0,54	54	295	0,59
Psicodélicas	115	0,23	23	410	0,82
Sem drogas	90	0,18	18	500	1,00
Total	500	1			

A frequência absoluta acumulada (nai) coleta o número de vezes que um valor é repetido mais qualquer outro inferior a ele. A frequência relativa acumulada (Pai) é obtida pela divisão da frequência absoluta acumulada pelo número total de casos ($Pa_i = na_i/n$). E a frequência percentual acumulada (%ai) obtida multiplicando por 100 a frequência relativa acumulado ($\%a_i = 100Pa_i$). As frequências absolutas (n_i) formam o ponto de referência de uma tabela de frequências: todas as outras frequências são calculadas a partir de frequências absolutas. Portanto, embora um tabela específica pode

incluir um tipo ou outro de frequências dependendo da informação que curiosamente, é recomendado que as frequências absolutas (incluindo o total n) sempre estão presentes.

Usando o menu Explore, é possível encontrar as seguintes estatísticas descriptivas:

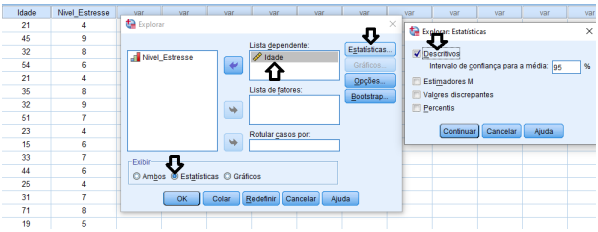


Figura 3.3: Estatísticas descritivas usando o Explore

Temos o seguinte resultado no SPSS:

Explorar

Resumo de processamento do caso

	Casos					
	Válido		Ausente		Total	
	N	Porcentagem	N	Porcentagem	N	Porcentagem
Idade	16	100,0%	0	0,0%	16	100,0%

Descritivos

			Estatística	Erro Padrão
Idade	Média		34,50	3,775
	95% Intervalo de Confiança para Média	Limite inferior	26,45	
		Limite superior	42,55	
	5% da média aparada		33,56	
	Mediana		32,00	
	Variância		228,000	
	Desvio Padrão		15,100	
	Mínimo		15	
	Máximo		71	
	Intervalo		56	
	Intervalo interquartil		23	
	Assimetria		,975	,564
	Curtose		,716	1,091

Figura 3.4: Output das estatísticas descritivas

3.2 Resumo das técnicas descritivas

Na Figura 3.5 apresentamos algumas estatísticas para dados univariados e bivariados. Os gráficos são alternativas ao uso de tabelas para descrever os dados. Podem indicar proporções, possibilitando uma análise exploratória e revelando-se úteis para apresentar dados categóricos e discretos.

Alguns gráficos são fáceis de interpretar, sem requerer, uma formação prévia, já que são intuitivos e acessíveis a todos. Outros requerem um conhecimento adicional sobre significado exato dos distintos elementos representativos. Por último, existe uma terceira categoria de gráficos, em número crescente, cuja complexidade interpretativa requer um conhecimento mais detalhadoda da técnica

utilizada. Para muitos na área de psicologia não é possível interpretar adequadamente esses gráficos sem haver compreendido suficientemente um método estatístico concreto que se relaciona aos gráficos, sendo nesses uma parte importante do mesmo. O enfoque mais importante possa ser usado pela comunidade acadêmica de saúde pública na análise gráfica de seus dados, possibilitando assim uma visualização e interpretação dos resultados analíticos com mais confiabilidade e representatividade do objeto em estudo. As interpretações é talvez o aspecto mais difícil, em alguns gráficos relacionados com técnicas avançadas já que depende sempre dos resultados que podem ser distintos em cada aplicação.

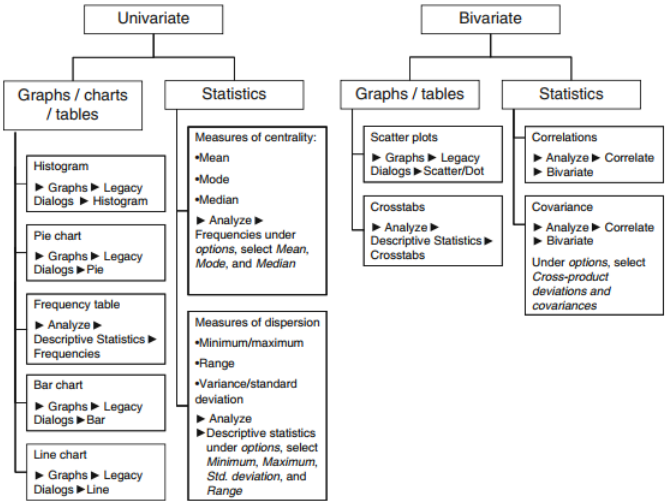


Figura 3.5: Estatísticas e gráficos dados univariados e bivariados

Os principais gráfico são:

1. Gráficos de barras. É utilizado para representar a frequência ou o percentual de dados categóricos. Diferencia-se do his-

tograma em sua apresentação por ter espaço entre as colunas. Quanto a visualização pode ser horizontal ou vertical.

2. Gráfico circular, também denominado gráfico de setores, torta ou piza. É indicado quando temos um conjunto de dados de categorias porporcionais que somam-se 100%, de forma que sua representação permite visualizar a distribuição dos dados. Em psicologia, seu uso é indicado quando se pretende representar dados nominais.

3. Histograma de frequência. Apresenta a distribuição dos valores dos dados de uma amostra, ou seja, é uma maneira de apresentar a quantidade de ocorrências de que cada valor da amostra teve. É útil para representar graficamente dados métricos contínuos.

4. O gráfico de 5 cinco medidas denominado de box plot também denominado diagrama de caixa e bigodes. Nesse diagrama indica também os valores estremo, bem como o modo como esses valores estão distribuídos.

5. Gráfico bidimensional. Pode ser representado também pelo cruzamento entre uma variável quantitativa e uma variável qualitativa. Tamém os gráficos de dispersão entre duas variáveis quantitativas com objetivo de verificar a correlação entre as duas variáveis quanto a intensidade e sentido da correlação.

6. Gráfico tridimensional. Esse tipo de gráfico, também denominado histograma bivariada. Objetos tridimensionais, portanto, são aqueles em que é possível medir comprimento, largura e altura (ou profundidade). São exemplos desses objetos

os cubos. O espaço tridimensional aplicado na aplicada na psicologia, por exemplo, um dimensão é o gênero, outra dimensão é a estatura e a terceira dimensão, o nível de estresse.

Capítulo 4

Medidas de tendência central ou posição

4.1 Medidas de tendência central ou posição

Quando se deseja representar os dados de uma distribuição de uma forma mais simples, por meio de um valor único, a melhor opção é a escolha de uma medida de tendência central. Essas medidas, que representam os parâmetros ou estimativas em torno dos quais ocorre a maior concentração dos valores observados no estudo, têm por objetivo mostrar o ponto central de equilíbrio de uma distribuição de dados.

Vimos até agora a sintetização dos dados sob a forma de tabelas, gráficos e distribuições de frequências usando o SPSS. Agora, vamos aprender o cálculo de medidas psicológicas que possibilitem representar um conjunto de dados relativos à observação de determinado fenômeno de forma resumida.

As medidas de tendência central são também chamadas de me-

didas de posição, e estabelecem o valor em torno do qual os dados se distribuem. Vale a pena chamar a atenção que, para o cálculo dessas medidas, é necessário que a variável seja quantitativa.

A média aritmética, ou simplesmente média, é a medida de tendência central mais comumente utilizada em cálculos que envolvam análise descritivas para comparações e inferências estatísticas entre amostras e populações. De cálculo simples e fácil, a média corresponde a um valor único que representa o ponto de equilíbrio entre todos os valores de uma série de dados numéricos coletados a partir de uma variável continua, além de apresentar propriedades matemáticas que permitem o desenvolvimento de cálculos estatísticos avançados. A média de uma população ou amostra é a soma de todos os elementos da população (amostra) dividida pelo número de elementos. Esta medida apresenta a mesma unidade dos dados.

Definição 4.1.1. *Média Aritmética Simples: É dada pelo quociente entre a soma dos valores observados e a Frequências total(o número total de observações).*

Sejam $x_1, x_2, x_3, \cdots, x_n$, portanto os n valores da variável X será representada por $\bar{x}$, O cálculo é definido pela seguinte forma:

$$\bar{X} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Sendo x_i valor genérico da observação e n o tamanho da amostra.

Exemplo 4.1.1. *Considere os 4 tipos de atenção e a quantidade de jovens os apresentam em 8 turmas:*

- *Atenção focada: uma resposta de curto prazo, que pode chegar a oito segundos, para estímulos auditivos, táteis ou visu-*

ais muito específicos. Quantidade de jovens em oito turmas: 2,3,5,4,5,6,7,5

- *Atenção prolongada: uma grau de atenção que produz resultados consistentes que envolvem uma tarefa contínua e repetitiva realiazada ao longo tempo. Quantidade de jovens em oito turmas: 3,3,8,4,5,6,10,8*

- *Atenção dividida: prestar atenção a várias coisas ao mesmo tempo. Essa é a capacidade limitada e afeta a quantidade de informação que é processada. Quantidade de jovens em oito turmas: 6,3,8,4,4,6,10,9*

- *Atenção seletiva: prestar atenção a coisas específicas, enquanto filtra as outras. Quantidade de jovens em oito turmas: 2,6,8,4,1,6,10,5*

Qual tipo de atenção tem a menor média? A menor média tem na atenção focada com valor 4.625.

Definição 4.1.2. *A média aparada (tri-média) é obtida eliminando do conjunto de dados as T observações menores e T observações maiores. O valor de V corresponde a uma percentagem entre 5% e 25% do número total de observações. Esta eliminação dos valores extremos é para eliminar o efeito de observações discrepantes, conhecidas como outliers, no cálculo da média aritmética.*

Considere os seguintes números de estressados em 25 clínicas psicológicas:

1	3	5	5	6
6	6	7	7	8
8	8	8	8	8
9	9	9	10	10
12	12	12	12	70

Tabela 4.1: Quantidades de estressados

Retirando os valores extremos 1 e 70, a tri-média será:

$$\bar{x} = \frac{3+5+\cdots+12+12}{23} = 8,17$$

Já a média aritimética de todos os valores seria 10,36, porém a média aparada será 8,17, pois eliminou dos dados o primeiro valor e o último.

Definição 4.1.3. *Média Aritmética Ponderada: É a média aritmética calculada quando os dados estiverem agrupados em distribuições de Frequências. Os valores $x_1, x_2, x_3, \cdots, x_n$ serão ponderados pelas respectivas frequências absolutas ou pesos $p_1, p_2, p_3, \cdots, p_n$.*

Sabendo que $\sum p_i = n$. Temos s seguinte média aritmética ponderada:

$$\bar{x} = \frac{x_1 p_1 + x_2 p_2 + x_3 p_3 + \cdots + x_n p_n}{\sum p_i} = \frac{\sum_{i=1}^{n} x_i p_i}{\sum p_i}$$

Na Tabela abaixo mostra cada uma das notas parciais obtidas por um candidato classificado em um concurso público na área de psicologia, com suas respectivas ponderações. Qual a média final do candidato?

Avaliação	Notas	Pesos
Escrita	8,5	5
Didática	9,1	4
Prática	8,8	3
Curricular	7,4	2
Escrita	6,0	1
Total	39,8	15

Tabela 4.2: Notas parciais do candidato e suas respectivas ponderações

$$\bar{x} = \frac{8,5.5 + 9,1.4 + \cdots + 7,421 + 6,0.1}{15} = \frac{126,1}{15} = 8,41.$$

Assim, a média final do candidato é igual a 8,41.

Exercício 4.1.1. *Calcular a média ponderada das notas de 1 aluno, que fez uma prova que tem peso 5, um trabalho que tem peso 3 e uma lista de exercícios que tem peso 2. O aluno conseguiu 8,5 na prova, 9,0 no trabalho e 6,0 na lista de exercício.*

Exercício 4.1.2. *Você está assistindo a um curso de psicologia no qual sua nota é determinada a partir de cinco fontes: 50% da média de seus testes, 15% de seu exame no meio do curso, 20% de seu exame final, 10% de seu trabalho no laboratório de psicologia e 5% do trabalho feito em casa. As suas notas são 86 (média dos testes), 96 (exame no meio do curso), 82 (exame final), 98 (laboratório de psicologia) e 100 (trabalho de casa). Qual é a média ponderada de suas notas?*

4.1.1 Média geométrica

Utiliza-se em psicologia com dados relativos como porcentagens ou taxa de incremento e se calcula por meio da seguinte forma:

$$M_G = \sqrt[n]{x_1 x_2 \cdots x_n}$$

Exemplo 4.1.2. *Um jovem, durante três meses, teve a seguinte variação de grau de ansiedade, considere que o grau varia de 0 a 100%. O primeiro mês foi de 20%, o segundo de 10% e o terceiro de 30%. Um psicólogo deseja conhecer o aumento médio percentual ao final desse período.*

Para facilitar os cálculos considere que inicialmente foi 100%, logo no primeiro mês foi para 120%, que, na sua forma decimal, escreve-se 1,2. Esse raciocínio será o mesmo para os três aumentos, então queremos a média geométrica entre: 1,2; 1,1; e 1,30.

$$M_G = \sqrt[n]{x_1 x_2 \cdots x_n} = \sqrt[3]{1.2 * 1.1 * 1.30} = \sqrt[3]{1.716} = 1.19721$$

Assim O aumento do grau de estresse foi de 19,72% por mês em média.

4.1.2 Média harmônica

A média harmônica é utilizada em casos em que é necessário ter a média de variação em relação ao tempo. É calculada da seguinte forma:

$$M_H = \frac{1}{\frac{1}{n}\left(\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}\right)}$$

A média harmônica é sempre menor ou igual a média geométrica e, por tanto, também inferior ou igual a média aritmética. Sendo n a quantidade de elementos.

Exemplo 4.1.3. *Uma jovem ansioso realiza um percurso pra universidade duas vezes. Na ida, ele faz o percurso com uma velocidade $v_1 = 90km/h$ e na volta, ele realiza o mesmo percurso com velocidade de $v_2 = 110km/h$. Qual foi a velocidade média ao juntar-se ida e volta?*

Quando se usa a média harmônica? Note que a distância é a mesma, para a ida e para a volta desse jovem ansioso que vai a universidade, o que muda é a velocidade causada pela ansiedade do jovem, consequentemente, está relacionado ao tempo. Se aumentar a velocidade, é óbvio que o tempo que se leva para percorrer uma mesma distância diminuirá, logo, essas grandezas são inversamente proporcionais. Quando se usa grandezas inversamente proporcionais, pode-se usar a média harmônica.

$$M_H = \frac{1}{\frac{1}{n}\left(\frac{1}{x_1} + \frac{1}{x_2}\right)} = \frac{1}{\frac{1}{2}\left(\frac{1}{90} + \frac{1}{110}\right)} = 99km/h$$

4.2 Média de variáveis discretas (sem intervalo de classe) de uma distribuição de frequência

A média é calculada usando a fórmula:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i f_i}{n}$$

Sendo $n = \sum_{i=1}^{n} f_i$.

Exercício 4.2.1. *Calcule a média da variável discreta (sem intervalo de classe) de uma distribuição de frequência.Considerando os números de crianças nascidas dos funcionários e professores do curso de psicologia no ano 2017.*

Número de crianças	f_i
0	2
1	8
2	10
3	12
4	4
Total	36

Tabela 4.3: Uma frequência do números de crianças

Exercício 4.2.2. *Na Tabela abaixo temos as frequências relativas do número de idosos com transtorno de ansiedade por Covid 19 em 5 grandes municípios. Calcular a mediana.*

Municípios	Número de idosos
A	2913
B	4500
C	4826
D	4928
E	5000

Tabela 4.4: Frequência do números de idosos

Exercício 4.2.3. *Na Tabela abaixo temos as frequências relativas do número de jovens que tem inteligencia relativa (é a capacidade*

de distinguir e tornar visível o invisível, permitindo-nos escolher) em 6 grandes municípios. Calcular a média.

Municípios	Número de idosos
1	2980
2	2800
3	2789
4	2890
5	2900
6	2800

Tabela 4.5: Frequência do números de crianças em seis municípios

4.3 Média de variável contínua (com intervalos de classe)

Quando os dados estiverem agrupados numa distribuição de Frequências, usaremos a média aritmética dos valores $x_1, x_2, x_3, \cdots, x_N$ ponderados pelas respectivas frequências absolutas: $f_1, f_2, f_3, \cdots, f_n$ vezes respectivamente, a média aritmética será:

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_i f_i}{f_1 + f_2 + \cdots + f_i} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i}$$

Sendo $n = \sum_{i=1}^{n} f_i$.

Exercício 4.3.1. *Determinar a média da distribuição:*

Rendas	Número de famílias
$2 \vdash 4$	5
$4 \vdash 6$	12
$6 \vdash 8$	13
$8 \vdash 10$	6
$10 \vdash 12$	4

Tabela 4.6: Valores fictícios

Sendo que $a \vdash b$ significa: fechado em a e aberto em b. Sendo números inteiros, o intervalo $2 \vdash 4$ tem os número 2 e 3, porém o número 4 não está presente nesse intervalo, pois está aberto.

4.4 Mediana

A mediana de um conjunto de dados corresponde ao valor que, no conjunto de dados, separa-o em dois subconjuntos de mesmo número de elementos, quando estes estão ordenados segundo uma ordem de grandeza. É, portanto, o valor que ocupa a posição central quando todos os valores observados estão dispostos em ordem crescente ou decrescente de magnitude.

Colocando os valores em ordem crescente, a mediana é o valor que divide a amostra, ou população, em duas partes iguais. Assim:

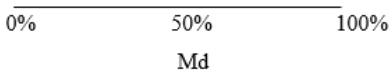


Figura 4.1: Mediana

Variável discreta (dados não agrupados, porém, ordenados):

- Se n for ímpar, a mediana será o elemento central (de ordem $\frac{n+1}{2}$)
- Se n for par, a mediana será: $X_{\frac{n}{2}} + X_{\frac{n}{2}+1}$.

A mediana é frequentemente utilizada em séries de dados que tem uma distribuição muito assimétrica, já que em este caso nem sempre em dados psicológicos é adequado utilizar a média aritmética.

Exemplo 4.4.1. *Uma variável psicológica X toma os seguintes 7 valores distintos: 1,2,5,6,7,8 e 12. Determinar a mediana.*

Pode-se observar que o valor da variável $x_i = 6$ deixa o mesmo número de observações, um total de 3 para cada lado. Assim, o valor da mediana é: $M_d = x_i = 6$.

Exemplo 4.4.2. *Uma variável psicológica X toma os seguintes 6 valores distintos: 9,2,5,3,6 e 8. Determinar a mediana.*

Neste caso, o primeiro que se deve fazer é organizar os dados em ordem crescente, ou seja, 2,3,5,6,8 e 9. O valor da variável que deixa o mesmo número de observações a ambos os lados, a mediana, situa-se entre 5 e 6. Assim, $M_d = \frac{5+6}{2} = 5,5$.

4.5 Cálculo da mediana – variável continua ou dados agrupados

1. Calcula-se a ordem $n/2$. Senho variável continua não se preocupe se n é par ou ímpar
2. Pela f_{ac} identifica-se a classe que contém a mediana (classe md)

3. Utiliza-se a fórmula:

$$Md = l_{Md} + \frac{(\frac{n}{2} - \sum f_{an})}{f_{Md}}.A$$

Sendo: l_{Md} o limite inferior da classe Md; n o tamanho da amostra ou população; A é a amplitude e f_{Md}, frequência da classe da mediana e $\sum f_{an}$ a soma das frequências da classe da mediana.

Para determinar o valor no intervalo mediano considere a seguinte figura:

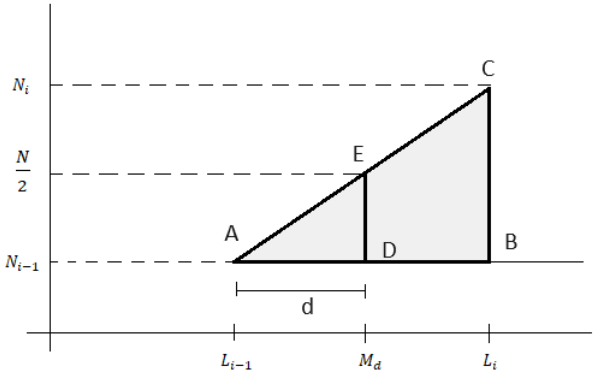


Figura 4.2: Determinação da mediana para dados agrupados em intervalos

O valor médio será: $M_d = L_{i-1} - d$. Determina-se d usando uma propriedade de semelhança dos triângulos: $\triangle ABC$ e $\triangle ADE$, ou seja,

$$\frac{DE}{AD} = \frac{CB}{AB}$$

$$\frac{\frac{N}{2} - N_{i-1}}{d} = \frac{N_i - N_{i-1}}{L_i - L_{i-1}}$$

Substituindo o valor de $d = L_{i-1} - M_d$ na equação 4.5 tem-se:

$$\frac{\frac{N}{2} - N_{i-1}}{M_d - L_{i-1}} = \frac{N_i - N_{i-1}}{L_i - L_{i-1}}$$

Fazendo $N_{i-1} = \sum f_{an}$, $N = n$, $n_i = F_{Md}$ e $L_{i-1} = l_{Md}$ temos:

$$M_d = l_{Md} + \frac{\left(\frac{n}{2} - \sum f_{an}\right)}{f_{Md}}.A_i$$

Sendo A_i a amplitude igual a $L_i - L_{i-1}$.

Exemplo 4.5.1. *Dado o seguinte intervalo de estresse. Calcule o valor da Mediana (Md).*

Nível de Estresse	Frequência	Frequência Acumulada
$1500 \vdash 2000$	10	10
$2000 \vdash 3000$	40	50
$\rightarrow 3000 \vdash 4500$	62	112
$4500 \vdash 7000$	24	136
$7000 \vdash 10000$	4	140

Tabela 4.7: Distribuição de frequência do nível de estresse

Sendo $A = 4500 - 3000$ (amplitude da classe da mediana).

$$M_d = 3000 + \frac{\left(\frac{140}{2} - 50\right)}{62}.1500 = 3483,87$$

4.6 Distribuição de frequência não unitárias

Quando a distribuição de frequência não é unitária, utiliza-se o seguinte critério para determinar o valor da mediana: seja N_i a primeira frequência absoluta acumulada igual ou superior a N/2, então:

- Se $N_{i-1} < \frac{N}{2} < N_i \rightarrow M_d = x_i$
- Se $N_i = \frac{N}{2} \rightarrow M_d = \frac{x_i + x_{i+1}}{2}$

Exemplo 4.6.1. *Obter a mediana da seguinte distribuição de frequência*

Nível de Estresse	Frequência	Frequência Acumulada
2	3	3
3	2	5
5	3	8

Tabela 4.8: Distribuição de frequência do nível de estresse

A metade das observações corresponde a $N/2 = 4$. O valor da variável que contém uma frequência acumulada de 4 é $x_2 = 3$, com $N_2 = 5$.

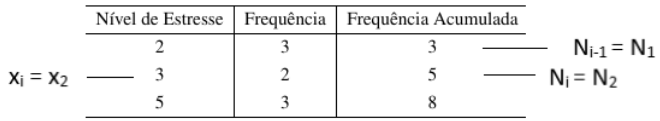


Figura 4.3: Passos para calcular a mediana

Portanto, como $N_1 < \frac{N}{2} < N_2 \rightarrow 3 < 4 < 5$, então a mediana será: $M_d = x_2 \rightarrow M_d = 3$.

4.7 Cálculo da moda– variável continua ou dados agrupados

Denota-se por a e b as distâncias respectivas dos intervalos anterior e posterior da classe modal como se observa na figura abaixo.

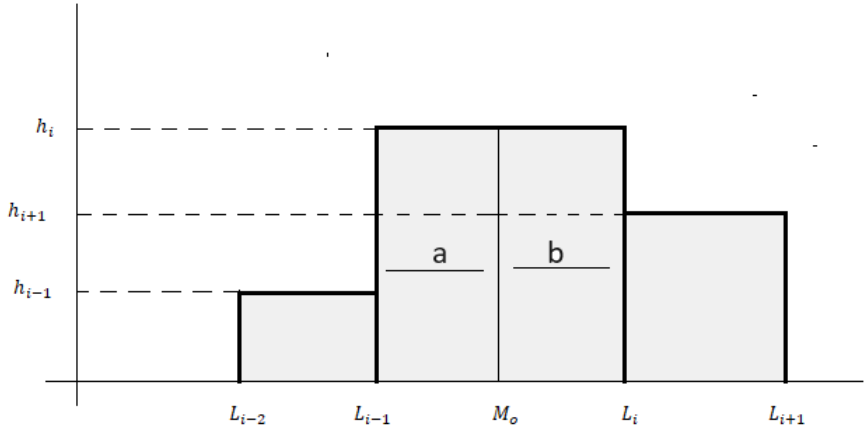


Figura 4.4: Determinação da moda para dados agrupados em intervalos

Observa-se que $ah_{i-1} = bh_{i+1}$. Assim, temos:

$$\frac{a}{h_{i+1}} + \frac{b}{h_{i-1}} = \frac{a+b}{h_{i+1}+h_{i-1}}$$

$$a = \frac{h_{i+1}(a+b)}{h_{i-1}+h_{i+1}} = \left(\frac{h_{i-1}}{h_{i-1}+h_{i+1}}\right) a_i$$

Como $M_o = L_{i-1} + a$, tem-se:

$$M_o = L_{i-1} + \left(\frac{h_{i+1}}{h_{i-1}+h_{i+1}}\right) a_i$$

Exemplo 4.7.1. *Considere a seguinte distribuição de frequência encontra a Moda (M_o).*

Níveis de ansiedade	Frequência
$0,9 \vdash 1,3$	20
$1,3 \vdash 1,6$	40
$1,6 \vdash 2,2$	25
$2,2 \vdash 3,0$	7,5
$3,0 \vdash 5,0$	2,5

Tabela 4.9: Distribuição de frequência de níveis de ansiedade

$$M_o = 1,3 + \left(\frac{25}{20+25}\right) 0,3 = 1,467$$

Exemplo 4.7.2. *Dado o seguinte intervalo de estresse. Calcule o valor da Moda (Mo).*

Nível de Estresse	Frequência	Frequência Acumulada
$2 \vdash 4$	5	
$4 \vdash 6$	12	
$6 \vdash 8$	13	
$8 \vdash 10$	6	
$10 \vdash 12$	4	

Tabela 4.10: Distribuição de frequência de níveis de estresse

4.8 Percentis e outros fractis

Defina-se quantis de ordem k como os valores da variável, ordenada de menor ao maior que divide em k partes com a mesma frequência de observações. O primeiro quantil de ordem k deixa a

sua esquerda a fração $1/k$ de frequência de observações. O r-ésimo quantil de ordem k deixa a sua esquerda a fração r/k de frequência de observações. Por exemplo, o quantil 15 de ordem 100 deixa por debaixo a 15% dos valores do total da série completa de valores. As quantidades mais utilizadas são os percentis, quartis e decis, o quais os quais se descrevem da seguinte forma:

- Os percentis são os 99 pontos que dividem a distribuição em 100 partes, tais que dentro de cada uma está incluido a 1% dos valores da distribuição.
- Os quartis são os três valores da variável que dividem a distribuição em 4 partes iguais, em 4 intervalos, dentro de cada qual está incluido 25% dos valores da distribuição. O percentil 25 P_{25} seria igual a Q_1, o percentil P_{50} seria igual a $P_{1/2}$ (igual a mediana) etc.
- Os decis são 9 pontos que dividem a distribuição em 10 partes iguais, tais que dentro de cada uma 10% dos valores da distribuição. O percentil P_{10} seria igual a D_1, o percentil 20 P_{20} seria igual ao decil 2 (D_2) etc.

Além de usar quartis para especificar uma medida de posição, podemos também usar os percentis e os decis. Esses fractis comuns são resumidos a seguir.

Quartis	Divide em 4 partes iguais:	Q_1, Q_2 e Q_3
Decis	Divide em 10 partes iguais:	$D_1, D_2, \cdots, D_{10}$
Percentis	Divide em 100 partes iguais:	$P_1, P_2, \cdots, P_{100}$

Tabela 4.11: Quartis, decis e percentis

Os percentis são geralmente usados nas áreas relacionadas a psicologia para indicar como um invidíduo se compara a outros em um conjunto. Eles tmabém podem ser usados para identificar valores excepcionalmente altos ou baixos. Por exemplo, as notas de uma avaliação psicológica e as medidas de crescimento de crianças são normalmente expressas em percentis. Por exemplo, as notas ou medidas no 95% percentil ou acima são excepcionalmente altas, enquanto aquelas no quinto percentil ou abaixo são excepcionalmente baixas.

Observe que existem três valores da variável que dividem as observações em quatro partes iguais.

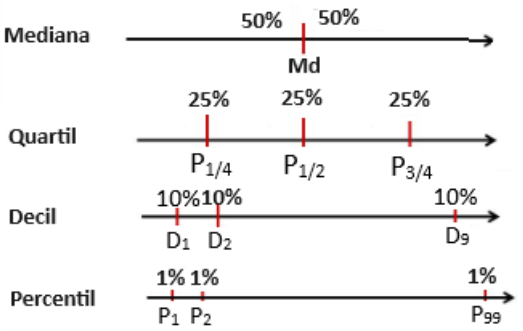


Figura 4.5: Medidas de quartis, decis e percentis numa reta

Exemplo 4.8.1. *Encontre os quartis da seguinte variável (x_i):*

Num. de ansiosos (n_i)	Freq. (f_i)	Freq. Acumulada (f_{ac})
1	3	3
5	2	5
6	5	10
8	3	13
10	3	13
11	8	24
12	11	35
14	7	42

Tabela 4.12: Distribuição de frequência do número de ansiosos

- O primeiro quartil será:

$$\frac{N}{4} = \frac{42}{4} = 10,5 \rightarrow 10 < 10,5 < 13$$

  Assim, tem-se que $P_{1/4} = 8$

- O segundo quartil (mediana) será:

$$\frac{N}{2} = \frac{42}{2} = 21 \rightarrow 16 < 21 < 24$$

  Portanto, representando o diagrama das frequeências acumuladas tem-se: $P_{1/2} = 11$

- Terceiro quartil

$$\frac{3N}{4} = \frac{3.42}{4} = 21 \rightarrow 31,5 < 24 < 35$$

  Portanto, representando o diagrama das frequeências acumuladas tem-se: $P_{3/4} = 12$

4.9 Representação de dados

Os investigadores em psicologia precisa da estatística para obter conclusões válidas a partir dos dados. O método estatístico se tem convertido em parte essencial para geração do conhecimento científico, e em praticamente todas as publicações especializadas em quase qualquer campo de conhecimento as técnicas estatísticas tem um papel muito importante.

Frequente problema em psicologia surge na hora de mostra graficamente os resultados obtidos. Às vezes, é necessário mais de dois eixos de coordenadas. Um plano bidimensional o máximo que se pode representar são três eixos. As coordenadas polares permitem representar num gráfico bidimensional qualquer número de eixo de coordenadas.

Um passo prévio para o cálculo das coordenadas polares é indicar ângulos das variáveis consideradas. A indicação do ângulos para as variáveis é diferente se existe valores negativos ou não, se os valores forem padronizados numa escala de 0 a +1, não existirão valores negativos.

4.9.1 Representação dos dados em coordenadas polares

As coordenadas polares de cada elemento da amostra ou população se calcula determinando a resultante para cada um dos eixos ou variáveis, ou seja, mudando sucessivamente o ponto na direção de cada eixo numa distância igual ao valor da variável correspondente: o ângulo resultante e a distância a origem são as coordenadas polares. Em geral, é mais fácil calcular e representar as correspondentes coordenadas retangulares mediante a seguinte transformação:

$$X = \sum_{i=1}^{n} |z_i| \cos(\alpha)$$

$$Y = \sum_{i=1}^{n} |z_i| \sin(\alpha)$$

Em que X e Y são coordenadas retangulares para cada no gráfico polar, z é o valor desse caso para a variável i, α é o ângulo em graus indicado para a variável i e n é o número da variável. O valor de n sempre seria o dobro do número de variáveis que realmente existem.

É necessário passar de ângulo a radianos e para isso utiliza as seguintes variáveis:

$$X = \sum_{i=1}^{n} |z_i| \sin(\alpha \frac{\pi}{180})$$

$$Y = \sum_{i=1}^{n} |z_i| \sin(\alpha \frac{\pi}{180})$$

As coordenadas polares nos permite ver de forma gráfica as lacunas que tem uma composição mais semelhantes para os parâmetros que se tem considerado na variável psicológica depressão.

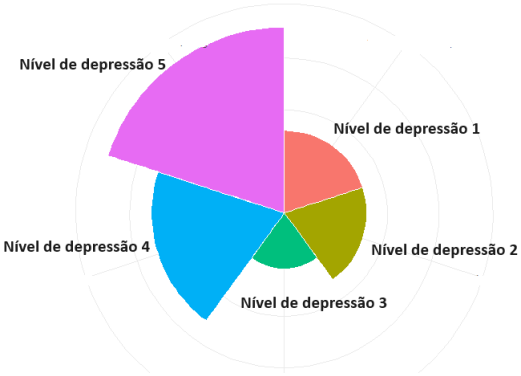


Figura 4.6: Gráfico de coordenadas polares representando 5 níveis de depressão

4.9.2 Comentários gerais

Os parâmetros de posição mais frequentes em psicologia são a moda, mediana e média. Em um número particular, eles frequentemente tomar valores diferentes, mas em certos casos (amostras simétricas como 4,4,8,8,8,8,9,9) podem coincidir. A moda se utiliza poucas vezes na literatura científica e quase nunca forma parte da inferência estatística. A média é a medida de posição mais frequente em estatística, radicando sua principal ventagem a facilidade no tratamento teórico. Entretanto, como medida descritiva, apresenta a desvantagem de ficar fortemente afetada pelos valores extremos da amostra, coisa que não existe com a mediana. Considere o seguinte dados:

1,2,4,5	$\bar{x} = 3,25$	$\tilde{x} = 3,5$
1,3,4,12	$\bar{x} = 5,00$	$\tilde{x} = 3,5$

Observa-se que a mudança de 5 pra 12 faz mudar drasticamente a média, mas a mediana fica inalterada.

4.10 Exercícios

1. Dado os seguintes níveis de estresse (variando de 0 a 10) de 12 jovens: 6,4,8,4,5,6,5,4,3,2,1,4. Encontre:

   (a) A média, a mediana, a moda dos níveis de estresse dos 12 jovens.

2. A tabela a seguir contém o Peso (em kg), Altura (em cm.), Idade e Gênero (homem, mulher) de trinta e seis indivíduos (Ind.),que tem inteligência intelectualmente dotada (habilidades em um ou mais domínios intelectuais)

Ind.	P	A	I	G	Ind.	P	A	I	G
1	75	173	21	H	19	55	160	22	M
2	81	178	22	H	20	72	174	21	H
3	56	162	22	M	21	56	161	23	M
4	68	180	21	M	22	84	182	22	H
5	79	182	24	H	23	61	163	24	M
6	89	185	22	H	24	76	172	22	H
7	62	157	21	M	25	68	169	21	M
8	59	165	22	M	26	58	162	24	H
9	83	180	23	H	27	78	178	22	H
10	62	157	21	M	28	77	165	21	H
11	71	182	22	M	29	55	163	24	M
12	78	174	25	H	30	77	181	23	H
13	81	181	22	H	31	75	174	23	H
14	60	167	21	M	32	68	178	21	H
15	89	185	22	H	33	82	190	21	H
16	59	155	22	H	34	63	173	21	M
17	75	177	24	H	35	62	174	21	M
18	81	183	22	H	36	68	182	22	M

   Com a ajuda do SPSS:

(a) Calcular as medidas descritivas de posição associadas à distribuição de frequência de cada variável

(b) Calcular as medidas descritivas da variável peso de acordo com o código da variável gênero

(c) Calcular as distribuições de frequência associadas à variável peso, altura e idade

3. A faixa etária de pessoas que procuram apoio psicológico em uma clínica. Encontre as porcentagens de cada faixa etária.

Faixa etária	Frequência	Porcentagens
≤ 25	242	
$25 \vdash 34$	627	
$35 \vdash 44$	679	
$45 \vdash 54$	481	
$55 \vdash 64$	320	
> 65	479	
Total	2828	100%

4. A quantidade de pessoa que fazem terapias comportamentais condiconamento aversivo em cada uma das 20 cidade de um estado do sudeste foram:

30,21,24,35,46,57,56,67,76,79,45,67,54,45 67
78,67,72,73,14,81,15,11,15,89,76,67,77,78,17

Suponha que se precisa padronizar os dados. Pede-se usando o SPSS:

(a) Encontrar o valor do coeficiente de variação de Pearson dos dados padronizados com média 15 e desvio padrão 2.

(b) Calcular e interpretar o 58^o percential

(c) Calcular e interpretar o 30^o decil

(d) Calcular e interpretar o 70^o decil

(e) Qual o valor da curtose dos dados brutos?

(f) Representar graficamente os dados brutos com a construção de um gráfico de barras (colunas) e um de pizza.

(g) Construir o box plot dos dados brutos

5. A quantidade de jovens que tem inteligência naturalista em 10 escolas são: 5, 4, 4, 5, 5, 7, 3, 7, 5, 7. Pede-se:

   (a) Faça uma tabela de distribuição frequência para essa quantidade

   (b) Encontre as estatísticas de posição unidimensionais comuns.

6. Sejam as notas de uma determinada prova em psicologia social: 3, 4, 8, 5, 5, 7, 3, 9, 5, 7. Pede-se:

   (a) Faça uma tabela de distribuição frequência para cada nota

   (b) Encontre as estatísticas de posição unidimensionais comuns.

7. As 15 informações a seguir representam as notas da disciplina psicologia da saúde de uma universidade: 3,4 3,5 4,5 5,0 5,5 5,5 4,6 4,5 4,5 5,5 6,0 6,6 3,5 3,5 8

(a) Insira os dados em uma variável chamada NOTAS no SPSS

(b) Faça um resumo estatístico de medidas de posição

(c) Gráfico de caixa (box plot)

(d) Recodifique a variável em 3 categorias como segue:

- De 3 a 5 com o valor 1 (falhou)
- De 5 a 7 com o valor 2 (aprovado)
- De 7 a 9 com o valor 3 (notável)

(e) A nova variável com valores 1, 2 e 3 (visto acima) denomine de QUALIFICAÇÃO

(f) Construa o gráfico de barras e pizza para variável QUALIFICAÇÃO

(g) Ordenar as informações contidas nas colunas NOTAS e QUALIFICAÇÃO

8. Os seguintes dados são o número de dias que tem faltado ao trabalho os alunos na pandemia de Covid 19:

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

Tabela 4.13: Distribuição de frequências do número de faltas

Pede-se:

(a) Distribuição de frequência.

(b) Um diagrama de barras.

9. Uma amostra de 654 representa a população que tiveram diagnóstico de transtorno depressivo maior (TDM). Perguntaram se havia sofrido algma das 10 ações que apresentam maus tratos psicológico em três meses anteriores a entrevista; o mesmo se fez em relação das 10 ações que representam maus tratos físicos. Os dados estão na tabela abaixo. Comparar os dois tipos de maus tratos com base nas medidas: media, mediana e percentis 10,25,75 e 90.

Tabela 4.14: População que tiveram diagnóstico de transtorno depressivo maior

Quant.	0	1	2	3	4	5	6	7	8	9	10
Psicológicos	397	104	78	30	10	10	12	5	3	1	4
Físicos	518	76	31	10	11	3	2	0	0	1	2

10. Seque a quantidade de pessoas que tiveram crise de ansiedade no período da Covid 19 em 30 cidades num estado brasileiro.

    175,165,167,172,176,178,179,185,169,178,179,180,175,173,180
    178,167,172,173,174,181,165,181,165,189,176,167,177,178,167

    Pede-se:

    (a) Distribuição de frequência agrupando em intervalos de amplitude 5

    (b) Qual a média de pessoas que tiveram crise?

    (c) Fazer um histograma

11. Completar a seguinte tabela de distribuição de frequências.

I_i	f_i	Fac_i
$12 \vdash 15$		45
$15 \vdash 21$		87
$21 \vdash 32$		158
$32 \vdash 36$		101
$36 \vdash 40$		29
$40 \vdash 48$		11
$48 \vdash 53$		10

12. Completar a seguinte tabela de distribuição de frequências. Sendo n_i a frequência, h_i amplitude, f_i percentual da frequência relativa e $F_{acum.}$ percentual da frequência relativa acumulada.

	n_i	f_i	$F_{acum.}$	x_i	h_i
$0 \vdash 50$	2			(0+50)/2 =25	50
$50 \vdash$	8		0.04		
$100 \vdash$			0.12		
		0.13		175	
$200 \vdash$	119				50
$\vdash 300$		0.11			
$300 \vdash$	171				
$\vdash 400$	188				
$\vdash 450$	200			425	

13. Realizou-se uma investigação que proporcionou os seguintes resultados:

x_i	0	1	2	3	4	5	6
n_i	1	12	22	34	26	14	11

Pede-se:

(a) Tabela da distribuição de frequências

(b) Porcentagem de menores ou iguais a 3

(c) Porcentagem de maiores que 5

14. Em um estudo realizado sobre o hábito de fumar por ansiedade, entrevistou 100 pessoa. A variável X mede o número de cigarros consumido diariamente:

Número cigarros (X)	Número pessoas
$1 \vdash 4$	5
$4 \vdash 6$	8
$6 \vdash 10$	15
$10 \vdash 15$	35
$15 \vdash 20$	24
$20 \vdash 40$	11
$40 \vdash 80$	2

Pede-se:

(a) Frequências relativas e cumulativas

(b) Qual o número de classe e largura das classes(amplitude)?

(c) Histograma de frequências e curva cumulativa

15. Considere a seguinte distribuição de frequência:

Y	Frequência
$0 \vdash 100$	13
$100 \vdash 200$	15
$200 \vdash 300$	20
$300 \vdash 400$	8
$400 \vdash 500$	4

Pede-se:

(a) Frequências absolutas e cumulativas

(b) Frequências relativas e cumulativas

(c) A classe modal

(d) Qual a porcentagem dos Y...
...entre 100-200?
...menos de 300?
...menos de 280?
...acima de 220?

16. A tabela a seguir classifica um grupo de pessoas de acordo com a frequência com que lêem jornais e assistem televisão

Período	TV	Jornais
Todos os dias	20	70
Algumas vezes	60	70
Nunca	30	30

Tabela 4.15: Distribuição de frequências

Pede-se:

(a) Insira os dados e obtenha a tabela que mostra a distribuição conjunta em percentuais

17. Numa pesquisa de psicologia perguntou-se aos indivíduos se no último mês tiveram dor de cabeça por causa do estresse, anotou-se os resultados por gênero e por classe social (4 níveis, do mais baixo ao mais alto). Represente graficamente no SPSS esses dados.

		Classe 1		Classe 2		Classe 3		Classe 4	
		H	M	H	M	H	M	H	M
Dor	Sim	197	195	129	141	40	35	8	5
Dor	Não	789	825	635	675	361	395	84	110
Totais		986	1020	764	816	401	430	92	115

Capítulo 5

Medidas de dispersão

5.1 Medidas de dispersão ou variabilidade

Estas medidas de variabilidades são utilizadas para quantificar o grau de variabilidade dos valores de uma amostra de dados em torno da sua média. Para avaliar o grau de variabilidade ou dispersão dos valores de um conjunto de números, lançaremos mão das estatísticas denominadas medidas de dispersão. Essas nos proporcionarão um conhecimento mais completo do fenômeno a ser analisado, permitindo estabelecer comparações entre fenômenos da mesma natureza e mostrando até que ponto os valores se distribuem acima ou abaixo da medida de tendência central.

A informação fornecida pelas medidas de posição ou tendência central necessita, em geral, ser complementada pelas medidas de dispersão. Estas servem para indicar o quanto os dados se apresentam dispersos em torno da região central (média mediana e a moda). Caracterizam, portanto, o grau de variação existente na serie de valores e servem para medir a representatividade das medidas

de tendência central. As medidas de dispersão que nos interessam são:

- Amplitude
- Desvio médio
- Variância
- Desvio padrão
- Coeficiente de variação
- Erro padrão da média
- Amplitude interquartil

5.1.1 Amplitude

É a diferença entre o maior e o menor dos valores da amostra de dados psicológicos. A sua utilização, além de mostrar o maior desvio, serve para uma avaliação preliminar dos dados, verificando-se a possibilidade de possíveis erros nas coletas destes ou nas digitações, já que as variáveis podem apresentar extremos conhecidos. A fórmula para encontrar a amplitude será:

$$A = Valor_{max.} - Valor_{min.}$$

Exercício 5.1.1. *Considere a seguinte amostra qualquer formada pela idade de jovens que apresenta estresse causado pela Covid 19. Idades=[17,23,21,20,19,19,21,21,20,25]. Qual o valor da amplitude?*

5.1.2 Desvio Médio

Desde que se deseja medir a dispersão os dados em relação à média, parece interessante a análise dos desvios em torno da média. Isto é, analisar o desvio médio absoluto de um conjunto de dados $x_1, x_2, \cdots, x_n$ é definido por:

$$DM = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$$

Sendo que as barras verticais representam o valor absoluto ou módulo, por exemplo, $|-3| = 3$. Note que nesta definição estamos trabalhando com o desvio médio, isto é, tomamos a média dos desvios absolutos. Isso evita interpretações equivocadas, pois, se trabalhássemos apenas com a soma dos desvios absolutos, um conjunto com um número maior de observações tenderia a apresentar um resultado maior para a soma devida apenas ao fato de ter mais observações.

Exercício 5.1.2. *Considere o número de enfermos em três clínicas de uma pequena cidade do interior da Paraíba (3, 4, 5). Encontre o desvio médio.*

Abaixo temos 4 fórmulas para encontrar desvio médio para dados discretos ou contínuos, para dados agrupados ou para dados não agrupados.

discretas ou contínuas / amostrais ou populacionais	Dados brutos ou discretos	Dados agrupados ou contínuos
Para dados amostrais	$DM = \frac{\sum \lvert x_i - \bar{x} \rvert}{n-1}$	$DM = \frac{\sum \lvert x_i - \bar{x} \rvert f_i}{n-1}$
Para dados populacionais	$DM = \frac{\sum \lvert x_i - \bar{x} \rvert}{N}$	$DM = \frac{\sum \lvert x_i - \bar{x} \rvert f_i}{N}$
Sabe-se que n é uma amostra da população N. Assim, N > n.		

Figura 5.1: Quatro fórmulas para encontrar desvio médio

5.1.3 Variância

A variância de uma variável x (S^2) mede a dispersão dos valores entorno da média. Obtém-se (S^2) pela soma de quadrados dos desvios de cada valor $x_1, x_2, \cdots, x_n$ em relação a média amostral ou populacional $\bar{x}$ e μ, respectiovamente, dividida pelo número $n-1$(amostral) ou n(populacional). Desse modo a S^2 é a média dos n - 1 desvio quadráticos e independentes.

É possível definir a variância usando o divisor n-1 no lugar de n; essa é a diferença entre os conceitos de variância amostral e variância populacional comentado acima, que será mais relevante na inferência estatística. Para dados discretos ou valores não agrupados a fórmula será:

- Variância amostral

$$S^2 = \frac{\sum_{i=1}^{n} (X-\mu)^2}{n-1} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right]$$

Exemplo 5.1.1. *Calcular a variância e o desvio padrão amostral dos seguintes das idades de 4 crianças com problemas psicológicos. As idades foram 4,6,8, e 10.*

$$\begin{aligned} S^2 &= \frac{1}{3}\left[4^2+6^2+8^2+10^2-\frac{(4+6+8+10)^2}{4}\right] \\ &= \frac{1}{3}\left[216-\frac{28^2}{4}\right] = \frac{20}{3} = 6,667 \end{aligned}$$

O desvio padrão seria então: $S = \sqrt{S^2} = \sqrt{6,667}$.

- Variância populacional

$$\sigma^2 = \frac{\sum_{i=1}^{n}(X-\mu)^2}{n} = \frac{1}{n}\left[\sum_{i=1}^{n}x_i^2 - \frac{(\sum_{i=1}^{n}x_i)^2}{n}\right]$$

Exemplo 5.1.2. *Calcular a variância e o desvio padrão populacional dos seguintes das idades de 4 crianças com problemas psicológicos. Suponha que as idades foram 4,6,8, e 10 venham de uma população (e não de uma amostra).*

$$\begin{aligned} \sigma^2 &= \frac{1}{4}\left[4^2+6^2+8^2+10^2-\frac{(4+6+8+10)^2}{4}\right] \\ &= \frac{1}{4}\left[216-\frac{28^2}{4}\right] \end{aligned}$$

O desvio padrão populacional seria então: $\sigma = \sqrt{\sigma^2}$.

O desvio padrão é calculado da seguinte forma: $S = \sqrt{S^2}$. Valores grande no desvio padrão significa que os valores amostrais estão bem distribuídos em torno da média, enquanto que um desvio padrão pequeno indica que eles estão condensados próximos da média. Em poucas palavras, quanto menor o desvio padrão, mais homogênea é a amostra.

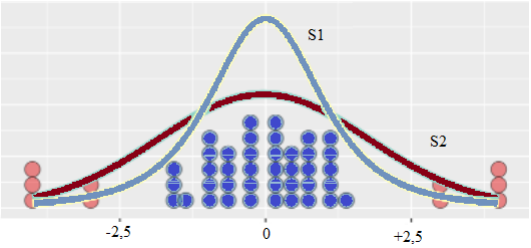


Figura 5.2: Dispersões dos desvios padrão em relação a média

Cálculo da variância e desvio padrão para dados amostrais agrupados

A fórmula para encontrar a variância para dados agrupados será:

$$S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 F_1 - \frac{(\sum_{i=1}^{n} x_i F_i)^2}{n}\right]$$

A fórmula para encontrar o desvio padrão será:

$$S = \sqrt{S^2}$$

Para melhor interpretar a dispersão de uma variável, calcula-se a raiz quadrada da variância, obtendo-se o desvio padrão que será expresso na unidade de medida original.

5.1.4 Coeficiente de Variação (CV)

O Coeficiente de variação é uma medida de variabilidade padronizada, ou seja, expressa percentualmente a variação dos dados em relação à média. Quando as medidas de duas ou mais variáveis são expressas em unidades diferentes como peso/altura, capacidade/comprimento, etc. não se pode compara-las através do desvio padrão, por este ser uma medida absoluta de variabilidade. Usa-se então o CV, que é uma medida relativa, que expressa o desvio padrão como uma porcentagem da média aritmética. Quanto mais próximo de zero, mais homogênea é a distribuição. Quanto mais distante, mais dispersas.

Esse coeficiente mede a dispersão em relação à média. É a razão entre o desvio padrão e a média. O resultado obtido dessa operação é multiplicado por 100, para que o coeficiente de variação seja dado em porcentagem. O CV fornece uma ideia de precisão experimental: quanto menor o CV, menor a variabilidade e melhor a precisão experimental. Por outro lado, quanto maior o CV, maior será a variabilidade experimental e pior será a precisão experimental.

Ou seja, é utilizada quando se pretende comparar o grau de dispersão de duas distribuições que não vem dadas pela mesma unidad, se utiliza uma estatística, devido a Karl Person, denominada de coeficiente de variação.

O CV é extremamente afetado pela escala da variável resposta. Por esse motivo ele é, em geral, apenas um bom indicador para comparar variáveis semelhantes. Podemos calcular o coeficiente de variação populacional ou amostral: se for populacional ou se for amostral.

$$CV = \frac{S}{\bar{x}} * 100$$

Sendo S o desvio padrão e $\bar{x}$ a média aritmética. Eis algumas regras empíricas para interpretação do coeficiente de variação de Pearson:

- Se $CV < 15\%$ há baixa dispersão
- Se $15\% \leq CV \leq 30\%$ há baixa dispersão
- Se $CV \geq 30\%$ há elevada dispersão

Exemplo 5.1.3. *O quadro mostra à média e o desvio padrão de duas pessoas que apresentam transtornos de ansiedade e compulsão alimentar. Dessas duas quem apresenta maior variação em relação às medidas peso (Kg) e altura (m)?*

Ind.	variáveis	$\bar{x} \pm s$	$C.V.\%$
A	Peso	$55,4 \pm 9,1$	$\frac{9,1}{55,4}.100\% = 16,43\%$
	Altura	$1,70 \pm 0,02$	$\frac{0,02}{1,76}.100\% = 1,17\%$
B	Peso	$68,2 \pm 13,6$	$\frac{13,6}{68,2}.100\% = 19,95\%$
	Altura	$1,80 \pm 1$	$\frac{1}{1,80}.100\% = 55,5\%$

Observa-se que a menor variação no peso e na altura está no indivíduo A, pois o C.V. são menores na variáveis peso e altura.

A utilização do C.V. assume a hipótoses de que o desvio padrão é proporcional a média, o que nem sempre é assim. O C.V. é uma medida sem dimensões, pois o numerador se elimina com o denominador independentemente das unidades de medidas. O C.V. tem fundamentalmente um valor comparativo para comparar um método de medida com outro.

5.1.5 Coeficiente de variação médio (C.V.M.)

Antes de definir o C.V.M. vamos definir desvio médio (D.M.) em relação a média p da seguinte forma:

$$DM_p = \frac{|\sum_{i=1}^{n} x_i - \bar{x}|n_i}{N}$$

O valor p pode ser: $p = \bar{x}$ ou $p = Md(mediana)$. Assim, tem-se:

$$C.V.M._{\bar{x}} = \frac{DM_{\bar{x}}}{|\bar{x}|} \qquad C.V.M._{Md} = \frac{DM_{Md}}{|Md|}$$

Alguns autores omite o uso de valores absolutos no denominador, às vezes incorrendo em coeficiente de variação negativo, o que não tem sentido.

Exemplo 5.1.4. *As medidas de peso de altura de 6 pessoas com transtorno de ansiedade generalizada num determinado hospital estão na tabela abaixo. Qual é o valor do coeficiente de variação média (em relaçao a mediana)?.*

Pesos	65	60	65	63	68	68
Alturas	1,70	1,50	1,68	1,70	1,75	1,80

Tabela 5.1: Medidas de peso e altura

Pesos	f_i	f_{ac}	Alturas	f_i	f_{ac}
60	1	1	1,50	1	1
63	1	2	1,68	1	2
65	2	4	1,70	2	4
68	2	6	1,75	1	5
			1,80	1	6

Tabela 5.2: Distribuição de frequência das medidas de peso e altura

Pesos: $N/2 = 3$, como $2 < 3 < 4$, a mediana dos pesos será: $Md = 65$. Quando as alturas: $N/2 = 3$, como $2 < 3 < 4$, as mediana da altura será: $Md = 1,70$.

O cálculo do desvio medio dos pesos será:

$$DM_{peso} = \frac{|65-65|.2 + |60-65|.1 + |63-65|.1 + |68-65|.2}{6} = 2,17$$

Os cálculos do desvio medio das alturas será:

$$DM = \frac{|1,70-1,70| + \cdots + |1,80-1,70|}{6} = 0,06$$

O coeficiente de variação dos pesos e alturas serão:

$$CV_{pesos} = \frac{2,17}{65} = 0,033 \qquad CV_{alturas} = \frac{0,06}{1,70} = 0,035$$

A média ponderado dos pesos e a variância serão, respectivamente:

$$\bar{x}_{pesos} = \frac{60.63 + (65.2) + (68.2)}{6} = \frac{289}{6} = 64,8$$

$$\sigma_{pesos} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 n_i}{N}} = 2,8$$

De forma análoga a média da altura e variância serão, respectivamente: $\bar{x}_{altura} = 1,68$ e $\sigma_{alturas} = 0,09$.

5.1.6 Escore padronizado

Quando se pretende comparar duas distribuições estadísticas ou duas medidas que não sejam a média e o desvio padrão, é necessário eliminar a influência dessas medidas, o qual se faz considerando uma nova variável para cada uma das distribuições a comparar, que se denomina variável padronizada.

Seja X é um variável estatística de média $\bar{X}$ e desvio padrão S, a variável padronizada correspondente é Z.

$$Z = \frac{x_i - \bar{x}}{S} \sim N(\bar{x} = 0, S = 1)$$

Que é outra medida relativa de dispersão é o escore padranizado (ou normalizado) para uma medida x_i.

Em efeito,

$$\begin{aligned}
\bar{Z} &= \frac{\sum_{i=1}^{n} z_i n_i}{N} \\
&= \frac{\sum_{i=1}^{n} \left(\frac{x_i - \bar{X} n_i}{S} \right) n_i}{N} \\
&= \frac{1}{S} \frac{0}{N} \\
&= 0
\end{aligned}$$

Por outra parte, o desvio padrão terá:

$$
\begin{aligned}
S_z^2 &= \frac{\sum_{i=1}^{n}(z_i - \bar{Z})^2 n_i}{N} \\
&= \frac{\sum_{i=1}^{n} Z_i^2 2n_i}{N} \\
&= \frac{\sum_{i=1}^{n} \left(\frac{x_i - \bar{X}}{S}\right)^2 n_i}{N} \\
&= \frac{1}{S^2} \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2 n_i}{N} \\
&= \frac{S^2}{S^2} \\
&= 1
\end{aligned}
$$

Um valor de escore negativo indica que a observação x_i está a esquerda da média, enquanto um escore positivo indica que a observação está à direita da média, $\bar{x}$.

Exemplo 5.1.5. *Considere medias e desvios padrões de dois alunos na disciplina de psicologia e estatística:*

Ana	Pedro
$\bar{x_A} = 6,5$	$\bar{x_P} = 5,0$
$S_A = 1,2$	$S_P = 5,0$

Tabela 5.3: Valores fictícios da media e desvio padrão de dois alunos

Assim, os escores padronizados são:

$$Z_A = \frac{7,5 - 6,5}{1,2} = 0,83$$

$$Z_P = \frac{6,0 - 5,0}{0,9} = 1,11$$

A melhor performance relativa foi no discente Pedro, pois $Z_P > Z_A$. Observe que, em termos absolutos, Ana conseguiu melhor nota. As notas de Pedro estão mais dispersas em relação a média, ou seja, quanto mais dispersos os valores em relação a média maior o C.V..

O problema em psicologia da sáude é que as diferentes medições e caracterizações, frequentemente, possuem domínios e grandezas distintas. É possível, por exemplo, que um jovem seja caracterizado pelos atributos descritivos "tempo que tem extresse", "idade" ou "salário" etc. O primeiro terá valores pertencentes ao intervalo $[1,100]$*meses* (sendo bem otimista), já o segundo teremos uma idade entre $[0,100]$ e o terceiro pertencente ao intervalo $[0,30000]$. Para minimizar os efeitos causados em situações como essas, é necessário aplicar procedimentos de transformação de dados (também conhecidos como data transformation). Esse procedimento abrage a normalização de dados e a conversão dos mesmos.

Exercício 5.1.3. *Considere a idade de 8 jovens que tem distúrbios do neurodesenvolvimento. Padronize essas 8 idades de forma que tenha média 0 e desvio padrão 1.*

25	16	24	20
18	5	4	8

Tabela 5.4: Valores fictícios de 8 níveis de estresse

Exercício 5.1.4. *Padronize os níveis de estresse de 12 idosos de modo que tenha média 0 e desvio padrão 1.*

6	10	4	9
8	5	4	8
7	3	4	1

Tabela 5.5: Valores fictícios de 12 níveis de estresse que variam de 1 a 10

5.1.7 Detectando outliers em dados psicológicos

Método Z padrão

Um ponto discrepante (outlier) em qualquer gráfico de dados é uma observação individual que se afastya do paderão global do gráfico. Nos trabalhos de coletas de dados em psicologia, podem ocorrer observações que fogem das dimensões esperadas - os outliers. Para detectá-los, pode-se calcular o escore padronizado (Z_i) e considerar outliers as observações cujos escores, em valor absoluto (em módulo), sejam maiores que 3.

Exemplo 5.1.6. *Os dados de uma pesquisa psicoplógica revelaram média 0,251 e desvio padrão de 0,019 para determinada variável. Verifica-se que o nível de ansiedade 0,298 e 0,355 podem ser considerados observações da referida variável.*

- *Para* $x_i = 0,298$*, teremos* $Z_i = \frac{0,298-0,251}{0,019} = 2,63$
- *Para* $x_i = 0,355$*, teremos* $Z_i = \frac{0,355-0,251}{0,019} = 5,473$

Obaserva-se que o dado 0,298 pode ser considerado normal, por outro lado, 0,355 é um outliers, portanto pode ser descartável.

Exercício 5.1.5. *Considere as idades de 10 pessoas que sofreram transtorno alimentar na infância. A idade dessas pessoas foram*

8,45,34,57,89,56,45,46,57,97. Existe algum outliers nessas idades? Se sim, quantos?

Exercício 5.1.6. *Considere as idades de 12 pessoas que sofreram transtorno alimentar na adolescência. A idade dessas pessoas foram 16,15,19,17,18,17,19,18, 48, 17, 18 e 61. Existe algum outliers nessas idades? Se sim, quantos?*

5.2 Erro-padrão da média

O erro-padrão da média de uma variável X, $S_{\bar{x}}$, dá uma ideia da precisão ou da representatividade da estimativa obtida para a média. Ele é inversamente proporcional ao tamanho da amostra e diretamente proporcional ao $S_{\bar{x}}$. O Erro-padrão da média é calculado pela fórmula:

$$EP = S_{\bar{x}} = \frac{S_{\bar{x}}}{\sqrt{n}}$$

É usual apresentar a média e o erro-padrão da média da seguinte forma: $S \pm S_{\bar{x}}$.

Embora aparentemente difícil este conceito é bastante fácil de ser entendido quando se conhece a aplicação do erro padrão da média: medir a variabilidade de um conjunto de médias de uma mesma população, em vez da variabilidade das observações individuais, como o faz o desvio padrão. Ou seja, o erro padrão nos dá uma ideia de quão variável pode ser a média retirada de uma população. Por exemplo, responda intuitivamente: qual dos procedimentos teria mais chance de mostrar a verdadeira média populacional, se tomássemos uma amostra de n elementos ou se medíssemos toda população? É claro que seria medir toda a população como um

todo, procedimento este que, na maioria dos casos, não é possível de ser feito. Na prática, trabalhamos, quase sempre, com amostras.

A distribuição amostral das médias segue o padrão da curva normal gaussiana, a área total sob ela é igual a 1, com 68% das médias, aproximadamente, situadas no intervalo entre $\mu - EP$ e $\mu + EP$ ao passo que, aproximadamente 95% estão entre o intervalo $\mu - 2EP$ e $\mu + 2EP$.

Na pratica, a distribuição amostral das médias pode ser considerada como normal sempre que $n \geq 30$ e, quanto maior o tamanho da amostra n, menor será o erro padrão e melhor será a estimativa da média da população.

Diante disto, podemos afirmar que o erro padrão é um parâmetro que permite ao pesquisador fazer dois tipos de inferências: estimar o tamanho provável do erro ao redor dos estimadores estatísticos, como a média, por exemplo, e realizar testes de significância estatística para verificação de hipóteses.

Exemplo 5.2.1. *Os valores abaixo se referem o tempo (em h) em horas de uma amostra de cinco rapazes que frequentaram uma clínica psicológica. O desvio padrão dessa amostra é igual a 3,84h. Determinar o erro padrão da média da amostra considerada.*

Tempo (h)	178	180	185	176	184

Tabela 5.6: Tempo de cinco rapazes que frequentaram uma clínica psicológica

Fazendo o cálculo temos: $EP = \frac{3,84}{\sqrt{5}} = 1,71$.

5.2.1 Diferença entre desvio padrão e erro padrão da média

a) O desvio-padrão amostral é uma medida de dispersão que indica o quanto os valores de um conjunto de dados se afastam da média amostral $(\bar{x})$.

b) Já o erro-padrão é uma medida de quão bem a média de uma amostra representa a média da população (μ) da qual ela foi retirada.

5.2.2 Amplitude interquartil

O intervalo interquartil se calcula a partir dos percentis 75 (P_{75}) e 25 (P_{25}) da seguinte forma: $Q = P_{75} - P_{25} = Q_3 - Q_1$. A amplitude interquartil de um conjunto de dados é a diferença entre o terceiro (Q_3) e o primeiro (Q_1) quartil. Esta medida de dispersão se utiliza quando se expressa a posição central por meio da mediana. Essas duas medidas se observa na gráfica de box-plot.

Exemplo 5.2.2. *Considere as idades de 10 pessoas que sofrem de algum transtorno.*
Idades=(18,45,34,57,89,56,45,46,57,97). Qual o erro padrão da média?

Considerando o exemplo acima temos:

- Amplitude: $A = 39 - 22 = 17$
- Variância: $\sigma^2 = \frac{(22-29,3)^2+(23-29,3)^2+\cdots+(39-29,3)^2}{20} = 27,6$
- Variância amostral: $S^2 = \frac{(22-29,3)^2+(23-29,3)^2+\cdots+(39-29,3)^2}{20-1} = 29,05$

- Desvio-padrão: $S = \sqrt{S^2} = 5,3$
- Desvio absoluto em relação a média: $DM = \frac{|22-29,3|+|23-29,3|+|22-29,3|+\cdots+|39-29,3|}{20} = 4,8$
- Coeficiente de variação: $CV = \frac{5,3.100}{29,3} = 18,3\%$
- Erro padrão da média: $EP = \frac{5,3}{\sqrt{20}} = 1,21$
- Amplitude interquartil: $AI = 34 - 25 = 9$

5.3 Exercícios

1. A quantidade de pessoa que fazem terapias comportamentais condiconamento aversivo em cada uma das 20 cidade de um estado do sudeste foram:

   30,21,24,35,46,57,56,67,76,79,45,67,54,45 67
   78,67,72,73,14,81,15,11,15,89,76,67,77,78,17

   Suponha que se precisa padronizar os dados. Pede-se usando o SPSS ou outro programa:

   (a) os escores padronizados de forma que tenham média 15 e desvio padrão 2

   (b) os escores padronizados de forma que tenham média 40 e desvio padrão 5

   (c) os escores padronizados de forma que tenham média 0 e desvio padrão 1

   (d) os escores padronizados de forma que tenham média 1 e desvio padrão 2

(e) verificar se existe outlers (considere o resultado do item c)

(f) construir o box plot dos dados brutos

(g) encontrar o valor do coeficiente de variação de Pearson dos dados padronizados com média 15 e desvio padrão 2.

(h) a amplitude total dos dados brutos

(i) a distribuição acima é assimétrica (considere os dados padronizados com média 0 e desvio padrão 1)?

(j) construir o histograma dos dados padronizados com média 0 e desvio padrão 1

(l) o 58^o percential (interprete).

(n) o 30^o decil

(m) o 70^o decil

(o) ovalor da curtose dos dados brutos?

(p) representar graficamente os dados brutos com a construção de um gráfico de barras (colunas) e um de pizza.

2. Com os dados da tabela seguinte. Encontre:

x_i	n_i
15	400
20	350
25	400
30	150
40	10

Tabela 5.7: Distribuição de frequência

(a) A média e a variância

(b) Uma representação gráfica dessa distribuição

3. A seguinte distribuição representa as idades de um grupo de indivíduos que assistiram uma palestra sobre transtorno de personalidade: paranoide (A), narcísica(B) e antissocial (C)

Idade	Frequência	Idade	Frequência
15	20	21	48
16	25	25	52
17	30	30	32
18	42	31	44
19	30	50	20
20	37	60	10

Tabela 5.8: Distribuição de frequência das idades

- Aqueles entre 15 a 19 (incluso) assistirão a palestra A
- Aqueles entre 20 a 30 (incluso) assistirão a palestra B
- Aqueles entre 31 a 60 assistirão a palestra C

Pede-se:

i. Construir uma coluna de variável no SPSS que os estratifique. Posteriormente encontre a amplitude de idade dos indivíduos em cada grupo.

ii. Construir alguns gráficos para os três grupos formados.

4. Determina-se 20 vezes o nível de glicose no sangue de uma mesma amostra de pessoas com ansiedade por meio de dois

métodos, A e B. Quais dos dois métodos tem maior dispersão?

Tabela 5.9: Níveis de glicose nos métodos A e B

A	140	141	142	127	138	136	135	142	126	148	139	142	141	151	144	146	145	148	147	136
B	130	132	146	138	145	148	147	135	136	137	141	146	138	131	134	146	139	140	148	146

5. Para ser emocionalmente saudável, o indivíduo precisa aceitar a realidade, mesmo quando essa realidade é desagradável. Terapeutas em 30 cidades ajudam indivíduos a alcançarem aceitação incondicional de si mesmo. Suponha que a quantidade de pessoas em cada uma das 30 cidade foram:

   175,165,167,172,176,178,179,185,169,178,179,180,175,173,180
   178,167,172,173,174,181,165,181,165,189,176,167,177,178,167

   Pede-se:

   (a) Determinar os escores padronizados de forma que tenham média 15 e desvio padrão 2

   (b) Determinar os escores padronizados de forma que tenham média 40 e desvio padrão 5

   (c) Determinar os escores padronizados de forma que tenham média 0 e desvio padrão 1

   (d) Determinar os escores padronizados de forma que tenham média 1 e desvio padrão 2

   (e) Verificar se existe outliers usando o método Z padrão (considere o resultado dos dados padronizados do item c)

   (f) Construir o box plot dos dados brutos

(g) Encontrar o valor do coeficiente de variação de Pearson dos dados padronizados com média 15 e desvio padrão 2

(h) Qual é a amplitude total dos dados brutos?

(i) A distribuição acima é assimétrica?

(j) Construir o histograma dos dados padronizados com média 0 e desvio padrão 1

(l) Representar graficamente os dados brutos com a construção de um gráfico de barras (colunas) e um de pizza

Use o SPSS!

6. Segue a quantidade de pessoas que tiveram transtorno de humor (depressão e mania) causa cognitiva em 100 municípios de um estado.

26	39	26	29	34	27	28	30	29	32
34	30	29	32	21	24	23	29	30	36
31	37	34	30	27	28	33	28	30	34
27	34	29	31	27	32	33	36	32	30
33	23	29	27	30	29	30	31	37	27
30	32	26	30	27	36	33	31	28	33
33	29	30	24	30	28	30	27	30	30
31	33	30	32	30	33	27	27	31	33
27	33	31	27	31	28	27	29	31	24
28	30	27	30	31	30	33	30	33	34

Usando o SPSS. Encontre:

(a) O Caule (ramo) e folha

(b) A amplitude

(c) Um Histograma

(d) Determinar o desvio pádrão

(e) Septuagéssimo quarto percentil

(f) A variância e o desvio padrão

(g) O coeficiente de variação (C.V.)

(h) O Box plot

(i) Se existem algum outliers?

(j) A amplitude semi-interquartélica

Capítulo 6

Medidas de forma

6.1 Medidas de assimetría

Denomina-se assimetria o grau de afastamento da simetria de uma distribuição de dados. Numa distribuição simétrica, as frequências mais altas ocorrem nos valores mais centrais de uma variável X, diminuindo gradualmente e de maneira simétrica em relação aos valores extremos e originando, aproximadamente, um mesmo número de valores menores e maiores que a média, cujo valor é semelhante aos da mediana e da moda. Em uma distribuição simétrica, há igualdade dos valores da média, mediana e moda.

Numa distribuição assimétrica negativa (inclinada para a esquerda) existirão mais valores da amostra maiores que a média, tendo a curva da distribuição uma cauda mais longa em relação aos valores menores que a média, cujo valor é menor que a mediana, que é menor que a moda. Neste caso, a média não se localiza no centro dos dados e a frequência diminui gradualmente em relação aos valores menores e, de forma mais abrupta, aos valores maiores

que a média.

Numa distribuição assimétrica positiva (inclinada para a direita) existirão mais valores da amostra menores que a média, tendo a curva da distribuição uma cauda longa em relação aos valores maiores que a média, cujo valor é maior que a mediana, que é maior que a moda. Neste caso, a média não se localiza no centro dos dados e a frequência diminui gradualmente em relação aos valores maiores e, de forma mais abrupta, aos valores menores que a média. Observação: média ($\bar{x}$) = a mediana ($\tilde{x}$) = a moda (M_o).

- $M_o < \tilde{x} < \bar{x}$ (Assimétrica à direita)
- $\bar{x} < \tilde{x} < M_o$ (Assimétrica à esquerda)
- $\bar{x} = \tilde{x} = M_o$ (Simétrica)

Eis uma ilustração gráfica de uma distrinuição simétrica e distribuições assimétricas.

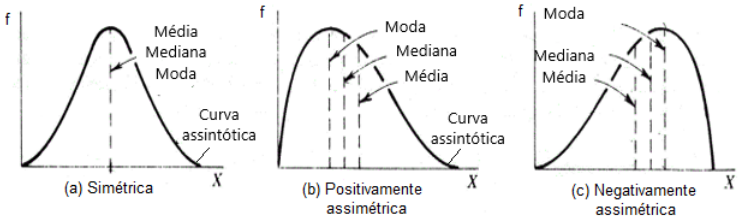


Figura 6.1: Ilustração gráfica de distribuições simétrica e assimétricas

Usaremos duas fórmulas para o calculo do coeficiente de assimetria.

- Primeiro coeficiente de Pearson

Para populacional:

$$AS = \frac{\mu - M_o}{\sigma}$$

Para amostral:

$$AS = \frac{\bar{x} - Md}{S}$$

- Fazendo $Md = \tilde{x}$ segundo coeficiente de Pearson (Coeficiente de Bowley) tem-se:

$$P_{3/4} - \tilde{x} \neq \tilde{x} - P_{1/4} \Rightarrow P_{3/4} - P_{1/4} - 2\tilde{x}$$

Logo, temos:

$$AS = \frac{P_{3/4} - P_{1/4} - 2\tilde{x}}{P_{3/4} - P_{1/4}}$$

A assimetria relativa será: $-1 \leqslant S \leqslant 1$. O valor S será Positivo a medida que o terceiro quartil se afasta de $\tilde{x}$, enquanto que o primeiro quartil se aproxima da mesma, tendo como limite: $P_{1/4} - P_{2/4}$, a assimetria assume o valor máximo positivo quando $S = 1$, ou seja, $AS = \frac{P_{3/4} - P_{1/4}}{P_{3/4} - P_{1/4}} = 1$. Ela será negativa a medida que o primeiro quartil afasta-se da mediana, enquanto o terceiro quartil aproxima-se da mesma, dando como limite $P_{3/4} - P_{2/4}$, e a assimetria assume valor máximo negativo quando $AS = \frac{-P_{3/4} + P_{1/4}}{P_{3/4} - P_{1/4}} = -1$.

Para corrigir parte do inconveniente de se desprezar a metade das ocorrências, Kelley aconselha o uso dos Centis equidistantes da mediana ($\tilde{x}$), tais como $P_{0,10} = D_1$ e $P_{0,90} = D_9$ (per-

centis 10 e 90, respectivamente), para cálculo da assimetria, assim temos:

$$AS = \frac{P_{0,90} - P_{0,10} - 2\tilde{x}}{P_{0,90} - P_{0,10}}$$

Os limites de S variam também de -1 a +1.

$$AS = \frac{P_{3/4} + P_{1/4} - 2P_{2/4}}{P_{3/4} - P_{1/4}}$$

Se:

1. $AS = 0$, diz-se que a distribuição é simétrica ($\bar{x} = \tilde{x} = M_o$)
2. $AS > 0$, diz-se que a distribuição é assimétrica positiva.
3. $AS < 0$, diz-se que a distribuição é assimétrica negativa.

Sendo $P_{2/4}$ o primeiro quartil, $\tilde{x} = P_{2/4}$ a mediana e $P_{3/4}$ o terceiro quartil.

Ainda pode-se encontrar o coeficiente de assimetria de Fisher usando o terceiro momento em relação a média (ou centro) definido pela seginte fórmula:

$$m_3 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3 n_i}{N}$$

Assim, o coeficiente de assimetria de Fisher será:

$$AS_F = \frac{m_3}{\sigma^3}$$

Sendo σ igual a:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 n_i}{N}}$$

Exemplo 6.1.1. *Considere a seguinte tabela que informa o número de filhos com o número de família:*

Número de filhos	0	1	2	3	4	5	6	7
Número de família	2	3	10	10	5	0	5	0

Tabela 6.1: Número de filhos e o número de família:

Para facilitar vamos considerar a seguinte tabela de cálculo:

x_i	n_i	N_i	$x_i n_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2 n_i$	$(x_i - \bar{x})^3 n_i$
0	2	2	0	-2,94	17,28	-50,82
1	3	5	3	-1,94	11,29	-21,90
2	3	5	3	-0,24	8,83	-8,30
3	10	25	30	0,06	0,03	0,002
4	5	30	20	1,06	5,61	5,95
5	0	30	0	2,06	0	0
6	5	35	30	3,06	46,81	143,26
7	0	35	0	4,06	0	0
			103		89,85	68,192

Tabela 6.2: Tabela para auxiliar os cálculos

Substituindo esses valores tem-se:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 n_i}{N}} = \frac{89,85}{35} = 1,60$$

Logo, $\sigma^3 = 1,60^3 = 4,096$

$$AS_F = \frac{68,192}{35}\frac{1}{4,096} = \frac{68,192}{143,36} = 0,47$$

Logo, existe assimetria a direita ou positiva, pois $AS_F > 0$

6.1.1 Teste de assimetria

Esse teste de hipótese de assimetria contrasta a normalidade da distribuição da que se tem extraído os dados mediante a consideração do coeficiente de assimetria amostral. Pretende-se testar as seguintes hipóteses: $H_0 : X \sim Normal$ *versus* $H_1 : X \sim Nonormal$.

A estatística de teste será o coeficiente de assimetria amostral definido por:

$$A = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2)^{3/2}}$$

Se a hipótese nula $H_0 : X \sim Normal$ é certa, a estatística tem uma distribuição assintótica normal de média zero e variância $\sigma^2 = \frac{6}{n}$. Assim, $A \sim N(0; \frac{6}{n})$. O teste de hipótese acima pode ser expresso da seguinte forma: H_0: X tem simetria normal (assimetria = 0) e H_1 : X não tem simetria normal.

Se a hipótese nula é certa, o coeficiente de assimetria amostral estima um parâmetro da população que é zero (o coeficiente de assimetria de uma distribuição normal é zero).

Rejeita-se H_o a nível α para valores grandes da estatística A. O teste também pode ser resolvido mediante o coeficiente de assimetria amostral padronizado definido como: $A_p = \frac{A}{\sqrt{6/n}}$.

Rejeita-se H_o a nível α para valores grandes da estatística A_p, ou seja, se A_p não está no intervalo de -2 a 2.

Exemplo 6.1.2. *Aplicou-se um teste a quinze alunos de estatística e psicologia II e verificou-se um nível de vocabulário em psicologia social. Os valores do teste foram os seguintes: 7.1; 5.2; 6.4; 6.7; 3.9; 7.0; 6.2; 7.1; 6.3; 7.3; 5.8; 4.1; 6.7; 5.0 e 7.7.*

$$A = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2)^{3/2}} = -0.7477$$

No programa R seria:

```
X1<-c(7.1, 5.2, 6.4, 6.7, 3.9, 7, 6.2,
7.1, 6.3, 7.3, 5.8, 4.1, 6.7, 5,7.7)
print(skewness(X1))
n<-length(X1);n
B1<-(1/n)*sum((X1-mean(X1))^3);B1
B2<-((1/n)*sum((X1-mean(X1))^2))^(3/2);B2
As<-B1/B2;As # As = -0,7477
```

Por tanto, o coeficiente de assimetria amostral padronizado será: $A_p = \frac{A}{\sqrt{6/n}} = -0.7477\sqrt{6/15} = -1.1822$, posto que, $A_p \in [-2;2]$, logo, aceita a hipótese de normalidade.

6.2 Medidas de achatamento ou curtose

Curtose nada mais é do que o grau de achatamento da curva de uma distribuição de Frequências. Isto considerando que uma curva pode apresentar-se mais achatada ou mais afilada em relação a uma curva considerada curva padrão ou curva normal.

Estudam a distribuição dos dados na zona central da série. A maior ou menor concentração de frequência próximo da média e na zona central da distribuição dará lugar a uma distribuição mais ou menos apuntada. Por essa razão, as medidas de curtose se chama também de apuntamento ou concentração central.As medidas de

curtoses se aplicam a distribuições próxima da normal, ou seja, unimodal simétricas ou ligeiramente assimétrica.

$$K = \frac{n(n+1)\sum_{i=1}^{n}(x_i-\bar{x})^4 - 3\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)(n-1)}{(n-1)(n-2)(n-3)\sigma^4}$$

Quando o valor é positivo diz-se que a distribuição está mais apontada que a normale se denomina leptocúrtico. Quando o valor é zero ou próximo a zero, a distribuição tem o mesmo apontamento que a distribuição normal e se denomina mesocúrtica. Por último, quando o valor é negativo, o apontamento é menor que o da distribuição normal e se denomina platicúrtica.

A curtose ou achatamento é mais uma medida com a finalidade de complementar a caracterização da dispersão em uma distribuição. Esta medida quantifica a concentração ou dispersão dos valores de um conjunto de dados em relação às medidas de tendência central em uma distribuição de frequências. Denomina-se Curtose o grau de achatamento da distribuição. Para medir o grau de curtose utiliza-se o coeficiente:

$$K = \frac{Q_3 + Q_1}{2(P_{90} - P_{10})}$$

Se:

1. $k = 0,263$, diz-se que a distribuição de frequências é mesocurtica.

2. $k > 0,263$, diz-se que a distribuição de frequências é platicútica.

3. $k < 0,263$, diz-se que a distribuição de frequências é leptocúrtica.

Se o valor da Kurtosis for k = 0, então tem o mesmo achatamento que a distribuição normal. Se o valor é $k > 0$ então a distribuição em questão é mais alta (afunilada) e concentrada que a distribuição normal.

Exemplo 6.2.1. *A quantidade de estressados em 20 escolas de um grande cidade são: 22,23,25,22,25,25,24,26,26,29,27,32,34,34,33, 35,33,36,36 e 39.*

- A média vale: $\bar{x} = \frac{22+23+\cdots+39}{20} = 29,3$
- Assimetria vale: $A = \frac{20((22-29,3)^3+\cdots+(39-29,3)^3)}{(20-1)(20-2)(20-3)5,39^3} = 0,197$
- A curtose vale: $k = \frac{20(20+1).24677,4-3.552,2.552,2.(20-1)}{(20-1)(20-2)(20-3).5,39^4} = -1,43$

6.2.1 Teste de curtose

Esse teste de hipótese de assimetria contrasta a normalidade da distribuição da que se tem extraído os dados mediante a consideração do coeficiente de curtose amostral. Pretende-se testar as seguintes hipóteses: $H_0 : X normal \; versus \; H_1 : X nonormal$.

A estatística de teste será o coeficiente de curtose amostral definido por:

$$B = \frac{B_4}{B_2^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4}{(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2)^2} - 3$$

Se a hipótese nula $H_0 : X \sim Normal$ é certa, a estatística tem uma distribuição assintótica normal de média zero e variância $\sigma^2 =$

$\frac{24}{n}$. Assim, $B \sim N(0; \frac{24}{n})$. O teste de hipótese acima pode ser expresso da seguinte forma: H_0: X tem curtose normal (curtose = 0) e H_1 : X não tem curtose normal.

Se a hipótese nula é certa, o coeficiente de curtose amostral estima um parâmetro da população que é zero (o coeficiente de curtose de uma distribuição normal é zero).

Rejeita-se H_o a nível α para valores grandes da estatística B. O teste também pode ser resolvido mediante o coeficiente de curtose amostral padronizado definido como: $B_s = \frac{B}{\sqrt{24/n}}$.

Rejeita-se H_o a nível α para valores grandes da estatística B, ou seja, se B não está no intervalo de -2 a 2.

Exemplo 6.2.2. *Aplicou-se um teste a quinze alunos de estatística e psicologia II e verificou-se um nível de vocabulário em psicologia social. Os valores do teste foram os seguintes: 5.1; 5.2; 5.1; 6.7; 3.7; 7; 6.2; 7.5; 6.3; 5.3; 5.8; 4.1; 6.7; 5 e 5.7.*

$$B = \frac{B_4}{B_2^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4}{(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2)^2} - 3 = -0,27$$

Portanto, o coeficiente de curtose amostral padronizado vale:

$$B_s = \frac{B}{\sqrt{24/15}} = \frac{-0,27}{1,26} = -0,21$$

Posto que, $B_s \in [-2, +2]$, aceita-se a hipótese de normalidade.

6.3 Gráfico Box Plot (box-and-whisker plot)

É uma medida de dispersão, de posição e de forma. É uma representação de uso crescente pelas suas interessantes propriedades descritivas. Consiste em um retângulo, que é uma caixa, e um prolongamento vertical ou horizontal, que são os bigodes (whiskers). O gráfico Box Plot (ou desenho esquemático) é uma análise gráfica que utiliza cinco medidas estatísticas: valor mínimo, valor máximo, mediana ou segundo quartil, primeiro e terceiro quartil da variável quantitativa. Este conjunto de medidas oferece a ideia da posição, dispersão, assimetria, caudas e dados discrepantes. A posição central é dada pela mediana e a dispersão pelo desvio interquartílico $dq = Q_3 - Q_1$. As posições relativas de Q_1 , Q_2 e Q_3 dão uma noção da assimetria da distribuição. Os comprimentos das caudas são dados pelas linhas que vão do retângulo aos valores atípicos.

Um outlier ou ponto discrepante é um valor que se localiza distante de quase todos os outros pontos da distribuição. À distância a partir da qual considera-se um valor como discrepante é aquela que supera 1,5dq. De maneira geral, são considerados outliers todos os valores inferiores $L_I = Q_1 - 1,5(Q_3 - Q_1)$ ou os superiores a $L_s = Q_3 + 1,5(Q_3 - Q_1)$.

Exemplo 6.3.1. *Dado as seguintes idades de pessoas de uma grande empresa que tem desenvolvimento moral: moralidade pré-convencional segundo a teoria de Kohlberg*

18	18	19	20	20
20	20	20	20	21
21	22	23	24	25
25	25	26	29	30
35	37			

Tabela 6.3: Valores fictícios para cálculo da gráfica box plot

Conhecendo os valores:

- $Md = 21,50$
- $Q_1 = 20$
- $Q_3 = 25,75$
- $dq = 5,75$
- $LI = Q_1 - 1,5 * dq = 11,375$
- $LS = Q_3 - 1,5 * dq = 34,375$

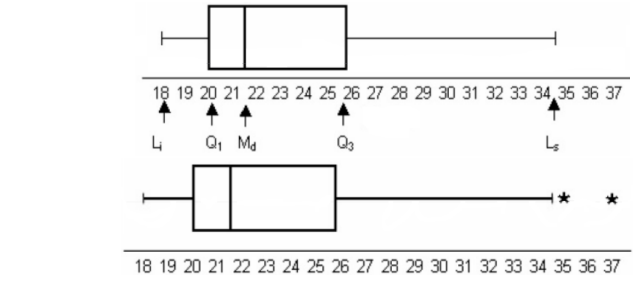


Figura 6.2: Ilustração gráfica de um box plot com e sem outliers

Muitos conjuntos de dados de vida real em psicologia têm distribuição que são aproximadamente simétricas e têm curvas em

forma de sino (normal). Você pode usar um gráfico caixa-e-bigodes (box plot) para determinar a forma da distribuição.

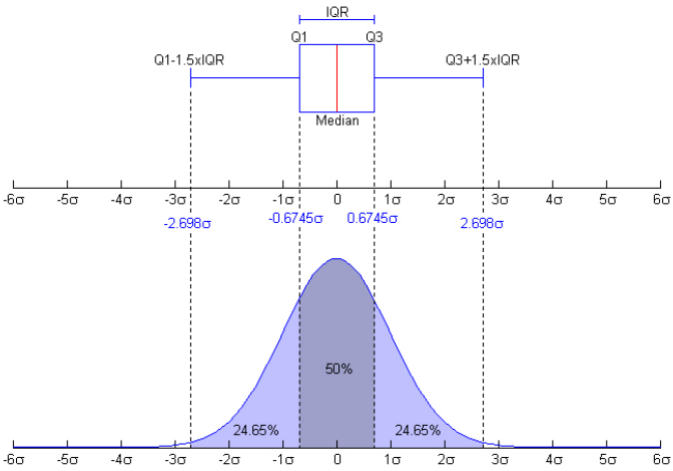


Figura 6.3: Ilustração gráfica do box plot numa distribuição normal

O SPSS dá os valores discreapantes alto (high) e baixo (low) separados do ramo-e-folhas, em vez de incluí-los nele.

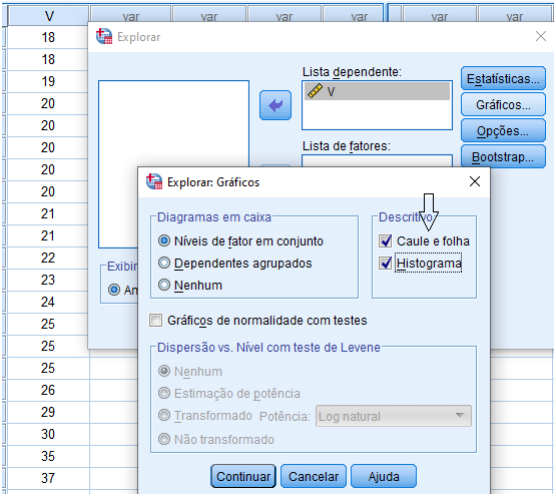


Figura 6.4: Ramo e folhas (Caule e folhas) e o histograma

Observa-se que 1,2 e 3 são os ramos e o segundo algorítmo são as folhas como vemos na Figura 6.5. Existem dois valores elevados, a saber, os números das posições 20 e 21, respectivamente.

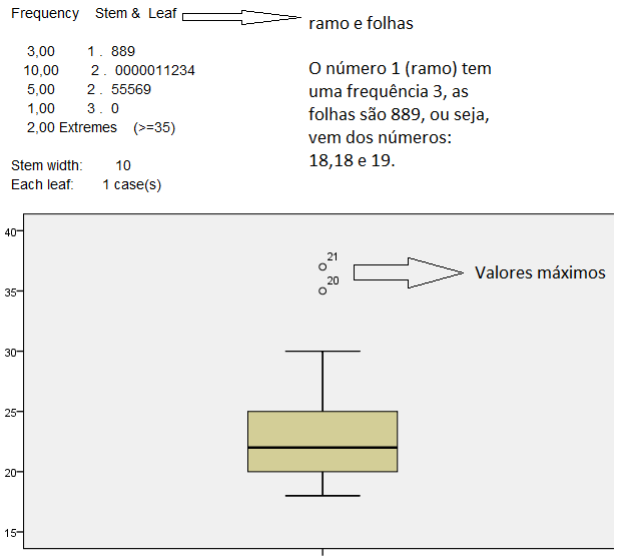


Figura 6.5: Output do SPSS: Ramo e folhas (Caule e folhas) e o box plot

Segue os passos para verificar as medidas de assimetria no SPSS. Selecione a opção Analisar, depois estatística descritiva. Slecione a variável V e clique em continuar para transportá-la para o campo Variável. Clique sobre o botão Opções ... e, em seguida, marque as opções Curtose (Kurtosis)e Assimetria (Skewness). As fórmulas para encontrar assimetria e erro padrão da simetría (EPA) são, respectivamente.

$$AS = \frac{n\sum F_i(x_i - \bar{x})^3}{(n-1)(n-2)S^3}$$

$$EPA = \frac{6n(n-1)}{(n-1)(n-2)(n-3)} \cong \sqrt{\frac{6}{n}}$$

A medida curtose do SPSS é a seguinte:

$$Curtose = \frac{n(n-1)\sum F_i(x_i-\bar{x})^4 - 3S^4(n-1)}{(n-1)(n-2)(n-3)S^4}$$

E o erro padrão da curtose é dado por:

$$EPC = \frac{24n(n-1)}{(n-2)(n+1)(n+3)} \cong \sqrt{\frac{24}{n}}$$

Encontre essas estatísticas no SPSS da seguinte forma. Segue o passos e o resultado no SPSS.

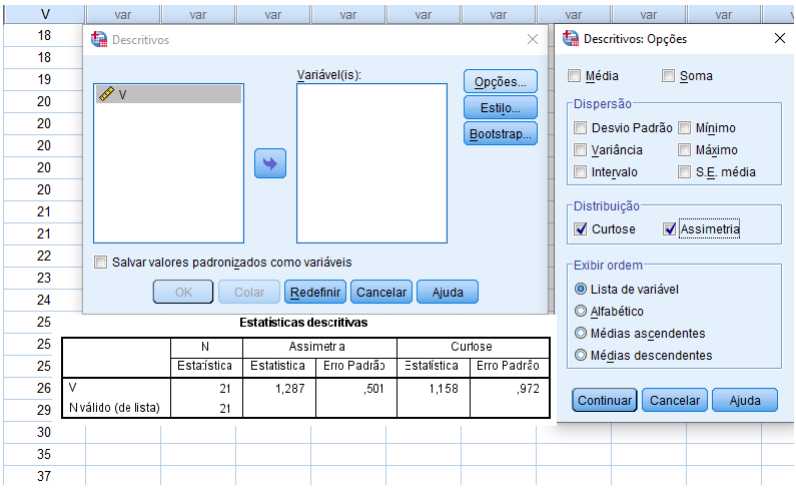


Figura 6.6: Medidas de formas no SPSS

6.4 Exercícios

1. Considere as idades de 10 pessoas que sofrem de algum transtorno. $Idades = (18, 45, 34, 57, 89, 56, 45, 46, 57, 97)$. Encontre: O valor da curtose, o valor da assimetria e o erro padrão da curtose (EPC) e o erro padrão da simetría (EPA). Use o SPSS.

2. Considere as idades de 10 pessoas que sofrem de algum transtorno. $Idades = (34, 41, 37, 51, 86, 46, 55, 36, 51, 82)$. Encontre: O valor da curtose, o valor da assimetria e o erro padrão da curtose (EPC) e o erro padrão da simetría (EPA). Use o SPSS.

3. Considere as idades de 10 pessoas que sofrem de algum trans-

torno. $Idades = (34,41,37,51,86,46,55,36,51,82)$. Pede-se o coeficiente de assimetria e a curtose. Use o SPSS.

4. Segue a quantidade de pessoas que tiveram transtorno de humor (depressão e mania) causa cognitiva em 100 municípios de um estado.

26	39	26	29	34	27	28	30	29	32
34	30	29	32	21	24	23	29	30	36
31	37	34	30	27	28	33	28	30	34
27	34	29	31	27	32	33	36	32	30
33	23	29	27	30	29	30	31	37	27
30	32	26	30	27	36	33	31	28	33
33	29	30	24	30	28	30	27	30	30
31	33	30	32	30	33	27	27	31	33
27	33	31	27	31	28	27	29	31	24
28	30	27	30	31	30	33	30	33	34

Usando o SPSS. Encontre:

(a) O coeficiente de assimetria. O que se conclui?

(b) A amplitude

(c) Um Histograma

(d) A curtose. O que se conclui?

Capítulo 7

Medidas correlacionais

7.1 Correlações

Indica o grau e sentido da variação concomitante de duas ou mais série de dados. Quanto a classificação temos:

Segundo critério quantitativo . A correlação pode ser:

- Perfeito $(r = \pm 1)$
- Imperfeita $(0 < r < 1; -1 < r < 0)$
- Nula $(r = 0)$

Segundo critério qualitativo . A variação pode ser:

Positiva (variação no mesmo sentido)

Negativo (variação em sentido contrário)

Para o caso de variável linear os coeficientes de correlação para duas variáveis são:

- r_{xy} (coeficiente de Pearson) para duas variáveis quantitativas (contínuas)

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{S_{xy}}{S_x S_y}$$

Ou ainda,

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Esse coeficiente é adimensional, e encontra-se cotado da seguinte maneira: $-1 \leq r_{xy} \leq 1$.

- r_s (coefciente de Sperman) para duas variáveis ordinais

$$r_s = 1 - \frac{6\Sigma D^2}{n(n^2 - 1}$$

- r_{bp} (bisserial pontual) (uma variável contínua e outra dicotômica

$$r_{bp} = \frac{\bar{X}_p - \bar{X}_q}{S_t}\sqrt{p(1-p)}$$

$$r_{bp} = \frac{\bar{X}_p - \bar{X}_t}{S_t}\sqrt{p/q}$$

Sendo:

p = porporção de indivíduos na categoria P da variável dicotômica

q = porporção de indivíduos na categoria Q da variável dicotômica

$\bar{X}_p$ = média na variável contínua dos indivíduos da categoria P

$\bar{X}_q$ = média na variável contínua dos indivíduos da categoria Q

$\bar{X}_t$ = média na variável contínua de todos os indivíduos

$\bar{S}_t$ = desvio padrão da variável contínua

- r_b (bisserial) uma variável contínua e outra dicotomizada

$$r_b = \frac{\bar{X}_p - \bar{X}_q}{S_t}\sqrt{\frac{pq}{y}}$$

$$r_b = \frac{\bar{X}_p - \bar{X}_t}{S_t}\sqrt{\frac{p}{y}}$$

Sendo:

p = porporção de indivíduos na categoria P da variável dicotomizada

q = porporção de indivíduos na categoria Q da variável dicotomizada

$\bar{X}_p$ = média na variável contínua dos indivíduos da categoria P

$\bar{X}_q$ = média na variável contínua dos indivíduos da categoria Q

$\bar{X}_t$ = média na variável contínua de todos os indivíduos

$\bar{S}_t$ = desvio padrão da variável contínua

y = probabilidade na ordenada

- Φ (quádruplo) para duas variáveis dicotômicas

$$\Phi = \frac{(A.D) - (B.C)}{\sqrt{[(A+B)(A+C)(C+D)(B+D)]}}$$

A	B	A+B
C	D	C+D
A+C	B+D	N = A+B+C+D

Observa-se que o somatório de A + B + C + D é igual a N. Sendo que A,B,C e D são as frequências em cada um dos elementos da matriz quadrada.

- r_t (tetracórico) para duas variáveis dicotomizadas

Tabela de Davidoff y Goheen:

$$r_t = \frac{(A.D)}{(B.C)}$$

$$r_t = \cos \frac{180^\circ \sqrt{BC}}{\sqrt{BC} + \sqrt{AD}}$$

Para o caso de variável curvilínea:

- η (coeficiente Eta)

Para distribuições multidimensionais:

- parcial:

Primeira ordem $r_{12.3}$

Segunda ordem $r_{12.34}$

- Semiparcial $r_{1(23)}$
- Múltiplo $R_{1,23}$ $R_{1,234}$

Exemplo 7.1.1. *Considere um banco de dados adaptado formado por 36 indivíduos. A base de dados apresenta 15 variáveis de vários tipos para exemplificar os cálculos de medidas correlacionais.*

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
Ind1	1	1	1	17	1	1	28	30	69	30	29	22	17	8	1
Ind2	2	1	1	18	1	2	27	43	68	20	30	28	16	9	1
Ind3	3	1	1	7	2	3	14	18	38	30	10	15	9	6	1
Ind4	4	1	1	12	2	2	23	23	50	26	21	25	14	5	1
Ind5	5	1	1	15	3	1	24	19	57	37	24	19	8	3	1
Ind6	6	1	1	6	3	3	14	22	33	15	11	20	11	4	2
Ind7	7	1	2	10	1	3	14	26	30	21	12	27	13	7	1
Ind8	8	1	2	7	1	2	18	20	38	27	8	19	12	6	2
Ind9	9	1	2	9	2	1	14	19	39	20	7	16	10	4	2
Ind10	10	1	2	4	2	3	10	18	22	32	5	12	8	5	2
Ind11	11	1	2	2	3	2	5	12	20	16	3	9	4	1	2
Ind12	12	1	2	7	3	1	14	17	37	27	6	14	9	3	2
Ind13	13	2	1	20	1	1	30	29	62	30	28	25	15	7	1
Ind14	14	2	1	9	1	3	18	22	35	30	12	21	11	8	2
Ind15	15	2	1	10	2	2	15	20	41	25	9	24	6	7	1
Ind16	16	2	1	13	2	2	20	25	50	22	20	22	16	4	1
Ind17	17	2	1	11	3	1	16	15	43	21	16	11	8	2	2
Ind18	18	2	1	2	3	3	8	11	19	12	2	8	10	1	2
Ind19	19	2	2	16	1	1	22	33	55	29	28	26	18	5	2
Ind20	20	2	2	8	1	2	18	28	30	16	11	19	16	7	1
Ind21	21	2	2	3	2	3	7	24	24	10	4	23	12	5	2
Ind22	22	2	2	9	2	2	12	20	47	19	7	20	7	2	1
Ind23	23	2	2	2	3	3	9	13	21	22	3	10	5	4	2
Ind24	24	2	2	4	3	1	8	14	23	30	5	12	6	1	2
Ind25	25	3	1	19	1	2	27	27	68	34	25	21	13	7	1
Ind26	26	3	1	14	1	3	24	22	60	32	23	16	8	6	1
Ind27	27	3	1	15	2	1	24	28	70	39	22	18	13	4	1
Ind28	28	3	1	9	2	3	17	19	31	24	10	16	6	7	1
Ind29	29	3	1	8	3	2	12	18	28	20	7	13	5	2	2
Ind30	30	3	1	11	3	1	18	20	46	19	15	17	7	2	2
Ind31	31	3	2	10	1	3	14	20	50	12	6	14	8	5	1
Ind32	32	3	2	13	1	1	18	24	48	19	18	20	14	4	2
Ind33	33	3	2	10	2	2	17	21	37	26	18	19	12	3	1
Ind34	34	3	2	4	2	3	10	18	35	14	6	10	5	2	1
Ind35	35	3	2	10	3	1	16	10	42	18	10	7	3	1	1
Ind36	36	3	2	3	3	2	12	14	17	22	2	9	4	2	2

Tabela 7.1: Tabela adaptada para os cálculos correlacionais

Suponha que as variáveis foram definidas da seguinte forma:

1. Identificação do indivíduo: De 1 ao 36

   Foram considerados 36 indivíduos.

2. Método psicológico aplicado : M_1, M_2 e M_3 (Variável independente principal)

   Trata-se de uma variável experimental em que cada uma das quais representam três métodos psicológicos: M_1, M_2 e M_3 é o método misto, mesclando os métodos 1 e 2.

3. Teve assistência psicológica : 1 (Sim) e 2 (Não)

   Definida pelo fato de ter ou não assistência de uma clínica psicológica.

4. Rendimento no trabalho (supunha que varia de 0 a 24 pontos).

   Esssa pontuação hipotética varia de 0 a 24 refletindo o rendimento no trabalho do indivíduo.

5. Ambiente familiar : (1: grande, 2: médio e 3: pequena)

   Interação dentro da família do indivíduo.

6. Status socioeconómico : (1: alto, 2: médio, 3: baixo). Para descrever os três lugares em que uma família ou um indivíduo pode se enquadrar.

   Classificado em três níveis: alto, médio e baixo sobre dados de ingresso familiares. Suponha que essa classificação foi obtida também através de questionários.

7. Nível de ansiedade: Suponha que varia de 0 a 30 pontos

   Suponha que o teste consiste de 30 itens que recebem pontuação de 0 e 1. Se todos forem 1 a soma total vale 30.

8. Nível intelectual : 0 a 50 pontos

   A inteligência geral também foi medida através de alguma escala de medição de inteligência. A pontuação máxima é 50.

9. Velocidade de leitura do ansioso: 0 a 80 pontos

   Suponha que foi realizado o teste de Angel Lázaro de velocidade de leitura. A pontuação máixima é de 80 pontos.

10. Memoria geral : 0 a 30 pontos

    Supoha que a pontuação obtida foi através da subescala de memória da bateria McCarthy de aptidões e habilidades psicomotoras. Esta subescala media a capacidade de memorizar conteúdos de curto alcance e é composta por tarefas que requerem reconhecimento de memória de: sequência de acertos, palavras e números. A pontuação máxima era de 30 pontos.

11. Compreensão de leitura do indivíduo ansioso : 0 a 30 pontos

    Mede compreensão, precisão e velocidade de leitura. A pontuação máxima é de 30 pontos.

12. Nível de estresse: 0 a 30 pontos

    Suponha que foi utilizado um teste para medir o nível de estresse. O nível máximo é 30 (alto estresse) e 0 (nenhum estresse).

13. Estilo cognitivo (suponha que varia de 0 a 18 pontos)

    A pontuação máxima é 18 pontos e a mínima 0.

14. Adaptação do ansioso na sociedade : 0 a 10 pontos

    Suponha que a máxima pontuação é 10 e a mínima é 0.

15. Os indivíduos tem níveis aspirações? (1: Sim, 2: Não)

Exemplo 7.1.2. *Qual é o grau de sentido da relação entre as variáveis: velocidade de leitura* (X_9) *e compreensão de leitura do indivíduo ansioso* (X_{11})*?*

A natureza das duas variáveis é contínua. O coeficiente mais adequado é o de Pearson.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{S_{xy}}{S_x S_y}$$

ou ainda,

$$r_{xy} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}}$$

Sabendo que $\sum X_9 = 1483$, $\sum X_{11} = 473$, $\sum X_9 c X_{11} = 69.29 + \cdots + 17.02 = 23561$, $\sum X_9^2 = 69361$, $\sum X_{11}^2 = 8815$ e $n = 39$ indivíduos. Façamos $X_9 = x$ e $X_{11} = y$ para visualizar melhor as variáveis na coeficiente de Pearson. Substituindo na equação acima temos:

$$r_{xy} = \frac{36.23561 - 1483.473}{\sqrt{36.69361 - (1483)^2}\sqrt{36.8815 - (473)^2}} = 0,88$$

Exemplo 7.1.3. *Deseja-se saber qual relação existe entre a velocidade de leitura* (X_9) *e compreensão de leitora* (X_{11})*, na subamostra de indivíduos que tiveram assistência psicológica.*

Para o cálculo da correlação, considerando que as variáveis são contínuas, poderia utilizar a correlação de Pearson, porém como a

amostra é pequena (n=18) é aconselhável o uso do r_s de Spearman. Dado por:

$$r_s = 1 - \frac{6\sum D^2}{n(n^2-1)}$$

Sendo D a diferença entre os postos indicado em cada pontuação dentro de cada uma das duas séries y n é o número de pares.

Passos para encontrar a correlação de Sperman:

1. Ordenam-se as pontuações X
2. Coloca-se junto a cada pontuação em X seu correspondente Y
3. Indica-se um número segundo a ordem ou lugar que ocupa a pontuação dentro da série X (posto na variável X)
4. Indica-se um número segundo a ordem que ocupa a pontuação dentro da série Y (posto na variável Y)
5. Claculam-se as diferenças entre os postos (valor da coluna 3 menos valor da coluna 4)
6. Por último, elevam-se ao quadrado essas diferenças

(1)	(2)	(3)	(4)	(5)	(6)
X	Y	R_x	R_y	D	D^2
70	22	1	7	-6	36
69	29	2	2	0	0
68	30	3,5	1	2,5	6,25
62	25	3,5	4	-0,5	0,25
60	23	6	6	0	0
57	24	7	5	2	4
50	20	8,5	9	-0,5	0,25
50	21	8,5	8	0,5	0,25
46	15	10	11	-1	1
43	16	11	10	1	1
41	9	12	16	-4	16
38	10	13	14,5	-1,5	2,25
35	12	14	12	2	4
33	11	15	13	2	4
31	10	16	14,5	1,5	2,25
28	7	17	17	0	0
19	2	18	18	0	0
					$\sum D^2 = 81,5$

Tabela 7.2: Calculos para encontrar a correlação de Spearman

Assim, o coeficiente de correlação de Spearman será:

$$r_s = 1 - \frac{6.81,5}{18(18^2-1)} = 1 - 0,084 = 0,961$$

Existe uma alta relação positiva, como a correlação de Spearman.

Exercício 7.1.1. *Uma clínica de psicólogos se aplicam dois testes*

(A e B) a um grupo de 10 indivíduos. Ao término dos testes, os pesquisadores ordenaram do maior ao menor as pontuações obtidas. Pergunta-se se existe correlação entre ambos os testes. Resposta: A correlação entre os dois testes tem um valor $\rho = 0,57$

Pontuações A	B	C	D	E	F	G	H	I	J	
Teste X	1	2	4	3	6	7	9	8	10	5
Teste Y	4	3	6	1	7	2	5	9	10	8

Tabela 7.3: Valores dos dois testes nos 10 indivíduos

Exemplo 7.1.4. *Deseja-se saber qual relação existe os que tiveram ou não assistência psicológica* (X_3) *e o rendimento no trabalho* (X_4) *dos indivíduos considerados.*

O coeficiente de correlação aqui é o biserial pontual, pois temos uma variável contínua e outra dicotômica, logo aplicaremos uma correlação biserial-pontual (r_{bp})

Com os dados das variáveis assistência psicológica (X_3) e rendimento no trabalho (X_4).

1	X3	2	X4
1	18	2	7
1	17	2	10
1	1	2	9
1	12	2	4
1	15	2	2
1	6	2	7
1	20	2	16
1	9	2	8
1	10	2	3
1	13	2	9
1	11	2	4
1	2	2	2
1	19	2	10
1	14	2	13
1	15	2	10
1	9	2	4
1	8	2	10
1	11	2	3
$\sum parcial =$	216	$\sum parcial =$	131

Tabela 7.4: Variáveis utilizadas para cálculo correlacional biserial-pontual

Pode-se encontrar: $\sum X = 347$, $\bar{X} = 9,639$ e $S = 4,98$

p = proporção de assistência psicológica = 18/36 = 0,5

q = proporção de não assistência psicológica = 18/36 = 0,5

$\bar{X}_p$ = 216/18 = 12 (média no rendimento no trabalho dos que tiveram assistência psicológica)

$\bar{X}_q$ = 131/18 = 7,28 (média no rendimento no trabalho dos que não assistência psicológica)

$\bar{X}_t$ = 9,639 (média na variável contínua de todos os indivíduos)

$\bar{S}_t$ = 4,98 (desvio padrão da variável contínua)

$$r_{bp} = \frac{\bar{X}_p - \bar{X}_q}{S_t}\sqrt{p(1-p)} = \frac{12-7,28}{4,981}\sqrt{0,5(1-0,5)} = 0,474$$

Substituindo na segunda fórmula temos o mesmo resultado como se observa na equação abaixo.

$$r_{bp} = \frac{\bar{X}_p - \bar{X}_t}{S_t}\sqrt{p/q} = \frac{12-9,639}{4,981}\sqrt{0,5/0,5} = 0,474$$

A correlação entre X_3 e X_4 é moderado positiva.

Exemplo 7.1.5. *Deseja-se saber se existe relação entre o nível de ansiedade: (X_7) (alto ou baixo) e a compreensão de leitura (X_{11}) que tem os indivíduos.*

O coeficiente de correlação aqui é a correlação biserial, pois temos uma variável contínua e outra dicotomizada, logo aplicaremos uma correlação biserial (r_b).

Observe que depois de dicotomizar a variável nível de de ansiedade (0-30 pontos) utilizou-se como critério a média aritmética para a indicação dos indivíduos para cada categoria (ALTO para a pontuação acima da média; e BAIXO para pontuação abaixo da média). Assim, pode-se forma a seguinte tabela:

Variável nível de estresse		Compreensão de leitura
Continua	dicotomizada	
28	Alto	29
27	Alto	30
14	Baixo	10
23	Alto	21
24	Alto	24
14	Baixo	11
14	Baixo	12
18	Alto	8
14	Baixo	7
10	Baixo	5
5	Baixo	3
14	Baixo	6
30	Alto	28
18	Alto	12
15	Baixo	9
20	Alto	20
16	Baixo	16
8	Baixo	2
22	Alto	28
18	Alto	11
7	Baixo	4
12	Baixo	7
9	Baixo	5
8	Baixo	3
27	Alto	25
24	Alto	23
24	Alto	22
17	Alto	10
12	Baixo	7
18	Alto	15
14	Baixo	6
18	Alto	18
17	Alto	18
10	Baixo	6
16	Baixo	10
12	Baixo	2

Tabela 7.5: Variáveis utilizadas para cálculo correlacional biserial

Pode-se encontrar: $\sum X = Alto + Baixo = 342 + 131 = 473$, $\bar{X} = 9 = 13,139$ e $S = 8,619$

p = proporção de indivíduos com nível de estresse alto = 17/36 = 0,472

q = proporção de indivíduos com nível de estresse baixo = 19/36 = 0,528

$\bar{X}_p$ = $\bar{X}_p = 342/17 = 20,118$ (média da compreensão de leitura de todos os sujeitos com nível de estresse alto)

$\bar{X}_q$ = $\bar{X}_q = 131/19 = 6,895$ (média da compreensão de leitura de todos os sujeitos com nível de estresse baixo)

$\bar{X}_t$ = 13,139 (média de todos os sujeitos)

$\bar{S}_t$ = desvio padrão da variável contínua

y = 0,398 (obtém-se da tabela de probabilidade da distribuição normal entrando por p = 0,472 na área menor ou q = 0,528 na área maior da curva normal, que corresponde a umo mesmo Z normal padrão, cuja probabilidade na ordenada é y).

Substituindo esses valores na fórmula tem-se:

$$r_b = \frac{\bar{X}_p - \bar{X}_q}{S_t}\sqrt{\frac{pq}{y}} = \frac{20,118 - 6,895}{8,619}\sqrt{\frac{0,472.0,528}{0,398}} = 0,96$$

Observe que usando a fórmula abaixo se pode encontrar o mesmo resultado.

$$r_b = \frac{\bar{X}_p - \bar{X}_t}{S_t}\sqrt{\frac{p}{y}} = \frac{20,118 - 6,895}{8,619}\sqrt{\frac{0,472}{0,398}} = 0,96$$

A correlação entre as duas variáveis consideradas é alta e positiva.

Exemplo 7.1.6. *Deseja-se saber se existe relação entre a assistência psicológica ou não (X_3) (variável dicotômica) e o nível de aspirações (X_{15}) (variável também dicotômica). Conhecendo essas duas variáveis usa-se o coeficiente quádruplo Φ (que é usado quando as duas variáveis são dicotômicas).*

$$\Phi = \frac{(A.D) - (B.C)}{\sqrt{[(A+B)(A+C)(C+D)(B+D)]}}$$
$$= \frac{(12.11) - (6.7)}{\sqrt{[(18)(19)(18)(17)]}} = \frac{90}{323,45} = 0,278$$

	1	2	
1	12	6	12+6
2	7	11	7+11
	12+7	6+11	N =30

Os somatórios da linha (X_3) e da coluna (X_{15}), denominada também por frequências marginais, nos indica que um total de18 (12+6) indivíduos tiveram assistência psicológica.

Entre as variáveis consideradas existe uma correlação baixa e positiva para a amostra de 36 indivíduos.

Exemplo 7.1.7. *Deseja-se saber se existe relação entre as variáveis: nível intelectual (8) (Baixo e alto) e estilo cognitivo (13) (alto e baixo).*

Variáveis 8		Variável 13		Variáveis 8		Variável 13	
	Código		Código		Código		Código
43	1	17	1	11	2	10	2
30	1	16	1	33	1	18	1
18	2	9	2	28	1	16	1
23	1	14	1	24	1	12	1
19	2	8	2	20	2	7	2
22	1	11	1	13	2	5	2
26	1	13	1	14	2	6	2
20	2	12	1	27	1	13	1
19	2	10	1	22	1	8	2
18	2	8	2	28	1	13	1
12	2	4	2	19	2	6	2
17	2	9	2	18	2	5	2
29	1	15	1	20	2	7	2
22	1	11	1	20	2	8	2
20	2	6	2	24	1	14	1
25	1	16	1	21	2	12	1
15	2	8	2	18	2	5	2
				10	2	3	2
				14	2	4	2

Tabela 7.6: Variáveis utilizadas para cálculo tetracórico (duas variáveis dicotomizadas

Como temos duas variáveis dicotomizadas, o coeficiente mais apropriado é o tetracórico. Toma-se como critério para dividir cada variável em duas categorias a média aritmética. Assim, obtemos que por cima da média no nível intelectual se encontram 15 indivíduos abaixo da média encontram-se 21 indivíduos. A variável estilo cognitivo as frequências se distribuem por igual: tem 18 pontuações superiores e 18 inferiores. Temos então uma tabela de dubla entrada da seguinte forma:

	$> \bar{X}$	$< \bar{X}$	
$> \bar{X}$	14	4	18
$< \bar{X}$	1	17	18
	15	21	N =36

Utilizando as tabelas de Davidoff e Gohhen, o coeficiente de correlação tetracórico é obtido diretamente de tabelas, entrando nas mesmas com o valor do quociente:

$$T = \frac{(A.D)}{(B.C)} = \frac{(14.17)}{(4.1)} = \frac{238}{4} = 59,5$$

Que está entre 58,8 e 70,95 nas tabelas. A este intervalo corresponde em tabelas um valor de coeficiente de correlação: $r_t = 0,94$.

Existe forma de calcular o coeficiente de correlação tetracórico, ainda são sem duvida algo mais laborioso que o exposto anteriomente, porém tem que utilizá-lo quando não se dispõe de tabelas. Os procedimentos dos que falamos são os seguintes:

a) Quando os elementos A e B da tabela de dubla entrada representam a correlação positiva e, portanto, os elementos B e C, a correlação negativa.

	+	-
+	A (+ +)	B (+ -)
-	C (- +)	D (- -)

$$
\begin{aligned}
r_t &= \cos\left(\frac{180^o\sqrt{BC}}{\sqrt{BC}+\sqrt{AD}}\right) \\
&= \cos\left(\frac{180^o\sqrt{4.1}}{\sqrt{4.1}+\sqrt{14.17}}\right) \\
&= \cos\left(\frac{180^o 2}{2+15,427}\right) = \cos\left(\frac{360}{17,427}\right) = \cos(20,658) = 0,936
\end{aligned}
$$

Valor aproximadamente igual ao obtido pelo procedimento anterior.

b) Igualmente pode-se obter resolvendo a equação do segundo grau obtendo assim:

$$\frac{AD-BC}{N^2 yy'} = r_t + r_t^2 \frac{z.z'}{2}$$

Sendo os valores y e y' e z e z' obtido usando a tabela da distribuição normal, nas colunas da ordenada e da pontuação típica, respectivamente, com os valores de p e p', ou proporções sobre o total de cada uma das categorias das variáveis.

7.2 Exercícios

1. Considere um banco de dados adaptado formado por 36 indivíduos. A base de dados apresenta 15 variáveis de vários tipos para exemplificar as diversas medidas correlacionais. Encontre as medidas correlacionais estudadas.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
Ind1	1	1	1	20	1	1	29	30	71	32	29	22	17	7	1
Ind2	2	1	1	18	1	2	27	43	68	20	30	28	16	9	1
Ind3	3	1	1	7	2	3	14	18	38	30	10	15	9	6	1
Ind4	4	1	1	12	2	2	23	23	50	26	21	25	14	5	1
Ind5	5	1	1	15	3	1	24	19	59	37	24	19	8	3	1
Ind6	6	1	1	6	3	2	14	21	33	15	11	20	11	4	2
Ind7	7	1	2	12	1	2	14	26	32	21	12	27	13	7	1
Ind8	8	1	2	7	1	2	18	20	36	27	8	19	11	6	2
Ind9	9	1	2	9	2	1	14	19	39	20	7	16	10	4	2
Ind10	10	1	2	4	2	3	10	18	24	32	5	12	8	5	2
Ind11	11	1	2	2	3	2	9	12	20	16	3	9	4	1	2
Ind12	12	1	2	7	3	1	14	17	37	27	6	14	9	3	2
Ind13	13	2	1	20	1	1	30	29	62	30	28	25	15	7	1
Ind14	14	2	1	8	1	3	18	22	35	32	12	21	11	8	2
Ind15	15	2	1	11	2	2	15	20	41	27	9	25	6	7	1
Ind16	16	2	1	13	2	2	20	25	50	22	20	22	16	4	1
Ind17	17	2	1	15	3	1	16	15	43	21	16	11	8	2	2
Ind18	16	2	1	2	3	3	8	11	19	12	2	8	10	1	2
Ind19	19	2	2	16	1	1	22	33	55	29	28	26	18	5	1
Ind20	20	2	2	8	1	2	18	28	30	16	11	19	16	7	1
Ind21	21	2	2	3	2	3	7	24	24	10	4	23	14	5	2
Ind22	22	2	2	9	2	2	12	22	47	19	7	20	7	2	1
Ind23	23	2	2	2	3	3	9	13	21	22	3	10	5	4	2
Ind24	24	2	2	4	3	1	8	14	23	30	5	12	6	1	2
Ind25	25	3	1	19	1	2	27	27	68	34	25	21	13	7	1
Ind26	26	3	1	14	1	3	24	22	60	32	23	16	8	6	1
Ind27	27	3	1	15	2	1	24	28	70	39	22	18	13	4	1
Ind28	28	3	1	9	2	3	17	19	31	24	10	16	6	7	1
Ind29	29	3	1	8	3	2	12	18	28	20	7	13	5	2	2
Ind30	30	3	1	11	3	1	18	20	46	19	15	17	7	2	1
Ind31	31	3	2	10	1	3	14	20	50	13	6	14	8	5	1
Ind32	32	3	2	12	1	1	18	24	48	19	18	20	14	4	2
Ind33	33	3	2	10	2	2	17	21	37	26	18	19	12	3	1
Ind34	34	3	2	4	2	1	10	18	34	14	6	10	5	2	1
Ind35	35	3	2	10	3	1	16	10	42	18	10	7	3	1	1
Ind36	36	3	2	3	3	1	13	15	17	23	2	9	4	2	1

Tabela 7.7: Variáveis utilizadas para cálculos correlacionais

2. Determina-se 20 vezes o nível de glicose no sangue de uma

mesma amostra de pessoas com ansiedade por meio de dois métodos, A e B. Encontre o coeficiente de correlação entre A e B.

Tabela 7.8: Níveis de glicose nos métodos A e B

A	140	141	142	127	138	136	135	142	126	148	139	142	141	151	144	146	145	148	147	136
B	130	132	146	138	145	148	147	135	136	137	141	146	138	131	134	146	139	140	148	146

3. A seguinte tabela tem a variável Idade (X) e o coeficiente intelectual (Y) de 10 indivíduos.

Tabela 7.9: Idade e coeficiente intelectual

Idade	56	42	72	36	63	47	55	49	38	42
Coef.	148	126	159	118	149	130	151	142	114	141

Pede-se:

a) O coeficiente de correlação linear entre a Idade e do coeficiente intelectual

b) Usando o SPSS encontre a reta de regressão de Y sobre X.

4. Pretende-se averigua se o grau de ansiedade dos alunos de um determinado centro está relacionado com seus coeficiente de rendimento.

Tabela 7.10: Ansiedade e coeficiente de rendimento

Ans.	8	10	12	16	14	18	22	23	19	26	28	30	27	37
Rend.	7	10	5	8	9	7	6	7	9	7	8	9	6	4

Capítulo 8

Distribuições para variáveis contínuas

Cada variável aleatória vem identificada por sua função de probabilidade (discreta) ou por sua função densidade (contínua), tendo cada uma dela uma função de probabilidade ou de densidade que lhe é devida. Tal circunstância parece impedir um estudo sistemático das variáveis aleatórias mas, afortunadamente, grande maioria dos fenômenos da natureza seguem exato ou aproximadamente umas poucas leis bem conhecidas que são chamadas leis ou distribuições de probabilidade teórica. Cada uma delas é em realidade uma família de leis que, tendo a mesma forma, diferem umas das outras apenas em seus parâmetros. No que segue se estudam superficialmente as mais importantes na práticas. Primeiramente as distribuições contínuas e posteriormente, as discretas.

8.1 Distribuição Normal

A distribuição normal tem uma função densidade que se indica na Figura 8.1, cujas principais características são:

1. É simétrica sobre o eixo das abscissas. Qualquer valor de x entre $-\infty$ e $+\infty$ é teoricamente possível.

2. Tem uma forma de sino, o que junto ao nome do seu descubridor (Gauss) faz que às vezes se denomine curva de Gauss.

3. É simétrica em relação a vertical traçada em μ (sua média), o que acasiona que a esquerda e direita de μ fique uma área de 0,5 (50% da curva simétrica); neste caso, a média é também a mediana da distribuição.

4. A distância entre o eixo vertical em μ e o ponto de inflexão é σ (o desvio padrão de x). Quando mais grande seja σ, mas achatada é a curva de densidade.

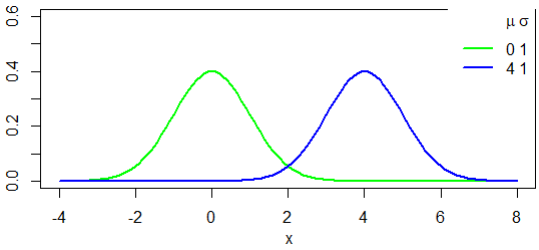


Figura 8.1: Curva de densidade da distribuição normal com diferentes médias e mesmo desvio padrão

- Uma variável X segue a distribuição Normal se sua curva de densidade tem forma de sino, simétrico em μ (média, moda e

mediana) e desvio padrão σ (a distância desde o eixo vertical em μ ao ponto de inflexão da curva). Portanto, como vimos, x toma valores entre $-\infty$ e $+\infty$. Indica-se por $x \sim N(\mu;\sigma)$, com μ e σ os dois parâmetros da distribuição.

- Padronização: Se X é uma variável qualquer de média μ e desvio σ, então $(X-\mu)/\sigma$ é a variável X padronizada com média 0 e desvio padrão 1, assim $Z = (X-\mu)/\sigma \sim N(\mu = 1;\sigma = 1)$.

- Tabela da distribuição normal: para cada probabilidade α a tabela de valor z_α de uma $N(0,1)$ tal que $P(-z_\alpha \leq z \leq +z_\alpha) = 1-\alpha$. Se $X \sim N(\mu;\sigma)$, os intervalos $X \in \mu \pm z_\alpha\sigma, X \leq \mu + z_\alpha\sigma$ e $X \geq \mu - z_{2\alpha}\sigma$ contem $(1-\alpha)100\%$ das observações de X.

- Teorema central do limite: Se X é uma variável qualquer de média μ e desvio padrão σ, se $\bar{X}$ é a média de uma amostra de tamanho $n \geq 30$, $\bar{X}$ se distribui aproximadamente como uma Normal: $\bar{X} \sim N(\mu;\sigma/\sqrt{n})$, com $\sigma/\sqrt{n}$ o erro padrão. Se X é Normal, o item anterior se verifica exatamente para qualquer valor de n.

Para calcular qualquer probabilidade é preciso dispor de uma tabela para cada $N(\mu,\sigma)$, algo que é impossível. Felizmente é possível converter qualquer variável $X \sim N(\mu,\sigma)$ em uma variável $Z \sim N(0,1)$ denominada normal padrão.

Exemplo 8.1.1. *As pontuações em um teste de aptidão escolar de uma escola geralmente são distribuído com uma média de 600 e uma variância de 10.000 (ou seja, desvio padrão $\sigma = 100$)?*

1. Que proporção de entrevistados tem uma pontuação abaixo de 300?

$$P(x<300)=P(\frac{x-\bar{x}}{s}<\frac{300-600}{100})=P(Z<-3)=1-0,9987=0,0013$$

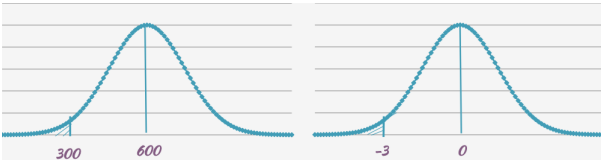


Figura 8.2: Proporção dos entrevistados que tem pontuação abaixo de 300

A proporção é de 0,0013.

2. Uma pessoa vai apresentar um teste, qual a probabilidade de obter uma pontuação de 850 ou superior?

$$P(x<300)=P(\frac{x-\bar{x}}{s}<\frac{850-600}{100})=P(Z<2,5)=1-0,9938=0,0062$$

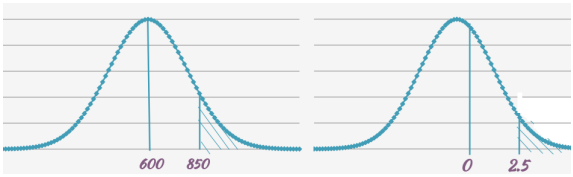


Figura 8.3: Proporção dos entrevistados que tem pontuação superior a 850

3. Qual proporção de pontuações estará entre 450 e 700?

$$P(450<x<700)=P(\frac{450-600}{100}<\frac{x-\bar{x}}{\sigma}<\frac{700-600}{100})=0,4332+0,3413=0,7745$$

8.1.1 Exercícios

1. A quantidade de homem de uma empresa com crise de ansiedade tem distribuição normal com média de 18,7 e desvio padrão 5. Calcule a probabilidade dessa quantidade seja menor que 21.

   Com $\mu = 18,7$, $\sigma = 5$ e a média amostra $\bar{x} = 21$. Resposta: $P(Z < 21) = 0,6772$

2. A quantidade de homem de uma empresa com crise de ansiedade tem distribuição normal com média de 18 e desvio padrão 4. Calcule a probabilidade dessa quantidade seja menor que 19.

3. O tempo médio de estudo de 500 estudantes homens de uma universidade é 151 horas semanais e o desvio padrão é de 15 horas semanais. Considere que o número de horas se distribuem numa normal. Encontre:

   a) A quantidade de estudente que estudam entre 120 e 155 horas semanais

   b) Quantos estudantes estudam mais de 185 horas semanais?

   c) Quantos estudantes estudam menos de 100 horas semanais?

   d) Quantos estudantes estudam entre 100 a 120 horas semanais?

4. O gasto médio de uma família que tem um filho autista segue uma distribuição normal com média 1500 reais e desvio padrão 167 reais. Encontre:

a) A porcentagem de gasto seja superior a 1700 reais. Resposta: 11,5%

b) A porcentagemde gasto esteja entre 1200 e 1600. Resposta: 68,9%

5. Um estudo experimental tem confirmado que as horas semanais de estudo dedicados por 500 estudantes do primeiro curso de uma faculdade se distribuem normalmente com média 25 e desvio padrão 9. Determinar:

   a) Que pocentagem de alunos dedica ao estudo entre 18 e 32 horas semanais? Resposta: 56%

   b) Quantos alunos estudam semanalmente mais de 35 horas? Resposta: 500 x 0,13 = 65 alunos.

6. Um professor realiza um teste de 100 itens a um curso de psicologia para 200 alunos. Suponha que as pontuações obtidas seguem uma distribuição normal com média 60 pontos e desvio padrão 10 pontos. Sendo Y o número de pontos obtidos ao fazer o teste. Pede-se:

   a) $P(Y \geq 70)$. (Resp. $P(Z \geq 1) = 0,1597$)

   b) $P(Y \leq 80)$. (Resp. $P(Z \leq 2) = 1 - P(Z \geq 2) = 1 - 0,0288 = 0,9772$)

   c) $P(Y \leq 30)$. (Resp. $P(Z \leq -3) = P(Z \geq 3) = 0,001352$)

   d) $P(Y \geq 46)$. (Resp. $P(Z \leq -1,4) = 1 - P(Z \geq -1,4) = 1 - P(Z \geq 1,4) = 1 - 0,080 = 0,9192$)

   e) $P(39 \leq Y \leq 80)$. (Resp. 0,9593)

f) $P(80 \leq Y \leq 82,5)$. (Resp. 0,9593)

g) $P(30 \leq Y \leq 40)$. (Resp. 0,02145)

h) $P((Y-60| \leq 20)$. (Resp. 0,9544)

i) $P((Y-60| \geq 20)$. (Resp. 0,0456)

j) Número de alunos que tiveram 70 pontos. Resp. 0,1587

8.2 t de Student

Quando uma variável segue uma distribuição normal, a média de uma amostra aleatória dessa variável também tem uma distribuição normal, e sua média é a média populacional desconhecida μ. Isso pode ser utilizado para estimar μ. Frequentemente, não se conhece o desvio padrão da população σ (apenas trabalha-se com uma amostra de indivíduos do total da população) e, ademas, pode ocorrer que o número de observações da amostra seja pequeno (menor de 30).

Nesse caso, pode-se utilizar o desvio padrão amostral (s) junto com a distribuição t de Student.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

A função densidade de probabilidade da distribuição t de Student vem dada pela seguinte expressão:

$$f(x) = \frac{1}{\sqrt{v\pi}} \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\frac{v}{2}} \left(1 + \frac{x^2}{v}\right)^{-\left(\frac{v+1}{2}\right)}$$

A distribuição t de Student pode ter diferentes formas dependendo dos graus de liberdade v. A aparencia geral da distribuição t tem caudas mais ampla que a normal, ou seja, a probabilidade dessas caudas é maior que na distribuição normal. A distribuição t se transforma em uma distribuição normal quando o número de dados tende a infinito.

Em psicologia, as aplicações da distribuição t de Student na inferência estatística são:

a) Para estimar,mediante intervalo de confiança, a média populacional.

b) Estimar e provar hipóteses sobre uma diferencia de médias.

As hipóteses ou premissas para poder aplicar a t de Student são que em cada grupo a variável estudada segue uma distribuição normal e que a dispersão em ambos grupos seja homogêneos (hipótese de homocedasticidade = igualdade de variânias) embora, pode-se usar sem assumir a igualdade de variâncias.

Com duas variáveis independentes X e Y, com distribuições: $X \sim N(0,1)$ e $Y \sim \chi^2_n$ define-se apartir dessas a variável aleatória (v.a.):

$$t = \frac{X}{\sqrt{\frac{Y}{n}}}$$

Com isso t tem uma distribuição t de Student com n graus de liberdade, ou seja, $t \sim t_n$.

Exemplo 8.2.1. *Dada em psicologia oito variáveis aleatórias* X_1, $X_2, \cdots, X_8$ *estocásticamente independentes e com distribuição* $N(0,1)$. *Define-se, a partir delas, três variáveis aleatórias X,Y e Z, de forma que* $X \sim t_5$, $Y \sim t_7$ *e* $Z \sim t_8$.

Para construir uma t de Student, atendendo a sua definição, terá que apartir de duas variáveis independentes, um com $N(0,1)$ e outra com distribuição χ^2 de Pearson com tantos graus de liberdade como se deseja que tenha a t de Student.

Assim, para definir uma v.a. X com distribuição t_5 serão necessárias, segundo exposto, uma $N(0,1)$ e uma χ^2_5, sendo ambas independentes. Para construir uma χ^2_5 utiliza-se as 5 primeiras variáveis $(X_1, X_2, X_3, X_4, X_5)$ (não poderia usar nenhuma delas como $N(0,1)$, pois são independentes).

Portanto, uma possiblidade seria indicar $X_8 \sim N(0,1)$ e $(X_1^2 + X_2^2 + X_3^2) + X_4^2 + X_5^2 \sim \chi^2_5$ a partir de ditas variáveis aleatórias (independentes).

$$X = \frac{X_8}{\sqrt{\frac{X_1^2+X_2^2+X_3^2)+X_4^2+X_5^2}{5}}} \sim t_5$$

Se $Y \sim t_7$ teríamos:

$$Y = \frac{X_1}{\sqrt{\frac{X_2^2+X_3^2+X_4^2+X_5^2+X_6^2+X_7^2+X_8^2}{7}}} \sim t_7$$

Por último, não é possível construir uma variável Z com distribuição t de Student com 8 graus de liberdade, já que são necessárias 9 normais reduzidas e independentes e só tem 8 variáveis.

8.2.1 Função de distribuição: tabelas

A função de distribuição de uma v.a. distribuida será dada por:

$$f(x) = P[X \leq x] = \frac{1}{\sqrt{n}\beta(\frac{1}{2}, \frac{n}{2})} \int_{-\infty}^{x} \left(1 + \frac{t^2}{n}\right)^{-\frac{(n+1)}{2}} dx$$

Esta função de distribuição tampouco pode se expressar explicitamente. Seus valores numéricos aparecem tabulados, junto com os de $p[X > x] = 1 - F(x)$. A tabela de cauda à direita recorre os valores $t_{n;\alpha}$ tal que:

$$P[X > t_{n,\alpha}] = \alpha$$

Sendo X uma v.a. com distribuição t_n. A estrutura tabela é similar a da distribuição χ^2. Na primeira coluna aparecem os graus de liberdade, enquanto que na primira fila aparecem valores da probabilidade α mais comens.

Exemplo 8.2.2. *Se* $t \sim t_{26}$*, calcular* $P(t > 1,706)$.

Para qualquer tipo de distribuição verifica-se que: $P(t > t_o) = 1 - P(t \leq t_o)$. Assim, obtem-se que:

$$P(t > 1,706) = 1 - P(t \leq 1,706) = 1 - 0,95 = 0,05$$

Exemplo 8.2.3. *Se* $t \sim t_8$*, calcular* $P(t \leq -0,889)$.

Tendo em conta que os valores negativos não aparecem na tabela, pode-se considerar que igual no caso da distribuição normal $N(0,1)$ por simetria da distribuição:

$$P(t \leq -0,889) = P(t \geq 0,889) = 1 - P(t \leq 0,889)$$
$$P(t \leq -0,889) = 1 - F(0,889) = 1 - 0,8 = 0,2$$

Exemplo 8.2.4. *Se* $t \sim t_7$*, calcular* $P(t \leq -0,889)$.

Exemplo 8.2.5. *Se* $t \sim t_{200}$*, calcular* $P(t \leq 0,9)$.

Como 200 > 30 não tem na tabela, deve-se usar uma aproximação da distribuição t de Student pela Normal: $P(t > 0,9) = P(T > 0,9)$. Sendo $T \sim N(0,1)$. Assim:

$$P(t > 0,9) = P(T \geq 0,9) = 1 - P(T \geq 0,9) = 1 - 0,8159 = 0,1841$$

Exercício 8.2.1. *Se $t \sim t_{180}$, calcular $P(t \leq 0,8)$.*

Exemplo 8.2.6. *Se $t \sim t_7$, obter t_o tal que $P(t \leq t_o) = 0,99$.*

Dado que n = 7 e $F(X) = 0,99$ está contido na tabela t de Student. Pede-se o valor da variável ao que corresponde dita função de distribuição. Assim $t_o = 2,998$.

Exercício 8.2.2. *Se $t \sim t_8$, obter t_o tal que $P(t \leq t_o) = 0,99$.*

Exercício 8.2.3. *Considere que uma variável psicológica aleatória segue uma distribuição t de Student. Pede-se calcular os pontos críticos:*

a) $t_{0,20;20}$ (Resp. 0,8660)

b) $t_{0,99;10}$ (Resp. -2,764)

c) $t_{0,25;10}$ (Resp. 0,711)

Exercício 8.2.4. *Considere que uma variável psicológica aleatória segue uma distribuição t de Student. Pede-se calcular as probabilidades:*

a) $P(t_{10} \geq 1,372)$

b) $P(t_8 \leq 1,2)$

c) $P(-0,5 \leq t_6 \leq 0,6)$

d) $P(|t_{24}| > 2)$

Exemplo 8.2.7. *Dado $n = 16$ (tamanho da amostra), $\mu = 12$ (Média populacional), $\bar{X} = 16,4$ (média amostra) e $S = 2.1$ (desvio padrão amostral). Encontre o valor t.*

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{16,4 - 12}{2,1/\sqrt{16}} = 8,38$$

Exercício 8.2.5. *Dado $n = 25$ (tamanho da amostra), $\mu = 10$ (Média populacional), $\bar{X} = 15,7$ (média amostra) e $S = 2,93$ (desvio padrão amostral). Encontre o valor t.*

Exercício 8.2.6. *Considere um gráfico da distribuição t de Student com 9 graus de liberdade. Encontre o valor de t_1 para que:*

a) A área sombreada da direita seja igual a 0,05. (Resp. 1,83)

b) Toda área sombreada seja 0,05. (Resp. 2,26)

c) Toda área não sombreada seja 0,99. (Resp. $t_{0,975} = 3,25$)

d) A área sombreada à esquerda seja 0,01. (Resp. $t_{0,99} = 2,82$)

e) A área à esquerda de t_1 seja 0,90. (Resp. $t_{0,90} = 1,38$)

Exercício 8.2.7. *Se X é uma v.a. com distribuição t_{25}, a abscissa que deixa a sua direita uma probabilidade 0,20 é $x = 0,856$. Para obter a abscissa que deixa a sua esquerda uma probabilidade 0,60, pode-se ter em conta que a sua direita deixará uma probabilidade 0,40 para obter que essa abscissa é 0,256. A simetria da distribuição t proporciona uma ferramenta adicional para manejar a tabela t.*

8.3 Qui-quadrado

A função densidade da distribuição qui-quadrada (χ^2) descreve-se por meio da seguinte expressão:

$$f(x) = \frac{1}{2^{\frac{v}{2}}\Gamma\left(\frac{v}{2}\right)} e^{-x/2} X^{\frac{v}{2}-1}$$

Em que V são os graus de liberdade e X não é negativo. A diferença do que ocorria com a distribuição normal, devido a que a distribuição χ^2 pode ter diferentes formas dependendo dos graus de liberdade.

O valor da variável que deixa a sua direita uma área α sob a curva de densidade chama-se o ponto crítico correspondente ao nível de significância α. São três as aplicações principais que tem a distribuição χ^2:

1. O teste de bondade de ajuste consiste na abordagem de até que ponto uma amostra pode-se considerar como pertencente a uma população com uma distrinuição teórica já conhecida. É um método que se utiliza frequentemente para determinar se uma serie de dados apresenta uma distribuição normal, de Poisson etc.

2. O teste de independencia determina se dois caracteres X e Y de uma população são dependentes ou independentes.

3. O teste de homogeneidade permite determinar se varias amostras que estudam o mesmo caráter U tem sido tomado ou não da mesma população, em relação a dita característica U.

Exemplo 8.3.1. *Preencher a tabela com valores críticos a 1% e a 5% com os graus de liberdades 2,5 10 e 20 de uma distribuição*

qui-quadrada.

Graus de liberdade (g.l.)	1%	5%
2	9,210	5,991
5	15,086	11,070
10	23,209	18,307
20	37,566	31,410

Exercício 8.3.1. *Preencher a tabela com valores críticos a 5% e a 10% com os graus de liberdades 3,5 10 e 20 de uma distribuição qui-quadrada.*

Graus de liberdade (g.l.)	5%	10%
3		
5		
10		
20		

Exercício 8.3.2. *Considere um gráfico de uma distribuição qui-quadrada com 8 graus de liberdade. Usando uma tabela da distribuição qui-quadrada. Pede-se:*

a) A área sombreada da direita seja igual a 0,05.

b) Toda área sombreada seja 0,20.

c) Toda área não sombreada seja 0,99.

d) A área sombreada à esquerda seja 0,01.

e) A área sombreada à esquerda seja 0,50.

e) A área sombreada à direita seja 0,50.

8.4 F de Fisher-Snedecor

A função de densidade de probabilidade da distribuição F de Fisher-Snedecor vem dada pela seguinte expressão:

$$f(x) = \frac{\Gamma\left(\frac{v+w}{2}\right)}{\Gamma\frac{v}{2}\Gamma\frac{w}{2}} v^{\frac{v}{2}} w^{\frac{w}{2}} \frac{x^{\frac{v}{2}-1}}{(w+vx)^{\frac{v+w}{2}}}$$

Em que v e w são os graus de liberdade do numerador e denominador respectivamente, sendo x não negativo. Ao depender dos graus de liberdade, a função densidade pode ter várias formas. Essa distribuição é usada principalmente em dois tipos de situações, exigindo em ambos casos que a distribuição das variáveis seja normal:

1. Para provar se duas amostras vir de populações que possuem variâncias iguais. Esta prova é útil para determinar se uma população normal tem uma maior variação que a outra e é importante já que na hora de cmparar médias, varias estatísticas representam como requisito de homogeneidade de variâncias.

2. Tmbém se aplica quando se trata de comparar simultaneamente varias médias populacionais (para variáveis qualitativas e para variáveis quantitativas).

Exemplo 8.4.1. *Se* $F \sim F_{9,10}$. *Calcular* $P(F > 3,02)$.

Dado que m =9 e n = 10 são valores que se encontram na tabela F de Snedecor. Precisa-se comprovar em que valor dela está o valor $F_o = 3,02$ da variável (para os graus de liberdade considerados para numeradores e denominadores). Para $\alpha = 5\%$, tem-se que: $P(F > 3,02) = \alpha = 0,05$.

Exercício 8.4.1. *Se* $F \sim F_{8,9}$. *Calcular* $P(F > 4)$.

Exercício 8.4.2. *Se* $F \sim F_{4,10}$. *Calcular* $P(F > 5)$.

Exemplo 8.4.2. *Se* $F \sim F_{9,40}$. *Calcular* F_o *de forma que* $P(F \geq F_o) = 0,95$.

Sabe-se que: $\alpha' = 1 - \alpha$ e

$$\frac{1}{F_{m,n,\alpha'}} = F_{m,n,\alpha} = F_{m,n,1-\alpha'}$$

Como $F_{40,9,0,05} = 2,83$ temos que:

$$F_o = F_{9;40;0,95} = \frac{1}{2,83} = 0,35336$$

Exercício 8.4.3. *Se* $F \sim F_{8,30}$. *Calcular* F_o *de forma que* $P(F \geq F_o) = 0,90$.

Exercício 8.4.4. *Se* $F \sim F_{7,20}$. *Calcular* F_o *de forma que* $P(F \geq F_o) = 0,99$.

Capítulo 9

Distribuições para variáveis discretas

9.1 Binomial

Uma variável apresenta uma distribuição binomial quando apenas tem dois possíveis resultados: sucesso e fracasso, sendo a probabilidade de cada um deles constante em uma série de repetições, ou seja, nem a probabilidade de sucesso nem a de fracasso mudam de um teste a outro, e ademais o resultado de cada teste é independente do resultado dos demais testes. A probabilidade de sucesso é representada por p e a probabilidade de fracaso por $q = 1 - p$.

No caso de variáveis discretas em lugar da função de densidade se utiliza a função de probabilidade ou de quantia, que dá uma probabilidade para cada valor da variável. A função de probabilidade binomial vem expressada pela seguinte forma:

$$f(x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

Exemplo 9.1.1. *A aplicação de um determinado tratamento de um transtorno de ansiedade a um grupo de indivíduos foi de 67% dos casos de uma empresa. Aplica-se o tratamento em 8 indivíduos.*

O valor de p = 0,67 e, portanto, o valor de q = 33, pois p + q =1.

- Qual é a probabilidade de que melhoram do transtorno de ansiedade 7 indivíduos?

$$\frac{8!}{7!(8-7)!} 0,67^7 0,33^{8-1} = 0,16$$

- Qual é probabilidade de que ao menos melhoram 3 indivíduos?

$$1 - \frac{8!}{7!(8-7)!} 0,67^7 0,33^{8-2} - \frac{8!}{1!(8-1)!} 0,67^7 0,33^{8-1} - \frac{8!}{0!(8-0)!} 0,67^0 0,33^{8-0} = 0,98$$

Exercício 9.1.1. *A aplicação de um determinado tratamento de depressão a um grupo de alunos foi de 72% dos casos de uma escola. Aplica-se o tratamento em 10 alunos. Pede-se:*

a) A probabilidade de que melhoram da depressão 8 alunos?

b) A probabilidade de que ao menos melhoram 5 alunos?

9.2 Hipergeométrica

Na distribuição hipergeométrica a variável também é aleatória e dicotômica como na distribuição binomial, mas se diferencia desta última em duas característica importantes: a população é finita, enquanto que na binomial pode ser infinita e, ademais, a probabilidade p muda - não é constante - já que o resultado de cada teste depende do resultado das anteriores. A função de probabilidade vem expressda pela seguinte forma:

$$f(x) = \frac{\frac{N_p!}{x!(N_p-x)!}\frac{N_q!}{(n-x)!(N_q-n+x)!}}{\frac{N!}{n!(N-n)!}}$$

Em que N_p e N_q são o número de elementos com probabilidade inicial p e q, respectivamente, ou seja, $p = N_p/N$ e $q = N_q/N$, N é o número total de elementos e n o número de elementos da amostra extraída dos N da população.

Exemplo 9.2.1. *Numa clínica 30 pacientes se tem comprovado que 8 estão com depressão. Se tem escolhido 4 pacientes nessa clínica.*

a) Qual a probabilidade de que ao menos um dos pacientes tenha depressão?

Assim, $N_p = 30 - 8 = 22$, $N_q = 8$, $n = 4$, $x = 4$ e $n = 4$. Substituindo os valores, temos:

$$\frac{\frac{22!}{4!(22-4)!}\frac{8!}{(4-4)!(8-4+4)!}}{\frac{30!}{4!(30-4)!}} = 0,27$$

b) Qual a probabilidade de que 3 dos pacientes estão com depressão?

$$\frac{\frac{22!}{1!(22-1)!}\frac{8!}{(4-1)!(8-4+1)!}}{\frac{30!}{4!(30-4)!}} = 0,04$$

Exercício 9.2.1. *Numa clínica 40 pacientes se tem comprovado que 11 estão com depressão. Se tem escolhido 5 pacientes nessa clínica. Pretende-se saber:*

a) A probabilidade de que ao menos um dos pacientes tenha depressão?

b) A probabilidade de que 3 dos pacientes estão com depressão?

Exercício 9.2.2. *Numa clínica 50 pacientes, 35 são homens e 15 mulheres. Forma-se um grupo composto de 8 pessoas escolhidas aleatoriamente.*

a) Tabular a distribuição do número de homens que compõe o grupo. Comprove que a soma das probabilidades vale 1.

```
N1<- 35  # Número de homens
N2<-15   # Número de mulheres
n<- 8   # Tamanho da amostra
x<- 0:8
hyper<- dhyper(x,N1,N2,n) # Probabilidades
cbind(x, hyper)
sum(hyper)
```

b) Representar graficamente as funções de quantidade e de distribuição do número de homens que formam o grupo.

```
par(mfrow=c(1,2))
barplot(hyper, names.arg=x, xlab="x",
        ylab="Pr(X=x)",
main = "Função de densidade",
        ylim=c(-0.01,0.4))
box(col="black")
hyper_acum<- phyper(x,N1,N2, n)
# Probabilidades acumuladas
barplot(hyper_acum, names.arg=x,
        xlab="x", ylab="F(x)",
        main="Função de distribuição",
        ylim = c(-0.01, 1.01))
box(col="black")

dhyper(4,N1,N2,n)
# 0,1331
hyper[x==4]
# 0,1331
```

c) Encontrar a probabilidade de que o grupo esteja formado por 4 homens e 4 mulheres

```
dhyper(4,N1,N2,n)
# 0,1331
hiper[x==4]
# 0,1331
```

9.3 Poisson

Um processo de Poisson é um processo de sucessos independentes que se caracteriza por:

- O número de sucessos em dois intervalos distintos sempre é independente
- A probabilidade de que um sucesso ocorra em um intervalo infinitamente é proporcional a longitude do intervalo.
- A probabilidade de que ocorra mais de um sucesso em um intervalo muito pequeno v é 0.
- Os sucessos são expressos por uma unidade de área, tempo etc.

A distribuição de Poisson descreve o número de sucessos em uma unidade de tempo de um processo Poisson. Muitos fenômenos se modelam como um processo de Poisson, por exemplo o número de estressados em um ambiente conflitante. As diferenças mais importantes em relação a distribuição binomial são que esta distribuição aplica-se a sucessos que podem ter uma probabilidade muito baixa e, ademais, o tamanho de n é infinito. A função de probabilidade da distribuição Poisson se expressa pela seguinte fórmula:

$$f(x) = \frac{\lambda^x}{x!e^\lambda}$$

Em que λ é a média de sucessos por unidade de tempo e x é a variável que indica o número de sucessos.

Exemplo 9.3.1. *A quantidade de pessoas ansiosas em uma zona de guerra é de 23 indivíduos por* $100m^2$. *Como trata-se de sucessos por unidade de área se utiliza uma Poisson.*

a) Qual é a probabilidade de não encontrar nenhum indivíduo estressado em $25m^2$?

$$\lambda = \frac{23.2m^2}{100m^2} = 5,75$$

A probabilidade que buscamos será:

$$\frac{\lambda^x}{x!e^{\lambda}} = \frac{5,75^0}{0!e^{5,75}} = 0,003$$

b) O número de pessoas deprimidas observados foi de 120 em 30 dias. Qual é a probabilidade de ver 5 pessoas deprimidas em 10 dias?

$$\lambda = \frac{120.10}{30} = 40$$

Logo, a probabilidade que buscamos será:

$$\frac{\lambda^x}{x!e^{\lambda}} = \frac{40^5}{5!e^{40}} = 3,61.10^{-12}$$

```
lambda = 5.75
lambda = 120*10/30; x<-0; x<-5
# a probabilidade será: P(X=x)
# a probabilidade será: P(X>=x)
# a probabilidade será: P(X>x)
# a probabilidade será: P(X<=x)
# a probabilidade será: P(X<x)
```

Exemplo 9.3.2. *Na tabela abaixo mostra a quantidade de psicólogo numa clínica e o número de pessoas não curadas psicologicamente.*

Tabela 9.1: Quantidade de psicólogos e pessoas não curadas

Número de psicólogos	0	1	2	3	4	5
Número de pessoas não curadas	120	200	140	20	10	2

Pede-se:

a) Ajustar a uma distribuição de Poisson

b) Calcular a probailidade com que chega $\lambda = 0,1,2,3,4,5$ número de pessoas não curadas psicologicamente.

A distribuição de Poisson apenas depende de um parâmetro λ que coincide com a média. Assim, para ajustar esta distribuição à Poisson é necessário calcular a média. Faça x o número de psicólogos e n_i a quantidade de pessoas não curados (considere num período de tempo).

Tabela 9.2: Tabela de distribuição

x	n_i	$x_i n_i$
0	120	0
1	200	200
2	140	280
3	20	60
4	10	40
5	2	10
	n =492	590

A média será: $\bar{x} = \frac{\sum x_i n_i}{n} = \frac{590}{492} = 1,2$. Logo, $\lambda = 1,2$ e, portanto,

$$P(x = k) = \frac{1,2^k}{k!}$$

As probabilidades com que chega k = 1,2,3,4,5 serão:

Tabela 9.3: Cálculos das probabilidades

x	0	1	2	3	4	5
p(x=k)=pk	0,3012	0,3614	0,2169	0,0867	0,0260	0,0062

Ou seja, a probabilidade que alcance 5 pessoas não curada psicologicamente é $p(x = 5) = 0,0062$. A probabilidade de que 1 seja curado será $p(x = 1) = 0,3614$.

Exercício 9.3.1. *Na tabela abaixo mostra a quantidade de psicólogo numa clínica e o número de pessoas não curadas psicologicamente.*

Tabela 9.4: Quantidade de psicólogos e pessoas não curadas

Número de psicólogos	0	1	2	3	4	5
Número de pessoas não curadas	100	210	150	25	8	3

Pede-se:

a) Ajustar a uma distribuição de Poisson

b) Calcular a probailidade com que chega $\lambda = 0,1,2,3,4,5$ número de pessoas não curadas psicologicamente.

Exemplo 9.3.3. *Uma editora edita 50000 exemplares de um mesmo livro de psicologia. Observando os exemplares anteriores sabe-se*

que a probabilidade de que tenha algum problema na impressão é de 0,0001. Pede-se encontrar a probabilidade de que:

a) não tenha nehum erro de impressão.

b) tenha 3 livros de psicologia com erro de impressão.

Os dados nos diz que n = 50000 e p = 0,0001. Cada livro é um sucesso independentemente do outro, o número n é grande e a probabilidade p é pequena. Utiliza-se a distribuição de Poisson de função de probabilidade.

$$P(X = x) = \frac{\lambda^x}{x!e^{\lambda}}$$

Sabe-se que $\lambda = np = 50000.0,0001 = 5$ e $P(X = x) = \frac{5^x}{x!e^5}$. Portanto,

$$P(X = 0) = \frac{\lambda^0}{0!e^5} = 0,00674$$

$$P(X = 3) = \frac{5^3}{3!e^5} = \frac{125}{6}0,00674 = 0,1404$$

Exercício 9.3.2. *Uma editora edita 40000 exemplares de um mesmo livro de psicologia. Observando os exemplares anteriores sabe-se que a probabilidade de que tenha algum problema na impressão é de 0,002. Pede-se encontrar a probabilidade de que:*

a) não tenha nehum erro de impressão.

b) tenha 4 livros de psicologia com erro de impressão.

Exercício 9.3.3. *A quantidade de pessoas ansiosas em um pequeno bairro é de 41 indivíduos por* $100m^2$. *Como trata-se de sucessos por unidade de área se utiliza uma Poisson.*

a) Qual é a probabilidade de não encontrar nenhum indivíduo estressado em $30m^2$?

b) O número de pessoas deprimidas observados foi de 130 em 35 dias. Qual é a probabilidade de ver 5 pessoas deprimidas em 12 dias?

Exercício 9.3.4. *O número de erros cometidos por uma pessoa que tem crise de ansiedade em uma jornada diária segue uma distribuição de Poisson de média 4.*

a) Qual é a probabilidade de que em uma jornada diária cometa ao mesmos um erro?

A distribuição de Poisson representa o número de ocorrências de um sucesso por unidade de tempo. A função de quantidade de uma variável aleatória de Poisson vem dada por:

$$f(x) = \frac{\lambda^x}{x!e^\lambda}$$

Para $x = 1,2,3,\cdots$. Sendo λ um parâmetro que representa tanto a média como a variância da distribuição. X é o número de erros que segue uma distribuição de Poisson de média 4 (parâmetro lambda). A probabilidade de cometer ao menos um erro é:

$$\begin{aligned} Pr(X \geq x) &= 1 - Pr(X < 1) = 1 - Pr(X = 0) \\ &= 1 - dpois(0, lambda = 4) \\ &= 0,9816 \end{aligned}$$

b) Qual a probabilidade de que em uma jornada cometa x = 0,1,2,3,4 e 5 erros?

Trata-se de avaliar a função de quantidade para diversos valores de x. Para isso devemos considerar no programa R:

```
x = 0:5
lambda = 4
Resposta = cbind(x, dpois(x,lambda))
```

Índice Remissivo

Referências Bibliográficas

Ávila, M. J. del M., García, J. M. T. *Técnicas Estadísticas Aplicadas*. Grupo Editorial Universitário, 2006.

Ávila, M. J. del M. *Estadística Matemática*. Grupo Editorial Universitário, 2006.

Díaz, M. J.F., Ramos, J.M.G., Vicente, A.F. e Muñoz, I.A. *Resolución de Problemas de Estadística Aplicada e las Ciencias Sociales. Guia práctica para profesores y alumnos*. Madrid: Editorial Sinteses, 1990.

Faceli, K., Lorena, A. C., Gama, J., Carvalho, A. C. P. L. F. *Inteligência artificial: uma abordagem de aprendizado de máquina*. LTC, 2011.

Gómez, Félix Calvo e Begoña Abad Miguélez *Ejercicios resueltos de estadística*. Bilbao: Universidad de Deusto, 1991.

Fernández-Abascal, H., Guijarro, M. M., Rojo, J. L., Sanz, J. A. *Cálculo de probabilidades y estadística*. Barcelona: Editorial Arial, S. A., 1994.

Gross, J., Ligges, U. *nortest: Tests for Normality*. R package version 1.0-4, 2015. http://CRAN.R-project.org/package=nortest

Ligges, U., Mächler, M. *Scatterplot3d - an R Package for Visualizing Multivariate Data*. Journal of Statistical, Software 8(11), 1-20, 2003.

Maindonald, J. H., Braun, W. J. *DAAG: Data Analysis and Graphics Data and Functions*. note = R package version 1.22, 2015. http://CRAN.R-project.org/package=DAAG

Paradis, E. *R for Beginners*. Institut des Sciences de l'Évolution, Université Montpellier II, France, 2005. https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna: Austria, 2015. http://www.R-project.org/.

Rmetrics Core Team., Wuertz, D., Setz, T., Chalabi, Y. *fBasics: Rmetrics - Markets and Basic Statistics*. R package version 3011.87, 2014, http://CRAN.R-project.org/package=fBasics.

Sarkar, D., Andrews, F. *latticeExtra: Extra Graphical Utilities Based on Lattice*. R package version 0.6-28, 2016. http://CRAN.R-project.org/package=latticeExtra

Van der Loo, M.P.J. *extremevalues, an R package for outlier detection in univariate data*. R package version 2.3, 2010.

Venables, W.N., Ripley, B.D. Package *MASS*. *Modern Applied Statistics with S*. Fourth Edition. Springer, New York, 2002. ISBN 0-387-95457-0

Organizador e Autores

Dr Edwirde Luiz Silva Camêlo (Brasil) - (Organizador) Professor Associado da Universidade Estadual da Paraíba (UEPB). Pós-doutorado em *Estadística Aplicada* (2016) e Doutor em *Estadística e Investigación Operativa* (2007) pela *Universidad de Granada*. Mestrado em Biometria e Estatística Aplicada (2001) pela Universidade Federal Rural de Pernambuco (UFRPE). Técnicas de estatística multivariada aplicada em diversas áreas tem sido sua principal linha de pesquisa. E-mail: edwirde@uepb.edu.br

Dra Dalila Camêlo Aguiar (Brasil) - Doutoranda em *Estadística Matemática y Aplicada* pela *Universidad de Granada* (UGR), Mestra em *Estadística Aplicada* (2016) pela UGR, Especialista em Estatística Aplicada (2011) pela Fundação de Apoio, Pesquisa e Extensão (FURNE) e Bacharela em Estatística (2010) pela Universidade Estadual da Paraíba (UEPB). Pesquisadora na área de estatística multivariada aplicada. E-mail: dalilacamelo@correo.ugr.es

Dr Ivan Olier (English) His is a *Reader in Artificial Intelli-*

gence and Data Science. My research interests lie in algorithms for Artificial Intelligence, with a specific focus on Causal AI, Digital Twins, and the modelling of large-scale, highly structured, and/or relational data (including multivariate time series, graphs, networks, etc.). My research finds its applications in various domains such as cardiovascular science, drug development, bioinformatics, healthcare, astrophysics, engineering, etc. Additionally, I hold a senior membership position at the Liverpool Centre for Cardiovascular Science (LCCS). E-mail:
I.A.OlierCaparroso@ljmu.ac.uk.

Dr Ramón Gutiérrez Sánchez (Espanha) - Professor da *Universidad de Granada* (UGR), Secretário do Departamento de *Estadística e IO*. Doutor desde 2005, pertence à linha de pesquisa de Análise Multivariada e Processos Estocásticos. Publicou mais de 40 artigos em *JCR* na área de Estatística. Também colabora com o Departamento de Parasitologia da UGR. E-mail: ramongs@ugr.es

A estatística básica e sua prática constitui uma introdução à estatística para estudantes de estatística e Psicologia I e II e também os tópicos especiais em estatística para o mestrado de Pós-Graduação em Psicologia da Saúde da Universidade Estadual da Paraíba. Em linhas gerais esse livro dar ênfase ao pensamento estatístico aplicado a psicologia; mais dados e conceitos, menos teorias e menos receitas; promover o ensino ativo que é caracterizado como métodos de ensino que fortalecem a formação de vínculos democráticos na relação entre professor, alunos e conteúdo.

30 anos
eduepb