Estatistica - Volume I

Teoria e exercicios passo-a-passo

O Volume I deste manual apresenta as estatisticas paramétricas e o Volume II as não-paramétricas.

Margarida Pocinho

01-01-2009

Índice Geral

I - INTRODUÇÃO 5

1. NOÇÕES GERAIS 9

2. POPULAÇÃO E AMOSTRA 11

3. MÉTODOS DE AMOSTRAGEM 12

3.1 Amostragens Probabilísticas e Não-Probabilísticas 12

3.2 Determinação do Tamanho da Amostra 16

3.3 Indivíduo ou Unidade Estatística 21

3.4 Variáveis 22

4. ESTATÍSTICA DESCRITIVA 27

4.1 Parâmetro e dado estatístico 28

4.2 Representação de uma variável estatística 28

4.3 Redução de uma variável estatística 31

Medidas de dispersão 39

5. CARACTERÍSTICAS DA DISTRIBUIÇÃO NORMAL 49

5.1 A CURVA NORMAL E OS DESVIOS-PADRÃO 51

6. TESTES ESTATISTICOS 55

6.1. TESTES PARAMÉTRICOS PASSO-A-PASSO 62

6.1.1 Teste t de Student (não relacionado) 62

6.1.2 Teste t de Student (relacionado) 65

6.1.3 Correlação momento-produto de brawais-pearson 68

6.1.4 Análise da variancia de um critério (ANOVA) 74

Índice de Ilustrações

FIGURA 1: POPULAÇÃO E AMOSTRA 11

FIGURA 2: AMOSTRA ESTRATIFICADA 14

FIGURA 3: CONVERSÃO DOS NÍVEIS DE CONFIANÇA EM DESVIOS PADRÃO 17

FIGURA 4: VARIÁVEIS PRIMARIAS E DERIVADAS 25

FIGURA 5: QUARTIS 39

FIGURA 6: CURVA SIMÉTRICA ACHATADA (PLATOCURTICA) 41

FIGURA 7: CURVA SIMÉTRICA MESOCURTICA 41

FIGURA 8: DISTRIBUIÇÃO NORMAL 50

FIGURA 9: IDENTIFICAR OS TESTES ESTATISTICOS 59

FIGURA 10: DIAGRAMAS DE DISPERSÃO DE PONTOS, SCATTERPLOT OU SCATTERGRAM 69

FIGURA 11: DIAGRAMAS DE DISPERSÃO : CAUSA & EFEITO 70

Índice de tabelas

TABELA 1: DISTRIBUIÇÃO DE FREQUENCIAS 29

TABELA 2: EXERCÍCIO DE DISTRIBUIÇÃO DE FREQUENCIAS 29

TABELA 3: CÁLCULO DO DESVIO MÉDIO 43

TABELA 4: CÁLCULO DO DESVIO MÉDIO PARA CLASSES 44

TABELA 5: CÁLCULO DA VARIÂNCIA 45

TABELA 6: EXERCICIO - CÁLCULO DA VARIÂNCIA 46

TABELA 7: GRELHA DE DECISÃO DOS TESTES 61

TABELA 8: CÁLCULO DO VALOR T 64

TABELA 9: VALORES CRÍTICOS T DE STUDENT 66

TABELA 10: CÁLCULO DO TESTE T EMPARELHADO 67

TABELA 11: CÁLCULO DO R DE PEARSON 72

TABELA 12: CÁLCULO DA ANOVA PARA TAMANHOS IGUAIS 76

TABELA 13: APRESENTAÇÃO DA ANOVA 78

TABELA 14: CÁLCULO DA DIFERENÇA MÍNIMA SIGNIFICATIVA - TUKEY 79

TABELA 15: CÁLCULO DA ANOVA PARA TAMANHOS IGUAIS 80

TABELA 16: TESTE POST-HOC -TUKEY 82

I - INTRODUÇÃO

Desde séculos o homem tem, muitas vezes, tomado notas de coisas e de pessoas, não com o único fim de acumular números, mas com a esperança de utilizar os dados do passado para a resolução de problemas do presente assim como para a previsão de acontecimentos futuros. No entanto, o sucesso quanto a este objectivo só foi possível em data muito recente: só no final do século XIX e, sobretudo, no princípio do século XX é que, com a aplicação de probabilidades aos problemas sobre a interpretação dos dados recolhidos, foi possível resolver alguns deles.

O jogo foi o motor de arranque e o primeiro beneficiado com as probabilidades. De facto, por volta de 1200 a.C. existiam dados com forma cúbica feitos a partir de ossos. No entanto, o jogo atingiu uma grande popularidade com os gregos e os romanos. Na Idade Média, a igreja católica era contra o jogo dos dados, não pelo jogo em si, mas pelo vício de beber e dizer palavrões que acompanhavam os jogos. Os jogadores inveterados do século XVI procuravam cientistas de renome para que estes lhes dessem fórmulas mágicas para garantir ganhos substanciais nas mesas de jogo.

O contributo decisivo para o início da teoria das probabilidades foi dada pela correspondência trocada entre os matemáticos franceses Blaise Pascal e seu amigo Pierre de Fermat, em que ambos, por diferentes caminhos, chegaram à solução correcta do célebre problema da divisão das apostas em 1654.

Quis o acaso que o austero Pascal conhecesse Méré, jogador mais ou menos profissional, que lhe contava as suas disputas com os adversários em problemas de resolução controversa sobre dados e apostas. Um desses problemas veio a interessar Pascal[1]. Depois de reflectir sobre ele, trocou uma interessante correspondência sobre o assunto com o matemático Fermat, seu amigo. Essas cartas históricas, que contêm as reflexões conjugadas de ambos, são os documentos fundadores da Teoria das Probabilidades.

[1] Em meados do século XVII, o jogador francês, o "Chevalier de Méré", que vinha calmamente ganhando a vida apostando o seu bom dinheiro em jogos de dados, decidiu oferecer a mesma quantia para uma aposta diferente. Vinha garantindo, de início, um seis em quatro jogadas de um só dado; passou, então, a apostar que conseguiria pelo menos um duplo seis em vinte e quatro jogadas de dois dados. Mas, percebeu que os seus lucros começaram a diminuir e sobre isso procurou aconselhar-se com o seu amigo Pascal. Este explicou a Méré que ele não estava a ser vítima de uma crise de má sorte mas, apenas, da acção imutável das probabilidades: enquanto a possibilidade de conseguir um 6 é uma em 3*8 jogadas de um só dado, a possibilidade para um duplo 6 é de uma em 24*61 jogadas de dois dados

Mais tarde, a Teoria das Probabilidades desenvolveu-se e através dos trabalhos de Jacques Bernoulli (1654-1705), Moivre (1667.1759) e Thomas Bayes(1702-1761). A Bernoulli deveu-se a publicação do livro "Ars Conjectandi" que foi publicado em 1713 e foi o primeiro a ser tratado inteiramente às teorias das probabilidades. Nesta obra inclui diversas combinações e das permutações, os teoremas binomial e polinomial e a lei dos grandes números (hoje chamado Teorema de Bernoulli). A lei dos grandes números pode enunciar-se do seguinte modo:

" A frequência relativa de um acontecimento tende a estabilizar-se nas vizinhanças de um valor quando o número de provas cresce indefinidamente"

Moivre introduziu e demostrou a lei normal. A Bayes deve-se o cálculo das chamadas probabilidades e das causas. Ou seja, este cálculo consistiu em determinar a probabilidade de acontecimentos perante certas condições iniciais.

Na segunda metade do século XVIII e na primeira metade do século XIX(1749-1827) elaborou uma posição concisa e sistemática dos acontecimentos probabilísticos e demonstrou uma das formas do "Teorema das Probabilidades".

Laplace escreveu: "A teoria das probabilidades, no fundo, não é mais do que o bom senso traduzido em cálculo, permite calcular com exactidão aquilo que as pessoas sentem por uma espécie de instinto.É natural como tal ciência, que começou com estudos sobre jogos de azar, tenha alcançado os mais altos níveis do conhecimento humano."

Em 1812, Laplace publicou uma importante obra de Teoria Analítica das Probabilidades, onde sistematizou os conhecimentos da época e onde se encontra definida a Lei de Laplace..

Destaca-se a participação de Gauss (1777-1855) no aprofundamento da "Lei Normal" e de Poisson na sua "Teoria da lei dos grandes números e da lei de repartição".

No século XIX e princípio do século XX a teoria das probabilidades tornou-se um instrumento eficaz, exacto e fiável do conhecimento.

Surge a célebre escola de S. Petersburgo. Desta escola resultaram grandes nomes, tais como: Tchébychev 81821-1894), Markov (1856-1922) e Liapounav (1857-1918).

À escola de S. Petersburgo sucedeu a escola soviética na qual destaca-se a participação de Kolmogorov (1903-1987) que axiomatizou correctamente a teoria das probabilidades.

A História regista censos, para fins de alistamento militar e de colheita de impostos, realizados há mais de 4000 mil anos, como é o caso do censo do imperador Yao na China, em 2200 A.C.. Nesta altura a estatística era simplesmente um trabalho de exibição e síntese dos dados referentes colhidos pelos censos. Esta estatística não envolvia nenhum trabalho probabilístico, pois todos os objectos do universo envolvido (a população) eram observados ou medidos.

Adolph Quéletet em 1850 foi o primeiro a utilizar uma amostra no seu estudo, e, a partir da análise probabilística, estender os resultados da amostra a toda a população.

A partir dele, rapidamente surgiu a ideia de dar um embasamento mais rigoroso para o método científico, a partir de uma fundamentação probabilista para as etapas da colecta e a da análise indutiva de dados científicos. Hoje esta concepção é essencial no trabalho científico, contudo só atingiu um nível prático no início do sec XX desenvolvendo-se em 3 grandes frentes:

A Estatistica estuda técnicas quem permitem quantificar probabilisticamente as incertezas envolvidas ao induzirmos para um universo observações feitas numa amostra do mesmo – Inferência Estatística. Os pais desta técnica são J. Neyman e Karl Pearson. Embora os estudos de Neyman e Pearson estivessem associados a questões de hereditariedade, os métodos e expressões que criaram, tais como "hipótese nula" e a "nível de significancia" fazem hoje parte da rotina diária de todo o estatístico e cientista.

Trata das precauções que o cientista deve tomar, antes de iniciar as suas observações ou medidas, de modo a que se possa dar uma boa probabilidade de que os objectivos pretendidos sejam atingidos – o delineamento das experimentações científicas. O pai desta técnica é R.A. Fisher que ao trabalhar na selecção genética de plantas agrícolas, desenvolveu uma imensa quantidade de resultados básicos sobre o delineamento de experimentações, divulgando-os em dois livros históricos: Statistical Methods for Research Wakers, 1925, e The Design of Experiments, publicado em 1935.

Suponhamos que um cientista faz simultaneamente a medida de duas ou mais variáveis: uma poderia ser a altura e a outra o peso de pessoas de uma população. Se ambas as variáveis (peso e altura) tendem a crescer ou decrescer simultaneamente, dizemos que são positivamente correlacionados. Dizemos que são negativamente correlacionados se uma variável tende a crescer e a outra a decrescer.

O cientista ao afirmar que duas ou mais variáveis são correlacionadas, pode utilizar uma série de técnicas (chamadas análise de regressão) para achar fórmulas expressando os valores de uma dessas variáveis em termos da outra, ou outras. Tudo isto dentro de uma margem de erro que o cientista poderá estimar probabilisticamente.

O pai da ideia da correlação entre variáveis foi Francis Galton, o qual no final do século passado a usou numa série de estudos de hereditariedade motivados pela teoria da evolução de Darwin e com objectivos decididamente eugénicos , contudo, a base matemática de Galton era precária, cabendo a Karl Pearson dar uma fundamentação mais matemática para a correlação.

A teoria das probabilidades, que começou com um jogo, transformou-se, hoje em dia, num dos ramos da matemática com mais aplicações nas outras ciências: exactas, naturais, sociais.

A Estatística conquistou, hoje, o seu lugar entre as ciências. O poder do seu método é, sobretudo, afirmado nas últimas décadas e aplica-se, agora, nos domínios mais variados. Até aqui, só um

pequeno número de pessoas se preocupou com estudos estatísticos, quer pela natureza das suas investigações, quer por causa da sua utilidade para as diferentes profissões. O valor e a importância do método estatístico residem no esforço para melhor compreender o nosso mundo, tão maravilhosamente complexo, tanto no ponto de vista físico como social, levam-nos a sonhar que ele se torne objecto de um conhecimento como as outras ciências. A vida corrente leva-nos a decisões para passar do conhecido ao desconhecido, da experiência à previsão.

Este manual tem por fim fornecer conhecimentos estatísticos (sem ter muitos conhecimentos matemáticos) e ajudar a interpretar os resultados que podem ser obtidos quer através do calculo manual, quer através de programas de computador.

1. NOÇÕES GERAIS

Para algumas pessoas, a Estatística não é senão um quadro de colunas mais ou menos longas de números que dizem respeito à população, à indústria ou ao comércio, como se vê frequentemente em revistas; para outras, ela dá gráficos mostrando a variação no tempo de um facto económico ou social, a produção ou os números relativos aos negócios de uma empresa, assim como se encontra nos escritórios de empresas privadas.

Tão diferenciados se apresentam os métodos estatísticos que não é possível estabelecer uma definição que os contenha a todos. Apesar disso, apresentamos a seguir uma definição que, embora necessariamente incompleta como qualquer outra, tem a vantagem de introduzir o aluno na matéria.

A Estatística tem como finalidade elaborar de uma síntese numérica que evidencie o que de mais generalizado e significativo exista num conjunto numeroso de observações.

O grande número de observações de que se parte reflecte uma diversidade tal que se torna ininteligível a sua interpretação. Para que, a partir dessa diversidade se possa começar a entender logo, torna-se necessário reduzir sucessivamente as observações, ganhando-se em generalidade o que se vai perdendo em individualidade.

A síntese implica, assim, que nos desprendamos do que é particular e individual para nos atermos ao que existe de mais geral no conjunto das observações; à medida que a síntese progride, vai-se perdendo o contacto com as particularidades imediatas.

Deste modo, a Estatística não se ocupa do que é excepcional, mas apenas do que é geral: não se interessa pelo indivíduo, mas por grupos de indivíduos; não se ocupa, em suma, de uma só medição, mas de um conjunto de medições.

Acrescente-se, ainda, que a síntese é numérica. Quer isto dizer que se prescinde inteiramente das palavras e dos recursos literários de mais ou menos efeito que elas possibilitam. Alcança-se a síntese pelo recurso exclusivo dos números.

Daí o afã com que frequentemente se escolhem os números de acordo com os argumentos. A Estatística é intrinsecamente uma disciplina não literária, manipula exclusivamente números e alcança a síntese ordenando-os e cooperando com eles.

"Estatística", deriva de "status" que em latim significa Estado, e que só por si demonstra a ligação que sempre existiu entre ambos;

O primeiro levantamento estatístico remonta a 3050 a.C., no Egipto, tendo como objectivo informar o estado sobre recursos humanos e económicos.

No séc. XVII d.C., a disciplina de Estatística era já leccionada nas universidades alemãs, continuando com a finalidade de descrever as populações e as riquezas do Estado.

Ainda no séc. XVII, dá-se a expansão dos seus campos de investigação a áreas como a Saúde pública; a Indústria; o Comércio e os Estudos Demográficos.

Os métodos de inferência estatística surgem com Jonh Graunt (1620-1674), um modesto comerciante, que tira conclusões válidas sobre uma população desconhecida por ele.

Fermat (1601-1665) e Pascal (1623-1662) permitem que o estudo do acaso tome uma expressão matemática, introduzindo o Cálculo das Probabilidades.

O Cálculo das Probabilidades e o aparecimento do Método dos mínimos quadrados, vêm credibilizar a Estatística conferindo-lhe a fundamentação matemática em que ela assenta hoje.

No séc. XVIII Lambert Quetelet (1796-1874) introduziu a Estatística nas análises da Meteorologia; da Antropometria; das Ciências Sociais; da Economia e da Biologia.

Aos contributos anteriores Francis Galton (1822-1911), acrescenta as noções de regressão e correlação; Karl Pearson (1857-1936) apresenta a mais bela e acabada teoria de Estatística, ficando também conhecido pelos seus coeficientes (r; c); Fisher com os seus trabalhos sobre inferência Estatística também deu um grande contributo ao desenvolvimento da Estatística.

Em 1943, dá-se uma grande reviravolta, uma vez que o tratamento de dados deixa de ser feito manualmente e passa a ser numa primeira fase apoiado por calculadoras potentes para mais Tarde ser feito quase exclusivamente de forma computadorizada.

O Método Estatístico, segundo a teoria de Cramer, pressupõe as seguintes fases:

Recolha de dados estatísticos: obtenção da amostra a partir da população, devendo depurar e rectificar os dados estatísticos, que no seu conjunto são denominados série estatística.

Descrição: conjunto de operações, numéricas ou gráficas, efectuadas sobre os dados estatísticos determinando a sua distribuição; procede-se à sua ordenação, codificação e representação por meio de quadros e tabelas.

Análise: consiste em tirar conclusões sobre a distribuição da população, determinar o seu grau de confiança e ainda formular hipóteses, tentando verificá-las, quanto ao fenómeno em estudo.

Predição: é uma previsão do comportamento do fenómeno em estudo, tendo em conta a definição da distribuição estatística.

2. POPULAÇÃO E AMOSTRA

População: somatório dos indivíduos ou elementos, com qualquer característica comum e que estão sujeitos a uma análise estatística, por terem interesse para o estudo. Quanto à sua origem pode ser: um conjunto de pessoas; um conjunto de objectos ou um conjunto de acontecimentos. Quanto à sua natureza pode ser: Existente ou real; Hipotética ou parcialmente existente. Pode ainda ser: um conjunto finito ou um conjunto infinito.

Amostra: é um subconjunto retirado da população, que se supõe ser representativo de todas as características da mesma, sobre o qual será feito o estudo, com o objectivo de serem tiradas conclusões válidas sobre a população.

Amostragem: é o procedimento pelo qual um grupo de pessoas ou um subconjunto de uma população é escolhido com vista a obter informações relacionadas com um fenómeno, e de tal forma que a população inteira nos interessa esteja representada (fig. 1)

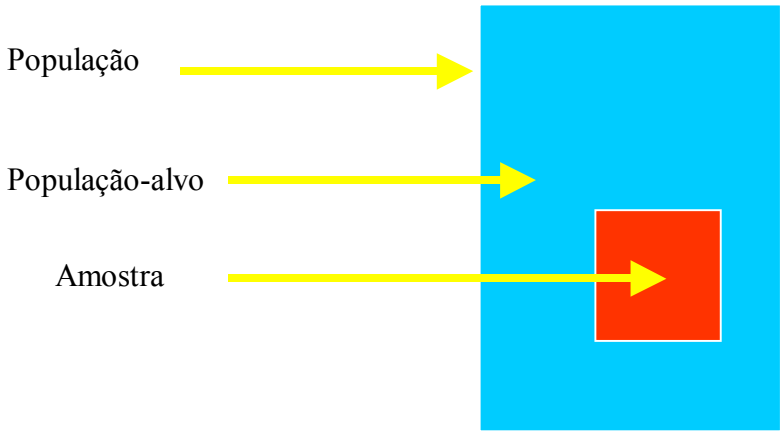


Figura 1: População e Amostra

O Plano de Amostragem serve para descrever a estratégia a utilizar para seleccionar a amostra. Este plano fornece os detalhes sobre a forma de proceder relativamente à utilização de um método de amostragem para determinado estudo.

Logo que o investigador delimite a população potencial para o estudo, ele deve precisar os critérios de selecção dos seus elementos, que podem ser de inclusão ou de exclusão dos sujeitos que farão parte do estudo:

Um investigador interessado pela readaptação após cirurgia de revascularização, pode concentrar-se somente nos sujeitos que tiveram uma única experiência deste tipo e excluírem os outros.

Uma amostra é dita representativa se as suas características se assemelham o mais possível às da população-alvo. É particularmente importante que a amostra represente não só as variáveis em estudo, mas também outros factores susceptíveis de exercer alguma influência sobre as variáveis estudadas, como a idade, o sexo, a escolaridade, o rendimento, etc.

A Representatividade avalia-se comparando as médias da amostra com as da população-alvo.

Como se ignora se todas as características da população estão presentes numa amostra dado que estas são muitas vezes desconhecidas, admite-se que existe sempre um grau de erro.

ERRO DE AMOSTRAGEM: é a diferença que existe entre os resultados obtidos numa amostra e os que teriam sido obtidos na população-alvo.

Duas soluções existem para reduzir ao mínimo o erro amostral:

1. Retirar de forma aleatória e um número suficiente de sujeitos que farão parte da amostra.

2. Procurar reproduzir o mais fielmente possível a população pela tomada em conta das características conhecidas desta.

3. MÉTODOS DE AMOSTRAGEM

3.1 AMOSTRAGENS PROBABILÍSTICAS E NÃO-PROBABILÍSTICAS

3.1.1 TIPOS DE AMOSTRAGENS PROBABILÍSTICAS

Os métodos de amostragem probabilística servem para assegurar uma certa precisão na estimação dos parâmetros da população, reduzindo o erro amostral.

A principal característica dos métodos de amostragem probabilística reside no facto de que cada elemento da população tem uma probabilidade conhecida e diferente de zero, de ser escolhida, aquando da tiragem ao acaso para fazer parte da amostra.

O objectivo desta abordagem é obter a melhor representatividade possível.

Tipos de Amostragem:

A Amostragem Aleatória Simples;

A Amostragem Aleatória Estratificada;

A Amostragem em Cachos;

A Amostragem Sistemática.

AMOSTRAGEM ALEATÓRIA SIMPLES

A Amostragem aleatória simples é uma técnica segundo a qual cada um dos elementos (sujeitos) que compõe a população alvo tem igual probabilidade de ser escolhido para fazer parte de uma amostra. A amostragem aleatória simples consiste em elaborar uma lista numérica de elementos de onde se tira, com a ajuda de uma tabela de números aleatórios, uma série de números para constituir a amostra.

AMOSTRAGEM ALEATÓRIA ESTRATIFICADA

A Amostragem aleatória estratificada é uma variante da amostra aleatória simples. Esta técnica consiste em dividir a população alvo em subgrupos homogéneos chamados «estratos» e a seguir tirar de forma aleatória uma amostra de cada estrato. A Amostragem aleatória estratificada é utilizada quando a população inteira é reconhecida por certas características precisas, tais como a idade, o sexo, a incidência de uma condição de saúde, tudo isto para assegurar a melhor representatividade possível.

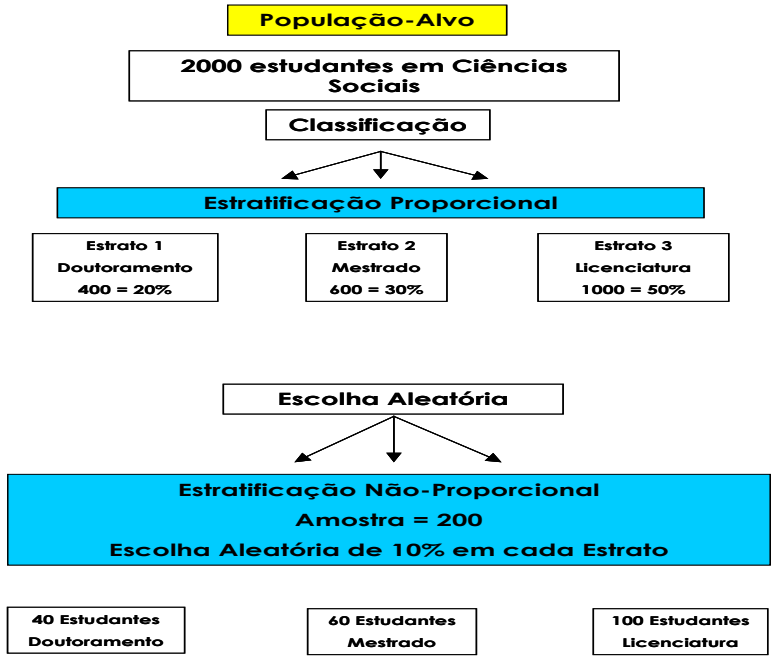


Figura 2: Amostra estratificada

AMOSTRAGEM EM CACHOS

Consiste em retirar de forma aleatória os elementos por cachos em vez de unidades. É útil quando os elementos da população estão naturalmente por cachos e por isso devem ser tratados como grupos ou quando não é possível obter uma listagem de todos os elementos da população-alvo.

AMOSTRAGEM SISTEMÁTICA

Consiste quando existe uma lista ordenada de elementos da população. Esta técnica consiste K elementos dessa lista sendo o primeiro elemento da amostra retirado ao acaso.

O intervalo entre os elementos corresponde à razão entre o tamanho da população e da amostra.

Exemplo: Se pretender uma amostra de 100 indivíduos e a população for de 1000 o sistema será 1000:100=10 (dez em dez é o sistema), isto é, será incluído um elemento da lista de 10 em 10 indivíduos a partir do 1.º n.º sorteado.

Importante

Se se utilizar uma amostragem por cachos ou outros tipos de agrupamentos, a amostra só é considerada probabilística se os grupos foram escolhidos ao acaso antes da repartição aleatória dos sujeitos nos grupos.

3.1.2 TIPOS DE AMOSTRAGENS NÃO PROBABILÍSTICAS:

É um procedimento de selecção segundo o qual cada elemento da população não tem a mesma probabilidade de ser escolhido para formar a amostra.

Este tipo de amostragem tem o risco de ser menos representativa que a probabilística no entanto é muitas vezes o único meio de construir amostras em certas disciplinas profissionais nomeadamente na área da saúde.

Tipos de Amostragens Não-Probabilísticas:

A Amostragem Acidental ou de Conveniência;

A Amostragem por Cotas;

A Amostragem de Selecção Racional ou Tipicidade;

A Amostragem por Redes ou Bola de Neve.

AMOSTRAGEM ACIDENTAL OU DE CONVENIÊNCIA

É formada por sujeitos facilmente acessíveis, que estão presentes num determinado local e momento preciso.

Exemplo: pessoas hospitalizadas. Um investigador pode ter acesso a uma unidade hospitalar para constituir uma amostra de pacientes hospitalizados.

Neste tipo de amostra tem a vantagem de ser simples em organizar e pouco onerosa, todavia este tipo de amostra provoca enviesamentos, pois nada indica que as primeiras 30 a 40 pessoas sejam representativas da população-alvo. São utilizadas em estudos que não têm como finalidade a generalização dos resultados.

AMOSTRAGEM POR COTAS

Idêntica à amostragem aleatória estratificada diferindo desta apenas pelo facto dos sujeitos não serem escolhidos aleatoriamente no interior de cada estrato ou de cada grupo.

AMOSTRAGEM POR SELECÇÃO RACIONAL OU POR TIPICIDADE

Tem por base o julgamento do investigador para constituir uma amostra de sujeitos em função do seu carácter típico.

Por exemplo: o estudo de casos extremos ou desviantes como uma patologia rara ou uma instituição.

AMOSTRAGEM POR REDES OU BOLA DE NEVE

Consiste em escolher sujeitos que seriam difíceis de encontrar de outra forma. Toma-se por base, redes sociais amizades e conhecimentos.

Por exemplo: Imigrantes de Leste.

Quando o investigador encontra sujeitos que satisfazem os critérios escolhidos pede-lhes que indiquem outras pessoas de características similares.

3.2 DETERMINAÇÃO DO TAMANHO DA AMOSTRA

Os tamanhos das amostras são relativos, isto é, depende do tamanho da população. Para determinar as amostras existem várias fórmulas, consoante o parâmetro em critério. As mais utilizadas na saúde são as que se baseiam na percentagem do fenómeno:

3.2.1 CÁLCULO DO TAMANHO DA AMOSTRA PARA POPULAÇÕES INFINITAS (>100.000 ELEMENTOS)

A amostra depende da:

1. Extensão do universo;
2. Do Nível de Confiança;
3. Do Erro Máximo permitido;
4. Da percentagem com que o fenómeno se verifica.

Formula $$n=\frac{\sigma^2 (pq)}{e^2}$$

n= Tamanho da amostra

σ = Nível de confiança escolhido expresso em n desvios padrão (s)

p = % com o qual o fenómeno se verifica

q = % complementar (100-p)

e = Erro máximo permitido

Se desejarmos um nível de confiança bastante alto – superior a 99% aplica-se a fórmula dos três desvios.

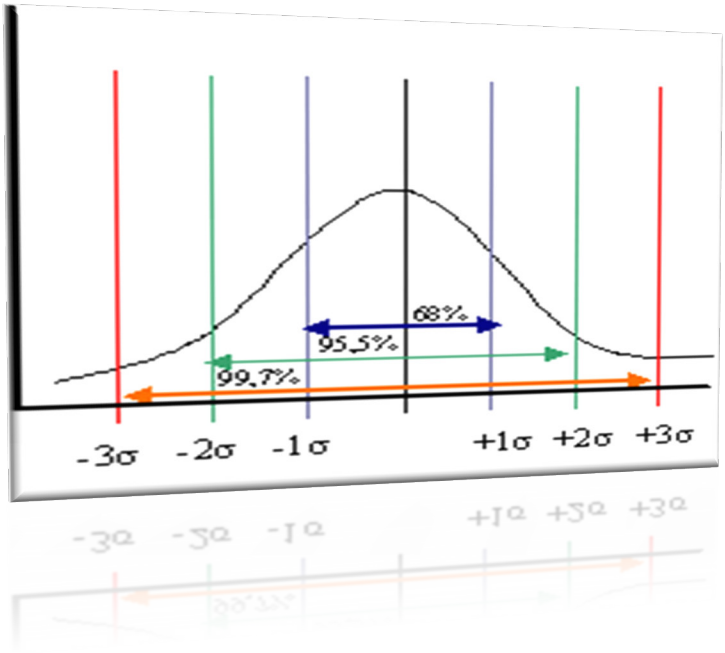


Figura 3: Conversão dos níveis de confiança em desvios padrão

Logo, o desvio $(s)^2$ seria igual a $3^2 = 9$

Se o erro máximo for de 2% o e^2 será igual a $2^2 = 4$

Exemplo: Se for possível admitir que o número de captações de água em profundidade se situam por volta dos 50%, não ultrapassando esta %, então p=50 e, consequentemente, q=100-50 ou seja 50. Assim, tem-se a equação

$$n = \frac{9\ (5050)}{4} = 5625$$

Isto é, para atender às exigências estabelecidas, o n.º de captações a analisar seria 5625.

Se todavia, for aceite o nível de confiança de 95% (2 desvios) e um erro máximo de 5% o n.º de elementos será bem menor.

os cálculos.

$$n = \frac{\sigma^2\ (pq)}{e^2}$$

$$n = \frac{4\ .\ (50.50)}{25} = 400$$

Convém lembrar que sempre que não seja possível estimar uma percentagem do fenómeno, deve utilizar-se sempre p=50

3.2.2 CÁLCULO DO TAMANHO DA AMOSTRA PARA POPULAÇÕES FINITAS (<100.000 ELEMENTOS)

1. A amostra depende da:
2. Extensão do universo;
3. Do Nível de Confiança;
4. Do Erro Máximo permitido;
5. Da percentagem com que o fenómeno se verifica.

Fórmula

$$n = \frac{\sigma^2\ p.q.N}{e^2\ (N-1) + \sigma^2\ p.q}$$

Onde:

n = Tamanho da amostra; N = tamanho da população

σ = nível de confiança escolhido, expresso em números de desvios padrão

p = percentagem do fenómeno; q = percentagem complementar

e^2 = erro máximo permitido

Exemplo: Verificar quantos dos 100 empregados de uma cantina cumprem correctamente as normas de higiene e segurança do trabalho.

Presume-se que esse n.º não seja superior a 30% do total; deseja-se um nível de confiança de 95% (2 desvios) e tolera-se um erro até 3%.

Então, n=90,4 Logo deverão ser pesquisados 90 empregados.

Confirme aplicando a fórmula $$n = \frac{\sigma^2 p.q.N}{e^2 (N-1) + \sigma^2 p.q}$$

Mas, se a população fosse de 10.000 empregados, com os mesmos critérios anteriormente referidos, então:

$$n = \frac{4 \,.\, (30.70).10000}{(9.9999)+4.(30.70)} = 853$$

O tamanho "óptimo" de uma amostra, não depende tanto do tamanho da população mas sim de dois parâmetros estatísticos: a margem de erro e o nível de confiança

Margem de erro – Uma amostra representa aproximadamente (e nunca exactamente) uma população. A medida deste "aproximadamente" é a chamada margem de erro, e é lido assim:

se uma pesquisa tem uma margem de erro de 2% e a Doença Cardíaca teve 25% de prevalência na amostra recolhida, podemos dizer que, naquele instante, na população, ela terá uma prevalência entre 23% e 27% (25% menos 2% e 25% mais 2%).

Nível de confiança – As pesquisas são feitas com um parâmetro chamado nível de confiança, geralmente de 95%.

Estes 95% querem dizer o seguinte: se realizarmos uma outra pesquisa, com uma amostra do mesmo tamanho, nas mesmas datas e locais e com o mesmo instrumento de recolha de dados, há uma probabilidade de 95% de que os resultados sejam os mesmos (e uma probabilidade de 5%, é claro, de que tudo difira).

Quando já se efectivou uma pesquisa e se deseja conhecer a margem de erro utilizada aplica-se a fórmula:

$$\sigma_p = \sqrt{\frac{(p.q)}{n}}$$

Onde:

n = Tamanho da amostra

σ_p = Erro padrão ou desvio da percentagem com que se verifica determinado fenómeno

p = percentagem do fenómeno

q = percentagem complementar

3.2.3 DETERMINAÇÃO DA MARGEM DE ERRO DA AMOSTRA

Exemplo: Numa pesquisa efectuada com 1000 adultos, verificou-se que 30% bebem café pelo menos uma vez por dia. Qual a probabilidade de que tal resultado seja verdadeiro para todo o universo.

$$\sigma_p = \sqrt{\frac{(30.70)}{1000}}$$
$$\sigma_p = 1{,}45$$

Como o valor encontrado (margem de erro) corresponde a um desvio, então para dois desvios (95,5%), temos 1,45 *2=2,90.

Para 3 desvios é o triplo (4,35).

Isto significa que, por exemplo, para um nível de confiança de 95% (2 desvios) o resultado da pesquisa apresentará como margem de erro 2,90 para mais ou menos.

É provável, portanto, que o n.º de consumidores de café esteja entre 27,10% (30%-2,90) e 32,90% (30%+2,90).

3.2.4 DETERMINAÇÃO DA AMOSTRA SEM CONHECER OS LIMITES DA POPULAÇÃO

Fórmula $$\frac{Z^2_{(\alpha/2)} * p(1-p)}{d^2}$$

Em que:

p= fenómeno]

α= erro

{Se IC (intervalo de confiança)=95%, α=5% (0,05)} Então α/2 = 0,05/2 =0,025

Z (α/2) = Z(0,025)= 1,96

d=número de desvios

Assim para um fenómeno que tenha uma prevalência de 25%, os resultados seriam:

$$n = \frac{1,96^2 * 0,25(1-0,25)}{(0,05)^2} = 288$$

Em termos estatísticos consideramos que uma amostra é: Pequena se n < 30 que o valor para a qual começa a tender à normalidade.

Cuidados a ter na escolha da amostra:

1. Imparcialidade: todos os elementos devem ter a mesma probabilidade e oportunidade de serem escolhidos;
2. Representatividade: deve conter em proporção todas as características que a população possui, qualitativa e quantitativamente, de modo a que não se torne tendenciosa;
3. Tamanho: suficientemente grande de modo a fornecer as principais características, por outro lado pequena para economizar tempo, dinheiro e pessoal.

3.3 INDIVÍDUO OU UNIDADE ESTATÍSTICA

O estudo Estatístico recai sobre a amostra, no entanto este é feito de modo pormenorizado a cada um dos elementos da amostra, que são designados por Indivíduo ou Unidade Estatística.

Unidade Estatística: é o factor elementar, o objecto de análise, que independentemente da sua natureza tem que possuir uma definição precisa.

As principais características de uma boa unidade Estatística são:

1. Propriedade ou adequação ao objectivo da investigação;
2. Clareza;
3. Mensurabilidade;
4. Comparabilidade.

No estudo de cada unidade Estatística, surgem resultados individuais com os quais são feitas as inferências sobre a população. Estes resultados têm o nome de Dado Estatístico.

Dado Estatístico: é o resultado do estudo efectuado a cada unidade Estatística tendo em conta a sua individualidade, sendo este depois tratado de modo a permitir inferir sobre a colectividade que a integra (população).

3.4 VARIÁVEIS

Propriedade em relação à qual os indivíduos de uma amostra variam. Note-se que as propriedades que não variam não são de interesse estatístico. Há muitos modos de dividir os diferentes tipos de variáveis.

Ao ser efectuada uma análise Estatística a uma população, os aspectos (características) que se têm em conta, um ou vários, são denominados por Variável Estatística.

Uma variável Estatística pode ser:

Qualitativa: se é a sua natureza que varia de elemento para elemento.

As variáveis qualitativas dividem:se em:

Variáveis nominais: quando o seu significado só se entende em função do nome e o número ou código que se lhe atribua não nos dá nenhuma informação (sexo, cor de olhos, grau de parentesco, tipo de patologia, presença/ausência de factores de risco, etc.).

Variáveis ordinais: quando existe uma ordenação possivel (gravidade de uma lesão, classe social, grau de escolaridade, etc.).

Quantitativa: se é a sua intensidade que varia de elemento para elemento, tornando-a mensurável ou referenciável.

As variáveis quantitativas dividem-se em:

Variáveis discretas: assume valores isolados, normalmente inteiros (n.º de filhos, n.º de factores de risco, n.º de dependentes, n.º de respostas, etc)

Variáveis contínuas: em que é possível qualquer operação aritmética, podendo assumir qualquer valor real (altura, peso, IMC, distância, etc).

Tendo em conta o número de atributos (características) que estão a ser estudadas, as variáveis podem ser:

Unidimensionais: se apenas corresponde a um atributo

Bidimensionais: se corresponde a dois atributos;

Pluridimensionais: se corresponde a vários atributos.

Modalidade: é toda a manifestação possível de uma variável, isto é, as várias hipóteses de resposta, podendo elas ser duas ou mais.

As modalidades têm obrigatoriamente que ser:

Incompatíveis: cada unidade Estatística não pode pertencer simultaneamente a duas ou mais modalidades;

Exaustivas: todas as unidades Estatísticas têm que ser inseridas numa modalidade.

A escolha das modalidades deve ser feita de acordo com as informações possuídas. No entanto, surgem situações em que há necessidade de se aumentar uma modalidade suplementar.

VARIÁVEIS DEPENDENTES E INDEPENDENTES

Gostaríamos agora de introduzir a terminologias variáveis independentes e variáveis dependentes. A variável manipulada pelo experimentador é conhecida como variável independente. Isto porque as situações experimentais que testam esta variável são definidas independentemente mesmo antes de a própria experiência se iniciar. A segunda variável, os resultados nos testes de estatística, é conhecida como variável dependente (os resultados de estatística **dependem** da utilização de um esquema de mnemónica), porque os resultados do teste são dependentes da maneira como o experimentador manipula a variável independente «esquema de mnemónica».

Assumindo que demonstramos que o esquema de mnemónica produz algum efeito, lembrar-se-á que a questão seguinte levantada pelo céptico tinha a ver com o facto dos alunos com menos dificuldades com os cálculos serem aqueles que beneficiariam mais do esquema do que aqueles que tinham maiores dificuldades com a estatística.

Uma forma de investigar esta possibilidade seria a de transformar a "facilidade em fazer operações matemáticas" em variável independente.

O investigador apresentaria então a todos os alunos um teste que avaliasse aquele facto, e seleccionaria de seguida dois grupos de estudantes, um grupo com facilidade em efectuar operações matemáticas e outro com dificuldades.

Se a ambos os grupos fosse apresentado o esquema de mnemónica, seria então possível avaliar o efeito da variável independente "facilidade em efectuar operações matemáticas" na outra variável "resultados do teste." Por outras palavras, seria o grupo de "bons estatísticos" ou o grupo de "maus estatísticos" que apresentava maiores progressos nos resultados do teste?

Um dos aspectos de que já se deve ter dado conta é de que não é possível manipular a variável independente "facilidade em fazer operações matemáticas" da mesma forma como manipulamos anteriormente variável dependente com ou sem esquema de mnemónica. Neste último caso é da inteira responsabilidade do experimentador decidir quais os alunos a quem dá o esquema de mnemónica e a quem não dá.

No que diz respeito à "facilidade em fazer operações matemáticas", não existe forma de o experimentador dar ou retirar a um aluno facilidade em fazer operações matemáticas. Ainda assim, o experimentador pode manipular essa variável criando dois grupos, um em que coloca os que têm dificuldade e outro em que coloca os que não têm, constituindo assim dois grupos experimentais.

A H_1 poderá ser: Apenas os alunos que têm maior facilidade em fazer operações matemáticas apresentam resultados superiores em estatística.

O esquema de mnemónica deixou de ser variável e passou a situação constante, já que neste caso todos usufruíram do mesmo. Por outras palavras, o investigador previra uma diferença entre os resultados do teste dos dois grupos de alunos após ter sido apresentado a ambos o esquema de mnemónica.

Uma outra variável independente do mesmo tipo é o sexo. Até mesmo o experimentador mais omnipotente não pode transformar um homem numa mulher e vice-versa. É até bastante comum formar grupos de homens e mulheres para se investigarem as diferenças de performance nas mais diversas tarefas, que possam ser devidas a esse factor.

Mas quando estamos perante um estudo cientifico, nem sempre é possível estabelecer relações de dependência e, existem mesmo alguns tipos de estudos em que esta denominação é contra-indicada,

por conterem apenas questões de investigação e serem, por isso, exploratórios (nível I), descritivos e em alguns casos descritivo-correlacionais (nível II). Nestes casos podemos definir as variáveis como primárias, secundárias e complementares, embora não seja obrigatório.

VARIÁVEIS PRIMÁRIAS, DERIVADAS OU SECUNDÁRIAS E COMPLEMENTARES

As variáveis primárias são as consideradas como principais no nosso estudo e as únicas que têm peso no momento da conclusão (variáveis incluídas nas hipóteses). Por exemplo, na pesquisa cuja pergunta é qual a qualidade de vida dos cuidadores de idosos acamados? A variável primária é a qualidade de vida.

As variáveis secundárias são importantes para avaliar a situação em estudo mas raramente são determinantes na conclusão do estudo.

As variáveis complementares são aquelas que utilizamos para caracterizar a nossa população ou amostra.

Em cada uma das variáveis deverá ser apresentado: a definição da variável, como, quem e quando será mensurada

Por exemplo, numa pesquisa cuja a pergunta é:

- Qual a prevalência de obesidade nos estudantes universitários?

A variável primária será a prevalência de obesidade; as variáveis secundárias serão a estatura, o peso, a circunferência abdominal e a qualidade de vida.

Os dados complementares serão a idade, sexo, curso de graduação, ano do curso de graduação.

As variáveis derivadas (ou variáveis secundárias) são novas variáveis que podem ser criadas a partir de operações lógicas e/ou matemáticas sobre variáveis existentes nas bases de dados (variáveis primárias)

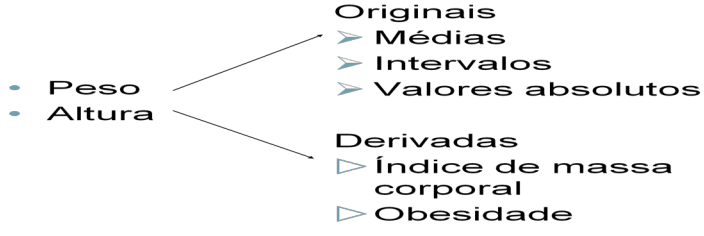


Figura 4: variáveis primarias e derivadas

Níveis de mensuração das variáveis

As variáveis diferem em "quão bem" elas podem ser medidas, ou seja, em quanta informação seu nível de mensuração pode gerar. Operacionalmente, muitas vezes pode-se estudar algo de diferentes maneiras.

Exemplificando, supondo que pretende estudar os hábitos tabágicos. Qual seria a escala? Haveria apenas 2 grupos: fumadores e não fumadores? Ou seria contado o número de cigarros consumidos durante determinado período? Utilizaria a Unidade Masso Ano (UMA)? Como seria definido o fumador? Quem fuma 1 cigarro por dia será considerado o quê? E que, fuma 1 maço de cigarros por dia? Pertencem à mesma categoria?

Assim, de acordo com sua escala de medição, as variáveis podem ser classificadas em 3 tipos:

ESCALA NOMINAL

São variáveis qualitativas. Os dados podem ser distribuídos em categorias mutuamente exclusivas. Seus valores só são registados como nomes, só permitindo classificação qualitativa, não existindo ordem entre as categorias existentes. Assim, pode-se dizer que dois indivíduos são diferentes em termos da variável analisada, mas não se pode dizer qual deles "tem mais" da qualidade representada pela variável.

Exemplos: sexo, estado civil, presença/ ausência de doença, patologia, causa de morte, etc.

As análises estatísticas mais comuns são o estudo de proporções e testes baseados no Qui-quadrado.

ESCALA ORDINAL

São variáveis qualitativas. Os dados podem ser distribuídos em categorias mutuamente exclusivas, mas que têm ordenação natural. São aquelas com possíveis resultados nominais, sem valores métricos, mas em que existe uma ordenação entre as categorias, com um resultado precedendo o outro (hierarquia ou grau). Portanto, permitem ordenar os itens medidos em termos de qual tem menos e qual tem mais da qualidade representada pela variável, mas não possibilitam que se diga "o quanto mais".

Exemplos: estágio da doença (inicial, intermédio, terminal); escolaridade (1.º CEB, 2.º CEB, 3.º CEB, Lic. MSC, PHD); peso, quando medido em 3 níveis (leve, médio, pesado); nível socioeconómico de

famílias residentes numa localidade (pobre, classe média, Alta); *classificação no teste* (muito bom, bom, satisfaz, medíocre, mau), , grau de estenose (ligeira, moderada, severa), etc.

As análises estatísticas mais comuns são o estudo de proporções, medianas, quartis, moda. Testes: Qui-quadrado, Kruskal-Wallis, regressão logística e outros testes não paramétricos.

ESCALAS INTERVALAR E PROPORCIONAL OU DE RAZÃO

A escala intervalar estabelecem-se intervalos iguais a partir de uma origem arbitrária, enquanto que na de razão existe um ponto zero a partir do qual se estabelecem intervalos iguais. Ambas são quantitativas e os seus dados são expressos por números. Permitem não apenas ordenar os itens que estão sendo medidos, mas também possibilitam quantificar e comparar o tamanho das diferenças entre eles. Os seus valores são medidos em uma escala métrica e por isso não são diferenciadas em alguns softwares estatísticos, como é exemplo o SPSS, em que são denominadas de *SCALE*.

Exemplos: Temperatura em °C; Idade, em anos; Peso corporal em quilos, *classificação no teste*: (0,..., 20), comprimento do segmento de recta desenhado etc.

É evidente que as variáveis *quantitativas* incluem mais informação, portanto permitem que sejam aplicadas provas estatísticas mais potentes.

4. ESTATÍSTICA DESCRITIVA

A Estatística Descritiva recolhe, organiza e analisa os dados de uma amostra, sem tirar qualquer conclusão sobre um grupo maior, enquanto que a Estatística Indutiva ou inferencial recolhe, organiza, analisa e estabelece relações entre os dados para fazer inferências sobre a população. Com base nos resultados obtidos sobre a amostra podemos inferir conclusões válidas sobre a população (este ramo da Estatística já exige a utilização de recursos matemáticos especiais, nomeadamente a Teoria das Probabilidades).

Assim, a Estatística Indutiva permite-nos fazer inferências sobre a população e chegar a leis e a teorias e a descritiva dá um apoio a esta tarefa.

4.1 PARÂMETRO E DADO ESTATÍSTICO

O parâmetro é toda a função definida a partir dos dados numéricos de uma população.

Exemplo: consideremos as seguintes notas em Estatística - 10 11 10 15 9

Média =Σxi/n = 55/5 = 11

O valor 11 é o parâmetro (resultado da média aritmética).

O dado estatístico é toda a função definida a partir dos dados numéricos de uma amostra.

Exemplo: consideremos a amostra: 10 10

Média =Σxi/n = 20/2 = 10 O valor 10 é o dado estatístico

4.2 REPRESENTAÇÃO DE UMA VARIÁVEL ESTATÍSTICA

Depois de termos definido algumas noções básicas de estatística, tratar-se-á, a seguir, da segunda fase de um estudo estatístico. Como já referimos, os dados numéricos recolhidos registam-se em séries estatísticas e, para serem analisados, devem ser ordenados e representados em quadros e em gráficos.

Quando trabalhamos com uma variável discreta ou descontínua falamos em seriação e quando trabalhamos com uma variável contínua falamos em classificação.

SERIAÇÃO DE UMA AMOSTRA

Como já referimos anteriormente, uma seriação implica que a variável seja discreta (exemplo: número de filhos de um casal, número de divisões de uma casa, etc.).

DISTRIBUIÇÃO DE FREQUÊNCIAS

É o arranjo dos valores e suas respectivas freqüências. Assim, a distribuição de freqüências para o exemplo será:

Valores Freqüência absoluta (Fi) Freqüência relativa (Fr) Percentagem (%)

FREQUÊNCIA ABSOLUTA (FI)

É o número de vezes que o elemento aparece na amostra, ou o número de elementos pertencentes a uma classe. A soma de todas as freqüências deve ser o número total de elementos do conjunto (N). Se

o número de elementos for muito grande ou pouco repetidos, podemos separar o conjunto em classes, que são intervalos numéricos a I— b ou a ≤ x ≤ b.

A diferença b – a chama-se amplitude das classes (h) e é utilizada a mesma amplitude para todas as classes com intervalos fechados à esquerda.

FREQUÊNCIA RELATIVA (FR)

A frequência relativa, para cada valor assumido por uma variável, é definida como a razão entre a frequência absoluta (Fi) e o número total de dados (N). Para calcularmos a percentagem de cada valor, basta multiplicar por 100 a frequência relativa.

Exercícios

Tabela 1: distribuição de frequencias

Valores	Frequência absoluta (Fi)	Frequência relativa (Fr)	Percentagem (%)
21	3	3/30 = 0.1	10
22	2	2/30 = 0.066	6.6
23	2	2/30 = 0.066	6.6
24	1	1/30 = 0.034	3.4
25	4	4/30 = 0.132	13.2
26	3	3/30 = 0.1	10
28	1	1/30 = 0.034	3.4
30	1	1/30 = 0.034	3.4
31	3	3/30 = 0.1	10
32	1	1/30 = 0.034	3.4
33	3	3/30 = 0.1	10
34	3	3/30 = 0.1	10
35	2	2/30 = 0.066	6.6
36	1	1/30 = 0.034	3.4
Total	30	1	100

Em uma pesquisa socioeconômica sobre itens de conforto, perguntou-se a cada um dos 800 entrevistados: Quantos aparelhos de TV há em sua casa? Os resultados aparecem na tabela:

Tabela 2: Exercício de distribuição de frequencias

Nº aparelhos	Frequência absoluta	Frequência relativa	Porcentagem
0	20		
1			
2		0.6	
3			7.5
4	30		

Complete a tabela.

Como já referimos anteriormente, uma classificação implica que a variável seja contínua (exemplo: a temperatura de um corpo, a altura de uma pessoa, a duração de certo fenómeno, etc. - variáveis relacionadas com o espaço, o tempo ou a massa).

Na primeira coluna temos as classes. Por convenção, as classes são abertas superiormente, com excepção da última classe, naturalmente.

Na segunda coluna temos as marcas da classe. Esta coluna pode ser também designada por x'i. A marca de uma classe é o ponto médio dessa classe, ou seja, é o ponto equidistante dos extremos de uma classe.

Exemplo: classe c0-c1 x'1 = (c0 + c1) / 2

Na terceira coluna apresentamos as frequências absolutas simples ou efectivas, ou seja, o número de vezes que os valores de determinada classe foram observados. Esta coluna pode ser também denominada por ni. Mais uma vez, Σ ni = n (número total de indivíduos pertencentes à população/ou amostra, a que se chama efectivo total).

Na quarta coluna apresentamos as frequências acumuladas, isto é, a soma das frequências absolutas correspondentes a valores inferiores a um determinado valor. Esta coluna pode ser também denominada por N(αi).

Na quinta coluna temos as frequências relativas simples. Esta coluna pode ser também denominada por fi, em que fi = ni/n e, de tal modo, que Σ fi =1.

Na sexta coluna apresentamos as frequências relativas acumuladas, isto é, a soma das frequências relativas correspondentes a valores inferiores a um determinado valor. Esta coluna pode ser também denominada por F(αi).

Numa classificação é habitual representarmos por K o número de classes (em geral K varia entre 5 e 20, inclusive) e por A a amplitude (em que A=x máximo - x mínimo).

Então,

- Se K é dado $\alpha = A/K$

- Se α é dado $K=A/\alpha$, sendo α a amplitude do intervalo de classe e A a amplitude do intervalo da amostra

Exemplo: Construa o quadro de frequências com os seguintes dados:

Classes [20-23[[23-26[[26-29[[29-32[[32-35[[35-38]

ni 2 5 7 10 4 2

Classes	ni	fi	N(αi)	F(αi)	x'i
[20-22]	2	2/30	2	2/30	21
[23-25]	5	5/30	7	7/30	24
[26-28]	7		14		
[29-31]	10		24		
[32-34]	4		28		
[35-38]	2		30		
$\alpha = 3$	$\Sigma ni = 30$	$\Sigma fi = 1$	$n = 30$		

4.3 REDUÇÃO DE UMA VARIÁVEL ESTATÍSTICA

CONCEITO DE REDUÇÃO E SUA CONVENIÊNCIA

Anteriormente definimos o conceito de variável estatística e construíram-se quadros e gráficos estatísticos com vista a uma descrição numérica e gráfica de uma variável estatística. Naturalmente, os gráficos permitem uma primeira síntese das informações registadas nos quadros.

Por outro lado, por simples aproximação das curvas de frequências absolutas ou relativas de duas ou mais variáveis, podemos fazer uma primeira comparação entre elas.

Há, porém, necessidade de sintetizarmos toda a informação respeitante a uma variável estatística, resumindo-se os dados a um pequeno número de elementos que bastam para caracterizá-la. Tal síntese consiste na redução de dados e os elementos numéricos obtidos designam-se por parâmetros da variável estatística.

Feita a representação dos dados estatísticos por meio de quadros e/ou de gráficos, importa fazer sobre os mesmos um estudo no sentido de se poder chegar a conclusões.

Para tal, impõe-se um trabalho de simplificação que consiste em proceder a sínteses, em reduzir grandes quantidades de dados a números simples que permitam uma análise rápida e uma fácil comparação com outras séries da mesma natureza ou de natureza diferente.

Tais números são designados, habitualmente, por características, parâmetros ou medidas e são agrupados em categorias conforme o tipo de informação que fornecem.

MEDIDAS DE TENDÊNCIA CENTRAL OU DE POSIÇÃO

Sob esta designação agrupam-se os parâmetros que, ou nos indicam algo de associável ao núcleo ou centro da distribuição, ou nos permitem compartimentá-la. Vamos considerar as seguintes medidas de tendência central ou de posição: média, mediana, moda e quantis.

MÉDIA

A média é o ponto de equilíbrio dos dados, isto é, tendo um conjunto de n valores x1, x2, ..., xn de uma variável X é o quociente entre a soma desses valores e o número deles.

A média aritmética simples (dados não agrupados) pode ser representada pela seguinte fórmula matemática:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Simplificando: $\bar{x} = \Sigma\, xi / n$

Numa amostra seriada os valores x1, x2, ..., xk ocorrem n1, n2, ..., nk vezes, respectivamente, a média aritmética ser $\bar{x} = \Sigma\, nixi / n = \Sigma\, fixi$

Exercicio: Para distribuição de frequência por variável discreta: Para os dados Populacionais calcule a Média, sabendo que

xi	4	6	7	8	10
ni	2	4	5	3	2

xi	ni	nixi
4	2	8
6	4	24
7	5	35
8	3	24
10	2	20
	Σ=16	Σ=111

Então, $\bar{x} = \Sigma\, nixi / n = 111/16 = 6{,}94$

Numa amostra classificada a fórmula definidora da média não se pode aplicar directamente porque não conhecemos os valores exactos da variável estatística, mas apenas o número de observações dentro de cada classe, isto é, quando os dados estão agrupados em classes, para o cálculo da média simples, devemos considerar o ponto médio de cada classe (marca) como representativo de todos os valores nela incluídos, pelo que aquela pode ser representada pela fórmula matemática seguinte:

$$\bar{x}' = \Sigma\, nix'i / n = \Sigma\, fix'i$$

Exemplo: Calcule a média aritmética, sabendo que

Classes	[3-5[	[5-7[	[7-9[	[9-11[	[11-13]
ni	2	4	5	3	2

Classes	ni	x'i	nix'i
[3-5[	2	3,5	7
[5-7]	4	5,5	22
[7-9[	5	7,5	37,5
[9-11[	3	9,5	28,5
[11-13]	2	12	24
	Σ=16		Σ=119

Então: $\bar{x}' = \Sigma\, nix'i / n = 119/16 = 7,43$

Temos de salientar que quando usamos a marca da classe estamos a colocar um certo erro de agrupamento, pelo que devemos considerar a Correcção de Sheppard, de tal modo que:

$\bar{x}$ é semelhante a $\bar{x}'$

Relativamente às propriedades da média aritmética podemos enunciar duas:

1. Somando ou subtraindo uma constante a todos os valores observados, a média resultante ficará aumentada ou diminuída, respectivamente, dessa constante;
2. Multiplicando ou dividindo os valores observados por uma constante diferente de zero, a média resultante ficará multiplicada ou dividida, respectivamente, por essa constante.

Média Aritmética Pesada ou Ponderada: É a média aritmética afectada por pesos (variável discreta e variável contínua).

Por outras palavras, associa-se a x1, x2, ..., xk certos factores de ponderação ou pesos p1, p2, ..., pk que dependem do significado ou importância atribuída às observações. Assim, a fórmula matemática da média será: $\bar{x}\, p = \Sigma\, pixi / pi$

Exemplo : Um professor de matemática quer saber a média poderada das suas avaliações nas quatro turmas em que lecciona, sabendo que o teste tinha uma ponderação de 30% e o trabalho uma ponderação de 70%:

Turma A - Média da nota do teste =65%	Média da nota do trabalho =78%
Turma B - Média da nota do teste =60%	Média da nota do trabalho =70%
Turma C - Média da nota do teste =40%	Média da nota do trabalho =28%
Turma D – Média da nota do teste =80%	Média da nota do trabalho =75%

Determine a média ponderada das quatro turmas em conjunto.

$\bar{x}p=\Sigma pixi/pi$ = [((65*30)/100) + ((78*70)/100) + ((60*30)/100) + ((70*70)/100) + ((40*30)/100) + ((28*70)/100) + ((80*30)/100) + ((75*70)/100)]/4 = 62,3

Se preferirmos em quadro a resolução será

xi	pi	pixi
65	30	1950
60	30	1800
40	30	1200
80	30	2400
78	70	5460
70	70	4900
28	70	1960
75	70	5250
Σ	400	24920

$\bar{x}\, p = \Sigma\, pixi / pi=24920/400=62,3$

Exercicio: Seja uma Amostra dos pesos de seis alunos de Administração. Encontre a média para: xi = 68, 56, 47, 66, 93, 56

Para além da média aritmética e da média aritmética ponderada, temos também a média geométrica, a média harmónica e a média quadrátrica (a estes três tipos de médias não iremos dar relevancia).

MEDIANA

MEDIANA (Md) é um valor que ocupa a posição central em uma série, logo, precisamos encontrar a posição média entre os dados.

A mediana de uma série de n observações x1, x2, ..., xn de uma variável X é o valor que ocupa a posição central quando as observações estão ordenadas por ordem crescente ou decrescente, isto é, a mediana de uma variável estatística é o valor dessa variável tal que a frequência dos valores que lhe são inferiores é a mesma que a frequência dos valores que lhe são superiores. Representa-se, habitualmente, por Md.

A mediana é usada quando na amostra há valores excêntricos em relação a outros valores.

Para o cálculo da mediana, temos de considerar duas situações: o caso em que N é ímpar e o caso em que N é par.

N é ímpar:

A mediana é um valor observado, de tal modo que o lugar que ocupa é dado pela fórmula

Md = (N +1) / 2

Exemplo: Determine a mediana para a seguinte série de dados

5 9 8 7 6

Ordenando por ordem crescente, vem 5 6 7 8 9

Como N é ímpar, então Md=(N+1) / 2 = (5+1) / 2 = 6/2 = 3 então a mediana ocupa a terceira posição ou terceiro termo, o seu valor é 7.

Exercicio: Determinar a Mediana da Amostra: X = 2, 20, 12, 23, 20, 8, 12.

Para n = 07 (ímpar) temos - Md = $(n+1)/2 =$

Interpretação: Podemos dizer que 50 % dos valores da série são menores ou iguais a _____ e que 50 % dos valores são maiores ou iguais a _____.

N é par:

A mediana não coincide com nenhum valor observado ficando compreendida entre dois valores centrais - classe mediana; convencionou-se tomar para mediana a média destes dois valores. A posição que a mediana ocupa é dada pela fórmula

Md = média dos valores que se encontram na posição N / 2 e (N / 2) + 1

Exemplo: Determine a mediana para a seguinte série de dados

5 6 12 9 8 7

Ordenando por ordem crescente, vem 5 6 7 8 9 12

Como N é par, então N/2 = 6/2 = 3; (N/2 + 1) = 3+1 = 4 Assim, a classe mediana é ocupada pelas posições 3 e 4, ou seja, pelos valores 7 e 8, pelo que a Md = (7+8) / 2 = 15/2 = 7,5

Exercicio: Determinar a Mediana para a Amostra X = 7, 21, 13, 15, 10, 8, 9, 13.

Para n = 08 (par) temos: Md 1 = e Md 2 =

Logo, a Md =

Interpretação: Podemos dizer que 50 % dos valores da série são menores ou iguais a _____ e que 50 % dos valores são maiores ou iguais a _____.

MODA

A moda (ou valor modal) de uma série de n valores x1, x2, ..., xn de uma variável X é o valor onde a frequência atinge o máximo (relativo). Representa-se, habitualmente, por Mo.

A moda é o valor da variável com maior efectivo, isto é, se uma variável é discreta, a(s) moda(s) é(são) o(s) valor(es) da variável estatística que se observa(m) com maior frequência.

Exemplo: Determine a moda para a seguinte série de valores

xi	4	6	8	10	2
ni	1	3	5	4	2

A moda é 8.

Exercicio: Seja uma Amostra aleatória dos pesos de seis alunos de Administração. Encontre a moda. se xi = 68, 56, 47, 66, 93, 56.

QUANTIS

Chama-se quantil de ordem p com $0 \leq p \leq 1$ e representa-se, habitualmente, por Cp ao valor de x tal que F(x)=p.

Alguns quantis têm denominações especiais:

Quartis:

Os quartis dividem a série ordenada em 4 partes iguais, contendo cada uma delas 1/4 ou 25% das observações.

Q1=1º quartil (corresponde ao quantil de ordem p=1/4)

Q2=2º quartil (corresponde ao quantil de ordem p=1/2)

Q3=3º quartil (corresponde ao quantil de ordem p=3/4)

Assim, Q1 é o valor da variável estatística que deixa atrás de si 25% das observações; Q2 é o valor da variável estatística que deixa atrás de si 50% das observações e Q3 é o valor da variável estatística que deixa atrás de si 75% das observações. A (Q1- Q3) chama-se intervalo interquartil e é o intervalo ao qual pertencem 50% das observações, deixando 25% para a direita e 25% para a esquerda.

É de notar que dizer que os quartis dividem a série em 4 partes iguais não significa que, por exemplo, os intervalos (Q1, Q2) e (Q2, Q3) têm a mesma amplitude, mas sim que contêm o mesmo número de observações.

Decis:

Os decis dividem a série ordenada em 10 partes iguais, contendo cada uma delas 1/10 ou 10% das observações.

D1=1º decil (corresponde ao quantil de ordem p=1/10)

D2=2º decil (corresponde ao quantil de ordem p=2/10)

Centis:

Os centis dividem a série ordenada em 100 partes iguais, contendo cada uma delas 1/100 ou 1% das observações.

C1=1º centil (corresponde ao quantil de ordem p=1/100)

C2=2º centil (corresponde ao quantil de ordem p=2/100)

RELAÇÕES ENTRE QUARTIS, DECIS, CENTIS E MEDIANA:

Como podemos observar na figura ao lado

Q1=C25

Q2=Md=D5=C50

Q3=C75

D1=C10

D2=C20

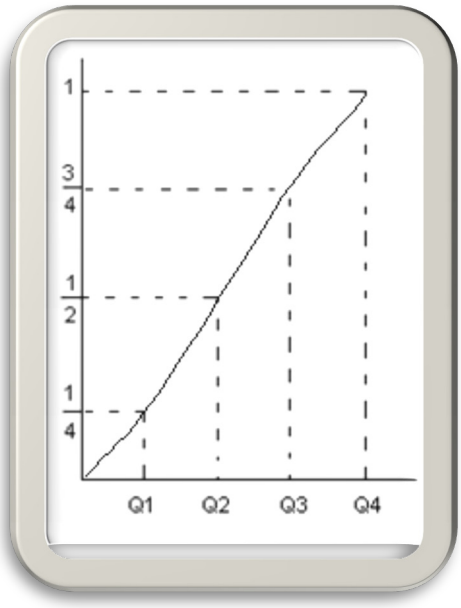


Figura 5: Quartis

MEDIDAS DE DISPERSÃO

Uma medida de tendência central não nos dá, só por si, uma informação exaustiva da distribuição considerada; pelo contrário, a capacidade que se lhe atribui de representar os elementos de uma distribuição depende do modo como estes se concentram ou dispersam em torno dela. Assim, podemos dizer que os parâmetros de tendência central não são suficientes para caracterizar uma série estatística, apesar de a mediana e os quantis darem já uma ideia sumária do modo como estão distribuídas as observações.

Consideremos o seguinte exemplo: Dois grupos de alunos com as seguintes classificações

A	2	3	10	16	19
B	8	9	10	11	12

A média e a mediana é 10 e, contudo, estas distribuições são muito diferentes. Com efeito, enquanto no grupo A as notas apresentam desvios muito grandes, na distribuição B todos os valores se aproximam de 10. A dispersão ou variabilidade da primeira série é mais acentuada do que na segunda.

Quer dizer: distribuições com a mesma tendência central podem apresentar aspectos bastante diferentes no que concerne à dispersão ou variabilidade, e à medida que esta dispersão aumenta, menos significativas da distribuição vão sendo as medidas de tendência central.

Assim, para melhor caracterizarmos uma distribuição, temos de considerar, além das medidas de tendência central, uma outra medida que exprima o grau de dispersão ou variabilidade dos dados.

Vamos considerar as seguintes medidas de dispersão: amplitude total, amplitude interquartis, desvio médio, variância, desvio padrão e coeficiente de dispersão ou de variação.

AMPLITUDE

A amplitude total é a diferença entre o maior valor e o menor valor, isto é, a amplitude total de uma variável estatística é a diferença entre o valor máximo e o valor mínimo dos valores observados. É a forma mais simples de avaliar a dispersão dos dados, de tal modo que quanto maior for a amplitude total maior é a dispersão dos dados.

A amplitude total pode ser também denominada de intervalo total ou campo de variação; representa-se, habitualmente, por A e apenas usa valores extremos.

Numa amostra seriada

A = xmáximo - xmínimo

Numa amostra classificada

A = extremo superior da última classe - extremo inferior da primeira classe

Se alguma destas classes for de amplitude indeterminada não é possível definir o intervalo de variação.

A amplitude total apresenta as seguintes desvantagens:

Embora seja fácil de calcular, a amplitude total depende somente dos valores extremos, que são, geralmente, os menos frequentes e os menos significativos de uma distribuição, desprezando-se os valores intermédios que são os mais frequentes. Além disso, os valores extremos são vulgarmente

anómalos e muito variáveis, consoante a amostra que se retire de uma população, de tal modo que duas distribuições podem ter a mesma amplitude total, mas dispersões muito diferentes.

Outro inconveniente da amplitude total é consequência de não tomar em consideração as frequências das observações.

Exemplo: Calcule a amplitude total do grupo G, sabendo que

G= 2 3 10 16 19

A = xmáximo - xmínimo = 19-2 = 17

Amplitude interquartis: Os quartis fornecem indicação quanto à forma como as observações se distribuem em torno da mediana.

Como o 1º e o 3º quartis representam valores abaixo dos quais estão, grosso modo, respectivamente, 25% e 75% das observações, entre eles existirão, assim, 50% das observações centrais. Consequentemente, quanto mais aproximados estiverem estes quartis, maior será a concentração das observações em torno da mediana.

A amplitude interquartis pode ser definida como a diferença entre o Quartil 3 e o Quartil 1. Esta medida de dispersão pode ser também denominada de intervalo interquartis ou intervalo quartílico. Como podemos observar na figura, quanto mais achatada é a curva maior é a amplitude e quanto maior é a amplitude interquartílica mais dispersa é a distribuição.

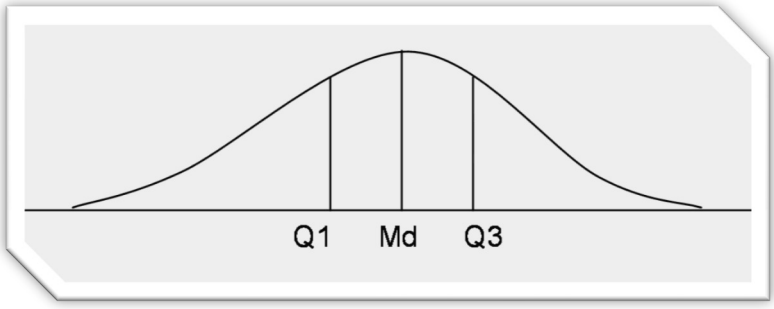


Figura 6: Curva simétrica achatada (platocurtica)

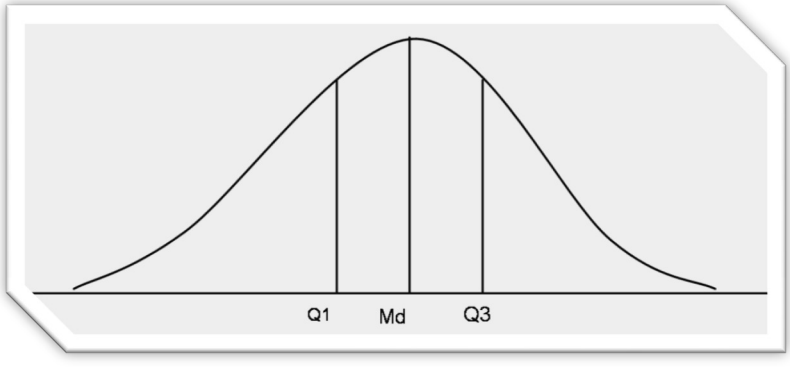


Figura 7: Curva simétrica mesocurtica

Exemplo: Calcule a amplitude interquartis, sabendo que Q3=177,46 e Q1=166,88.

Q = Q3-Q1 = 177,46-166,88 = 10,58

Podemos também calcular a amplitude semi-interquatis ou intervalo inter-quartílico ou intervalo semi-quartil ou desvio quartílico ou amplitude semi-interquartílico, que se representa, habitualmente, por Q e pode ser definida do seguinte modo:

$$Q = (Q3-Q1) / 2$$

A distribuição é tanto mais dispersa quanto maior for a amplitude interquartis.

Podemos enumerar as vantagens e as desvantagens do uso da amplitude interquartis e da amplitude semi-interquartis:

Estas medidas são mais completas do que a amplitude total, porque usam dois valores menos extremos (Q1 e Q3). No entanto, têm ainda a limitação de não entrarem em linha de conta com a disposição das frequências nos intervalos definidos pelos valores separados - exemplo: a amplitude interquartis será a mesma, quer as 50% das observações se acumulem num só ponto, quer estejam uniformemente distribuídas por esse intervalo interquartis.

As medidas de dispersão que passaremos a descrever não têm esta limitação, porquanto o seu cálculo depende de todos os valores da série.

DESVIO

Dados n valores x1, x2, ..., xn de uma variável X, chama-se desvio de cada valor **xi** em relação à constante **c**, a diferença de xi para c, isto é, **xi - c**

Note-se que os desvios da variável X em relação a c, isto é, (x1-c), (x2-c), ..., (xn-c) constituem os n valores da variável X-c.

DESVIO MÉDIO

Falamos em desvio médio quando consideramos os desvios de cada valor xi em relação à média aritmética, isto é:

$$xi- \bar{x}$$

O simples total destes desvios não pode ser utilizado como medida de dispersão, por ser identicamente nulo. De facto, para n valores singulares, ter-se-á:

$$\Sigma (xi- \bar{x})=0$$

No entanto, o quociente entre a soma dos módulos destes desvios e o número deles, já pode ser considerado como medida de dispersão

$$D.M.= \Sigma |xi- \bar{x}| / n$$

Exemplo: Calcule o desvio médio para A = 4 5 3

Tabela 3: Cálculo do Desvio Médio

xi	$xi- \bar{x}$	$\|xi- \bar{x}\|$
4	4-4=0	0
5	5-4=1	1
3	3-4=-1	1
Σ=12	3-4=-1	Σ = 2

$\bar{x} = \Sigma xi / n =12/3=4$

D.M.=Σ |xi-x| / n = 2/3=0,67

Observação: também se utiliza o desvio médio em relação a qualquer outra medida de posição central.

Numa amostra seriada temos:

D.M.=Σ ni |xi-x| / n = Σ fi |xi-x|

Se os valores da variável estiverem tabelados de modo que cada valor xi corresponda a frequência absoluta ni, o desvio médio é igual à soma dos produtos das frequências pelos valores absolutos dos respectivos desvios em relação à média, dividida pelo efectivo da distribuição.

Numa amostra classificada, os desvios em relação à média aritmética são calculados a partir dos pontos médios de cada classe, ou seja,

$$D.M.= \Sigma\ ni\ |x'i-\ \bar{x}'|\ /\ n = \Sigma\ fi\ |x'i-\ \bar{x}'|$$

Exemplo: Calcule o desvio médio para classes [4-6[[6-8]

ni 1 2

Tabela 4: Cálculo do Desvio Médio para classes

Classes	ni	x'i	x'i- $\bar{x}$'	|x'i- $\bar{x}$'|	ni| $\bar{x}$'i-x'|
[4-6[	1	4,5	4,5-6,2	1,7	1,7
[6-8]	2	7	7-6,2	0,8	1,6
				Σ=2,4	Σ=3,3

$\bar{x}' = \Sigma\ nix'i\ /\ n = [(1x4,5) + (2x7)]\ /\ 3 = 6,2$

Então, o desvio médio é

D.M.=Σ ni |x'i-x'| / n = 3,3 /3 = 1,1

O desvio médio apresenta a seguinte desvantagem: Embora dependa de todos os valores observados, o desvio médio tem a desvantagem de considerar os valores absolutos dos desvios, o que impede o seu tratamento algébrico.

VARIÂNCIA

Outra maneira de eliminarmos os sinais dos desvios, consiste em elevá-los ao quadrado. Por isso, em vez da média dos valores absolutos dos desvios considera-se a média dos quadrados dos desvios. Obtém-se, assim, uma outra medida de dispersão bastante usada - a variância.

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Dados n valores x1, x2, ..., xn de uma variável X, chama-se variância e representa-se, habitualmente, por s2 ou s2x a média aritmética dos quadrados dos desvios em relação à média dessas valores, isto é,

$$S^2 = \Sigma (xi- \bar{x})^2 / n-1$$

Exemplo: Calcule a variância para X=17,18,19,20,21

Resolução:

1.º passo: calcular a média $\bar{x} = \Sigma xi / n = (17+18+19+20+21) / 5 = 19$

Então, a variância é

Tabela 5: Cálculo da Variância

xi	$(xi- \bar{x})$	$(xi- \bar{x})^2$
17	17-19=-2	4
18	18-19=-1	1
19	19-19=0	0
20	20-19=1	1
21	21-19=2	4
N=5		Σ=10

$$S^2 = \Sigma (xi- \bar{x})^2/ n = 10/(5-1) = 2,5$$

Se x1, x2, ..., xn ocorrem n1, n2, ..., nk vezes, respectivamente, temos

Para uma amostra seriada:

$$S^2 = \Sigma ni \ (xi- \bar{x})^2 / n-1$$

Para uma amostra classificada:

$$S'^2 = \Sigma ni \ (x'i- \bar{x}')^2 / n-1$$

Exemplo: Calcule a variância para a distribuição cuja média é 103

Classes	[85-90[	[90-95[	[95-100[	[100-105[	[105-110[	[110-115[	[115-120]
ni	12	25	38	85	93	16	9

Tabela 6: Exercicio - Cálculo da Variância

Classes	ni	x'i	(nix'i)	(x'i- $\bar{x}$')	(x'i- $\bar{x}$')2	ni(x'i- $\bar{x}$')2
85-90	12	87	1044	-15,5	240,25	2883
90-95	25	92	2300	-10,5	110,25	2756,25
95-100	38	97	3686	-5,5	30,25	1149,5
100-105	85	102	8670	-0,5	0,25	21,25
105-110	93	107	9951	4,5	20,25	1883,25
110-115	16	112	1792	9,5	90,25	1444
115-120	9	117,5	1057,5	15	225	2025
	278		28500,5			12162,25

$\bar{x}' = \Sigma (nix'i) / n$ =28500,5/278 =102,5

Então, a variância é

$$S'^2 = \Sigma\, ni\, (x'i- \bar{x}')^2/ n = 12162,25/ (278-1) = 43,91$$

Podemos, agora, enumerar as propriedades da variância:

Somando ou subtraindo uma constante a todos os valores observados, a variância resultante permanecerá inalterada;

Multiplicando ou dividindo todos os valores observados por uma constante diferente de zero, a variância resultante virá multiplicada ou dividida, respectivamente, pelo quadrado dessa constante.

Correcção de Sheppard:

Ao calcular-se a média e a variância da amostra classificada através da distribuição empírica das marcas, comete-se um certo erro (erro de agrupamento), pois supomos que as observações agrupadas em cada classe têm todas o valor da respectiva marca. Todavia, existem fórmulas correctivas devidas a Sheppard, isto é,na variancia, ao valor calculado deve subtrair-se 1/12 ao quadrado da amplitude das classes (a)

$s^2x = s^2x' - a^2/12$ assim, no nosso exemplo anterior, a variancia corrigida era $s^2x = 43,91 - (5^2/12)$

$s^2x = 41,827$

DESVIO PADRÃO

O desvio padrão pode ser definido como a raiz quadrada da variância, representando-se, habitualmente, por sx, isto é,

$$sx = \sqrt{s^2x}$$

Ainda que a variância nos dê uma boa informação sobre a distribuição ou variabilidade dos valores observados em relação à sua média, apresenta, no entanto, a desvantagem de não se exprimir na mesma unidade a que estão referidos os dados iniciais. Contudo, esta desvantagem poderá ser eliminada se extrairmos a raiz quadrada da variância. A nova medida chama-se desvio padrão ou desvio quadrático.

Numa amostra seriada, temos:

$$sx = \sqrt{s^2x}$$

Exemplo: Calcule o desvio padrão, sabendo que a variância de uma amostra seriada é 2.

$sx = \quad s^2x = \quad 2 = 1{,}414$

Numa amostra classificada, temos

$$s'x = \sqrt{s^2x'}$$

Exemplo: Calcule o desvio padrão, sabendo que a variância corrigida de uma amostra classificada é 4327,16.

UTILIZAÇÃO DAS MEDIDAS ABSOLUTAS DE DISPERSÃO

Amplitude total

Utiliza-se quando:

Os dados forem muito raros ou demasiado dispersos para se justificar o cálculo de uma medida mais precisa de dispersão;

For apenas necessário o conhecimento dos resultados extremos;

Desejamos um índice muito rápido de dispersão.

Amplitude interquartílica

Utiliza-se quando:

A mediana é a medida de tendência central usada;

Existirem resultados extremos que poderiam afectar o desvio padrão de uma maneira desproporcionada;

A distribuição é truncada;

A distribuição apresenta uma forte assimetria.

Desvio médio

Utiliza-se quando:

Desejamos ponderar todos os desvios em relação à média de acordo com a sua grandeza;

Os desvios extremos influenciarem indeterminadamente o desvio padrão.

Desvio padrão e Variância

Utilizam-se quando:

Se procura uma medida de dispersão em relação com a curva normal;

Tiverem de ser calculados posteriormente coeficientes de correlação e outras estatísticas;

Se desejar obter uma medida que se revista de um máximo de estabilidade;

Se se trata somente de descrever uma distribuição prefere-se o desvio padrão à variância. A variância intervém sobretudo na análise estatística.

COEFICIENTE DE DISPERSÃO OU VARIAÇÃO

As medidas de dispersão a que anteriormente nos referimos são medidas que se exprimem na mesma unidade dos dados e, sendo assim, torna-se impossível comparar entre si as dispersões de duas distribuições cujos valores não se refiram à mesma unidade.

Exemplo: Distribuição A: $\bar{x}_A=30$ $s_A=10$

Distribuição B: $\bar{x}_B=600$ $s_B=20$

Qual é a distribuição mais dispersa? Se compararmos os desvios padrões é a B, porque tem maior desvio padrão. Mas a variação de 20 para 600 é muito maior do que 10 para 30. Assim, em vez de

compararmos os desvios padrões, aplicamos outra medida de dispersão relativa que é o coeficiente de variação ou de dispersão, que pode ser definido pela fórmula.

$$CV = \frac{S}{\overline{X}} \cdot 100\%$$

Exemplo 26: A distribuição dos pesos e das alturas de um grupo de estudantes de determinada Universidade conduziu aos seguintes resultados:

X: Pesos	Média=57,5Kg	Desvio Padrão=7,5Kg
Y: Alturas	Média=170cm	Desvio Padrão=7,1cm

Determine o coeficiente de dispersão para cada uma das distribuições e, depois, indique em qual delas a dispersão relativa é maior.

$V_x = sx / \overline{x} = 7,5Kg / 57,5Kg = 0,130 = 13\%$

$V_y = sy / \overline{Y} = 7,1cm / 170cm = 0,042 = 4,2\%$

Assim, podemos dizer que a dispersão relativa é mais acentuada na distribuição dos pesos (X).

A dispersão é maior na distribuição que tiver maior coeficiente de dispersão.

Se pretendermos estabelecer comparações entre dispersões absolutas, devemos usar o desvio padrão, de tal modo que quanto maior for o desvio padrão maior será a dispersão.

Se pretendermos estabelecer comparações entre dispersões relativas, devemos usar um coeficiente de dispersão, de tal modo que quanto maior for o coeficiente de dispersão V maior será a dispersão.

5. CARACTERÍSTICAS DA DISTRIBUIÇÃO NORMAL

A distribuição normal é simétrica e apresenta uma curva em forma de sino, como mostra a figura. A sua principal característica é a de as três medidas de tendência central - média, mediana e moda - Se encontrarem todas no mesmo ponto da curva, ou seja, todas terem o mesmo valor ou, pelo menos, valores muito próximos. Se os elementos que constituem uma distribuição estão muito próximos ou muito dispersos, encontraremos assimetrias positvas ou negativas, consoante a media seja inferior à mediana e moda (negativa) ou superior às mesmas (positiva).

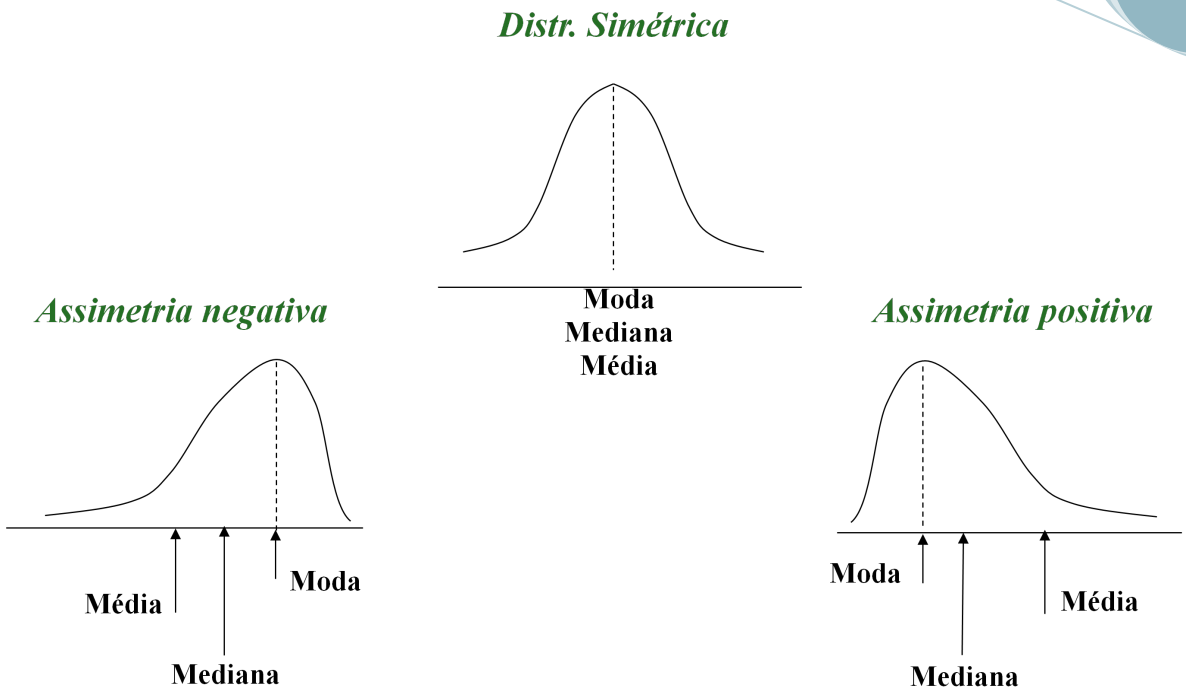


Figura 8: Distribuição normal

Estas não são distribuições normais, apesar de a média, a mediana e a moda se encontrarem todas no mesmo ponto (é isto que confere simetria à distribuição); a distribuição normal tem sempre a forma de um sino. Como foi «descoberta» pelo matemático Gauss, também lhe chamamos distribuição gaussiana.

A maior parte dos dados recolhidos com organismos vivos têm este padrão. Podemos observar que, devido à forma da curva, há poucos resultados muito baixos e poucos resultados muito eleva. dos (a curva «cai» nos extremos esquerdo e direito, o que se deve às baixas frequências encontradas), enquanto a maioria dos resultados se encontram junto à média. Vamos debruçar-nos sobre o padrão de resultados muito em breve, mas nesta fase referiremos outra característica muito importante da distribuição normal. Teóricamente, a curva nunca toca o eixo horizontal, mas aproxima-se dele infinitamente. Esta é uma propriedade matemática da distri' buição que não se reflecte na recolha de dados «real». Não nos cruzamos com seres humanos com dimensões gigantescas ou micros' cóspicas!

Então as propriedades da distribuição normal são as seguintes:

1) É simétrica;

2) Tem forma de sino;

3) A média, a mediana e a moda encontram-se no mesmo ponto da curva;

4) Tem duas pontas que nunca tocam o eixo horizontal.

Podemos perguntar-nos quão rígida é a distribuição normal relativamente aos elementos. Por outras palavras, quanto pode uma curva desviar-se da forma de sino e continuar a ser considerada normal? Geralmente, usamos duas abordagens para tomarmos este tipo de decisão; na verdade, este problema é mais importante do que o leitor pode pensar, pois existem testes estatísticos, descritos mais à frente, que só podem realizar-se se os elementos forem normalmente distribuídos. Uma das abordagens baseia-se na observação dos dados «por averiguação», para lhe darmos um nome mais respeitável. Se o conjunto de números for extenso, tornar-se-á mais fácil desenhar uma distribuição de frequências. A outra abordagem reside em seguir um dos procedimentos matemáticos para determinar se um conjunto de resultados é normalmente distribuído. A versão do teste do quiquadrado que incluímos neste manual é um desses procedimentos. Na verdade, é improvável que nesta fase da sua carreira estatística necessite de saber com grande precisão se uma distribuição é considerada normal ou não, pelo que o teste gráfico deve bastar. No entanto, deve ser sensível ao problema.

5.1 A CURVA NORMAL E OS DESVIOS-PADRÃO

Suponhamos que temos um conjunto de números cuja média é 50 e cujo desvio padrão é 5. Chamamos a este valor (centímetros, segundos, pontos numa escala, ou outra coisa qualquer) um desvio padrão. Dez centímetros, segundos, etc., seriam dois desvios padrão e quinze centímetros, segundos, etc., três desvios padrão ... sempre com referência ao conjunto através do qual obtivemos o valor 5. É como se pudéssemos tirar o desvio padrão e transformá-lo numa unidade de medida de uma escala; é como se disséssemos que uma polegada são 2,54 cm. Nunca misturaríamos polegadas e centímetros nos mesmos cálculos, mas poderíamos converter uma unidade na outra. Do mesmo modo, não misturamos resultados de desvios padrão com resultados reais, mas convertemos uma escala na outra.

Voltemos às proporções de números em diferentes partes da distribuição. Se retirarmos uma parte da curva entre a média, que esta marcada no eixo horizontal da figura que se segue como 50, e um desvio padrão, marcado no eixo horizontal como 55, sabemos que devemos ter cerca de um terço de todos os resultados neste conjunto, porque é o que acontece sempre com a distribuição normal. De uma forma mais precisa, a proporção exacta do número total de resultados que se encontram entre a média e um desvio-padrão acima da média (50 e 55 neste caso) é 34,13 %. Como a distribuição normal é simétrica, deve verificar-se a mesma coisa abaixo da média, isto é, devemos ter outros 34,13 % dos resultados entre os valores 50 e 45 - sendo 45 o valor da média menos um desvio-padrão de 5 pontos. Observemos as duas partes a sombreado. A aritmética elementar diz-nos que 68,26 % do total dos resultados se encontram entre os valores 45 e 55, pertencendo 31,74 % aos valores extremos que se

encontram nos outros dois lados. Mais uma vez, a simetria da curva significa que, para esta proporção, metade de 31,74 %, ou seja, 15,87 %, encontra-se de cada um dos lados da distribuição. Por outras palavras, cerca de 16 % de todos os números neste conjunto serão menores do que 45, sendo a mesma quantidade maior do que 55.

Exemplo: Suponhamos que o professor obtém os resultados de um teste de leitura feito a 200 crianças. Os resultados são normalmente distribuídos com uma média de 60 e um desvio-padrão de 8. A partir das propriedades da distribuição normal, sabemos que cerca de dois terços dos resultados, isto é, aqueles que foram obtidos com cerca de 136 crianças, encontrar-se-ão entre os 52 e os 68 pontos. Cerca de 32 crianças (16 %) terão resultados abaixo de 52 e cerca de 32 terão resultados acima de 68. Já os referimos todos. Suponhamos então que os pais de uma criança que obteve 68 gostariam de saber algo acerca do progresso do seu filho. Quando souberam que o resultado da criança estava «acima da média», ficaram contentes, mas gostariam de saber, posteriormente, quão acima da média se encontra, relativamente aos outros 50 % de crianças que também obtiveram resultados «acima da média». Por outras palavras, os pais pretendem saber qual a posição relativa do desempenho do filho. Se os resultados estivessem todos muito perto da média, sendo a nota máxima 68, os pais continuariam encantados. Ficariam, porém, menos satisfeitos se soubessem que a nota máxima tinha sido 90, com um grande conjunto de notas altas, acima de 70. No entanto, o professor sabe que o desvio-padrão das notas foi 8 e, por isso, um terço de todos os resultados estava entre 60 e 68. Sabendo que 50 % dos resultados obtidos estavam «abaixo da média», podemos perceber que a posição desta criança está ao nível de 84 % dos resultados, na parte superior de todos os resultados. Afinal, os pais têm razões para estarem contentes! Se a criança tivesse obtido 76, os pais teriam muito mais razões para estarem orgulhosos, pois saberiam que o seu filho estava acima de 98 % das outras crianças (nota 76 e dois desvios-padrão acima da média); uma nota 84 colocaria o menino na posição invejável de estar acima de 99,87 % das outras crianças - por outras palavras, num grupo de 200 crianças, estaria, muito provavelmente, no topo. Os desvios-padrão cortam proporções fixas da distribuição normal, a partir da média e até ao infinito (pelo menos teoricamente), nas duas direcções. Deve certificar-se de que percebeu como se obtém a posição relativa da nota 76 (isto é, 50 % + 33 % + 15 %) e como se calcula que neste grupo de crianças existem outras quatro com notas acima de 76. Veja se consegue calcular a nota que colocaria a criança na posição, menos invejável, de estar apenas a quatro lugares do fim.

A resposta é 44. Para obtermos esta nota necessitamos de saber que nota representa dois desvios-padrão abaixo da média ou que nota corresponde a 2 %. Partindo de 60, a média, se lhe subtrairmos o valor de dois desvios-padrão - 16, duas vezes o valor de 8, que é um desvio-padrão - obteremos 44. Devemos ter cuidado e não misturar os valores dos desvios-padrão com os resultados reais. Neste exemplo não subtraímos o valor 2 da média de 60, apesar de querermos o resultado que estava dois

desvios-padrão abaixo dele. Subtraímos 16 pontos, pois este é o número que corresponde a dois desvios-padrão para este conjunto de re sultados.

Resultados z Nos exemplos considerados os resultados encontravam-se sempre na média, ou exactamente um, dois ou três desvios-padrão acima ou abaixo dela. Temos, porém, de examinar resultados que não sejam tão facilmente convertíveis para desvios-padrão. Suponhamos, por exemplo, que uma criança com pais ansiosos obteve uma nota 64 num teste de leitura. A posição da criança na curva seria a metade da distância, no eixo horizontal, entre o resultado da média (60) e um desvio-padrão acima (68).

A posição da criança é exactamente a meio entre os pontos 60 e 68. Significará isto que a sua posição no grupo é o ponto central entre a média de 50 % e 84 % da nota 68? Isto é, encontrar-se-á a criança acima de 67 % dos colegas? Olhemos cuidadosamente para as duas porções da curva que está dividida pela linha ao nível da nota 64. Serão simétricas? Não - e aqui temos um problema que torna o cálculo de uma posição relativa muito mais complicado e cansativo do que gostaríamos. Quanto mais nos afastamos da média, menos resultados correspondem às diferentes proporções. Assim, se tivermos duas porções entre 60 e 64 e entre 64 e 68, haverá menos resultados neste último intervalo. Haverá ainda menos no intervalo séguinte, entre as notas 68 e 72, e assim sucessivamente. Isto também é verdadeiro para os resultados abaixo da média, mas, neste caso, são os resultados mais elevados, e não os mais baixos, que se encontram mais perto da média. Há muito menos resultados entre 44 e 48 do que entre 48 e 52, apesar de, em ambos os casos, a variação de notas ser de 4 pontos, ou seja, meio desvio-padrão. Quando olhamos para a forma de uma distribuição normal, o tamanho diferente das proporções que cada desvio-padrão compreende parece óbvio. No entanto, o problema de decidir a posição relativa de uma nota 64, quando comparada com os resultados, não desapareceu. Como podemos determiná-la? A resposta é dada através de resultados z. Os resultados z correspondem a desvios-padrão e, na verdade, são virtualmente a mesma coisa, excepto no facto de um resultado z se referir sempre à posição de um ponto em relação a média. Isto vai tornar-se claro em breve. Para já, pensemos que um resultado z de 1 é a mesma coisa que um desvio-padrão de 1, que um resultado z de 2 e um dp 2, e assim por diante. Como não há, virtualmente, nada numa distribuição normal depois do terceiro desvio-padrão ou resultado z - em qualquer das direcções -, é r,aro que os desvios-padrão ou os resultados z incluam o valor 4. E comum referirmo-nos aos resultados z como mais ou menos; aos desvios-padrão descrevemo-los como situando-se acima ou abaixo da média, em vez de mais ou menos. Um desvio-padrão tem um valor definido não variável, enquanto um resultado z se refere a uma posição relativa na curva e é referido em função da média. Como, até agora, um resultado z tem o mesmo significado que um desvio-padrão acima da média, podemos considerar que os resultados z e os desvios-padrão são iguais. No entanto, um desvio-padrão pode referir-se a um conjunto de resultados que distem um desvio-padrão de qualquer ponto da curva, enquanto os resultados z têm

posições fixas. Um resultado z de + 1 corresponde exactamente a um desvio-padrão acima da média, e não a qualquer conjunto de resultados que constituam um desvio-padrão. Voltemos ao problema do resultado de 64 e à sua posição relativa. Sabemos que a sua posição é exactamente metade de um desvio-padrão acima da média, pelo que lhe damos um resultado z de + 0,5.

Há tabelas que nos permitem ver muito facilmente onde os resultados z se situam na curva normal. Procure uma tabela estatística da distrbuição normal vejamos como utilizá-las. Utilizaremos o nosso exemplo de 64, cujo valor z é + 0,5.

Lemos o valor na primeira coluna da esquerda, encabeçada por z, até chegarmos ao valor 0,5. Olhamos para a coluna à direita e vemos o número 19,15. Temos de somar 50 %, de modo a obtermos o valor 69,15. Sabemos então que há 69,15 % dos resultados abaixo de 64 e 30,85 % acima. Devemos arredondar os valores para 69 % e 31 %, respectivamente. Consideremos outro exemplo, desta vez com o valor 65. Este valor está 5 pontos acima da média e o desvio-padrão para o conjunto é de 8. Um resultado de 5 pontos acima da média é 5/8 de desvio-padrão acima da média. Se fizermos as contas, sabemos que z é + 0,63. Como se encontra acima da média, o seu valor é positivo. Voltemos à tabela. Como z tem, desta vez, duas casas decimais, os procedimentos vão ser um pouco diferentes. O valor imediatamente à direita (22,7) é a percentagem correcta para um resultado z de 0,6. No entanto, o nosso resultado é 0,63, pelo que temos de andar três colunas da tabela até ao valor 0,03, no topo. Este valor, somado ao valor 0,6, dá-nos o z de 0,63 - ou seja, 23,57. Como o nosso z é positivo, devemos somar-lhe 50 % para obtermos o valor final de 73,57. Assim, a nota 65 está à frente de 74 % da escala. Podemos ver pela tabela que 49 % de todas as notas em cada um dos lados da curva estão incluídas num z de 2,33 ou um bocadinho mais abaixo, para sermos mais precisos). Notemos que, matematicamente, as caudas da curva nunca tocam o eixo horizontal, nem incluem todos os resultados possíveis.

Reparemos agora na posição relativa de uma pessoa que obtenha um resultado abaixo da média, digamos uma nota 41 na amostra original. Esta nota está 19 pontos abaixo da média, apenas um pouco menos do que dois desvios-padrão. Para sermos precisos, está 19/8 ou 2,375 abaixo. O seu z será - 2,375. Na tabela SI iem anexo vemos que um z de + 2,3 inclui 48,93 % dos resultados, mas o nosso resultado z é o valor um pouco superior de 2,375. A nossa tabela só pode ser usada com duas casas decimais, pelo que vamos arredondar este valor para 2,38. Paramos, desta vez, junto da coluna de 0,08 e obtemos o valor 49,13. Assim, um z de + 2,38 inclui 50 % + 49,13 % = 99,13 % de todos os resultados. Até agora tudo bem, mas o problema é que o nosso valor era negativo. Basta virarmos a nossa curva ao contrário e trabalharmos com a sua imagem ao espelho. Assim, com o nosso valor - 2,38 sabemos que 99,13 % de todas as notas da distribuição estão acima dele e apenas 0,87 % abaixo. Se considerarmos esta pequena proporção de 1 %, devemos esperar que, na nossa amostra de 200 indivíduos, 1 %, ou seja, dois indivíduos tenham notas inferiores a 41. No outro extremo das notas,

devemos esperar que apenas dois alunos tenham notas de 19 ou mais pontos acima da média, ou seja, notas que excedam os 79 %.

O modo de obter o valor z é dado pela expressão formal

Z = desvio da nota em relação à média

desvio-padrão

Se o desvio em relação à média tiver um sinal positivo ou negativo, se estiver acima ou abaixo da média, respectivamente, z ficará com o sinal correcto.

Nota: : tenha cuidado quando trabalhar com *z* e *dp*, de modo a usá-los sempre que os dados através dos quais foram obtidos sigam uma distribuição normal Ou aproximadamente normal. De outro modo, arranjará confusões

6. TESTES ESTATISTICOS

Estatística Paramétrica: calcula as diferenças numéricas exactas entre os resultados.

Estatística Não paramétrica: considera se certos resultados são superiores ou inferiores a outros resultados.

Requisitos para utilização de testes paramétricos

Quando se pretende empregar um teste t de Student ou uma análise da variância para fazer comparações entre amostras (testes paramétricos), existe uma lista de requisitos que inclui, entre outros:

que a variável tenha sido mensurada num nível mínimo intervalar;

que a distribuição seja simétrica e mesocurtica;

que a característica estudada (variável) tenha distribuição normal numa dada população.

Sempre que não se pode, honestamente, admitir a simetria e a normalidade de distribuição, ou os dados foram recolhidos num nível de mensuração inferior ao intervalar, devemos recorrer a testes que não incluem a normalidade da distribuição ou nível intervalar de mensuração. Esses testes chamam-se não paramétricos

Vantagens dos testes não-paramétricos

Podem ser utilizados, mesmo quando os seus dados só podem ser medidos num nível ordinal, isto é, quando for apenas possível ordená-los por ordem de grandeza) podem ser utilizados mesmo quando os seus dados são apenas nominais, i.e., quando os sujeitos podem apenas ser classificados em categorias.

Poder de um teste

O poder de um teste é a probabilidade de rejeitarmos a H0 quando ela é realmente nula

Os testes mais poderosos (os que têm maior probabilidade) de rejeição de H0, são testes que possuem pré-requisitos mais difíceis de satisfazer (testes paramétricos como t e F).

As alternativas não paramétricas exigem muito menos pré-requisitos mas produzem testes de significância com menos poder que os correspondentes paramétricos.

Em consequência

Ao rejeitar-se a H0 sem preencher as exigências mínimas dos testes paramétricos, é mais provável que essa rejeição seja falsa (se rejeitar a H0 quando ela é verdadeira comete um erro de tipo I; se aceitar a H0 quando ela é falsa comete um erro de tipo II). Quando os requisitos de um teste paramétrico são violados, torna-se impossível conhecer o seu poder e a sua dimensão (α)

É obvio que os investigadores querem, a todo o custo, rejeitar a H0 quando ela é mesmo falsa, evitando um erro de tipo I.

O teste ideal seria aquele que $\alpha=0$ e $\beta=1$, o que implicaria que o teste conduziria sempre à decisão correcta, contudo este teste ideal raramente existe.

A probabilidade do erro de 1ª espécie deve ser reduzida, fixando α teórico em 0,1; 0,05 ou 0,01. o valor fixado para α depende da importância que se dá ao facto de rejeitar a H0 quando esta é verdadeira.

Uma ilustração deste ponto de vista pode ser feita com o seguinte exemplo:

Uma pessoa é inocente até prova do contrário

H0: A pessoa é inocente

H1: A pessoa é culpada

Erro I: A pessoa é condenada mas está inocente

Erro II: A pessoa é absolvida mas é culpada

Naturalmente a justiça procura reduzir a possibilidade de ocorrer o erro de 1ª espécie, pois entende-se que é mais grave condenar inocentes que absolver criminosos.

Para certos sistemas judiciais um $\alpha = 0,1$ é demasiado elevado, optando por $\alpha=0,01$; noutros sistemas judiciais pode admitir que $\alpha= 0,05$ é um valor razoável.

ASSIM …

Fixada a probabilidade do erro de tipo I (dimensão do teste), o teste mais potente é aquele em que a escolha da região critica minimiza a probabilidade do erro de 2ª espécie. Diz-se também que esta região critica é a mais potente.

Facilmente se conclui que o teste mais potente é aquele que, uma vez fixada a probabilidade de rejeitar a H0, quando ela é verdadeira, maximiza a potência ou a capacidade para rejeitar a mesma hipótese quando esta é falsa.

Pressupostos

Para saber se uma variável é simétrica dividimos o coeficiente assimetria (Skewness) pelo erro padrão e se o resultado estiver entre 2 e -2 a distribuição é simétrica.

Para saber se uma variável é mesocurtica dividimos o coeficiente de achatamento (Kurtosis) pelo erro padrão e se o resultado estiver entre 2 e -2 a distribuição é mesocurtica.

Mas se os resultados de um teste paramétrico, não cumpriram com os requisitos (no mínimo dados intervalares; distribuição simétrica, mesocurtica e normal), então não têm interpretação significativa.

Quando acontecem estes factos, a maioria dos investigadores opta por testes de significância não-paramétricos.

Para escolher qualquer tipo de teste estatístico

Distinguir se a nossa amostra é constituída pelos mesmos sujeitos em todas as situações ou se é formada por diferentes sujeitos para cada situação

Inter-sujeitos ou design não-relacionado

este tipo de design é utilizado quando um indivíduo ou objecto é avaliado apenas uma vez. A comparação é efectuado entre os grupos de sujeitos/ objectos cujos resultados são não-relacionados.

Desvantagem: conjunto das diferenças individuais na forma como os sujeitos reagem ou respondem à tarefa.

Intra-sujeitos ou design relacionado

A comparação é feita entre os mesmos sujeitos (sujeitos do mesmo grupo).

A importância destes designs é a eliminação de quaisquer particularidades individuais, uma vez que ficam igualizadas em todas as situações.

Desvantagem: Efeito de memória e aprendizagem.

Amostras emparelhadas

Igualizam-se sujeitos diferentes mas emparelhados, em termos de idade, sexo, profissão e outras características gerais que parecem importantes para cada pesquisa em particular.

estes tipos de designs podem ser considerados de designs relacionados, uma vez que é controlado nas suas características relevantes.

Desvantagem: Dificuldade em encontrar sujeitos que permitam o emparelhamento de todas as características relevantes.

Dificuldades arranjar grandes amostras.

ESTRATÉGIAS ESTATISTICAS DE ANÁLISE DE DADOS

A maioria dos investigadores principiantes enfrenta sérias dificuldades quando tem de usar a análise estatística. É apontado como prováveis causas o ensino de Estatística que, frequentemente, tem um enfoque matemático ou de receita que não conduzem ao aproveitamento desta ferramenta e o consequente despoletar de uma "ansiedade matemática", que pode levar os estudantes a evitar o seu uso. Essa situação conduz, não raras vezes, à dependência de outros para seleccionar a estatística adequada ao seu projecto. O objetivo desta lição é ajudar a ter uma idéia da potencialidade da estatística apropriada a sua pesquisa.

Primeiro examine seu estudo, identifique o que quer com sua análise estatística, devendo, para isso, especificar claramente as várias questões a que quer que sua análise estatística responda (conhecer a

associação ou verificar as diferenças). Comece por escrever as suas questões de pesquisa e hipóteses. Depois identifique a variável dependente e independente bem como os seus níveis de mensuração. Apos estar na posse dessa informação consulte a figura que se segue e vai ver que tudo começa a ficar mais fácil.

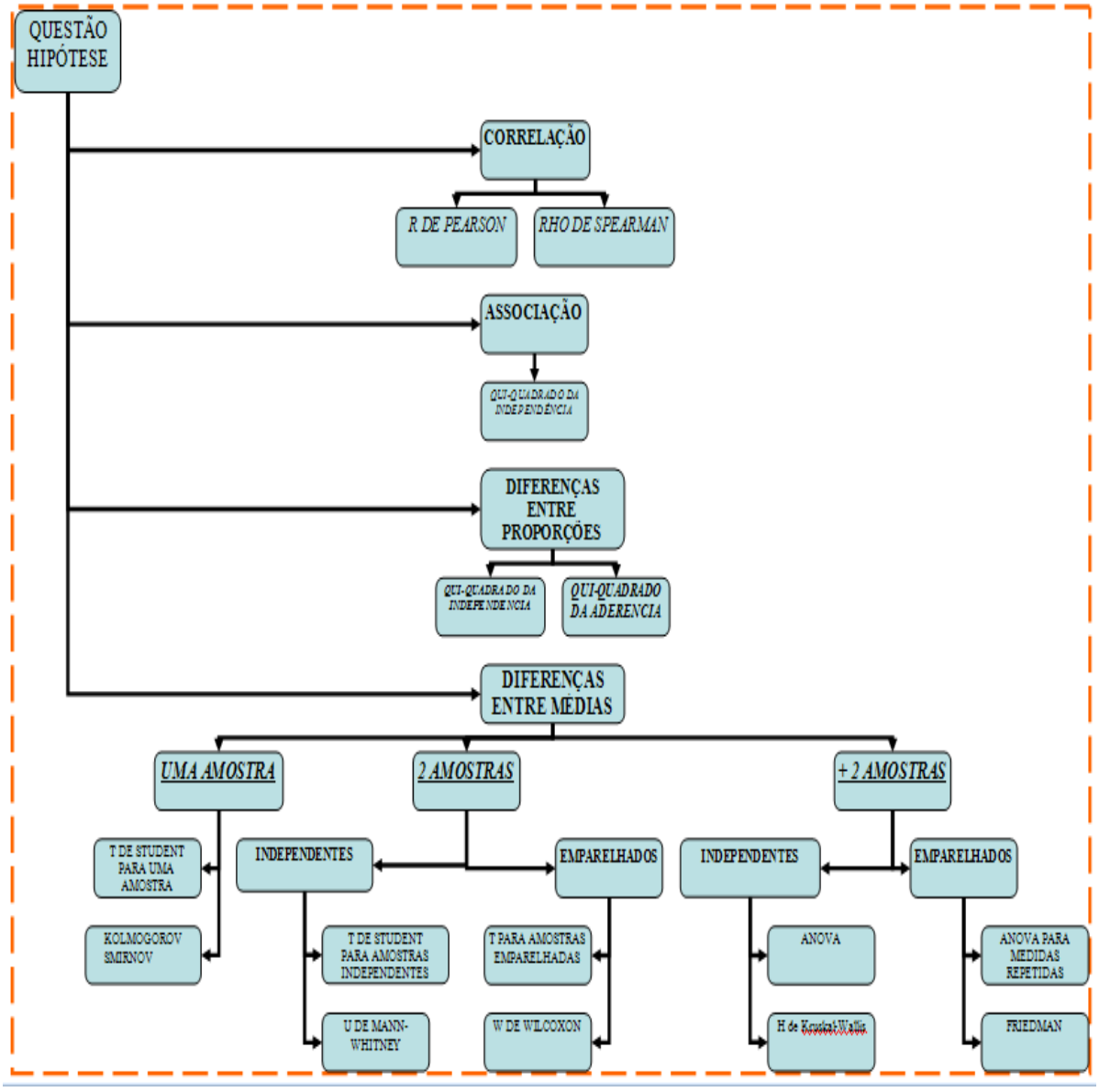


Figura 9: identificar os testes estatisticos

Como **segundo passo** na escolha da estatística apropriada, verifique se sua variável dependente é adequada para a estatística paramétrica. A **estatística paramétrica** envolve pelo menos dois pressupostos iniciais: o primeiro é se a variável dependente segue uma **distribuição normal** e, o segundo, é se os dados entre diferentes sujeitos são **independentes ou emparelhados/relacionados**.

Portanto, uma variável dependente qualitativa ou categórica não se enquadra neste tipo de estatística, devendo usar o enfoque da estatística não paramétrica.

Assim recorremos a e**statística paramétrica** quando analisamos **variáveis dependentes contínuas**. Se essas variáveis violam os pressupostos e não tem como corrigir essa violação, então deve utilizar a e**statística não paramétrica**. Só tem duas opções: ou aprende a lidar com a Estatística não paramétrica ou então aumenta o tamanho da amostra.

Examine cada variável dependente uma por uma nesse processo. Nem todas terão as mesmas características. Um **erro comum**, por exemplo, é assumir que pode usar sempre o mesmo teste estatístico se os grupos experimentais são equivalente em idade, género, anos de estudos e outras variáveis demográficas. Idade e anos de estudo são duas variáveis geralmente analisadas com estatística paramétrica. O género e a etnia são variáveis nominais e por isto devem ser analisadas com Estatística não paramétrica.

Definir quais as estratégias estatísticas a utilizar exige o conhecimento das lições anteriores. As mais robustas estratégias estatísticas exigem que as variáveis apresentem propriedades intervalares para que sejam obtidos resultados fidedignos. Contudo na investigação com seres humanos nem sempre é possível termos variáveis quantitativas, por isso para cada teste estatístico paramétrico existe um equivalente não paramétrico mas destes últimos existem vários que não tem equivalente paramétrico.

Por exemplo se tanto a nossa variável dependente (VD) quanto a independente (VI) forem nominais e quisermos conhecer a associação entre elas podemos recorrer ao qui-quadrado (x^2) da independência; se ambas forem ordinais podemos recorrer ao rho de spearman mas se forem quantitativas e cumprirem com os restantes pré-requisitos da estatistica paramétrica (simétricas, mesocurticas e distribuição normal) podemos utilizar o teste r de Pearson.

Se em vez de querermos ver umas associação ou correlação pretendermos verificar se existem diferenças na distribuição de uma variável (VD) em função de outra com nivel de mensuração nominal e dicotómica (VI) então podemos utilizar o teste t de Student para amostras independentes (caso estejam cumpridos os prerequisitos impostos à VD ié, quantitativa, simétrica e apresente distribuição aproximadamente normal) ou o seu equivalente não paramétrico u de Mann-Whitney (caso não estejam cumpridos os pré-requisitos da estatistica paramétrica mas a VD tenha um nivel de mensuração no minimo ordinal).

Se a figura anterior não o deixou muito esclarecido experimente consultar o quadro que se segue. Otestes estatísticos paramétricos estão assinados com um asterisco (*)

Tabela 7: Grelha de decisão dos testes

	NIVEIS DE MENSURAÇÃO		
	Nominal	**Ordinal**	**Quantitativa**
Testes para uma amostra	**TESTE DE QUI-QUADRADO DA ADERÊNCIA**	**TESTE DE KOLMOROGOV-SMIRNOV**	**-TESTE DE KOLMOROGOV-SMIRNOV** **-TESTE T PARA UMA AMOSTRA ***

		Variáveis Independentes		
		Qualitativas		Quantitativa
		Nominal/ dicotomica	**Ordinal/ Grupo**	
Variáveis Dependentes	**Nominal**	TESTE DE QUI-QUADRADO DA INDEPENDENCIA **KAPPA DE COHEN** **MACNEMAR** **Q DE COCHRAN**	TESTE DE QUI-QUADRADO DA INDEPENDENCIA	
	Ordinal	TESTE DE QUI-QUADRADO DA INDEPENDENCIA *TESTE DE U DE MANN-WHITNEY* *TESTE DE H DE KRUSKAL-WALLIS*	RHO DE SPEARMAN *W DE WILCOXON* **KAPPA DE COHEN** **MACNEMAR** TESTE DE QUI-QUADRADO DA INDEPENDENCIA	RHO DE SPEARMAN
	Quantitativa	*TESTE T DE STUDENT PARA DADOS INDEPENDENTES* * TESTE DE U DE MANN-WHITNEY *TESTE ANOVA DE UM CRITÉRIO E RESPECTIVO POST-HOC* * TESTE DE H DE KRUSKAL-WALLIS e RESPECTIVO POST-HOC (**Nemenyi**)	RHO DE SPEARMAN	*TESTE T DE STUDENT PARA N EMPARELHADOS* * *W DE WILCOXON* R DE PEARSON * RHO DE SPEARMAN *TESTE ANOVA PARA MEDIDAS REPETIDAS* * *TESTE FRIEDMAN*

Supondo que suas variáveis dependentes tivessem uma distribuição normal ou que sua amostra fosse suficientemente grande, deve verificar todas as possibilidades de análise: univariada , bivariada, múltipla e multivariada, se for o caso. A análise univariada é quando a variável é analisada *per se,* análise bivariada quando uma variável dependente é relacionada com uma única variável independente, análise múltipla quando se analisa uma variável dependente em função de várias variáveis independentes, e análise multivariada, quando se analisa várias variáveis dependentes contínuas em função de variáveis independentes categóricas ou quando se analisa a estrutura das variáveis, visando a redução do número de variáveis.

O quadro anterior não esgota as analises estatísticas, aliás existem outras tantas quantas as que apresentamos aqui, contudo mostra as mais utilizadas nas análises univariadas e bivariadas.

6.1. TESTES PARAMÉTRICOS PASSO-A-PASSO

6.1.1 TESTE T DE STUDENT (NÃO RELACIONADO)

CARACTERÍSTICAS E REQUISITOS DO TESTE T NÃO RELACIONADO OU INDEPENDENTE

1. Teste para a comparação de médias;
2. Distribuição com forma leptocúrtica, isto é, as caudas da distribuição são mais grossas do que na distribuição normal;
3. Escala de medida intervalar e Contínua;
4. Simétrica;
5. De forma campanular;
6. Varia de mais infinito a menos infinito;
7. desvio padrão da variável de acordo com n.
8. distribuição normal;
9. n >=30).

Utiliza-se para designs experimentais com duas situações testando uma variável independente, quando nessas situações se encontram sujeitos diferentes - designs não relacionados. O teste t não relacionado é o equivalente paramétrico do teste não paramétrico U de Mann-Whitney; ambos comparam diferenças entre dois grupos.

O objectivo deste teste é comparar a quantidade da variabilidade devida às diferenças previstas nos resultados entre dois grupos com a variabilidade total nos resultados dos sujeitos. As diferenças previstas são calculadas como uma diferença entre os resultados médios entre os dois grupos.

A estatística t representa o tamanho da diferença entre as médias para os dois grupos, tomando em consideração a variância total.

Para que o valor observado de t seja significativo terá de ser igual ou superior aos valores críticos de t apresentados na tabela.

Instruções passo-a-passo

1. Elevar ao quadrado cada resultado individual para ambos os grupos em separado
2. Adicionar os totais dos resultados ao quadrado para cada grupo
3. Elevar ao quadrado todos os resultados individuais para cada grupo
4. Calcular a média para cada grupo
5. Calcular t:

$$t = \frac{\bar{\chi}_1 - \bar{\chi}_2}{\sqrt{\frac{\sum x_1^2 - (\sum x_1)^2}{n_1} + \frac{\sum x_2^2 - (\sum x_2)^2}{n_2}}{(n_1 - 1) + (n_2 - 1)} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

em que

$\bar{\chi}_1$ = média do grupo 1

$\bar{\chi}_2$ = média do grupo 2

Σx_1^2 = soma dos quadrados para o grupo 1

Σx_2^2 = soma dos quadrados para o grupo 2

$(\Sigma x_1)^2$ = resultados totais do grupo 1 ao quadrado

$(\Sigma x_2)^2$ = resultados totais do grupo 2 ao quadrado

n_1 = número de sujeitos do grupo 1

n_2 = número de sujeitos do grupo 2

(n1 - 1) + (n2 - 1) = graus de liberdade (gl)

Se t observado ≥ t crítico rejeita-se H_0 Se t observado < t crítico aceita-se H_0

Exemplo: para verificar se duas dietes para emagrecer são igualmente eficazes, um médico separou ao acaso um conjunto de pacientes em dois grupos. Cada paciente seguiu a dieta designada para o seu grupo durante 4 meses. O médico registou a perda de peso em kg de cada paciente por grupo. Os dados estão apresentados no quadro que se segue:

Tabela 8: Cálculo do valor t

Grupo 1 (dieta 1)		Grupo 2 (dieta 2)	
Resultados	Resultados ao quadrado	Resultados	Resultados ao quadrado
10	100	2	4
5	25	1	1
6	36	7	49
3	9	4	16
9	81	4	16
8	64	5	25
7	49	2	4
5	25	5	25
6	36	3	9
5	25	4	16
$\Sigma x_1 = 64$	$\Sigma x1^2 = 450$	$\Sigma x2 = 37$	$\Sigma x2^2 = 165$

Calcule o valor de t observado[2] e verifique se é igual, superior ou inferior ao valor crítico e intrepete o resultado.

[2] Solução: $t_{Obs}=3,1$ $t_{crit}(18)=2,9$

6.1.2 TESTE T DE STUDENT (RELACIONADO)

CARACTERÍSTICAS E REQUISITOS DO TESTE T RELACIONADO OU EMPARELHADO

Utiliza-se para designs experimentais com duas situações testando uma variável independente, quando os mesmos sujeitos (ou emparelhados) se encontram em ambas as situações - design relacionado. O teste t relacionado é equivalente ao teste não paramétrico de Wilcoxon.

O objectivo é comparar as diferenças entre as duas situações experimentais com a variabilidade total nos resultados. Quando os mesmos sujeitos são usados em ambas as situações podem comparar-se pares de resultados obtidos por cada indivíduo quando sujeito a ambas as situações.

A estatística t apresenta o tamanho das diferenças entre os resultados dos sujeitos para as duas situações. Para que seja significativo o valor de t terá de ser igual ou superior aos valores críticos da tabela

Instruções passo-a-passo

1. Calcular as diferenças entre os resultados dos sujeitos subtraindo os resultados da situação B para a situação A
2. Elevar essas diferenças ao quadrado
3. Calcular o somatório das diferenças obtidas ($\sum d$)
4. Calcular o somatório do quadrado das diferenças ($\sum d^2$)
5. Elevar ao quadrado as diferenças totais $(\sum d)^2$
6. Calcular t:

$$t = \frac{\Sigma d}{\sqrt{\frac{N \Sigma d^2 - (\Sigma d)^2}{N - 1}}}$$

em que

Σ d = soma das diferenças dos resultados A e B

Σd^2 = soma dos quadrados das diferenças

$(\Sigma d)^2$ = soma das diferenças elevadas ao quadrado

N = número de sujeitos

N – 1= gl

Por fim consulta-se a tabela dos valores críticos e,

Se t observado ≥ t crítico rejeita-se H0 Se t observado < t crítico aceita-se H0

Tabela 9: Valores críticos t de student

grau de liberdade	p				
	0,10	**0,05**	**0,025**	**0,010**	**0,005**
1	3,078	6,314	12,706	31,821	63,656
2	1,886	2,920	4,303	6,965	9,925
3	1,638	2,353	3,182	4,541	5,841
4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,355
9	1,383	1,833	2,262	2,821	3,250
10	1,372	1,812	2,228	2,764	3,169
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977

Exemplo: Para verificar se a eficácia de uma dieta era influenciada pelo ministrar de um fármaco, um médico decidiu administrar, a um grupo de individuos que o tinham procurado para perder peso, um placebo em conjunto com uma dieta que já havia administrado um mês antes aos mesmos sujeitos. Referiu aos seus casos que aquele medicamento servia para perder apetite e ajudava a queimar gorduras.

Registou a perda de peso que tinha ocorrido nos 30 dias antecedentes à tomada de placebo e trinta dias após o placebo. Os resultados estão no quadro que se segue

Tabela 10: Cálculo do teste t emparelhado

Sujeito	**Situação A (com placebo)**	**Situação B (só com dieta)**	**d (A-B)**	**d^2**
1	10	2	8	64
2	5	1	4	16
3	6	7	-1	1
4	3	4	-1	1
5	9	4	5	25
6	8	5	3	9
7	7	2	5	25
8	5	5	0	0
9	6	3	3	9
10	5	4	1	1
Total	64	37	Σ d = 27	Σ d^2 = 151

Resolução 50:

Instruções Passo-a-Passo:

1. construir tabela
2. calcular as médias
3. Σ d = 27
4. Σ d^2 = 151
5. $(\Sigma d)^2 = 27 \times 27 = 729$
6. proceder aos calculos
7. g.l. = N - 1 = 10 - 1 = 9

calcule o valor observado[3] de t e verifique se é superior, igual ou inferior ao valor crítico de e interprete os resultados.

6.1.3 CORRELAÇÃO MOMENTO-PRODUTO DE BRAWAIS-PEARSON

Quando estudamos um grupo relativamente a dois caracteres vemos, como já dissemos, que pode existir uma relação entre eles.

Se medirmos os raios de várias circunferências e também os seus perímetros verificamos que existe uma relação entre eles que é constante; neste caso temos "dependência funcional". Isto quer dizer que existe uma fórmula exprimindo a medida do segundo em função da do primeiro: P=2Πr.

Suponhamos agora que registamos, durante todos os dias de um certo período de tempo, o numero de alunos que frequentam a biblioteca do Instituto Superior Miguel Torga e o número de passageiros dos SMTUC da linha 6 (CHC-HUC). Vê-se bem que entre as duas estatísticas assim obtidas não é esperada nenhuma relação. Diremos que os dois caracteres são "independentes". Mas espera-se que exista uma dependência estatística entre as pessoas que tentam o suicídio e a depressão. Diremos que estes caracteres estão correlacionados.

Desde que os dois caracteres sejam tais que as suas variações sejam sempre no mesmo sentido, ou em sentidos contrários, pressentimos que os caracteres estejam ligados entre si: dizemos, então, que existe uma correlação entre eles.

Estes métodos de correlação foram criados por Sir Francis Galton, que trabalhou juntamente com Pearson, nos fins do século XIX. A correlação e a regressão são dois aspectos que andam sempre muito ligados, pertencendo à Estatística correlacional. Assim, importa fazermos a distinção entre eles:

A correlação pode ser definida como o grau de semelhança no sentido das variações entre os valores correspondentes dos dois caracteres, isto é, a correlação preocupa-se quer com a descrição da relação entre variáveis quer com a sua direcção (directa ou inversamente proporcional, positiva ou negativa).

Já a regressão é usada quando queremos conhecer as variáveis preditoras de uma outra conhecida.

[3] $t_{obs}=2,90$; $t_{crit(0,010)}=2,821$

TIPOS DE COEFICIENTE DE CORRELAÇÃO

Basicamente, podemos considerar dois tipos de coeficientes de correlação:

- Coeficiente de correlação momento-produto de Brawais-Pearson, cujo símbolo é "r", e que é uma técnica de estatística paramétrica;

- Coeficiente de correlação Rho de Spearman-Rank, cujo símbolo é "ρ", e que é uma técnica de estatística não paramétrica.

Devemos salientar que, para o cálculo das correlações, é necessário termos sempre duas medidas para cada sujeito.

REPRESENTAÇÃO GRÁFICA

À representação gráfica da correlação chamamos diagrama de dispersão de pontos ou scatterplot ou scattergram e, genericamente, toma a seguinte forma:

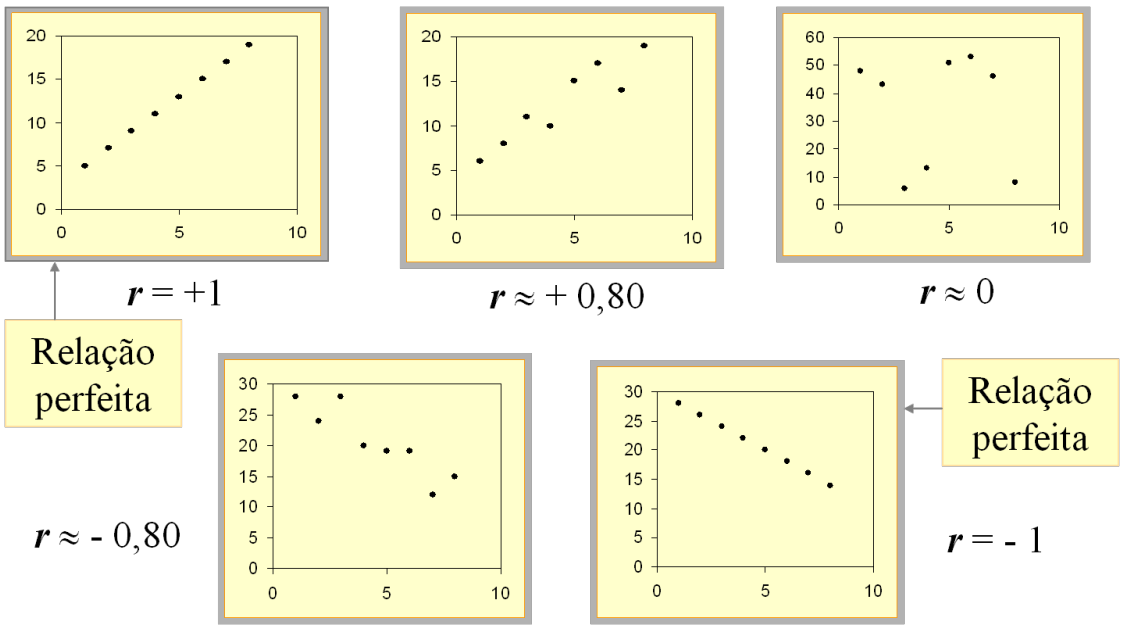


Figura 10: Diagramas de dispersão de pontos, scatterplot ou scattergram

- A análise de *r* deve vir acompanhada do diagrama de dispersão, pois a associação pode não ser linear.

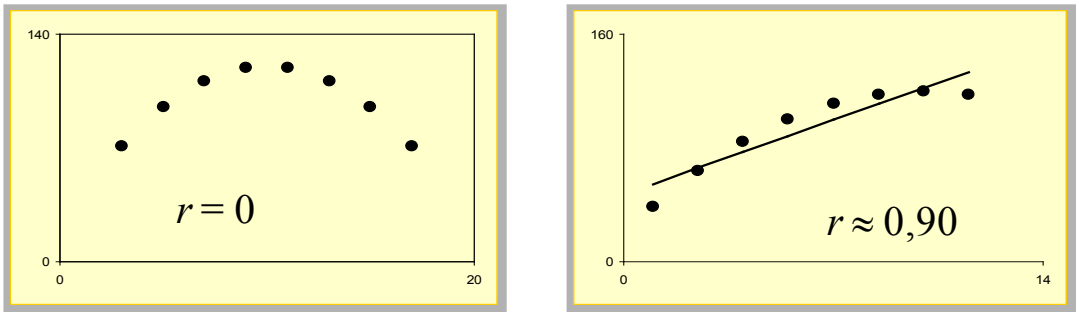


Figura 11: Diagramas de dispersão : causa & efeito

Suponhamos que temos duas séries estatísticas formadas pelos valores xi e yi de dois caracteres. Podemos fazer uma representação gráfica dos dados representando todos os pontos (xi, yi) e obtendo a nuvem de pontos.

Segundo os dados, a nuvem de pontos pode apresentar diversos aspectos.

Por exemplo os pontos podem distribuir-se na semelhança de uma linha recta ou de uma curva: isto sugere a existência de uma relação funcional entre X e Y.

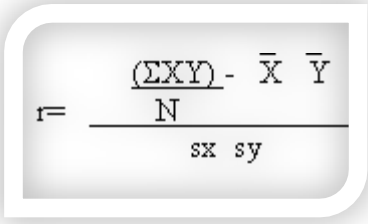
Os pontos podem ser dispersos e colocados ao acaso no plano; pode acontecer que os pontos cubram uma porção do plano da qual se pode definir o contorno; esta forma sugere que as duas variáveis estão ligadas. Limitemo-nos ao caso mais simples em que a nuvem tem uma forma alongada lembrando uma elipse e suponhamos que a sua orientação é tal que desde que X cresça, a variável Y também cresce. A forma desta nuvem sugere a possibilidade da existência de uma recta tal que os valores estimados por esta recta, a partir dos valores de xi, sejam boas aproximações dos valores de yi. Nós podemos determinar pelo método dos mínimos quadrados uma recta tal que a soma dos quadrados dos desvios seja mínima. Esta recta é chamada recta de regressão de Y em X ou recta de estimação de Y em X.

Mas, poderíamos, de um modo semelhante, procurar uma recta tal que os valores de x estimados ao longo desta recta, a partir de yi, constituam igualmente boas aproximações de xi. Esta recta é chamada recta de regressão de X em Y ou recta de estimação de X em Y.

Normalmente, estas rectas são distintas uma da outra. Elas serão confundidas quando existe ligação funcional linear e são perpendiculares ao eixos quando há independência. Compreendemos, assim, que a correlação entre os caracteres é tanto maior quanto maior as rectas de regressão estejam mais próximas uma da outra.

1. Este tipo de coeficiente de correlação utiliza-se quando:
2. As duas variáveis são contínuas;
3. A distribuição se aproxima da distribuição normal;
4. É preferível para distribuições unimodais;
5. Escala intervalar de medida.

Fórmula



Então $-1 \leq r \leq 1$

Interpretação:

O coeficiente de correlação obtido pode se interpretado com base em:

Para Cardoso:

$r \leq 0,2$	Correlação muito baixa (valores desprezíveis)
$0,2 < r \leq 0,5$	Correlação baixa
$0,5 < r \leq 0,7$	Valores significativos
$0,7 < r \leq 0,9$	Alta correlação
$0,9 < r \leq 1$	Muito alta correlação

Para Borg:

$0,20 < r \leq 0,35$	Ligeira relação entre as variáveis, embora já possam ser estatisticamente significativas
$0,35 < r \leq 0,65$	Correlação estatisticamente significativa para além do nível de 1%

$0,65 < r \leq 0,85$ Correlações que tornam possíveis predições do grupo de que são dignas

$r > 0,85$ Íntima relação entre as variáveis correlacionadas

Para Byrman e Cramer,

se Eta, r, Rho, phi:

$\leq 0,2$ Correlação muito fraca e sem significância

$0,2 < r \leq 0,39$ Correlação fraca

$0,4 < r \leq 0,69$ Correlação moderada

$0,7 < r \leq 0,89$ Correlação forte

$0,9 < r \leq 1$ Correlação muito elevada

Coeficiente de correlação dá-nos:

A direcção que é indicada pelo sinal + ou -

A intensidade ou força que é dada pelo valor que varia entre -1 e 1. Se a correlação for zero não existe correlação entre as variáveis (exemplo: cor dos olhos e inteligência).

Exemplo: Considere as classificações (numa escala de 0 a 100) obtidas por 10 alunos nas disciplinas Estatistica I (STAT I), Estatistica II (STAT II),Portugues (PORT) e Franças (FRA):

Tabela 11: Cálculo do r de Pearson

Estudante	STAT I (X)	STAT II (Y)	PORT (Z)	FRA (W)	XY	XZ	XW
1	75	75	45	45	5625		
2	70	70	50	50	4900		
3	70	70	50	50	4900		
4	65	65	55	55	4225		
5	60	60	60	60	3600		
6	60	60	60	60	3600		
7	55	55	65	65	3025		
8	50	50	70	70	2500		
9	50	50	70	70	2500		
10	45	45	75	75	2025		
Σ	600	600	600	600	36900		

1. Com base dos dados que se seguem calcule o coeficiente de correlação[4] entre X e Y

   Sabe-se que:

   $\bar{X}=60$ $\bar{Y}=60$ $\Sigma XY=36900$

   $s^2x=90$ $s^2y=90$

qual o valor de r? _____

que conclusão retira dos resultados?

2. recorrendo ao valores da tabela precedente calcule os valores necessários à obtenção do coeficiente de correlação entre X e Z.

qual a média das variáveis?

qual o valor de r? _____

que conclusão retira dos resultados?

[4] a)=1 b)=-1

6.1.4 ANÁLISE DA VARIANCIA DE UM CRITÉRIO (ANOVA)

CARACTERÍSTICAS E REQUISITOS DA ANOVA

O ponto 6.1.1 explica como comparar médias de duas populações, com base em amostras dessas populações. Mas às vezes é preciso comparar médias de mais de duas populações. Por exemplo, para verificar se pessoas com diferentes níveis socioeconómicos, isto é, alto, médio e baixo têm, em média, o mesmo peso corporal, é preciso comparar médias de três populações.

Para comparar médias de mais de duas populações aplica-se a ANOVA (o teste F), na forma que a seguir se descreve, desde que a variável em estudo tenha distribuição normal ou aproximadamente normal. Mas antes de mostrar como se faz esse teste, convém apresentar um exemplo.

6.1.4.1 ANÁLISE DA VARIÂNCIA COM IGUAL TAMANHO

Se a variável em estudo tem distribuição normal ou aproximadamente normal, para comparar mais de duas médias aplica-se o teste *F*.

Primeiro, é preciso estudar as *causas de variação*. Por que é os dados variam? Uma explicação é o facto de as amostras provirem de populações diferentes. Outra explicação é o acaso, porque até mesmo os dados provenientes de uma mesma população variam.

O teste *F* é feito através de uma *análise de variância,* que separa a variabilidade devido aos "tratamentos" (no exemplo, devido às amostras terem provindo de populações diferentes) da variabilidade residual, isto é, devido ao acaso. Para aplicar o teste *F* é preciso fazer uma série de cálculos, que exigem conhecimento da notação.

Para fazer a análise de variância é preciso proceder aos seguintes calculos:

1. Graus de liberdade

gl dos grupos: k – 1

gl do total: n-1

gl dos residuos: (n-1)-(k-1) = n-k

2. calcular o valor de Correcção (C) que é dado pelo total geral ao quadrado e dividido pelo número de dados.

$$C = \frac{(\Sigma x)^2}{n}$$

3. calcular a Soma dos Quadrados Total (SQT)

$$SQT = \Sigma x^2 - C$$

4. calcular a Soma do Quadrado do Total de cada repetição (SQTr)

$$SQTr = \frac{\Sigma T^2}{r} - C$$

5. calcular a Soma dos Quadrados dos Resíduos (SQR)

$$SQR = SQT - SQTr$$

6. calcular o Quadrado médio do Total de cada repetição (QMTr)

$$QMTr = \frac{SQTr}{k-1}$$

7. calcular o Quadrado médio do Total do Residuo (QMR)

$$QMR = \frac{SQR}{n-k}$$

8. finalmente calcular o valor de F

$$F = \frac{QMTr}{QMR}$$

SeF observado ≥ F crítico rejeita-se H0 Se F observado < F crítico aceita-se H0

para interpretar os resultados necessitamos de comparar o F calculado com o valor dado na tabela de F, ao nivel de significância estabelecido, observando os k-1 graus de liberdade no numerador e os n-k graus de liberdade no denominador (coluna da esquerda).

Exemplo: Um profissional de saúde recém contratado para acompanhar um conjunto de atletas de alta competição, verificou, pelos registos clínicos deixados pelo seu antecessor, que alguns atletas com o mesmo tipo de lesão (em grau e extensão) tinham mais recidivas que outros, apesar das condições de treino e o tempo de recuperação ser o mesmo. Colocou a hipótese de que tal acontecimento se podia dever às diferentes terapêuticas que eram utilizadas para tratar as mesmas lesões. Os resultados podem ser observados no quadro que se segue:

Tabela 12: Cálculo da ANOVA para tamanhos iguais

	Tratamento A	Tratamento B	Tratamento C	Tratamento D
	11	8	5	4
	8	5	7	4
	5	2	3	2
	8	5	3	0
	8	5	7	0
Σ	40	25	25	10
$\bar{\chi}$	8	5	5	2

1.º passo:

os graus de liberdade (gl)dos grupos: k – 1 = 4-1=3

gl do total: n-1 = 20-1=19

gl dos residuos: n-k = 20-4=16

calcular o valor de Correcção (C) que é dado pelo total geral ao quadrado e dividido pelo número de dados.

$$C = \frac{(\Sigma x)^2}{N} = \frac{(11+8+5+4+8+5+7+4+5+2+3+2+8+5+3+8+5+7)^2}{20} = \frac{100^2}{20} = 500$$

calcular a Soma dos Quadrados Total (SQT)

$$SQT = \Sigma x^2 - C = 11^2+8^2+5^2+4^2+8^2+5^2+7^2+4^2+5^2+2^2+3^2+2^2+8^2+5^2+3^2+8^2+5^2+7^2 - 500 = 658-500 = 158$$

calcular a Soma do Quadrado do Total de cada repetição (SQTr)

$$SQTr = \frac{\Sigma T^2}{R} - C = \frac{40^2+25^2+25^2+10^2}{5} - 500 = 590-500 = 90$$

calcular a Soma dos Quadrados dos Resíduos (SQR)

$$SQR = SQT - SQTr = 158-90 = 68$$

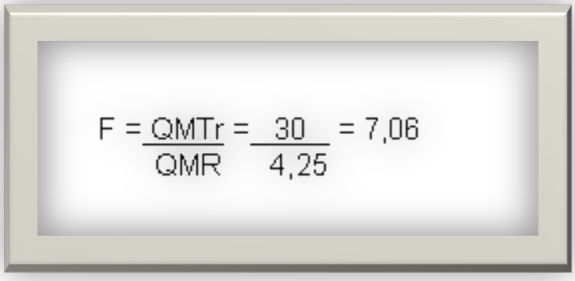
calcular o Quadrado médio do Total de cada repetição (QMTr)

$$QMTr = \frac{SQTr}{k-1} = \frac{90}{3} = 30$$

calcular o Quadrado médio do Resíduo (QMR)

$$QMR = \frac{SQR}{n-k} = \frac{68}{16} = 4,25$$

calcular o valor de F



Finalmente ir à tabela F para um nível de significância (p) de 5% (0,05) e observar qual o F teórico para 3 e 16 graus de liberdade.

Como o valor calculado (7,06) é maior que o da tabela (3,24), concluímos que as médias das recidivas diferem em função do tratamento, para um nível de significância de 0,05.

A acompanhar este comentário, os valores calculados devem ser apresentados num quadro, da seguinte forma:

Tabela 13: Apresentação da ANOVA

Causas de variação	gl	SQ	QM	F	p
Tratamentos	3	90	30	7,06	<0,05
Resíduo	16	68	4,25		
Total	19	158			

Mas, como se pode observar, apesar da tabela mostrar que existem diferenças significativas, não nos informa, que tratamentos é que produzem diferenças e quais são semelhantes. Sempre que as diferenças são significativas, e só nesse caso, temos que proceder às comparações à posteriori (Post-Hoc). Podemo-nos socorrer de diversos testes (LSD; Bonferroni; Sidak; Scheffe; SNK; Tukey; etc.), a grande diferença entre eles reside no tipo de distribuição em que assentam e no tipo de ajustamento).

Apresentaremos de seguida apenas o teste de Tukey, por ser dos mais utilizados e o mais simples de calcular, quando recorremos ao cálculo manual.

6.1.4.1.1 TESTE DE TUKEY PARA COMPARAÇÃO ENTRE AS MÉDIAS

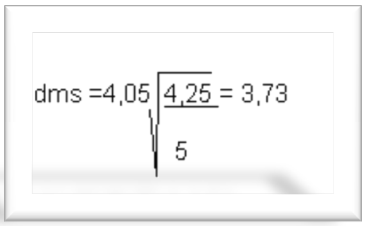
O teste Tukey permite estabelecer a diferença mínima significante, ou seja, a menor diferença entre as médias que deve ser tomada como significativa em determinado nível de significância. Essa diferença (dms) é dada por:

Onde q é um valor dado em tabela

QMR é o quadrado médio do residuo da ANOVA

r é o número de repetições

assim, se consultarmos a tabela verificamos que o q para comparar quatro tratamentos com 16 gl no residuo é de 4,05. como QMR=4,25 e r=5, temos:



De acordo com o teste de Tukey, duas médias são estatisticamente diferentes sempre que o valor absoluto da diferença entre elas for igual ou superior ao valor da dms.

Passemos então à observação dos valores:

Tabela 14: Cálculo da diferença mínima significativa - Tukey

Pares de médias	Valor absoluto da diferença	dms	p
A-B	(8-5) 3	3,73	ns
A-C	(8-5) 3		ns
A-D	(8-2) 6		<0,05
B-C	(5-5) 0		ns
B-D	(5-2) 3		ns
C-D	(5-2) 3		ns

É fácil de observar que só existem diferenças entre a média dos tratamentos A e a média dos tratamentos D, em que o tratamento D é aquele com que se obtém, significativamente, menos recidivas

6.1.4.2 ANÁLISE DE VARIÂNCIA COM DIFERENTES TAMANHOS

O pesquisador, nem sempre tem amostras do mesmo tamanho, mesmo assim é possivel conduzir uma análise da variância (ANOVA). Aliás todos os cálculos, com excepção SQTr, são feitos da mesma forma em ambas as situações.

Assim em vez de fazer a soma dos quadrados pela fórmula

$$SQTr = \frac{\Sigma T^2}{r} - C$$

Utiliza:

$$SQTr = \frac{T_1^2}{r_1} + \frac{T_2^2}{r_2} + \ldots + \frac{T_k^2}{r_k} - C$$

Para se certificar de que entendeu faça o seguinte exercicio:

Tabela 15: Cálculo da ANOVA para tamanhos iguais

	Tratamento A	Tratamento B	Tratamento C
	15	23	19
	10	16	15
	13	19	21
	18	18	14
	15		16
	13		
$\bar{\chi}$	84	76	85

O resultasdo do valor de F term de lhe dar 3,96. Confira e interprete

Não se esqueça que as diferenças foram significativas por isso tem de proceder às comparações à posteriori (Post-Hoc) e também aqui a fórmula mudou, por isso vamos ver como se calcula o teste de Tukey quando temos tamanhos diderentes:

6.1.4.2.1 TESTE DE TUKEY PARA COMPARAÇÃO ENTRE AS MÉDIAS

O teste Tukey para amostras com tamanhos diferentes é dada pela seguinte fórmula:

$$dms = q\sqrt{\left(\frac{1}{r_i} + \frac{1}{r_j}\right)\frac{QMR}{2}}$$

No caso do exemplo, para comparar a média de A com a média de B tem-se:

$$dms\ (A;B) = 3{,}77\sqrt{\left(\frac{1}{6} + \frac{1}{4}\right)\frac{8}{2}}.$$

dms (A;B) =4,87

De forma análoga faça os calculos para comparar a média de A com a média de C

e de B com a média de C.

Qual a sua conclusão?

De acordo com o teste de Tukey, duas médias são estatisticamente diferentes sempre que o valor absoluto da diferença entre elas for igual ou superior ao valor da dms.

Passemos então à solução dos exercício proposto e observação dos valores:

Tabela 16: Teste post-hoc -Tukey

Pares de médias	Valor absoluto da diferença	dms	p
A-B	|14-19| = 5	4,87	<0,05
A-C	|14-17| = 3	4,57	ns
B-C	|19-17| = 2	5,06	ns

Conclui-se que em média A é significativamente diferente de B, ao nível de significância de 0,05.