CAPÍTULO 1 INTRODUÇÃO

- Objectivos do capítulo
- Definir dado
- Diferenciar dado qualitativo de quantitativo
- Definir População e Amostra
- Diferenciar amostra de população
- Definir Variável
- Diferenciar variável qualitativa de quantitativa
- Definir variável independente e dependente
- Diferenciar variável independente de variável dependente

1. INTRODUÇÃO

A Estatística, é uma ciência que trata fundamentalmente da manipulação de dados, por se considerar dado como a unidade mínima para a obtenção de informações. Em Estatística, esses dados são manipulados através de métodos, dos quais se destacam os "Métodos Estatísticos", que nos permitem responder a questões como as seguintes:
1- O que é que os Cidadãos pensam da postura camarária de Maputo sobre os resíduos?
2- Os rendimentos das famílias nos últimos anos tem sofrido realmente algum aumento?
3- Qual será o comportamento da despesa pública nos próximos dez anos, tendo em conta os dados dos últimos 20 anos?
4- O que nos diz o inquérito sobre a tendência do voto em relação aos candidatos para as presidenciais?
Estas são algumas das várias questões que a Estatística tem como "obrigação" responder para satisfação dos concidadãos ou redefinição de políticas.

2. NOÇÃO DE ESTATÍSTICA

A noção "Estatística" é original da palavra "Estado", já que foi sempre o papel fundamental de qualquer governo colectar, organizar, resumir, analisar e interpretar um conjunto de dados da população, para deles tirar conclusões. São casos de Natalidade, Mortalidade, Taxas, Percentagens, Índices e muitas outras espécies de informações e actividades.

A medição (com ou sem maior aproximação da grandeza) e a contagem dessas quantidades ou qualidades, gera um conjunto de dados numéricos que são úteis para o desenvolvimento de muitos tipos de funções administrativas, pedagógicas e/ou governamentais, para o fortalecimento e formulação de políticas.

Estes dados recolhidos são sempre susceptíveis de algum tratamento para facilitar a emissão de pareceres.

Exemplo:
Suponhamos que Moçambique, entre os anos de 1990 e 1999, tenha exportado as seguintes quantidades de Mariscos para a União Europeia:

Tabela 1.1 – Quantidades de mariscos exportadas entre 1990 e 1999

Ano	Quantidades (em toneladas)
1990	3000
1991	2700
1992	2850
1993	1000
1994	1350
1995	1410
1996	1850
1997	1450
1998	2100
1999	1380
Total	19090

Como se pode observar, após o ano de 1990, as exportações desses produtos foram baixando duma forma irregular, tendo atingido o pior mínimo em 1993.

Quanto a estes números, que poderiam resultar de dados concretos do Ministério das Pescas, qualquer Analista poderia questionar:
- Será que em 1993 os Parceiros da União Europeia diminuíram as preferências pelos Mariscos Moçambicanos?
- Teria havido mudança de Políticas Fiscais ao nível da União Europeia?

O leitor ter-se-á se apercebido, que muitas e várias perguntas poderiam ser levantadas em benefício da dúvida.

Foi fácil notar que, com os dados apresentados na tabela 1, muitas perguntas e conclusões poderiam ter sido tiradas pelos observadores independentes do processo.

Os dados numéricos apresentados são a unidade mínima da Estatística, sendo assim considerados como matéria prima que precisa de ser transformada usando "Métodos Estatísticos" para posterior análise.

A Estatística é um conjunto de todos os passos dum projecto de experiências (resultantes de medição ou observação), organização dos dados, interpretação e conclusões daí resultantes.

Actualmente a Estatística é muito mais usada em fenómenos imprevisíveis (aleatórios), como o de comportamento da despesa pública, níveis de aprovações da 7ª a 8ª classes em todo o território nacional, etc. Face a estas situações, o governo pode pensar em

conter a despesa pública, captar mais impostos e investir na criação de mais escolas, etc.

Em suma, a Estatística pode ser apresentada estrategicamente em três áreas: (a) Estatística Descritiva, (b) Estatística Inferencial e (c) Teoria de Decisões. Esta apresentação é estratégica porque em muitos escritos a Estatística é subdividida em duas grandes áreas: (a) A Descritiva ou Dedutiva e (b) A Inferencial ou Indutiva (onde se incorpora a teoria de decisões).

Ao analisarmos um conjunto de dados, devemos determinar em primeiro lugar se se trata de uma amostra ou de uma população completa. Essa determinação afectará não somente os métodos utilizados, mas também as conclusões a que chegamos. Utilizamos métodos de estatística descritiva para resumir ou descrever as características importantes de um conjunto conhecido de dados populacionais, e recorremos a métodos de estatística inferencial quando utilizamos dados amostrais para fazer inferências (ou generalizações) sobre uma população.

3. CONCEITOS BÁSICOS

3.1 Dado e Informação

Dado

Definição 1: É a unidade mínima da informação

Definição 2: O que é recolhido e preparado para produzir algum resultado

Exemplo: Notas dos estudantes duma turma

Informação

Definição: É o resultado do processamento dos dados

Exemplo: Percentagens dos admitidos, excluídos e dispensados da turma

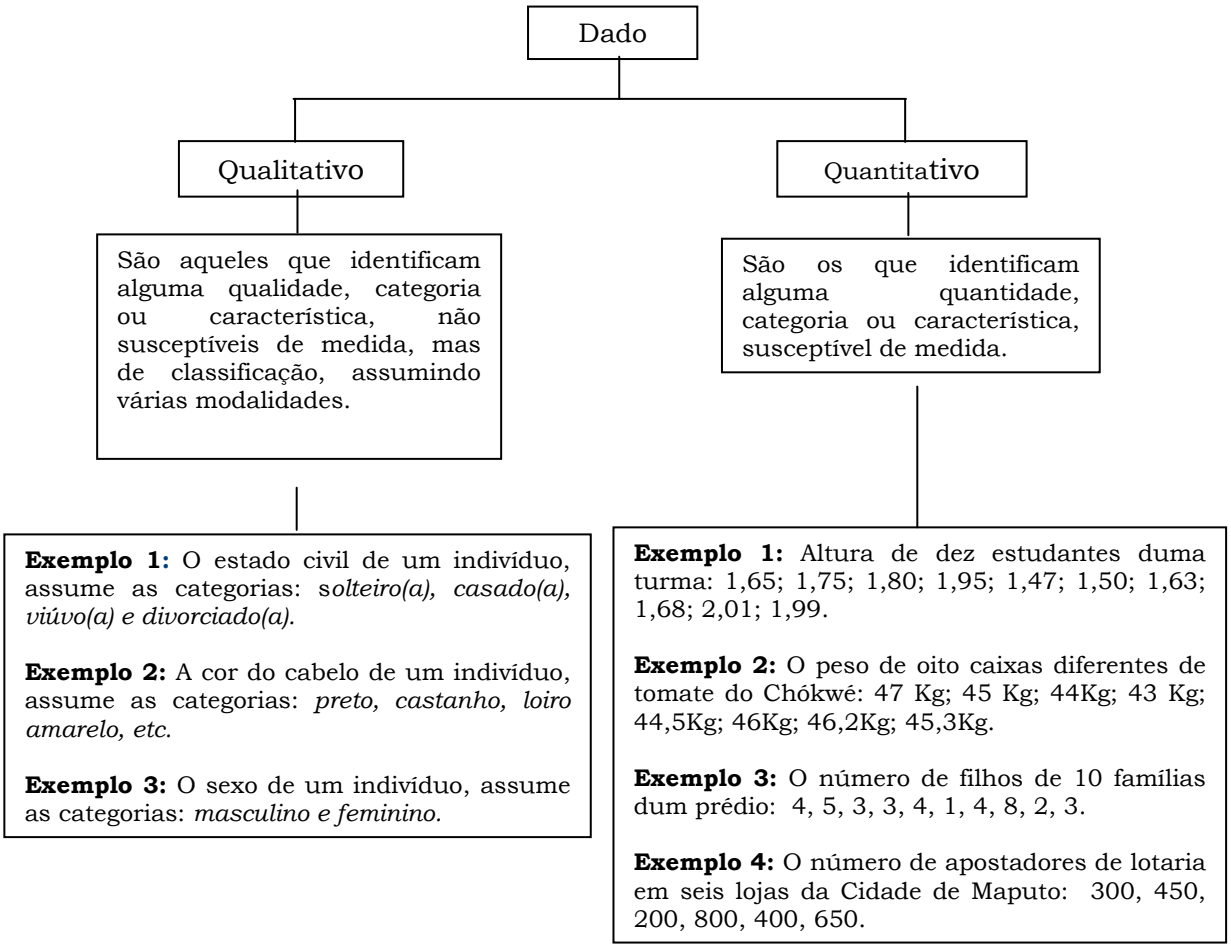


Fig 1.1

3.2 Populaçao e Amostra

População

Uma noção fundamental em Estatística é a de conjunto ou agregado, conceito para o qual se usam, indiferentemente, os termos População ou Universo. Pode-se considerar como População a colecção de unidades individuais, que podem ser pessoas ou resultados experimentais, com uma ou mais características comuns, que se pretendem estudar.

Também se define como População o conjunto de indivíduos, objectos ou resultados experimentais acerca dos quais se pretende estudar alguma característica em comum.

Pode-se tomar como exemplo, a população constituída pelos alunos da 10ª classe matriculados na Escola Secundária Manyanga em Maputo, da qual podemos estar interessados em estudar as seguintes características populacionais:

- Altura (em cm) dos alunos
- Notas (entre 0 e 20) dos alunos

Depois de medir a altura de cada aluno, obteríamos um conjunto de dados com o seguinte aspecto:

Tabela 1.2 – Altura e Notas de estudantes

Nome	Altura	Nota
Alberto	145	10
Alexandre	161	15
Adelaide	158	13
Gabriel	156	16
Valentim	146	9
Carla	150	11
Júlia	163	10
Manuela	157	11
Mimai	140	18
Júlio	139	13
Sululu	162	8

Nem sempre é possível estudar exaustivamente todos os elementos da população porque:

A população pode ter dimensão infinita; exemplo: População constituída pelas pressões atmosféricas, nos diferentes pontos da cidade de Maputo.

O estudo da população pode ser um processo destrutivo da mesma população; exemplo: População dos fósforos de uma caixa (pretendendo ter a certeza, se o palito acende ou não);

Há rapidez no apuramento dos resultados – o tempo para estudar uma população pode ser bastante elevado; exemplo: Estudo da tendência de voto numa data pré eletoral;

Há problemas de economia – estudar a população pode ser muito despendioso; exemplo: Características que apresenta cada doente doente do Sida;

Precisa-se de maior precisão – nalguns casos é mais difícil estudar uma população por esta não ser precisa;

Há problemas de acesso – pode não existir vias de acesso para chegar à população;

Existe instabilidade política – casos de guerra ou tumultos;

Há problemas demográficos – quando a população está mais geograficamente dispersa; etc.

Definição 2: População é um conjunto de elementos com pelo menos uma característica em comum.

Exemplo 1: Se pretendemos saber o número de eleitores que podem votar nas próximas eleições municipais de Maputo, a População em estudo é "Os eleitores da cidade de Maputo com cartão de eleitor ".

Exemplo 2: Se pretendermos saber o número médio de aprovações em Estatística, dos estudantes nas instituições do ensino superior em Moçambique após a independência, a população em estudo será "Os inscritos a Estatística desde a independência até à data".

Nota: Se toda a população puder ser pesquisada estamos perante um censo.

Exemplo 3:

Num Restaurante dois amigos conversam:
- Mas que rica matapa eu estou a comer!
- Não estarás a tirar conclusões precipitadas? Ainda só comeste um bocadinho!
- A amostra que comí já é suficiente…
- Talvez tenhas razão, já que não é necessário comer matapa de todas folhas da mandioqueira para concluir que ela é boa.

Amostra

Quando não é possível estudar, exaustivamente, todos os elementos da população, estudam-se só alguns elementos, a que damos o nome de Amostra.

Definição: São dados ou observações, recolhidos a partir de um subconjunto da população, que se estuda com o objectivo de tirar conclusões para a população da qual foi recolhida.

Exemplo: Relativamente à população das alturas dos alunos do 1º ano matriculados na Escola Secundária de Jécua, na província de Manica, consideremos a seguinte **amostra**, constituída pelas alturas (em cm) de 20 alunos escolhidos ao acaso:
145, 163, 157, 152, 156, 149, 160, 157, 148, 147, 151, 152, 150, 148, 156, 160, 148, 157, 153, 162.

Quando a amostra não representa correctamente a população diz-se enviesada ou viciada e a sua utilização pode dar origem a interpretações erradas.

3.3 Variável

Definição: Características observáveis em cada elemento pesquisado. Descreve o fenómeno.

Para cada variável, para cada elemento pesquisado, num dado momento, há APENAS UM resultado possível, que pode ser classificado pelo nível de mensuração e pela forma de manipulação. O nível de mensuração é que vai determinar se ela refere a uma quantidade discreta ou contínua ou a uma qualidade nominal ou ordinal.

Pelo nível de mensuração ou natureza as variáveis classificam-se em: Qualitativas e Quantitativas, como ilustrado na figura seguinte.

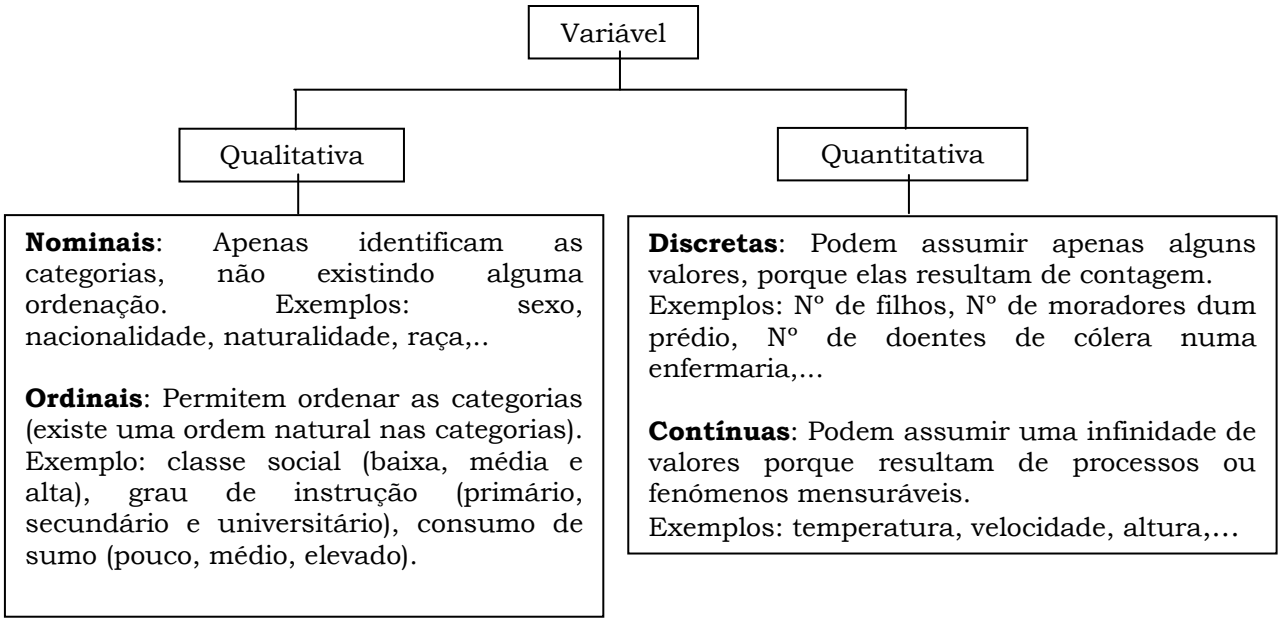


Fig 1.2

Variáveis Quantitativas (intervalares)

Suas realizações são números resultantes de contagem ou mensuração. As variáveis quantitativas subdividem-se em:

Classificação funcional de variáveis

As variáveis quanto à funcionalidade classificam-se em: independentes e dependentes

Variáveis Dependentes

Definição: São aquelas cujos efeitos são esperados de acordo com causas. Elas se situam, habitualmente, no fim do processo causal e são sempre definidas na hipótese ou na questão de pesquisa.

Variáveis Independentes

São aquelas cujos efeitos queremos medir. Podem ser assinaladas as "causas" do fenómeno que se quer estudar. Quando um estudo tem mais de uma hipótese, podem ser definidas diversas variáveis dependentes. Elas podem ser independentes umas das outras ou constituir uma ordem hierárquica, na qual certas variáveis dependentes podem ter um efeito sobre outras variáveis dependentes.

Nota: Atente-se para o facto de que as pesquisas determinam o tipo de variáveis a usar. A pesquisa experimental usa as variáveis independentes e as dependentes. A pesquisa sintética, não necessita de classificação, pois as variáveis se relacionam em rede. As pesquisas de desenvolvimento, não distinguem as variáveis, pois o objectivo é estabelecer e validar uma intervenção ou um instrumento de medida de uma construção.

Observação: A variável constitui o primeiro nível de operacionalização de uma construção teórica e, para cada uma, se deve dar em seguida, uma descrição operacional. Para algumas variáveis a descrição é simples, porém em outros casos, essa definição é mais complexa.

Vejamos um exemplo com resumo mais detalhado:

Tabela 1.3 – Classificação de variáveis

	Idade	Consumo de Bebida Tradicional	Classe Social
Nominal		Sim, Não	
Ordinal	Criança, Jovem, Velho	Pouco, Médio, Muito	Baixa, Média, Alta
Discretas	Número de Anos Completos	Número de Copos de Álcool ingeridos	Número de Salários mínimos completos
Contínuas	Idade em anos, meses, dias, ...	Quantidade de Álcool presente no sangue	Renda Familiar em Meticais

Exercícios resolvidos

1- Com os dados a seguir, que mostram a distribuição salarial de 18 trabalhadores de uma firma,

Tabela 1.4 – Distribuição salarial de 18 trabalhadores duma firma

Nome	Sexo	Anos Serviço	Função	Salário	Nacionalidade
Luís	Masculino	15	Gerente	5,700,000.00Mt	Moçamb.
Mateus	Masculino	16	Escritório	4,200,000.00Mt	Moçamb.
Santos	Masculino	12	Limpeza	2,145,000.00Mt	Moçamb.
Armando	Masculino	8	Escritório	2,190,000.00Mt	Moçamb.
Dinís	Masculino	15	Limpeza	4,500,000.00Mt	Moçamb.
Valentim	Masculino	15	Escritório	3,210,000.00Mt	Moçamb.
Alberto	Masculino	15	Vendas	3,600,000.00Mt	Moçamb.
Adelaide	Feminino	12	Escritório	2,190,000.00Mt	Moçamb.
Sululu	Masculino	15	Tradução	2,790,000.00Mt	Moçamb.
Luisa	Feminino	12	Escritório	2,400,000.00Mt	Moçamb.
Celeste	Feminino	16	Vendas	3,030,000.00Mt	Moçamb.
Júlio	Masculino	8	Tradução	2,835,000.00Mt	Moçamb.
Alex	Masculino	15	Escritório	2,775,000.00Mt	Moçamb.
Gabriel	Masculino	15	Tradução	3,510,000.00Mt	Moçamb.
Helena	Feminino	12	Escritório	2,730,000.00Mt	Moçamb.
Trevour	Masculino	12	Escritório	4,080,000.00Mt	Estrangeira
Kelvin	Masculino	15	Vendas	4,600,000.00Mt	Estrangeira
Xantelo	Feminino	16	Gerente	10,375,000.00Mt	Moçamb.

Identifique as Variáveis.

Resolução: As variáveis são as que fazem parte do cabeçalho *Nome, Sexo, Anos Serviço, Função, Salário e Nacionalidade*. A nacionalidade, nome, sexo e função são variáveis qualitativas. *Anos Serviço* é uma variável quantitativa discreta e *salário*, uma variável quantitativa contínua (porque resulta de todos ponderáveis e imponderáveis).

2) Classifique as variáveis que se seguem em - qualitativas, quantitativas discretas ou quantitativas contínuas.
- Altura de um indivíduo;
- Peso de um indivíduo;
- Idade de um indivíduo;
- Zona de origem;

Resolução:
- Altura de um indivíduo: É quantitativa contínua.
- Peso de um indivíduo: É quantitativa contínua.

- Idade de um indivíduo: É quantitativa discreta.
- Zona de origem : É qualitativa nominal.

Exercícios Propostos

1) Classifique as variáveis que se seguem em - qualitativas, quantitativas discretas ou quantitativas contínuas.
- Cor do cabelo de um idividuo;
- Situação sócio económica de um indivíduo;
- Número de filhos que uma família possui e o respectivo sexo;
- Tipo de cereais que uma família consome e as respectivas quantidades;
- Programa televesívo com maior audiência e o número de vezes que vai ao ar por semana

2) Um dos grandes problemas quando se faz um estudo estatístico, consiste em como seleccionar os casos em estudo. Imagine que se pretende fazer um estudo afim de se verificar se o uso de preservativo reduz os efeitos graves de contrair HIV.
Observe os exemplos que a seguir se apresentam e indique a alternativa que escolheria se tivesse que fazer um estudo sobre a utilização de preservativos.
- Usar dados de todos os vendedores de preservativos ou apenas dos jovens que usam preservativos;
- Usar dados de todos os jovens que usam preservativos ou apenas dos vendedores de preservativos;
- Usar os dados de todos os jovens que usam preservativos da mesma maneira ou categorizar por níveis de contaminação;

- usar os dados dos contaminados agrupando-os por sexo ou analizar todos em conjunto.

3) Num determinado ano lectivo, uma universidade admitiu 3.700 homens dos 8.300 candidatos e 1500 mulheres das 4.300 condidatas. Os dados apresentam-se na tabela abaixo.

Tabela 1.5 Numero de Candidatos e Admitidos por Faculdade

	Homens		Mulheres	
Faculdade	N° de candidatos	N° de admitidos	N° de candidatas	N° de admitidas
Letras	2300	700	3200	900
Ciências	6000	3000	1100	600
Total	8300	3700	4300	1500

a) Qual a diferença entre as taxas de admissão entre os dois sexos? Haverá alguma evidência de descriminação por sexo?

b) Haverá alguma diferença nas taxas de admissão entre os homens e mulheres na Faculdade de Ciências? E na Faculdade de letras?
c) Como se pode explicar a contradição verificada nas duas alíneas anteriores?

4- Classifique as variáveis que se seguem em qualitativas, quantitativas discretas ou quantitativas contínuas.

- Altura de um indivíduo;
- O seu peso;
- Idade;
- Nacionalidade;
- Cor dos olhos;
- Situação sócio económica de um indivíduo;
- Número de filhos de famílias residentes num prédio
- Número de transistores numa Caixa
- Quantidades ingeridas por família dum Bairro durante uma semana
- Programas televisivos com maior audiência
- Produto interno bruto de 15 paises da região

Um dos grandes problemas quando se faz um estudo estatístico, consiste em como seleccionar os casos em estudo. Imagine que se pretende fazer um estudo afim de se verificar se o não encandeamento entre automóveis reduz os efeitos graves provocados pelos acidentes:
Usar todos automóveis ou apenas os transportes semi colectivos?
Tomar apenas as distâncias focais de cada farol ou apenas a potência das lâmpadas colocadas?
Analisar os dados dos acidentes agrupando-os em Transportes semi colectivos e outros veículos ou analisar todos em conjunto?
Analisar dado de todos acidentes da mesma maneira ou categorizá-los

CAPÍTULO 2 ESTATÍSTICA DESCRITIVA

Objectivos do capítulo:
- Criar tabelas com base na definição de variáveis
- Definir frequência relativa e absoluta
- Definir frequências acumuladas
- Diferenciar as frequências
- Implementar tabelas e gráficos
- Diferenciar gráficos
- Aplicar gráficos em situações concretas
- Descrever prinicipais medidas
- Diferenciar medidas de Tendência Central, Dispersão e Posição
- Determinar as principais medidas
- Aplicar diferentes medidas em exemplos
- Definir os principais indicadores genéricos
- Diferenciar os principais indicadores
- Saber definir economia
- Estudar o conceito rácio duma forma geral
- Definir Correlação
- Criar diagrama de dispersão
- Discutir regressão
- Determinar coeficientes
- Interpretar o coeficiente de correlação

A análise estatística começa quando um conjunto de dados torna-se disponível de acordo com a definição do problema da pesquisa. Um conjunto de dados, seja de uma população ou de uma amostra, contém muitas vezes um número muito grande de valores. Além disso, esses valores, na sua forma bruta, encontram-se muito desorganizados. Eles variam de um valor para outro sem qualquer ordem ou padrão, por isso os dados precisam ser organizados e apresentados numa forma sistemática e sequencial por meio de uma tabela ou gráfico. Quando fazemos isso, as propriedades dos dados tornam-se mais aparentes e tornamo-nos capazes de determinar os métodos estatísticos mais apropriados para serem aplicados no seu estudo.

1. FREQUÊNCIA

Tomemos a variável idade da tabela 1.4, para distribuir em frequências e fazer alguma análise.

Para apresentarmos a distribuição de frequência absoluta, colocamos os valores que a variável toma na primeira coluna e o número de vezes que o dado aparece (repetido) na segunda coluna.

Tabela 2.1 – Distribuição de frequência absoluta da variável anos de serviço da tabela 1.4

Anos de Serviço (x_i)	Frequência (f_i)
8	2
12	5
15	8
16	3
Total	18

A tabela 2.1, é uma tabela que representa a distribuição de frequência absoluta de dados discretos (Anos de Serviço), em virtude de resultar duma contagem.

Quando se pretende analisar o impacto da frequência absoluta relativamente ao número total de observações, recorre-se a frequência relativa (percentual). Vejamos a tabela a seguir.

Tabela 2.2 – Distribuição de frequências relativas da tabela 2.1

Anos de Serviço (x_i)	Frequência (f_i)	frequência Relativa % (f_r)
8	2	$\frac{2}{18}\times 100 = 11{,}11$
12	5	$\frac{5}{18}\times 100 = 27{,}78$
15	8	$\frac{8}{18}\times 100 = 44{,}44$
16	3	$\frac{3}{18}\times 100 = 16{,}67$
Total	18	100,00

Neste caso, pode-se dizer por exemplo que 44,44% dos trabalhadores tem quinze anos de serviço e que mais de metade dos trabalhadores tem 15 ou mais anos de serviço (16).

Frequências absoluta acumulada e relativa acumulada

Para o exemplo representado pelas tabelas 1.3, 1.4 e 2.1 podemos calcular a frequência relativa (tabela 2.2) e as respectivas frequências absolutas e relativas acumuladas. Observe a tabela seguinte:

Tabela 2.3 – Distribuição de frequências acumuladas

Variável (Anos de Serviço)	Frequência Absoluta (f_i)	Frequência Absoluta Acumulada (F_i)	Frequência Relativa % (f_r)	Frequência Relativa Acumulada % (F_r)
8	2	2	11,11	$\frac{2}{18}\times 100 = 11{,}11$ ou 11,11
12	5	$2+5=7$	27,78	$\frac{7}{18}\times 100 = 38{,}89$ ou $11{,}11+27{,}78=38{,}89$
15	8	$7+8=15$	44,44	$\frac{15}{18}\times 100 = 83{,}33$ ou $38{,}89+44{,}44=83{,}33$
16	3	$15+3=18$	16,67	$\frac{18}{18}\times 100 = 100{,}00$ ou $83{,}33+16{,}67=100{,}00$
Total	18		100,00	

A frequência relativa é dada pela fórmula $f_r = \frac{f_i}{n}$ (I) e a Acumulada $F_r = \frac{F_i}{n}$ (II)

Cálculo de frequência absoluta acumulada (F_i)
F_1 é igual a f_1
$F_i = F_{i-1} + f_i$, para qualquer i>1, i é número natural

A **frequência relativa acumulada** pode ser calculada de duas maneiras. Na primeira, tal como é feito na tabela acima, dividindo o valor da frequência absoluta acumulada pelo total de observações. Na segunda maneira, acumulamos o valor da frequência relativa. Este último método pode levar a um acumular de erros, de forma que o último valor da frequência relativa acumulada se distancie consideravelmente de 1.

Observação: Daqui em diante onde se lê frequência entenda-se como sendo frequência absoluta e que a frequência relativa pode aparecer em muitos casos de forma percentual, onde: f_i representará a frequência absoluta da i-ésima ocorrência, f_r representará a frequência relativa e as F_i e F_r as respectivas frequências acumuladas.

Dados brutos

São aqueles obtidos directamente da pesquisa, isto é, ainda não sofreram qualquer processo de síntese ou análise. Em geral são apresentados em tabelas e frequentemente omitidos na maioria das publicações, quase sempre por uma questão de espaço.
O conjunto de dados constitui uma amostra cujo tamanho denota-se por n.

Exemplo: Foi realizado um estudo piloto com objectivo de verificar como os hábitos de vida das pessoas influenciam o risco de desenvolvimento de doenças cardíacas. Hoje em dia muitos resultados estão completamente integrados à prática cardiológica.

A tabela a seguir mostra as medidas em mg/dl de colesterol, referente a um exame realizado em 1952 em 80 pacientes. Os pacientes que não compareceram estão representados por espaço em branco. Observando estes dados, na maneira como estão apresentados, é muito difícil saber o valor em torno do qual as medidas estão agrupadas, a forma da distribuição e a extensão da variabilidade.

Tabela 2.4 - Nível de colesterol (mg/dl) em 80 indivíduos (dados brutos)

278	182	247	227	277	194	196	276	244	192
118	219	255	201		209	219	228	209	209
171	213	233	226	209	200	200	363	209	200
179	167	192	277	317	146	217	292	217	255
212	233	250	243	150	209	184	199	250	479
175	194	221	233		184	217	150	167	265
242	180	255	170	209	161	196	165	234	179
248	184	291	185	242	276	243	229	242	250

Tabela 2.5 – Distribuição de frequências dos dados da Tabela 2.4

Colesterol mg/dl	Frequência absoluta		Frequência relativa	
	Simples (f)	Acumulada (F)	Simples (f_r)	Acumulada (F_r)
$100 \vert - 150$	2	2	0.0256	0.0256
$150 \vert - 200$	24	26	0.3077	0.3333
$200 \vert - 250$	35	61	0.4487	0.7821
$250 \vert - 300$	14	75	0.1795	0.9615
$300 \vert - 350$	1	76	0.0128	0.9744
$350 \vert - 400$	1	77	0.0128	0.9872
$400 \vert - 450$	0	77	0.0000	0.9872
$450 \vert - 500$	1	78	0.0128	1.0000
Total	78		1	

$\vert -$ Significa que o intervalo está fechado do lado esquerdo e aberto do lado direito

Etapas para a construção de tabelas de frequências para dados agrupados e com intervalos de classes iguais:

1) Encontrar o máximo (maior) e o mínimo (menor) valores do conjunto de dados;

2) Escolher um número de subintervalo ou classes (mais adiante abreviado pela letra k), em geral de mesma amplitude (tamanho, abreviado por c), que englobem

todos os dados sem haver superposição dos intervalos. Os extremos dos intervalos são chamados de limites de classes;

3) Contar o número de elementos que pertencem a cada classe; este número é denominado frequência absoluta, em geral representada por f_i;

4) Determinar a frequência relativa de cada classe, dividindo a frequência da classe pelo número total de observações.

Observação: Algumas regras práticas podem ser usadas na construção de tabelas, como por exemplo variar o número de classes entre 5 e 15; o número de classes pode ser $\sqrt{n}$; o tamanho de cada classe é escolhido como o quociente entre a amplitude (diferença entre o maior e o menor números) do conjunto e o número de classes escolhido. Por conveniência, é possível modificar esse valor para facilitar a construção e interpretação da tabela; o limite inferior da primeira classe e o limite superior da última classe devem ser um pouco menor e maior que a menor e maior observação respectivamente.

Exemplo: Foram levantadas tentativas de suicídio por intoxicação aguda registadas num centro de assistência a problemas Tóxicos. No período de Janeiro/1992 a Fevereiro/1993, 302 casos tiveram caracterização.

Tabela 2.6 -Distribuição de profissões entre pacientes potencialmente suicidas

Profissão	f_i	f_r
Serviços gerais*	75	24,8
Doméstica**	55	18,2
Do lar	53	17,5
Indeterminada	29	9,6
Emprego especializado***	23	7,6
Menor	20	6,6
Desempregado	15	5,0
Estudante	14	4,6
Lavrador	12	4,0
Autônomo	4	1,3
Aposentado	2	0,7
Total	302	1,00

Como é possível observar, na distribuição de pacientes potencialmente suicidas, a incidência foi observada entre profissionais mal remunerados e com sobrecarga de trabalho, sem perspectiva de ascensão social. O alto percentual entre menores e estudantes (6,6% + 4,6%) confirma o facto de que o índice de tentativas de suicídio é preocupante.

1.1 DIAGRAMAS

Com uma simples inspecção do histograma da figura a seguir, pode ver que a maioria dos indivíduos tem colesterol em torno de 225 mg/dl. Tem ainda uma boa descrição de como os diferentes níveis se distribuem em torno deste valor.

Histograma para o nível de colesterol da tabela 2.5

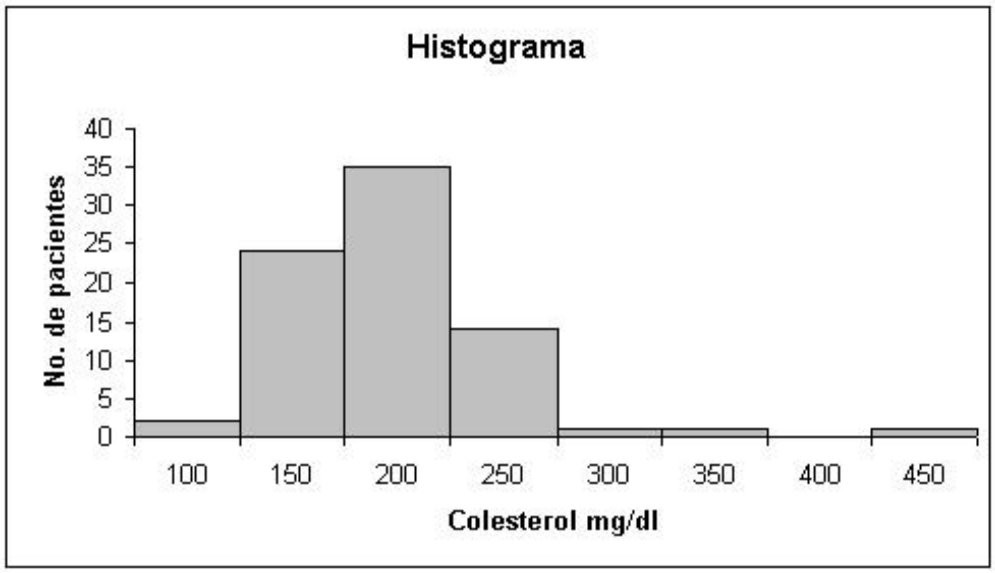


Fig 2.1

A partir do histograma pode-se construir o polígono de frequência, que consiste em unir através de segmentos de rectas as ordenadas correspondentes aos pontos médios de cada classe.

Polígono de frequência para o colesterol da tabela 2.5

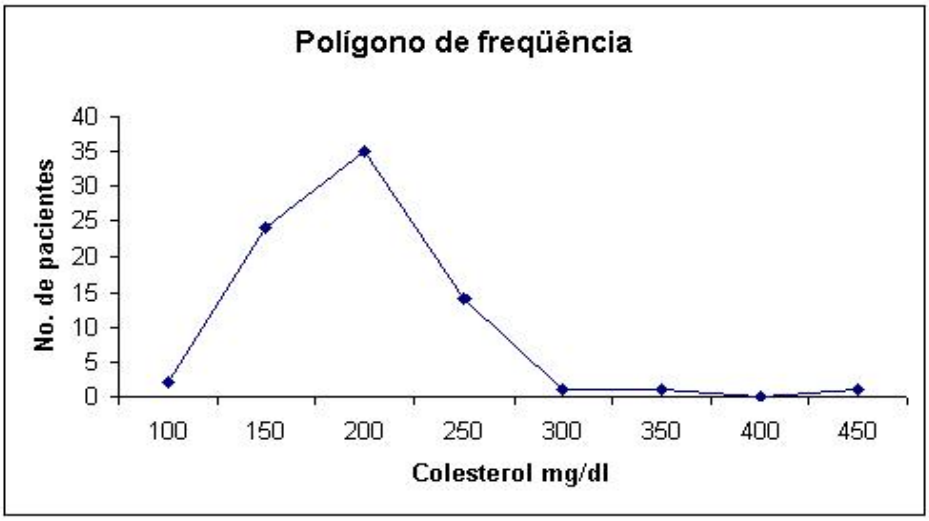


Fig 2.2

A **ogiva** é um gráfico de frequências acumuladas (usualmente relativas). Para construí-la coloca-se no eixo horizontal os intervalos de classe nos quais a variável em estudo foi dividida. Para cada limite de intervalo é assinalado no eixo vertical sua percentagem acumulada. Em seguida, os pontos marcados são ligados por segmentos de recta. Embora a palavra ogiva não seja bastante sugestiva, trata-se de uma poligonal ascendente.

Ogiva do colesterol da tabela 2.5

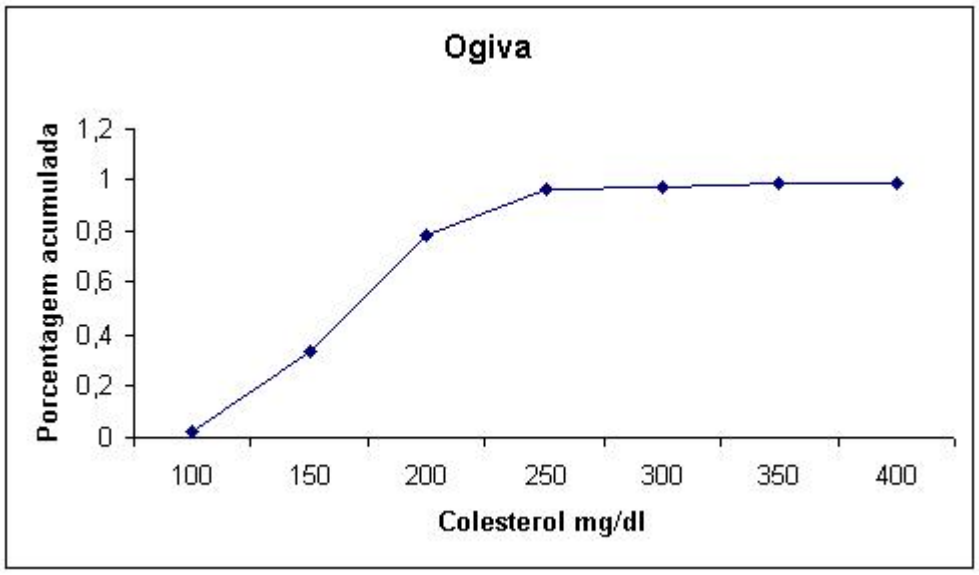


Fig 2.3

O histograma e o polígono de frequências servem para visualizar a forma de distribuição da variável em estudo. Através da ogiva podem-se estimar os percentís da distribuição, isto é, o valor que é precedido por certa percentagem de interesse pré-estabelecida. Por exemplo, estimar o valor da variável do qual se tem 50% dos indivíduos.

Diagrama Circular

Como o próprio nome sugere, esta representação é constituída por um círculo, em que se apresentam vários sectores circulares, tantos quantos as classes consideradas na tabela de frequências da amostra em estudo. Os ângulos dos sectores são proporcionais às frequências das classes. Por exemplo, uma classe com uma frequência relativa igual a 0,20, terá no diagrama circular um sector com um ângulo igual a $360^{o} \times 0{,}20 = 72^{o}$. É uma representação utilizada essencialmente para dados qualitativos.

Exemplo: Categoria profissional dos funcionários de uma Faculdade

Tabela 2.7 - Categoria profissional dos funcionários de uma Faculdade

Classes	Freq. Abs. (f)	Freq. Relat. (f_r)	
Professor Estagiário	20	0,47	$0{,}47 \times 100\% = 47\%$
Professor Auxiliar	12	0,29	$0{,}29 \times 100\% = 29\%$
Administrativo	7	0,17	$0{,}17 \times 100\% = 17\%$
Contínuo	3	0,07	$0{,}07 \times 100\% = 7\%$
Total	42	1,00	$1{,}00 \times 100\% = 100\%$

Gráfico em Sectores (Pizza)

Error!

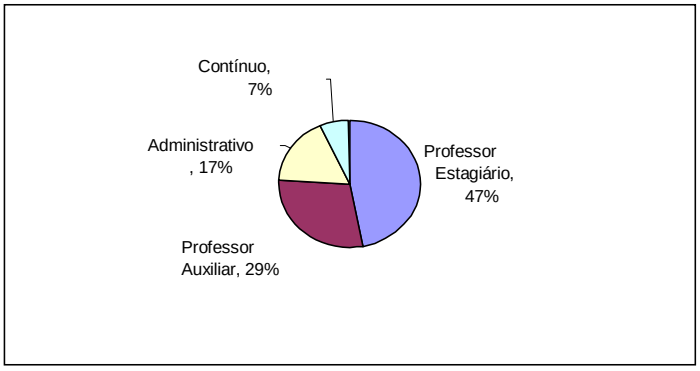


Fig 2.4

Diagrama de Barras para dados Qualitativos, representados na tabela 2.7

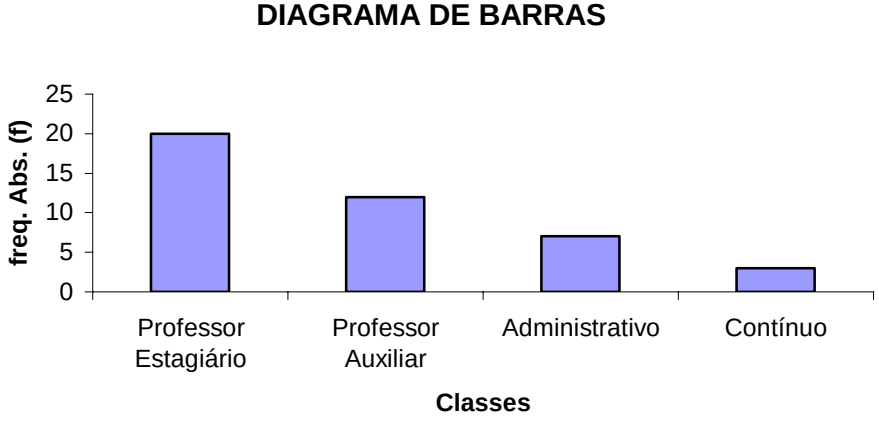


Fig 2.5

É de notar que dentre os dois diagramas (circular e o de barras), o circular representa uma mais valia em termos de facilidade na leitura de dados.

Exemplo:
Num determinado ano lectivo, uma Universidade admitiu 3700 homens dos 8300 candidatos e 1500 mulheres das 4300 candidatas. Os dados apresentam-se na tabela abaixo.

Tabela 2.8 – Distribuição de homens e mulheres por faculdades

	Homens		Mulheres	
Faculdade	Nº de Candidatos	Nº de Admitidos	Nº de Candidatas	Nº de Admitidas
Letras	2300	700	3200	900
Ciências	6000	3000	1100	600
Total	8300	3700	4300	1500

Diagrama de barras composto do tipo lado a lado (clustered) da tabela 2.8

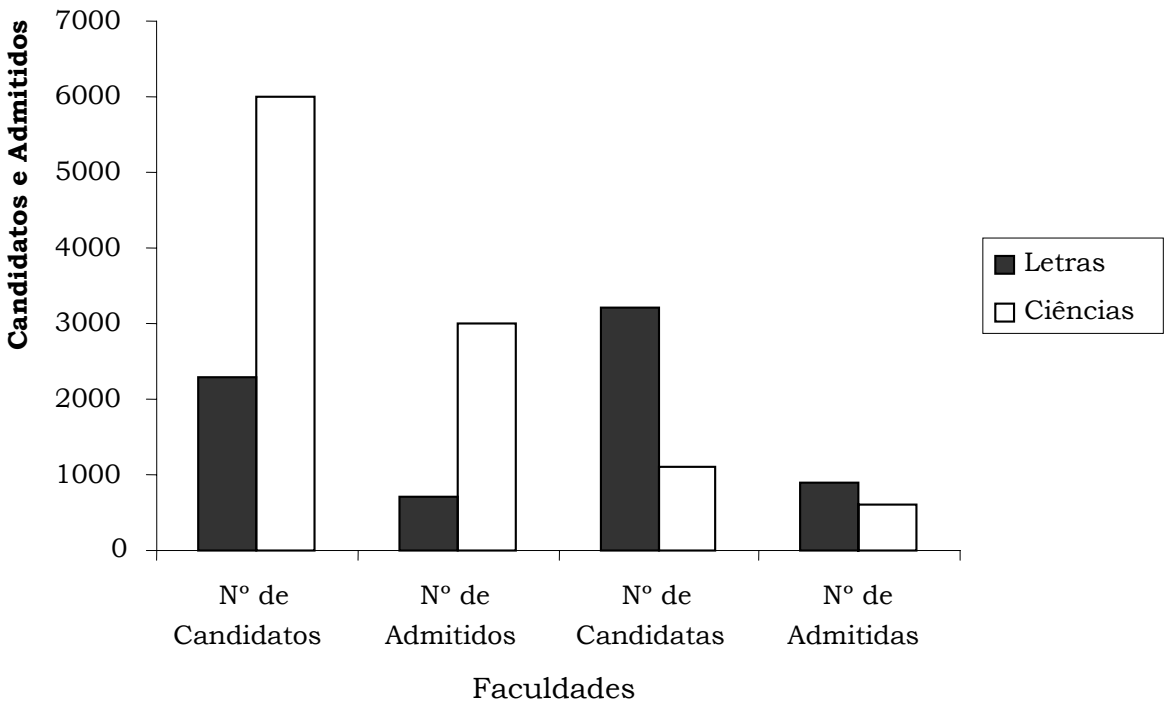


Fig 2.6

Este tipo de diagramas é mais usado para tipo de variável qualitativa cruzada. Atente-se para o facto da tabela de distribuição de frequências ser cruzada (tabela 2.8).

Nota: Um diagrama de barras apresenta legenda, enquanto que um histograma não apresenta legenda

Histogramas (Vejamos de novo mais detalhadamente)

Histograma é uma representação gráfica de uma tabela de distribuição de frequências. Desenhamos um par de eixos cartesianos em que no eixo horizontal (abcissas) colocamos os valores da variável em estudo, enquanto que no eixo vertical (ordenadas) colocamos os valores das frequências. O histograma tanto pode ser representado para as frequências absolutas como para as frequências relativas. No caso do exemplo apresentado através da tabela 4, o histograma seria:

Exemplo:
Suponhamos o seguinte conjunto de dados: 14, 12, 13, 11, 12, 13, 16, 14, 14, 15, 17, 14, 11, 13, 14, 15, 13, 12, 14, 13, 14, 13, 15, 16, 12, 12
A sua distribuição de frequências é a seguinte:

Tabela 2.9 – Distribuição de frequências

Variável	Freq. Abs. (f)	Freq. Relat. (f_r) %	F	F_r (%)
11	2	$\frac{2}{26}\times 100\% = 7{,}69$	2	7,69
12	5	$\frac{5}{26}\times 100\% = 19{,}23$	$2+5=7$	$7{,}69+19{,}23=26{,}92$
13	6	$\frac{6}{26}\times 100\% = 23{,}08$	$7+6=13$	$26{,}92+23{,}08=50{,}00$
14	7	$\frac{7}{26}\times 100\% = 26{,}92$	$13+7=20$	$50{,}00+26{,}92=76{,}92$
15	3	$\frac{3}{26}\times 100\% = 11{,}54$	$20+3=23$	$76{,}92+11{,}54=88{,}46$
16	2	$\frac{2}{26}\times 100\% = 7{,}69$	$23+2=25$	$88{,}46+7{,}69=96{,}15$
17	1	$\frac{1}{26}\times 100\% = 3{,}85$	$25+1=26$	$96{,}15+3{,}85=100{,}00$
Total	26	100,00%		

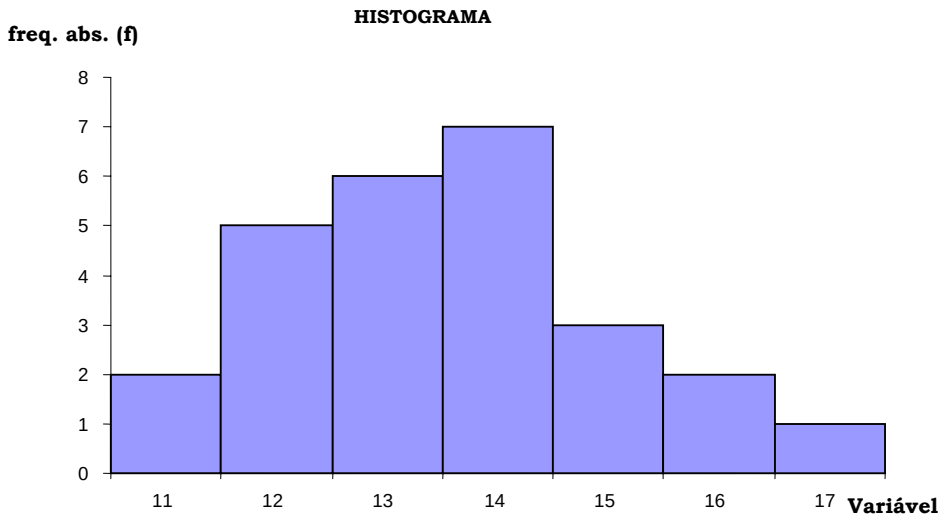


Fig 2.7

Para já a pergunta que se coloca é a seguinte: porquê os dados aparecendo em forma de pontos o gráfico está em forma de barras?

A resposta a essa pergunta é simples, basta ver mais adiante nos itens relativos a dados agrupados, facilmente ter-se-á a resposta. De referir que os Histogramas são construídos para dados agrupados (que se apresentam em classes)

Para o gráfico atrás apresentado considerou-se que 11 é ponto médio de [10,5; 11,5]; 12 é ponto médio de [11,5;12,5]; e assim sucessivamente.

Diagrama ramo-e-folhas

Tanto o histograma como os gráficos em barras dão uma ideia da forma da distribuição da variável sob consideração. Mas a forma da distribuição é tão importante quanto as medidas de posição e de dispersão. Por exemplo saber que a renda per capita de Moçambique é de tantos dólares pode ser um dado muito interessante, mas saber como essa renda se distribui é ainda mais importante. Um procedimento alternativo para resumir um conjunto de valores, com o objectivo de se obter uma ideia da forma de sua distribuição, é o diagrama de Ramo-e-folhas. Uma vantagem desse diagrama relactivamente histograma é que não perdemos (ou perdemos pouca) informação sobre os dados em sí.

Não existe uma regra fixa para construir o diagrama ramo-e-folhas, mas a ideia básica é dividir cada observação em duas partes: A primeira (ramo) é colocada à esquerda de uma linha vertical, a segunda (folha) é colocada à direita.
Observação: Um ramo com muitas folhas significa maior incidência daquele ramo (realização)

Exemplo: Tomemos as vendas do peixe Mapatana pela Empresa Finage Mar nos últimos 18 anos

Tabela 2.10 – Nível de vendas de Mapatana pela Empresa Finage Mar

Ano	Vendas (1000 t)	Ano	Vendas (1000 t)
1	280	10	365
2	305	11	280
3	310	12	375
4	330	13	380
5	310	14	400
6	340	15	369
7	310	16	390
8	340	17	400
9	369	18	369

Coloquemos as vendas em rol (ordem crescente ou decrescente). Para o caso concreto será pela ordem crescente. 280, 280, 305, 310, 310, 310, 330, 340, 340, 365, 369, 369, 369, 375, 380, 390, 400, 400. O respectivo diagrama de ramo-e-folhas ficaria:

Diagrama de Ramo e Folhas da Tabela 2.10

Ramo	Folhas
28	0 0
30	5
31	0 0 0
33	0
34	0 0
36	5 9 9 9
37	5
38	0
39	0
40	0 0

Fig 2.8

Repare-se que após a extracção dessa informação podemos tirar algumas conclusões:

- Não há um valor de maior destaque, isto é, um ramo que esteja mais solto do resto dos ramos contendo folhas;
- Todos valores estão razoavelmente concentrados entre 280 e 400
- Um valor mais ou menos mediano para este conjunto de dados é por exemplo, 340;
- Há uma leve assimetria em direcção aos valores pequenos; a suposição de que estes dados possam ser considerados como amostra de uma população com distribuição simétrica, em forma de sino (a chamada distribuição normal), pode ser questionada.

Observação: A escolha do número de linhas do diagrama ramo-e-folhas é equivalente à escolha do número de classes de um histograma. Um número pequeno de linhas (ou

classes) enfatiza a parte "Folha" da relação, resumido pela primeira coluna, enquanto que as restantes colunas enfatizam a parte "Ramo".

O Diagrama de Pontos (Scatter Plot) para a tabela 2.9

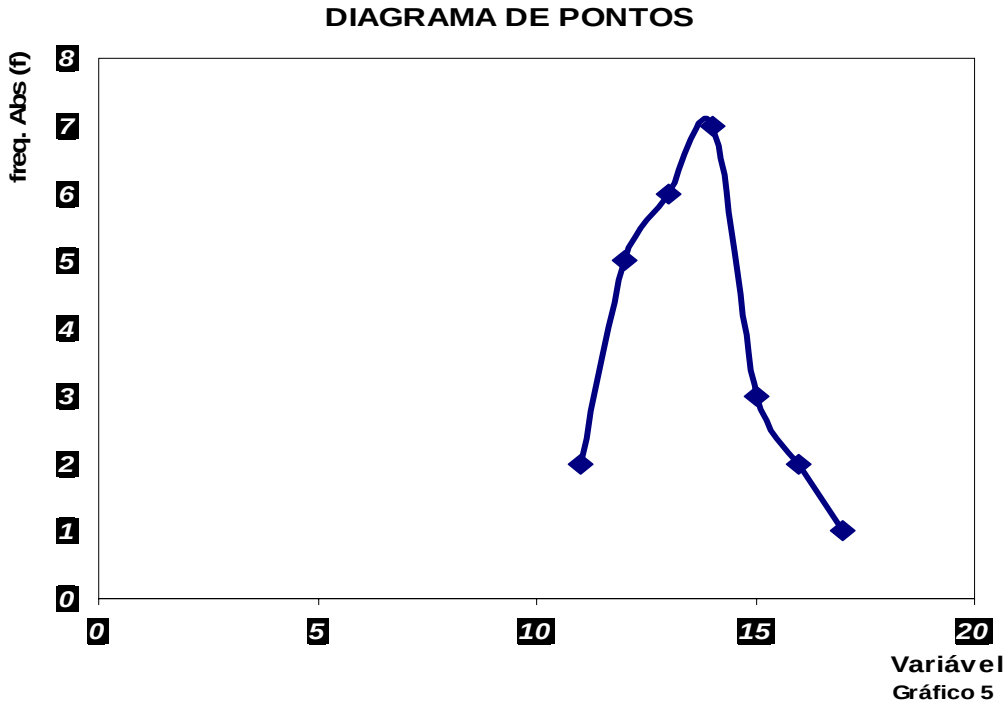


Fig 2.9

Este tipo de diagrama é muito usado para associação entre duas variáveis quantitativas. Também se usa este tipo de diagrama quando se pretende fazer Análise de Covariância. Mais adiante será usado para Análise de Correlação e de Regressão. De salientar ainda que cada ponto é um par ordenado de cada duas variáveis.

O Diagrama de Box Plot

Exemplo: Dada a tabela a seguir, faça um diagrama de Box-Plot.

Tabela 2.11 – Nível de temperatura da unha pós fricção

Nome	Idade	Temperatura
Luisa	12	5
Zicai	14	8
Kelvin	15	41
Trevour	18	13
Virgínia	47	14
Ivete	12	65
Santos	16	10
Acendino	19	13
Noé	89	32

Diagrama de Box – Plot

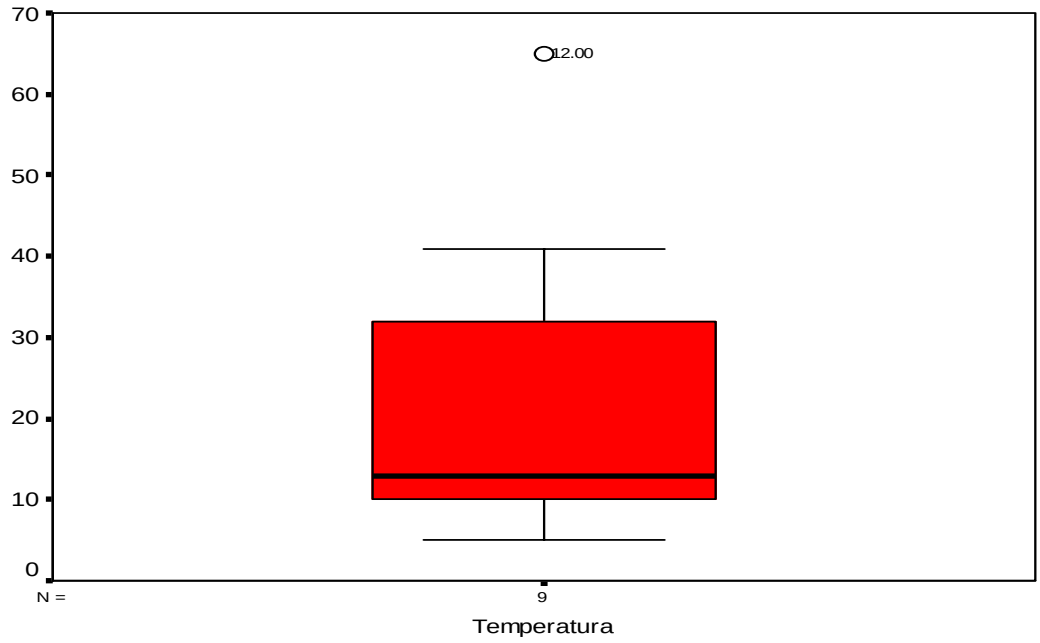
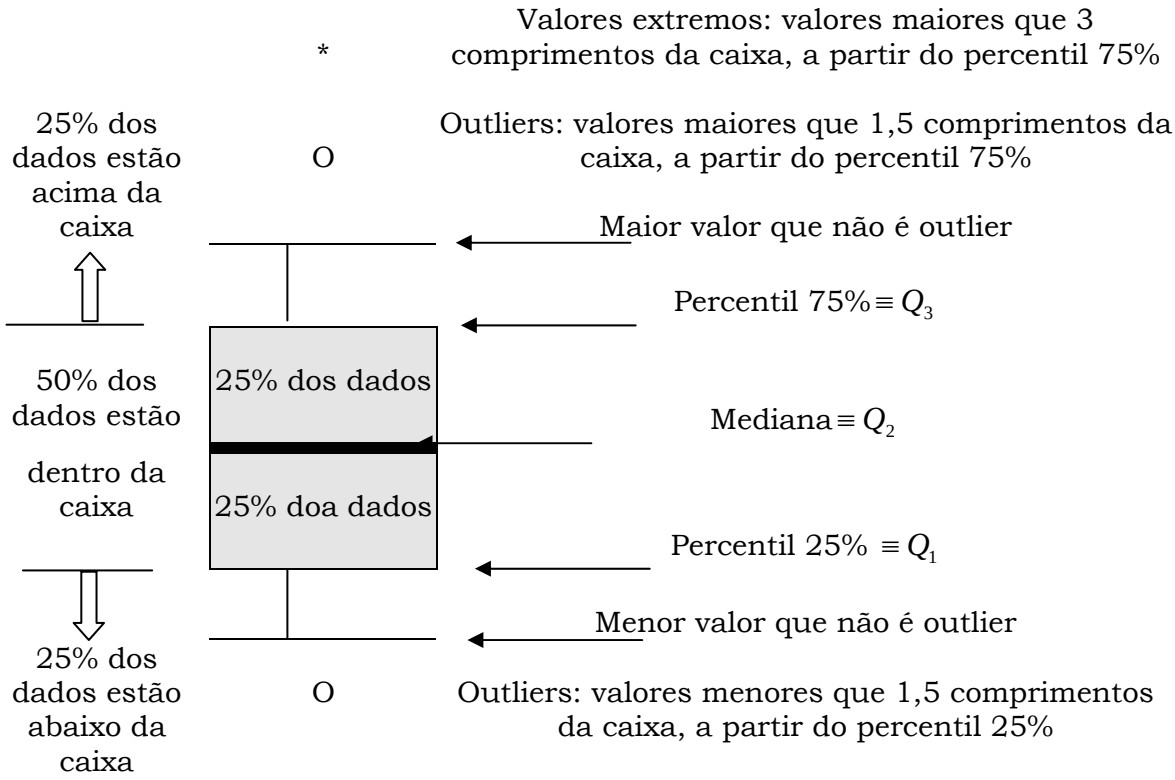


Fig 2.10

Este tipo de diagrama é mais usado para valores de uma variável quantitativa em função de uma qualitativa, para análise de variância. Repare-se que esta análise é para uma relação entre variáveis.

Como construir o diagrama de Box-Plot?



Comprimento da caixa = amplitude interquartílica = $Q_3 - Q_1$

Fig 2.11

Diagrama de Linhas

Diagrama de Linhas para dados da tabela Tabela 2.11

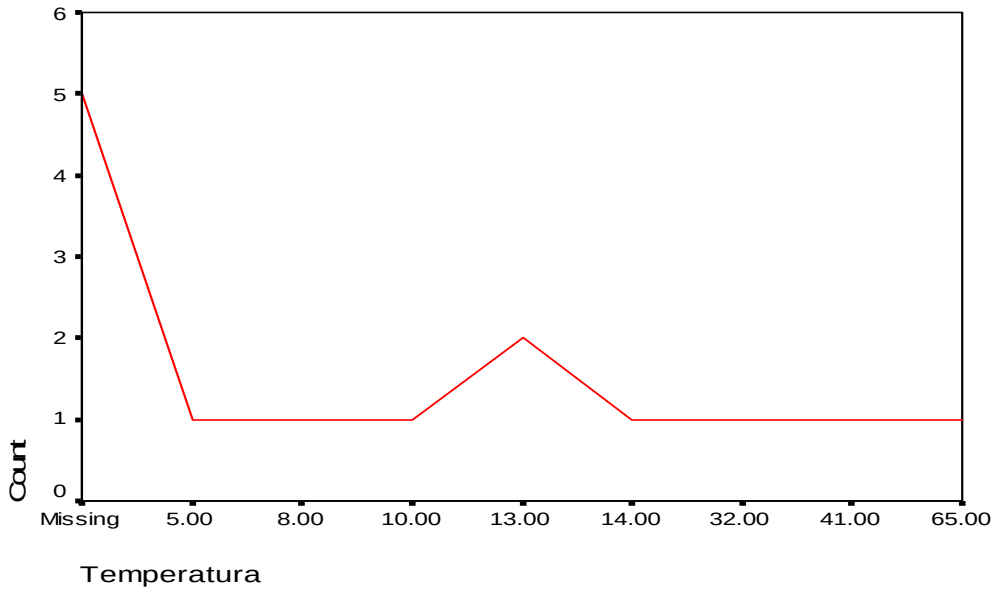


Fig 2.12

Este tipo de gráficos é usado para análise de séries temporais e cronológicas, quando se pretende analisar a trajectória de variáveis ao longo do tempo.

Diagrama de Áreas

Diagrama de Áreas para dados da tabela Tabela 2.11

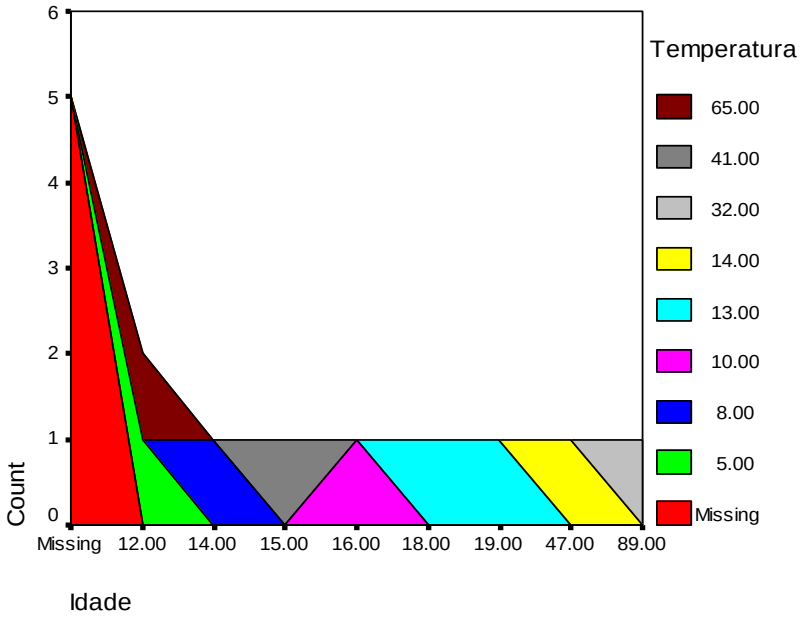


Fig 2.13

O gráfico de áreas é usado para Análise de Regressão e Correlação, cujos valores das variáveis que se pretendam são para trajectórias ao longo do tempo. Assim também o são os gráficos do tipo Drop-Line tratados a seguir.

Diagrama Drop Line

Diagrama Drop Line para dados da tabela Tabela 2.11

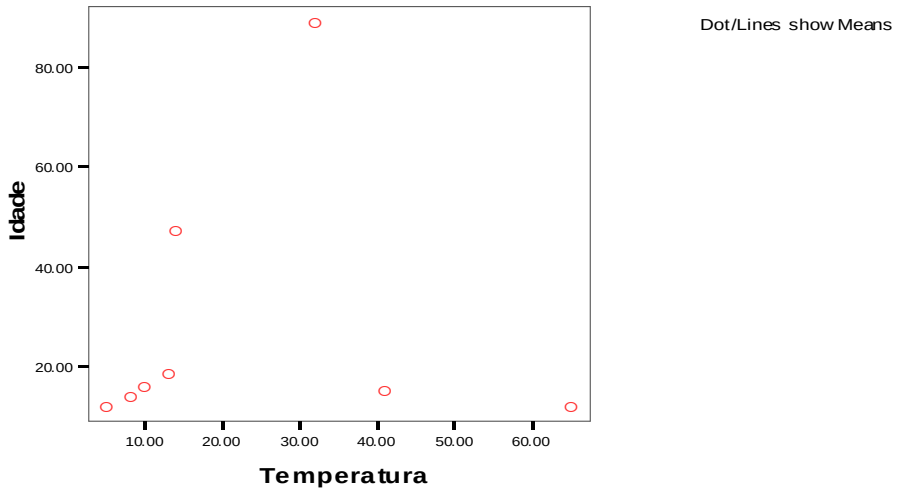


Fig 2.14

Vejamos resumidamente a seguir:

Quadro 2.1 – Procedimentos disponíveis para a apresentação de dados.

Tipo da variável	Valores da variável	Tipo de tabela	Tipo de estatística	Tipo de gráfico
Qualitativa Sexo Classe Grau de instrução do pai Tipo de escola Turno Repetente em História	Recomenda-se definir a variável como numérica e depois colocar os rótulos 1=Feminino; 2=Masculino. 1=1ª Classe do 1º Grau; 2=2ª classe do 1º Grau; 3=3ª classe do 1º Grau. 1=Analfabeto; 2=Primário; 3=Secundário; 4=Superior 1=Pública; 2=Privada. 1=Matutino; 2=Vespertino e 3=Noturno. 1=Sim; 2=Não.	Tabela de distribuição de frequências	Frequências absolutas e relativas	Barras simples Ou Circular
Qualitativa cruzada Repetente versus Classe		Tabela de distribuição de frequências cruzada	Frequência simples; relativa à linha e/ou coluna (valor esperado, teste chi-quadrado)	Gráfico de barras composto: Lado a lado (clustered) Superposto (stacked); Opostos
Quantitativa Discreta (que toma poucos valores) Número de filhos por mulher Número de reprovações por série Número de horas por dia que estuda matemática	0; 1; 2; 10 0; 1; 2; 3; 4 0; 1; 2; 3; 4	Tabela de distribuição de frequências	Frequências absolutas e relativas	Gráfico de bastão; Gráfico de barras simples

Observa-se que variáveis qualitativas ordinais podem ser tratadas como variáveis quantitativas, por exemplo, Classe em que estuda, que poderia ser interpretado como número de anos de estudo aprovados. Assim, o estudo da taxa de fracasso escolar por Classe pode ser trabalhado, tanto com o teste qui-quadrado, quanto com a análise de regressão e correlação.

Quadro 2.2 – Procedimentos disponíveis para a apresentação de dados.

Tipo da variável	Valores da variável	Tipo de tabela	Tipo de estatística	Tipo de gráfico
Quantitativa Discreta (que toma muitos valores) Número de alunos por turma Idade do pai (anos completos) Número de veículos que passam por um ponto movimentado Quantitativa Contínua Nota na prova de matemática Valor na escala de atitudes(*) Renda familiar Coeficiente de Inteligência Tempo gasto na prova	20; 21; ..., 50 30, 31,.....,70,... 0,1,2, ..., 500,... Intervalo fechado de 0 a 10: [0; 10] Intervalo fechado de 20 a 80: [20; 80]. Intervalo semifechado de 0 a M: [0; M[Intervalo fechado de 0 a 150: [0; 150] Intervalo fechado de 0 a 2 horas: [0; 2]	Tabela de distribuição de frequências desde que os dados tenham sido agrupados em faixas ou intervalos Tabela de distribuição de frequências desde que os dados tenham sido agrupados em faixas ou intervalos	Média; Mediana Moda Desvio padrão Coeficiente de variação Quartis ... Média; Mediana Moda Desvio padrão Coeficiente de variação Quartis ...	Diagrama de ramo e folha Histograma (pode usar a opção da distribuição normal, caso se esteja trabalhando sob esse pressuposto) Diagrama de ramo e folha Histograma (pode usar a opção da distribuição normal, caso se esteja trabalhando sob esse pressuposto)
Relação entre variáveis	Quando se quer analisar associação entre duas ou mais variáveis quantitativas		Análise de correlação Análise de regressão	Scatter plot ou diagrama de pontos
	Uma quantitativa em função de uma qualitativa		Análise de variância	Diagrama de ramo e folha, box-plot
	Uma quantitativa em função de variáveis qualitativas e quantitativas		Análise de covariância	Scatter plot ou diagrama de pontos
Séries temporais Número de alunos matriculados no período de 1980 a 1998	Quando se pretende analisar a trajetória de variáveis ao longo do tempo	Tabela contendo a variável tempo e as variáveis estudada	Análise de séries temporais Análise de regressão e correlação	Gráfico de linhas; De áreas Drop-line

(*) pela forma de construção, esta variável seria discreta

1.2 Tabulação de frequência com intervalos de classes iguais e histograma para variáveis contínuas

Até agora vimos como são calculadas as frequências (absolutas, relativas e acumuladas) para variáveis quantitativas discretas. Nesse caso a tabulação dos resultados é mais simples. Se tratamos de variáveis quantitativas contínuas, os valores observados devem ser apresentados em tabelas, agrupados em classes.

Para a determinação dessas classes não existe uma regra pré estabelecida, sendo necessário um pouco de tentativa e erro para a solução mais adequada. Nalguns casos pode-se determinar intervalos arbitrários, segundo o interesse do Analista, mas na maior parte dos casos usam-se os pressupostos que se seguem:

- Se n é tamanho da amostra, considera-se que a amostra é grande se $n > 30$ e pequena se $n \leq 30$;
- Se k for o número de classes que pretendemos, então $k = 5$ se $n \leq 30$ e $k = \sqrt{n}$ se $n > 30$;
- Se λ é amplitude total, será obtido através da diferença entre o maior e o menor valores da amostra.
- Se c é o intervalo de cada classe, será obtido pela fórmula $c = \frac{\lambda}{k}$;
- f é a frequência absoluta;

Se $X_{\min}$ e $X_{\max}$ forem os valores mínimos e máximos da amostra, teremos a distribuição de classes como se ilustra na tabela seguinte:

Tabela 2.12 – Formas de representação numa tabulação de frequências

Classes	f	...
$[X_{\min}$; $X_{\min}$ + c[		
$[X_{\min} + c$; $X_{\min}$ + $2 \times c$ [		
$[X_{\min}$ $2 \times c$; $X_{\min}$ + $3 \times c$ [		
.....		
$[X_{\max}$ - $2 \times c$; $X_{\max}$ - c [		
$[X_{\max}$ - c; $X_{\max}$]		

Exemplo: Com base nos dados constantes na tabela 2.1 – Níveis de Comercialização de Mapatana, Faça a Distribuição de frequências dos dados.

Resolução: As tonelagens por ano são dados quantitativos, defindindo desta forma uma variável quantitativa contínua. Vamos transformá-la em intervalar.

1º) Definir o número de classes: como $n = 18$, o que significa que $n \leq 30$, então k (número de classes) é igual a 5, isto é, $k = 5$;

2º) Definir a amplitude total (amplitude do conjunto das vendas) λ : $\lambda = \text{valor máximo - valor mínimo} = 430 - 280 = 150$;

3º) Definir Amplitude de Classe c: $c = \frac{\lambda}{k}$

$$\text{c} = \text{amplitude de classe} = \frac{\text{valor máximo - valor mínimo}}{\textit{número} \text{ de classes}} = \frac{430 - 280}{5} = 30$$

4º) Preparar as classes:

Tabela 2.13 – Determinação de classes da tabela 2.10

Classe	Limite inferior	Limite Superior	x_i
1	280	310	295
2	310	340	325
3	340	370	355
4	370	400	385
5	400	430	415

X_i é o ponto médio de cada classe, e é obtido por $X_i = \frac{l_i + l_s}{2}$

onde l_i é limite inferior da classe e l_s limite superior da classe.

5º) Distribuição (Tabulação) de Frequências

Tabela 2.14 – Distribuição de frequências da tabela 2.13

1	2	3	4	5	6	7	8
Classes (Tonelagem)	x_i	f_i	f_r	$X_i f_i$	$X_i - \overline{X}$	$\left(X - \overline{X}\right)^2$	$\left(X - \overline{X}\right)^2 f_i$
[280– 310[	295	3	0.12	885	531,67	282669,44	848008,3
[310– 340[	325	4	0.16	1300	946,67	896177,78	3584711,1
[340– 370[	355	4	0,16	1420	1066,67	1137777,78	4551111,1
[370– 400[	385	5	0.28	1925	1571,67	2470136,11	12350680,6
[400– 430]	415	2	0.20	830	476,67	227211,11	454422,2
Total		18	1.00	6360			21788933,3

6º) Como os dados constantes na tabela são representantes de uma variável quantitativa contínua, é claro que o respectivo gráfico será um Histograma a seguir

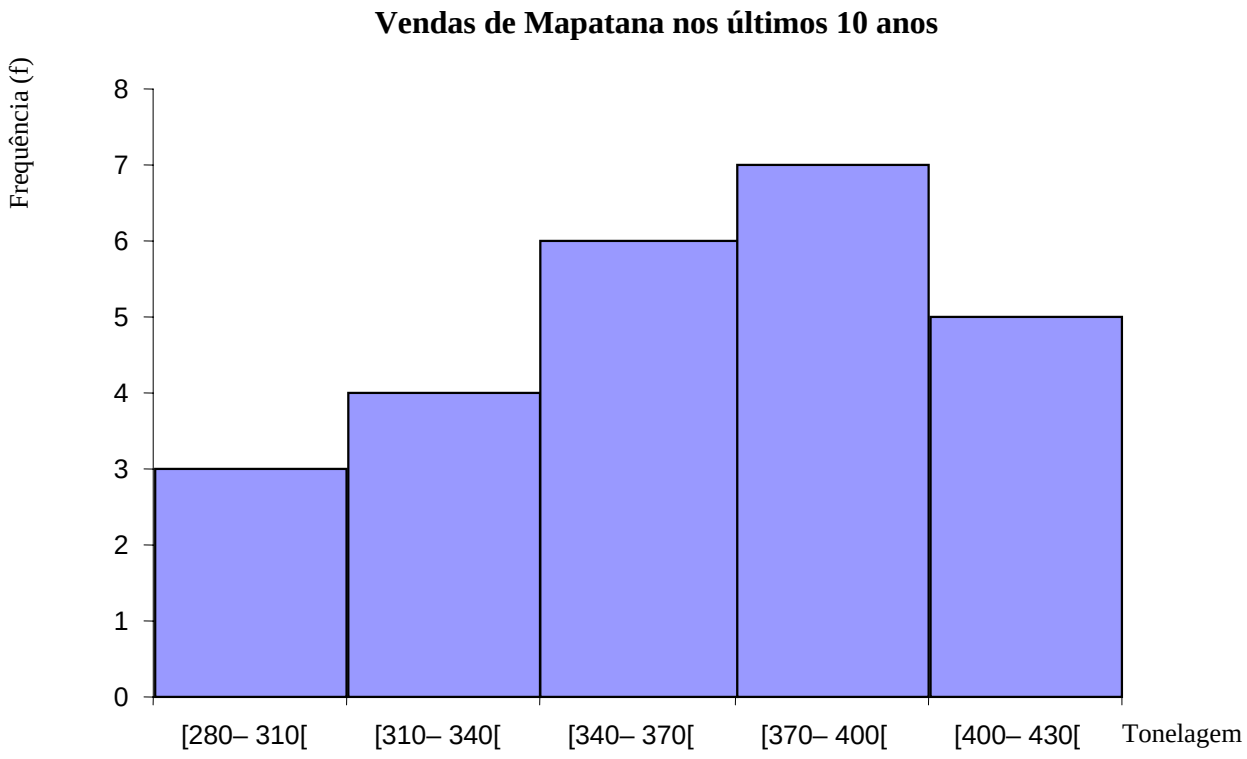


Fig 2.15

PRINCIPAIS ESTATÍSTICAS: DEFINIÇÃO E OPERACIONALIZAÇÃO

2. MEDIDAS DE TENDÊNCIA CENTRAL

Definição: Indicam onde se concentram a maioria dos dados.

Certamente que, se porventura existirem mais dados afastados das medidas de Tendência Central, os seus resultados teóricos(1) estarão mais afastados do valor central da respectiva distribuição.

(1) são os que resultam dum cálculo por meio duma fórmula.

2.1 Média

Definição: é o Centro de Gravidade do conjunto de dados. Ela é definida como a soma dos produtos dos valores da variável e respectiva frequência dividida, pelo número de observações

Dados não agrupados

Média amostral Média populacional

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \quad (V) \qquad \mu = \frac{\sum_{i=1}^{N} X_i}{N} \quad (VI)$$

A fórmula (V), pode ser usada para dados da tabela 2.13 se considerarmos a venda por ano como sendo X_i.
Dados agrupados

Média amostral Média populacional

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i f_i}{n} \quad (VII) \qquad \mu = \frac{\sum_{i=1}^{N} X_i f_i}{N} \quad (VIII)$$

Para os dados constantes na tabela 2.13, usamos a fórmula VII, e então a média seria

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i f_i}{n} = \frac{280\times2+305\times1+320\times1+330\times1+310\times2+340\times2+369\times1+365\times1+375\times1+380\times1+400\times2+371\times1+390\times1+370\times1}{18} = 346{,}39$$

Assim ficamos com cálculo longo, preferindo com que se utilize os dados da tabela 4.17 para o cálculo da mesma média. Sendo assim, pode-se ver que 9085,00 é o total da coluna 5 e 18 é total da coluna 3. Como a referida tabela possui classes, a média a usar só pode ser aquela que se refere a dados agrupados. Se tivéssemos usado a fórmula V, a média seria 346,4 .

Essa diferença não é relevante em virtude da fórmula **V** ser usada para dados não agrupados A 363,4 foi mais ocasionada pela correcção dos dados, que foram registados duma forma discreta para a notação intervalar (contínua).

A média e os valores extremos

A média apresenta um problema crítico ou mesmo bicudo; ela é fortemente influenciada pelos valores extremos. Por esta razão deve-se fazer uma análise cautelosa dos dados.

Exemplo:
Suponha que estejamos interessados em estudar a distribuição do rendimento médio de nove famílias, em número de salários mínimos (de 1.000.000,00Mt), com os seguintes valores:

X: Número de salários mínimos
X: 1, 1, 1, 1, 2, 2, 3, 5, 20

O rendimento médio dessas nove famílias é quatro. Mas o que acontece se a família com renda igual a 20 salários mínimos fosse retirada da amostra? O valor da média cai para dois salários mínimos, o que parece mais razoável, já que esse valor descreve melhor este conjunto de dados.

Vejamos a distribuição de renda das nove famílias da amostra diagramaticamente na tabela a seguir:

Tabela 2.15 – Distribuição de Rendas de nove famílias

*																			Valor Extremo
*																			
*	*																		
*	*	*		*															*
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Este exemplo ilustra como a média é vulnerável ao efeito de valores extremos. Neste caso é recomendado utilizar a mediana, por ser o valor mais ao centro da distribuição, não susceptível de influências.

Mas será que sempre usaremos a mediana se isso acontecer? Será que a mediana é a única medida fiável e nunca mais se usa a média?

Pode-se ver mais adiante que no tratamento das medidas de dispersão (dispersão relativa ou total) faz-se sempre em relação à média, excepto a amplitude total, amplitude interquartil e a amplitude de classe. As escolas, as empresas e sectores sociais usam em geral a média para emitir um juízo. A mediana só nos diz a posição do(s) elemento(s) central(is).

Média aritmética ponderada: a média aritmética ponderada do conjunto

$x_1, x_2,\ldots,x_k$, com pesos $w_1, w_2,\ldots,w_k$ é calculada por $\overline{X}_{ap} = \frac{x_1 w_1 + x_2 w_2 + \ldots + x_k w_k}{w_1 + w_2 + \ldots + w_k}$.

a) **Média geométrica**: a média geométrica dos valores positivos $x_1, x_2,\ldots,x_n$, é calculada por $\overline{X}_g = \sqrt[n]{x_1 \times x_2 \times \ldots \times x_n}$.

Média geométrica ponderada: a média geométrica ponderada do conjunto

$x_1, x_2,\ldots,x_k$, com pesos $w_1, w_2,\ldots,w_k$, é calculada por $\overline{X}_{gp} = \sum \sqrt[wi]{x_1^{w_1}.x_2^{w_2}\ldots x_k^{w_k}}$.

b) **Média Harmônica**: a média harmônica dos valores x_1, x_2,...,x_n é calculada por

É o inverso da média aritmética dos inversos.

$$\overline{X}_h = \frac{1}{\frac{\frac{1}{x_1}+\frac{1}{x_2}+\ldots+\frac{1}{x_n}}{2}} = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\ldots+\frac{1}{x_n}} = \frac{n}{\sum\frac{1}{x_i}}.$$

Exemplo - Calcular a média harmônica simples dos seguintes conjuntos de números:

a) { 10, 60, 360 } Resp: 3/(1/10+1/60+1/360) = 25,12

b) { 2, 2, 2, 2 } Resp: 4/(1/2+1/2+1/2+1/2) = 2

Média harmônica ponderada: a média harmônica ponderada do conjunto

x_1, x_2, ..., x_k, com pesos w_1, w_2, ...,w_k, é calculada por $\overline{X}_{hp} = \frac{\sum w_i}{\sum \frac{w_i}{x_i}}$.

Media Harmônica Ponderada : (para dados agrupados em tabelas de frequências)

$$\overline{X}_{hp} = \frac{\sum f_i}{\sum \frac{f_i}{x_i}}$$

Exemplo - Calcular a média harmônica dos valores da tabela abaixo:

classes	fi	xi	fi/xi
1 |--------- 3	2	2	2/2 = 1,00
3 |--------- 5	4	4	4/4 = 1,00
5 |--------- 7	8	6	8/6 = 1,33
7 |--------- 9	4	8	4/8 = 0,50
9 |--------- 11	2	10	2/10 = 0,20
total	**20**		**4,03**

Resp: 20 / 4,03 = 4,96

Propriedades da média harmônica

A média harmônica é menor que a média geométrica para valores da variável diferentes de zero. $\overline{X}h < \overline{X}g$ e por extensão de raciocínio podemos escrever : $\overline{X}h < \overline{X}g < \overline{X}$

OBS1: A média harmônica não aceita valores iguais a zero como dados de uma série.

A igualdade $\overline{X}g = \overline{X}h = \overline{X}$ só ocorrerá quando todos os valores da série forem iguais.

OBS2: Quando os valores da variável não forem muito diferentes, verifica-se aproximadamente a seguinte relação:

$$\overline{X}g = (\overline{X} + \overline{X}h) / 2$$

Demonstraremos a relação acima com os seguintes dados:

z = { 10,1 ; 10,1 ; 10,2 ; 10,4 ; 10,5 }

Média aritmética = 51,3 / 5 = 10,2600

Média geométrica = 10,2587

Média harmônica = 5 / 0,4874508 = 10,2574

Comprovando a relação: 10,2600 + 10,2574 / 2 = 10,2587 = média geométrica

c) **Média quadrática**: a média quadrática dos valores positivos $x_1, x_2,\ldots,x_n$, é calculada por $\overline{X}_q = \frac{x_1^2 + x_2^2 + \ldots + x_n^2}{n} = \frac{\sum x_i^2}{n}$. É a raiz quadrada da média aritmética dos quadrados

Média Quadrática Simples: (para dados não agrupados) $\overline{X}_q = \sqrt{\frac{x_1^2 + x_2^2 + ... + x_n^2}{n}}$

Exemplo - Calcular a média quadrática simples do seguinte conjunto de números:

a = { 2 , 3 , 4 , 5 } Resp: 3,67

Média Quadrática Ponderada: Quando os valores da variável estiverem dispostos em uma tabela de frequências, a média quadrática será determinada pela seguinte expressão: $\overline{X}_{qp} = \sqrt{\frac{\sum x_i^2 f_i}{\sum f_i}}$

Exemplo - Calcular a média quadrática dos valores da tabela abaixo:

classes	**fi**	**xi**	xi^2	xi^2 **. fi**
2 |--------- 4	5	3	9	45
4 |--------- 6	10	5	25	250
6 |--------- 8	12	7	49	588
8 |--------- 10	10	9	81	810
10 |-------- 12	5	11	121	605
total	**42**			**2298**

No excel =(2298/42)^(1/2) Resp: 7,40

ou aplica-se a raiz quadrada sobre 2298/42

OBS:

- Sempre que os valores de **X** forem positivos e pelo menos um dado diferente é válida a seguinte relação: $\overline{X}q > \overline{X} > \overline{X}g > \overline{X}h$
- A igualdade entre as médias acima se verifica quando os valores da variável forem iguais (constantes)
- A média quadrática é largamente utilizada em Estatística, principalmente quando se pretende calcular a média de desvios (x - $\overline{x}$) , em vez de a média dos valores originais. Neste caso, a média quadrática é denominada **desvio-padrão**, que é uma importante medida de dispersão.

2.2 Mediana

Mediana (Me): Divide o conjunto de dados em dois subconjuntos com o mesmo número de elementos, abaixo dela fica metade dos dados (50%) e acima a outra metade (50%).

Tomando os dados do exemplo anterior, relativos à distribuição do rendimento mensal, teremos:

Tabela 2.16 – Determinação da Mediana

Lugar / Posição	1º	2º	3º	4º	5º	6º	7º	8º	9º
Variável	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
Valores da Variável	1	1	1	1	2	2	3	5	20

50% dos dados = 4 dados abaixo do valor da mediana

Mediana=2

50% dos dados = 4 dados acima do valor da mediana

Observe que a mediana é independente dos valores extremos, porque ela só leva em consideração os valores de posição central.

Passos para determinar a mediana:

Quando o número de dados é impar

Colocar os dados em rol – ordená-los na forma ascendente (pode ser também na forma descendente, mas não é comum e pode atrapalhar na hora de calcular as medidas de posição)

O lugar ou posição que a mediana ocupa é: $\frac{n+1}{2}$, onde n é o número de elementos, e

O valor da mediana é o valor da variável que ocupa o lugar $\frac{n+1}{2}$. $Me = X_{\left(\frac{n+1}{2}\right)}$

Pelo exemplo: para $n = 9$, $\frac{n+1}{2} = 5$, logo o valor da mediana seria: $M_e = X_5 = 2$

Se n fosse igual a 21, então o valor da mediana seria: $M_e = X_{11}$

Se n fosse igual a 49, então o valor da mediana seria: $M_e = X_{25}$

Quando o número de dados é par

Ordenar os dados em ordem ascendente

O lugar ou posição que a mediana ocupa está entre: $\frac{n}{2}$ e $\left(\frac{n}{2}\right)+1$, e

O valor da mediana será a média simples dos valores que ocupam esses lugares

$$\text{Me} = \frac{X_{\frac{n}{2}} + X_{(\frac{n}{2}+1)}}{2}$$

Para dados agrupados em intervalos de classe da mesma amplitude, a mediana é dada por

$$Me = l_1 + \frac{\frac{N}{2} - \sum f_1}{f_{med}} \times c$$

Onde l_1 - limite inferior da classe mediana

$\frac{N}{2}$ - ponto médio onde se localiza a respectiva frequência acumulada

$\sum f_1$ - soma das frequências inferiores (anteriores) a classe mediana

c - intervalo de classe dado por $c = \frac{\lambda}{k}$

f_{med} - frequência da classe mediana

2.3 Moda

Moda (Mo): é o valor que ocorre (se repete) com maior frequência. A moda pode não existir, bem como pode ter mais de um valor, principalmente quando a variável toma muitos valores. No exemplo cujos dados estão representados na tabla da página anterior, a moda é igual a 1.

Para dados agrupados em intervalos de classe da mesma amplitude, a moda é dada por:

$$M_o = l_1 + \frac{d_1}{d_1 + d_2} \times c$$

Onde l_1 - limite inferior da classe modal

c - intervalo de classe dado por $c = \frac{\lambda}{k}$

d_1 - Diferença entre as frequências da classe modal e a imediatamente inferior

d_2 - Diferença entre as frequências da classe modal e a imediatamente superior

Vejamos os seguintes dados agrupados (em classes), na tabela a seguir e vamos calcular a mediana e moda

Tabela 2.17 – Tabulação de frequências para determinar as classes modal e mediana

Classes (Tonelagem)	x_i	f_i	f_r	F_i	F_r	
[280– 310[	295	3	0.12	3	3/25	
[310– 340[	325	4	0.16	7	7/25	
[340– 370[	355	6	0.24	13	13/25	Classe Mediana
[370– 400[	385	7	0.28	20	20/25	Classe Modal
[400– 430[	415	5	0.20	25	25/25	
Total	------	25	1.00			

Primeiro calculamos $\frac{N}{2}$, $\frac{N}{2}=\frac{25}{2}=12{,}5\approx 13$. Este valor pertence à classe onde $F_i=13$; sendo assim, a classe com $F_i=13$ é denominada classe mediana. Calculamos então a respectiva mediana

$$Me=l_1+\frac{\frac{N}{2}-\sum f_1}{f_{med}}\times c=340+\frac{12{,}5-7}{6}\times 30=340+27{,}5=367{,}5$$

A maior frequência é 7; sendo assim, a classe que possui a maior frequência é considerada a classe modal

$$M_o=l_1+\frac{d_1}{d_1+d_2}\times c=340+\frac{7-6}{(7-6)+(7-5)}\times 30=340+\frac{1}{3}\times 30=340+10=350$$

Observação: Um dos casos não tratados na moda é o referente a intervalos de classes diferentes. Para este caso usa-se a fórmula de King, com a seguinte forma

$m=l_i+\frac{\frac{f_{i+1}}{c_{i+1}}}{\frac{f_{i+1}}{c_{i+1}}+\frac{f_{i-1}}{c_{i-1}}}\times c_i$, com c_i como intervalo da classe modal e f_i freqüência absoluta da classe modal. A classe modal é aquela com maior peso. O peso é obtido pela fórmula $\frac{f_i}{c_i}$.

Remetemos a investigação do leitor para mais esclarecimentos embora haja um exercício do tema sobre "*Correlação e regressão*", onde numa das alíneas se mostra a utilidade dessa fórmula

Tabela 2.18 – Resumo da definição de principais medidas de tendência central

Medida	Notação	Definição, propriedades
Média	$\bar{X}$	É a soma dos produtos dos valores da variável e respectiva frequência, dividida pelo número de observações
Mediana	Me	É o valor que ocupa a posição central da série de observações de uma variável, dividindo o conjunto em duas partes iguais. 50% dos dados tomam valores menores ou iguais ao valor da mediana e os 50% restantes acima.
Moda	Mo	È definida como valor que ocorre com mais frequência, dos valores observados

3. MEDIDAS DE POSIÇÃO

Assim como as medidas de tendência central têm objetivo de fornecer indicadores do local onde a maioria dos dados se concentram, as medidas de posição tem de objetivo de indicar onde é o ponto de corte para uma certa posição. As medidas mais usadas são os quartis e suas versões mais gerais, decís e percentis.

3.1 Quartís (Quantís)

Quartil: Como atrás tratado nas medidas de tendência central, a mediana divide um conjunto de dados em duas partes iguais, os quartis dividem-no em quatro partes iguais. Assim, usando o exemplo das alturas dos 56 alunos duma Escola (Tabela 2.10), teremos :

Vejamos o que acontece com os quartís:
$n=56 \Rightarrow n/4=56/4=14$

Para se determinar a posição quartil sem agrupar os dados é necessário coloca-los em rol (ordem crescente ou descrescente), só após essa ordenação é que se pode determinar os quartís. Usando os dados da tabela Tabela 2.30, teremos:

Tabela 2.19 – Posição quartil

25%	25%	25%	25%
14	14	14	14
14	28	42	56

⇧ ⇧ ⇧

$Q_1=0,75*X_{14} + 0,25*X_{15}$
$Q_1=0,75*1,67 + 0,25*1,69$
$Q_1=1,675$

$Me=0,5*X_{28} + 0,5*X_{29}$
$Me=0,5*1,73+1,5*1,73$
$Me=1,73$

$Q_3=0,25*X_{42} + 0,75*X_{43}$
$Q_3=0,25*1,78 + 0,75*1,78$
$Q_3=1,78$

Exemplo:

1- Faça os correspondentes histogramas para os percentís, decís e quartís apresentados

2- Retomando o exemplo das alturas dos 56 alunos e feitos os cálculos dos respectivos Quartís tem-se:

Tabela 2.20 – Resumo de quartís para Box Plot

Estatísticas	Altura dos Alunos
Q_1	1,675
Q_3	1,78
$Q_3 - Q_1$	0,105
$1,5 *(Q_3 - Q_1)$	0,1575
$3,0 *(Q_3 - Q_1)$	0,315
Outliers inferiores $X_i \leq Q_1 - 1,5 *(Q_3 - Q_1)$	$\leq 1,5175$
Valores extremos inferiores $X_i \leq Q_1 - 3,0 *(Q_3 - Q_1)$	$\leq 1,36$
Outliers superiores $X_i \geq Q_3 + 1,5 *(Q_3 - Q_1)$	$\geq$ **
Valores extremos superiores $X_i \geq Q_3 + 3,0 *(Q_3 - Q_1)$	$\geq$ **

** valores menores que 1,55 ou maiores que 1,89, por tanto impossíveis.
Apresente o diagrama de Box – Plot para as alturas desses alunos

Tabela 2.21 – Resumo da definição de quartís

Estatística	Notação	Definição, propriedades
1º quartil	Q_1	É o valor que ocupa a posição tal que um quarto dos dados (25%) tomam valores menores ou iguais ao valor do primeiro quartil.
2º quartil (Mediana)	Q_2 Me	Coincide com o valor da mediana, ou seja 50% dos dados tomam valores menores ou iguais aos da mediana. Entre o primeiro quartil (Q_1) e a mediana (Me) ficam 25% dos dados.
3º quartil	Q_3	É o valor que ocupa a posição tal que um quarto dos dados (25%) tomam valores maiores ou iguais ao valor do terceiro quartil. Entre a mediana (Me) e o terceiro quartil (Q_3) ficam 25%

4.2 Percentís

De todos os percentis os mais importantes são:

Tabela 2.22 – Principais percentís

Percentil	Notação	Definição, propriedades
1º	P_1	1% dos dados tomam valores menores ou iguais
5º	P_5	5% dos dados tomam valores menores ou iguais
10º	P_{10}	10% dos dados tomam valores menores ou iguais
25º	P_{25}	25% dos dados tomam valores menores ou iguais (Q_1)
50º	P_{50}	50% dos dados tomam valores menores ou iguais (Q_2 = Me)
75º	P_{75}	25% dos dados tomam valores maiores ou iguais (Q_3)
90º	P_{90}	10% dos dados tomam valores maiores ou iguais
95º	P_{95}	5% dos dados tomam valores maiores ou iguais
99º	P_{99}	1% dos dados tomam valores maiores ou iguais

Em geral somos solicitados para calcular percentil: 1, 5, 10, 25, 50, 75, 90, 95 e 99.

Exemplo:

Percentil i	1	5	10	25	50	75	90	95	99
Posição $(\frac{in}{100})$	1	3	6	15	30	45	54	57	59
Valores de p_i	1,55	1.58	1.63	1.67	1.73	1.78	1.82	1.85	1.88

Os valores dos percentís aqui apresentados, representam o enquadramento das alturas da turma que começam de 1,55 e terminam em 1,89.

4.3 Decís

De todos os decís os mais importantes são:

Tabela 2.23 – Principais Decís

Decíl	Notação	Definição, propriedades
1º	D_1	10% dos dados tomam valores menores ou iguais
2º	D_2	20% dos dados tomam valores menores ou iguais
5º	D_5	50% dos dados tomam valores menores ou iguais
9º	D_9	90% dos dados tomam valores menores ou iguais

Tabela 2.24 – Determinação da Posição decil

Decil i	1	2	5	9
Posição $(\frac{in}{10})$	6	11	28	50
D_i	1,63	1,67	1,80	1,82

Resumo das medidas de Posição (separatrizes)

Quartis: dividem a série em 4 partes iguais

Q_1 = 1º quartil, deixa 25% dos elementos

1º) Calcular a posição: $posição = \frac{n}{4}$ (seja n ímpar ou par)

2º) Pela F_i identifica-se a classe que contém o Q_1

3º) Aplica-se a fórmula: $Q_i = L_{Q_1} + \frac{\frac{n}{4} - F_i}{f_{Q_1}} \times c$

sendo
* L_{Q1} = limite inferior da classe do Q_1
* n = tamanho da amostra ou nº de elementos
* F_i = frequência acum. anterior à classe do Q_1
* c = intervalo da classe do Q_1
* f_{Q1} = frequência simples da classe do Q_1

Q_2 = 2º quartil, é igual a mediana, deixa 50% dos elementos

Q_3 = 3º quartil, deixa 75% dos elementos

1º) Calcular a posição: $posição = \frac{3n}{4}$ (seja n ímpar ou par)

2º) Pela F_i identifica-se a classe que contém do Q_3

3º) Aplica-se a fórmula: $Q_3 = L_{Q_3} + \frac{\frac{3n}{4} - F_i}{f_{Q_3}} \times c$

sendo
* L_{Q3} = limite inferior da classe do Q_3
* n = tamanho da amostra ou nº de elementos
* F_i = frequência acum. anterior à classe do Q_3

* c = intervalo da classe do Q_3
* f_{Q3} = frequência simples da classe do Q_3

Decis: dividem a série em 10 partes iguais

1º) Calcular a posição: $posição = \frac{in}{10}$ (seja n ímpar ou par), onde i = 1, 2, 3, 4, 5, 6, 7, 8 e 9

2º) Pela F_i identifica-se a classe que contém o D_i

3º) Aplica-se a fórmula: $D_i = L_{D_1} + \frac{\frac{in}{10} - F_i}{f_{D_1}} \times c$

sendo
* L_{Di} = limite inferior da classe D_i , i = 1, 2, 3, ..., 9
* n = tamanho da amostra ou nº de elementos
* F_i = frequência acum. anterior à classe do D_i
* c = intervalo da classe do D_i
* f_{Di} = frequência simples da classe do D_i

Percentis: dividem a série em 100 partes iguais

1º) Calcular a posição: $posição = \frac{in}{100}$ (seja n ímpar ou par), em que i = 1, 2, 3, ..., 98, 99

2º) Pela F_i identifica-se a classe que contém o P_i

3º) Aplica-se a fórmula: $P_i = L_{P_1} + \frac{\frac{in}{100} - F_i}{f_{P_1}} \times c$

sendo
* L_{Pi} = limite inferior da classe P_i , i = 1, 2, 3, ..., 99
* n = tamanho da amostra ou nº de elementos

* F_i = frequência acum. anterior à classe do P_i
* c = intervalo da classe do P_i
* f_{Pi} = frequência simples da classe do P_i

Vimos que o resumo de dados por meio de tabelas de frequências, diagramas (de Barras, ramo e folhas, histogramas, circular e Box-Plot) fornecem muito mais informações sobre o comportamento de uma variável do que a própria tabela ou a organização original de dados. Muitas vezes é ainda mais cómodo resumir ainda mais esses dados, apresentando resultados que resumem toda a conjuntura de dados. É evidente, que um único número (de princípio real) que possa resumir todos os dados, pode nalguns casos fornecer informação relevante, casos das medidas de posição (localização/ tendência central) ou mesmo as de dispersão.

Resumo: Medidas de posição - quartis, percentis e decís. Os quartis dividem o conjunto de dados em quatro partes iguais, os percentis em cem partes iguais e os decís em dez partes iguais.

5. MEDIDAS DE DISPERSÃO

Medem o grau de variabilidade ou dispersão dos dados relativamente ao centro de distribuição.

5.1 Amplitude total (λ)

Amplitude (λ): mede a distância entre o valor máximo e o valor mínimo; ela é uma estatística rudimentar, porque embora ofereça uma noção de dispersão ela não diz qual é sua natureza. A amplitude interquartil (I_Q), ou comprimento da caixa (Num diagrama de Box Plot), é a distância entre o primeiro e terceiro quartil, é muito útil para detectar valores extremos.

$$\text{Amplitude} = \lambda = X_{\text{máximo}} - X_{\text{mínimo}}$$

A *amplitude*, é medida de dispersão mais simples por provir da diferença entre o maior e o menor valor nos dados. Para uma distribuição de frequências que usa intervalos de classe, a amplitude pode ser considerada como a diferença entre o maior e o menor limite de classe ou a diferença entre os pontos médios dos intervalos de classe extremos. Os preços de acções e de outros activos financeiros são frequentemente descritos em termos de sua amplitude, com a apresentação pelas Bolsas de Valores do maior valor e do menor valor da acção num determinado período de tempo.

Para algumas distribuições simétricas a média pode ser aproximada tomando-se a semi-soma dos dois valores extremos,[1] que é frequentemente chamada de semi-amplitude. Por

[1] Foi o que foi feito ao calcular a média para valores agrupados em classes de frequência. Nesse caso utilizou-se o ponto médio de cada intervalo de classe como representativo da média de cada intervalo. Assim, ao multiplicarmos a frequência de cada classe pelo valor do ponto médio, estamos calculando aproximadamente a soma das observações em cada intervalo, admitindo como hipótese que a distribuição dos dados em todos os intervalos é simétrica.

exemplo, é prática entre os meteorologistas derivar a média diária de temperatura tomando a média somente dos valores máximo e mínimo de temperatura ao invés, de digamos, a média das 24 leituras horárias do dia.

A amplitude tem alguns defeitos sérios. Ela pode ser influenciada por um valor atípico na amostra. Além disso o seu valor é independente do que ocorre no interior da distribuição, já que somente depende dos valores extremos. Este defeito é ilustrado na figura a seguir:

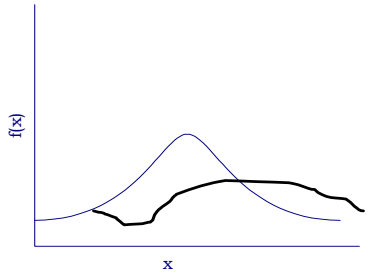


Fig 5.1

Na figura acima são mostradas duas distribuições com diferentes variabilidade, mas com mesma amplitude. A amplitude tende a crescer, embora não proporcionalmente à medida que o tamanho da amostra cresce. Por esta razão, não podemos interpretar a amplitude correctamente sem conhecermos o número de dados ou informações.

5.2 Desvio médio

O **Desvio Médio (DM)** é a média dos valores absolutos dos desvios e a variância (S^2) é a média dos quadrados dos desvios.

Ao calcular a variância elevou-se ao quadrado cada desvio, isto é, os desvios foram aumentados. Para retirar esse efeito, deve-se extrair a raiz quadrada da variância, dando origem ao desvio padrão (S).
Tomemos de novo o nosso exemplo para calcular desvio médio, a variância e o desvio padrão:

Tabela 2.25 – Determinação dos desvios médio e padrão

Nome	Teste 1	Teste 2	Teste 3	Média ($\overline{X}$)	$\sum(X-\overline{X})^2$	S^2	DM	S
Matorwa Ndapota	13,0	13,0	13,0	13,0	----------	-------	-----	------
Desvio ($X-\overline{X}$)	0,0	0,0	0,0	---------	0,0	0,000	-----	0,000

$\lvert X - \overline{X} \rvert$	0,0	0,0	0,0	--------	----------	-------	0,0	------
Nhamadzi Chingozi	14,0	12,0	13,0	13,0	----------	-------	-----	------
Desvio ($X - \overline{X}$)	1,0	-1,0	0,0	--------	2,0	0,667	------	0.816
$\lvert X - \overline{X} \rvert$	1,0	1,0	0,0	--------	----------	--------	0,667	------
Creva Vanduzi	18,0	12,0	9,0	13,0	----------	--------	------	------
Desvio ($X - \overline{X}$)	5,0	-1,0	-4,0	---------	42,0	14,000	------	3,742
$\lvert X - \overline{X} \rvert$	5,0	1,0	4,0	---------	----------	--------	3,333	------

Observa-se que o desvio padrão é sempre maior ou igual ao desvio médio, e isto devido ao facto de ter elevado ao quadrado cada desvio, aumentando desproporcionalmente o peso dos valores extremos. Lembrar que o facto de se ter extraído a raiz quadrada da variância não elimina completamente o efeito de se ter elevado ao quadrado cada desvio, uma vez que a raiz quadrada de uma soma não é igual à soma da raiz quadrada de cada parcela.

5.3 Variância e desvio padrão (S)

Construindo o desvio padrão:

Dada seguinte tabela, que representa o desempenho de três estudantes duma escola

Tabela 2.26 – Comparação dos desvios médio e padrão

Nome	Teste 1	Teste 2	Teste 3	Média
Matorwa Ndapota	13,0	13,0	13,0	13,0
Nhamadzi Chingozi	14,0	12,0	13,0	13,0
Creva Vanduzi	18,0	12,0	9,0	13,0

Qual dos três é mais regular?
Neste caso, a média não sugere diferença, sendo possível encontrar a diferença se se comparar os desvios padrão.

Para entender a construção do desvio padrão deve-se, primeiro, analisar a natureza dos desvios dos valores da variável em relação à sua própria média. No exemplo (sugerido acima), cujas médias dos três alunos são iguais, mas seus desempenhos diferentes, deve-se analisar os desvios para se ter a certeza de que os três alunos têm desempenhos diferentes. Pode-se ver que o aluno Matorwa é constante no seu desempenho, Nhamadzi vai progredindo aos poucos e o Creva tem uma queda abrupta no seu desempenho e não consegue se recuperar. Ou seja, apesar dos três alunos terem o mesmo desempenho médio, eles tem variabilidades diferentes.

Vejamos agora os desvios dos valores da variável em relação à média.

Tabela 2.27 – Desvio padrão de notas de três estudantes

Nome	Teste 1	Teste 2	Teste 3	Média ($\overline{X}$)	$\sum (X - \overline{X})^2$	S
Matorwa Ndapota	13,0	13,0	13,0	13,0	----------	----------
Desvio ($X - \overline{X}$)	0,0	0,0	0,0	----------	0,0	0,000

Nhamadzi Chingozi	*14,0*	*12,0*	*13,0*	*13,0*	----------	----------
Desvio ($X - \overline{X}$)	14−13=1	*-1,0*	*0,0*	----------	*2,0*	*0.816*
Creva Vanduzi	18,0	12,0	9,0	13,0	----------	
Desvio ($X - \overline{X}$)	5,0	-1,0	-4,0	----------	42,0	3,742

Vejamos o **DM** para Nhamadzi: $\mathrm{DM} = \frac{\sum_{i=1}^{3}\left|X_i - \overline{X}\right|}{n} = \frac{1-1+0}{3} = 0$

Poder-se-ia se pensar em construir um desvio médio, como sendo a soma dos desvios dividida pelo número de observações, porém, a soma dos desvios é igual a zero. Então, como construir uma medida de dispersão? Como o problema é a compensação dos valores positivos com os negativos, a pergunta é: como converter os valores negativos em positivos? De duas maneiras: tomando valor absoluto (distância) ou elevando ao quadrado cada desvio. Assim têm-se o desvio médio e a variância.

Desvio médio

$$\mathrm{DM} = \frac{\sum_{i=1}^{n}\left|X_i - \overline{X}\right|}{n}$$

Variância

$$S^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n}$$

Desvio padrão

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

Tabela 2.28 – Comparação dos desvios médio e padrão

Nome	Média ($\overline{X}$)	Mediana	Moda	λ	S^2	DM	S
Matorwa Ndapota	13,0	13,0	13	0,0	0,000	0,0	0,000
Nhamadzi Chingozi	13,0	13,0	14,12,13	2,0	0,667	0,667	0.816
Creva Vanduzi	13,0	12,0	18,12,9	9,0	14,000	3,333	3,742

Logo, conclui-se que apesar dos três alunos terem a mesma nota média, seus desempenhos tem diferentes graus de variabilidade, sendo que o aluno Matorwa Ndapota tem um desempenho perfeitamente homogéneo enquanto que o aluno Creva Vanduzi tem o mais disperso.

Observe-se que quanto mais disperso é o conjunto de dados maior será o desvio padrão, desvio médio e amplitude.

Entretanto, às vezes pode-se querer comparar o grau de dispersão de dois conjuntos de dados com unidades de medidas diferentes. Neste caso, deve-se usar o coeficiente de variação (C_v), que é uma medida de dispersão relativa, uma vez que não está afectada pelas unidades da medida da variável.

O coeficiente de variação (C_v) calcula-se pela fórmula: $C_v = \frac{S}{\overline{X}} * 100$ onde S é desvio e $\overline{X}$ a média.

Exemplo:

Tomemos casos em que tenha a média de rendimentos familiares de três países (com sistemas monetários diferentes) e respectivos desvios. Como é que poderíamos comparar e saber em que país a distribuição dos rendimentos é mais homogénea?

Tabela 2.29 – Rendimentos familiares de três famílas de três países

País	Moeda	Média	Desvio Padrão	Coeficiente De variação
A	Euro	5.000	1.000	20%
B	Dólar	10.000	1.000	10%
C	Metical	2.000	1.000	50%

Neste exemplo,os rendimentos familiares é que são iguais, isso não implica que elas tenham a mesma distribuição de renda. A renda é mais homogênea no país B em virtude da renda não possuir maior variação, determinada pelo coeficiente de variação.

Observações:
Tanto a média como o desvio padrão podem não ser medidas adequadas para representar um conjunto de dados, pois:

- São afectados, de forma exagerada, por valores extremos;
- Apenas com estes dois valores não temos ideia da simetria ou assimetria da distribuição dos dados.

Teorema de Chebyshev (para o desvio padrão)

Dado um número k maior do que um, então pelo menos $1-\frac{1}{k^2}$ dos valores de uma amostra ou população pertencerão ao intervalo de k desvios padrão antes e k desvios padrão além da média. Este intervalo tem extremos $m-k\times s$ e $m+k\times s$.

Exemplo:
Na tabela a seguir estão apresentados 56 valores de cada uma das seis variáveis que representam informações sobre alunos do sexo masculino fazendo graduação em Estatística, num certo ano.
Determine a média e variância
Faça o diagrama de ramo e folhas
c) Use o teorema de Tchebyshev para determinar os intervalos da percentagem de alturas para $k=2$ e para $k=3$.

Tabela 2.30 – Listagem de dados de alunos

Nº do aluno	Nº de irmãos	Altura	Peso	Idade	Origem	Grau Acadêmico do Pai
1	2	1,71	70,9	18	Zimpeto	EP2
2	3	1,72	76,2	20	Mahlazine	EP2
3	2	1,69	72,6	18	Polana	Superior
4	1	1,69	60,0	22	Matola C	EP2
5	3	1,77	71,3	19	Hulene	EP2
6	0	1,55	53,6	19	Mandimba	EP2
7	0	1,66	65,8	20	Nhamudima	EP2
8	5	1,63	65,0	19	Canongole	EP2
9	3	1,73	87,8	19	Chingodzi	Superior
10	5	1,70	73,8	22	Mapfunde	Superior
11	4	1,82	81,3	20	Manyati	EP2
12	3	1,73	72,2	19	Vumba	Superior
13	2	1,80	74,7	24	Chirambandine	EP2
14	3	1,77	73,4	19	Cacarue	EP2
15	2	1,73	69,1	21	Penhalonga	EP2
16	3	1,71	98,1	21	Messica	EP2
17	2	1,74	71,2	18	Gôndola	Superior
18	2	1,71	67,3	19	Macuti	EP2
19	3	1,74	69,0	21	Munhava	Superior
20	3	1,71	79,7	18	Nhamussue	EP2
21	2	1,88	85,7	18	Brandão	EP2
22	3	1,76	83,4	19	Matola 700	Superior
23	2	1,62	64,0	20	Namaacha	Superior
24	1	1,67	72,1	23	Mukuananda	Superior
25	3	1,64	63,5	19	Andrade	Superior
26	2	1,77	69,2	19	Manjacaze	EP1
27	2	1,73	76,8	23	Benfica	Superior
28	1	1,80	91,2	20	Chopal	EP2
29	2	1,73	64,8	21	Chidenguele	Nenhum
30	2	1,66	68,2	19	Tsangano	Superior
31	2	1,79	82,5	20	Pontagea	Superior
32	3	1,80	105,7	20	Matola Gare	EP1
33	3	1,63	61,8	21	Zimpeto	EP2
34	2	1,77	79,4	20	Matendeni	EP2
35	1	1,86	87,2	19	Vaz	Superior
36	0	1,66	59,9	25	Albasini	EP2
37	1	1,82	82,2	20	CMC	EP2
38	6	1,85	79,2	21	Aeroporto	EP2

39	2	1,69	69,4	22	Manga Mascarenha	Superior
40	3	1,58	62,0	22	Zonas Verdes	EP1
41	3	1,77	80,6	18	Soalpo	Superior
42	0	1,76	70,4	19	Nhamadjessa	Superior
43	4	1,67	65,9	18	Montalto	Superior
44	4	1,55	74,9	21	Inharrime	EP1
45	1	1,80	83,4	18	Morrumbene	EP2
46	2	1,71	77,4	18	Cacarwe	Superior
47	3	1,78	78,6	19	Chinhamacungo	Superior
48	2	1,70	78,6	24	Catembe	EP2
49	1	1,75	81,9	22	Polana Caniço	EP2
50	3	1,75	74,0	21	Maxaquene	EP2
51	1	1,81	77,2	23	Nhamatanda	Superior
52	4	1,71	70,0	22	Alto Maé	EP2
53	2	1,74	79,0	18	Vanduzi	Superior
54	1	1,78	83,4	21	Mavonde	EP2
55	5	1,89	92,2	21	Chicamba	Superior
56	2	1,82	94,6	20	Homoíne	EP2

Resolução:

a) $\lambda = 1{,}89 - 1{,}55 = 0{,}34 \qquad k = \sqrt{56} = 7{,}48 \approx 7{,}5 \approx 8 \qquad c = \frac{\lambda}{k} = \frac{0{,}34}{8} = 0{,}0425$

Tabela 2.31 – Distribuição de frequências para determinação da média e variância

Classes	X_i	f_i	F_i	$X_i f_i$	$(X_i - \overline{X})^2$
$[1{,}5500;1{,}5925[$	1,57125	3	3	4.714	0.025
$[1{,}5925;1{,}6350[$	1,61375	3	6	4.841	0.014
$[1{,}6350;1{,}6775[$	1,65625	8	14	13.250	0.005
$[1{,}6775;1{,}7200[$	1,69875	12	26	20.385	0.001
$[1{,}7200;1{,}7625[$	1,74125	10	36	17.412	0.000
$[1{,}7625;1{,}8050[$	1,78375	12	48	21.405	0.003
$[1{,}8050;1{,}8475[$	1,82625	4	52	7.305	0.009
$[1{,}8475;1{,}8900[$	1,86875	4	56	7.475	0.019
$\sum$		56		96.787	0.077

$$\overline{X} = \frac{\sum X_i f_i}{\sum f_i} = \frac{96.787}{56} = 1{,}73 \qquad s^2 = \frac{\sum (X_i - \overline{X})^2 f_i}{n-1} \quad \text{e} \quad s = \sqrt{s^2} = 0{,}036$$

b) Tabela 2.32 Distribuição de frequências a partir do diagrama de ramo-e-folhas

Ramo	Folhas	f_i	F_i
1,5	5 5 8	3	3
1,6	2 3 3 4 4 5 6 6 6 7 7 9 9 9	14	17
1,7	0 0 1 1 1 1 1 1 2 3 3 3 3 3 4 4 5 6 6 7 7 7 7 7 8 8 9	27	44

1,8	0 0 0 0 1 2 2 2 5 6 8 9	12	56
Total		56	

c) As alturas do exemplo da tabela têm $\overline{X} = 1{,}73$ e o desvio padrão $S = 0{,}070$, respectivamente. Seja o intervalo $1{,}73 \pm k0{,}070$. Pelo Teorema de Chebyshev têm-se: Se $k = 2$, pelo menos $1 - \frac{1}{4} = \frac{3}{4}$ (75%) dos valores estão no intervalo $]1{,}73 - 2 \times 0{,}070; 1{,}73 + 2 \times 0{,}070[=]1{,}59; 1{,}87[$. Na realidade esse intervalo contém 93,3% das alturas, como pode ser verificado na tabela. Se $k = 3$, pelo menos $1 - \frac{1}{9} = \frac{8}{9}$ (88,9%) das alturas estão no intervalo $]1{,}73 - 3 \times 0{,}070; 1{,}73 + 3 \times 0{,}070[=]1{,}52; 1{,}94[$. Na realidade esse intervalo contém 100% das alturas.

5.4 Amplitude Iinterquartil $I_Q = Q_3 - Q_1$

A $\text{Amplitude interquatil} = I_Q = Q_3 - Q_1$ dá a concentração de 50% dos dados que se encontram dispersos relativamente ao centro. Este conjunto de dados vai desde o primeiro até ao terceiro quartil.

Tabela 2.33 – Resumo das principais definições sobre medidas de dispersão

Estatística	Notação	Definição, propriedades
Amplitude	λ	É a distância entre o valor mínimo e máximo e da variável $\lambda = l_{\max} - l_{\min}$
Amplitude Interquartílica	I_Q	É a distãncia entre o valor do primeiro e do terceiro quartil $I_Q = Q_3 - Q_1$
Desvio médio	DM	É a média dos valores absolutos dos desvios dos valores da variável em relação à média
Variância	S^2	É a média dos quadrados dos desvios dos valores da variável em relação à média
Desvio padrão	S	É a raiz quadrada da variância
Coeficiente de variação	C_v	É uma medida de dispersão relativa. É definida como o quociente entre o desvio padrão e a média, multiplicado por 100, para expressar porcentagem, isto é: $C_v = \frac{S}{\overline{X}} \times 100\%$

5.5 Posições relativas da média,mediana e moda em função da assimetria das distribuições

5.4.1 Medidas de Dispersão, Assimetria e Curtose

Muitas séries estatísticas podem apresentar a mesma média, mas no entanto, os dados de cada uma dessas séries podem distribuir-se de forma distinta em torno de cada uma das médias dessas séries. Na análise descritiva de uma distribuição estatística é fundamental, além da determinação de uma medida de tendência central, conhecer a dispersão dos dados e a forma da distribuição. Duas séries de dados podem possuir a mesma média, mas uma pode apresentar valores mais homogéneos (menos dispersos em relação à média) do que a outra. Um país, por exemplo, com uma distribuição de rendimento familiar mais equânime, terá uma dispersão de suas rendas menor do que um país com estrutura de rendimento familiar mais diferenciada em diversos estratos ou categorias sociais. Uma máquina que produz parafusos e que estiver menos ajustada do que outra, produzirá medidas de parafusos com distribuição mais dispersa em torno de sua média.

A inequação das médias

A importância das médias é com frequência exagerada. Se dizemos que o rendimento familiar médio de um determinado país é de 5.000.000,00Mt por ano não sabemos muita coisa sobre a distribuição do rendimento familiar desse país. Uma média, como um simples valor adoptado para representar a tendência central de uma série de dados é uma medida muito útil. Porém, o uso de um simples e único valor para descrever uma distribuição abstrai-se de muitos aspectos importantes.

Em primeiro lugar, nem todas as observações de uma série de dados têm o mesmo valor da média. Quase sem excepção, as observações incluídas numa distribuição distanciam-se do valor central, embora o grau de afastamento varie de uma série para outra. Muito pouco pode ser dito a respeito da dispersão mesmo quando diversas medidas de tendência central são calculadas para a série. Por exemplo, não podemos dizer que distribuição tem maior ou menor grau de dispersão da informação dada pela tabela abaixo.

Tabela 2.33 – Medidas de tendência central de duas distribuições

	Distribuição A	Distribuição B
Média	15	15
Mediana	15	12
Moda	15	6

Uma segunda consideração é que as formas de distribuição diferem de um conjunto de dados para outro. Algumas são simétricas e outras não. Assim, para descrever uma distribuição precisamos também de uma medida do grau de simetria ou assimetria. A estatística para esta característica é chamada de *medida de assimetria*.

Finalmente, existem diferenças no grau de achatamento entre as diferentes distribuições. Esta propriedade é chamada de *curtose* (em inglês, *kurtosis*). Medir a curtose de uma distribuição significa comparar a concentração de observações próximas do valor central com a concentração de observações próximas das extremidades da distribuição.

Medidas de Assimetria

Duas distribuições também podem diferir uma da outra em termos de assimetria ou achatamento, ou ambas. Como veremos, assimetria e achatamento (o nome técnico utilizado para esta última característica de forma da distribuição é *curtose*) têm importância devido a considerações teóricas relativas à inferência estatística que são frequentemente baseadas na hipótese de Populações distribuídas normalmente. Medidas de assimetria e de curtose são, portanto, úteis para se precaver contra erros aos estabelecer esta hipótese.

Diversas medidas de assimetria são disponíveis, mas introduziremos apenas uma, que oferece simplicidade no conceito assim como no cálculo. Esta medida, a medida de assimetria de Pearson, é baseada nas relações entre a média, mediana e moda. Recorde que estas três medidas são idênticas em valor para uma distribuição unimodal simétrica, mas para uma distribuição assimétrica a média distancia-se da moda, situando-se a mediana numa posição intermediária à medida que aumenta a assimetria da distribuição. Consequentemente, a distância entre a média e a moda poderia ser usada para medir a assimetria. Precisamente,

Assimetria = média - moda

Quanto maior é for esta distância, seja negativa ou positiva, maior é a assimetria da distribuição. Tal medida, entretanto, tem dois defeitos na aplicação. Primeiro, porque ela é uma medida absoluta, o resultado é expresso em termos da unidade original de medida da distribuição e, portanto, ela muda quando a unidade de medida muda. Segundo, a mesma grandeza absoluta de assimetria tem diferentes significados para diferentes séries de dados com diferentes graus de variabilidade. Para eliminar estes defeitos, podemos medir uma medida relativa de assimetria. Esta é obtida pelo *coeficiente de assimetria de Pearson*, denotado por S_{Kp} e dado por: $S_{Kp} = \frac{\overline{X} - X_m}{S}$

A aplicação desta expressão envolve outra dificuldade, que surge devido ao facto do valor modal da maioria das distribuições ser somente uma distribuição, enquanto que a localização da mediana é mais satisfatoriamente precisa. Contudo, em distribuições moderadamente assimétricas, a expressão $X_m = \overline{X} - 3(\overline{X} - X_5)$ é adequada (não envolve imprecisão muito grande). A partir disto, vemos que:

$$\overline{X} - X_m = \overline{X} - \left[X - 3(\overline{X} - X_5)\right] = 3(\overline{X} - X_5)$$

Com este resultado, pode-se rescrever o coeficiente de assimetria de Pearson como:

$$SK_p = \frac{3(\overline{X} - X_5)}{S}$$

Esta medida é igual a zero para uma distribuição simétrica, negativa para distribuições com assimetria para a direita e positiva para distribuições com assimetria para a esquerda. Ela varia dentro dos limites de ± 3.

Aplicando SK_P aos dados agrupados de gastos com consumo de alimentos das famílias, temos:

$$SK_p = \frac{3(170,25 - 167,92)}{23,71} = +0,295$$

Este resultado revela que a distribuição de gastos com consumo de alimentos tem assimetria moderadamente positiva (o que significa maior concentração de famílias nas classes de menor gasto). É muito comum encontrar distribuições positivamente assimétricas em dados económicos, particularmente na produção e séries de preços, os quais podem ser tão pequenos quanto nulos, mas podem ser infinitamente grandes. Distribuições assimetricamente negativas são raras em análises sociais.

Assimetria positiva

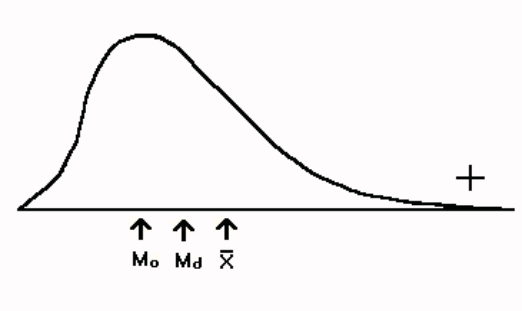


Fig 2.16 A

Distribuição Simétrica

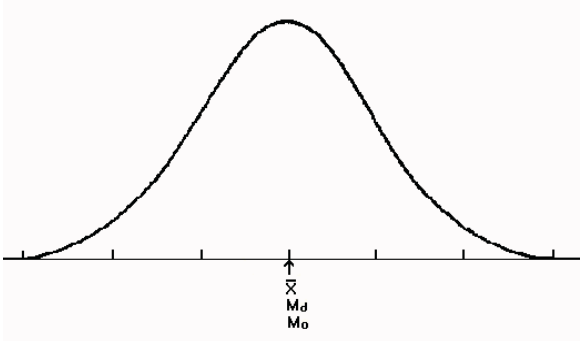


Fig 2.16 B

Assimetria Negativa

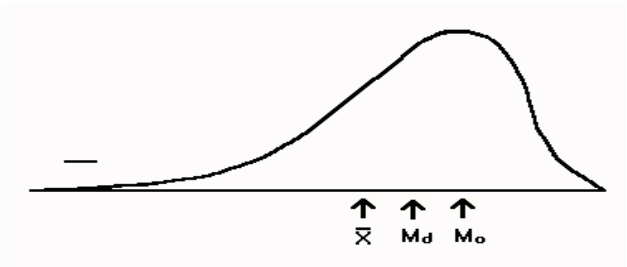


Fig 2.16 C

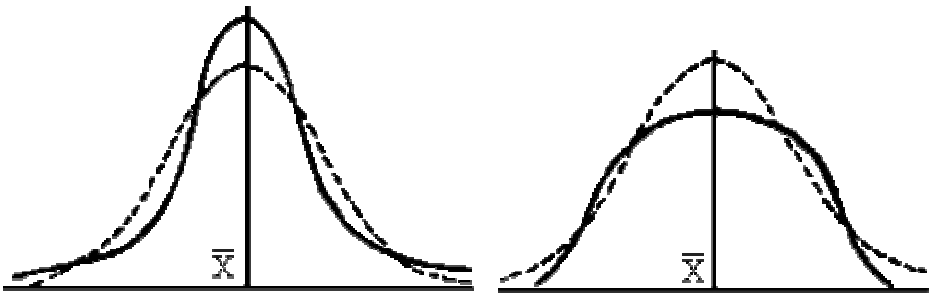
5.4.2 Curtose: uma medida de achatamento

Apresentaremos agora uma medida de achatamento das distribuições, o *coeficiente de curtose*, denotado por K. Esta medida é algebricamente tratável e geometricamente interpretável. É definida como a relação entre o desvio semi-interquartílico, ou seja, a metade do valor do desvio interquartílico, e o intervalo entre o decil 9 e o decil 1:

$$K = \frac{\frac{1}{2}(Q_3 - Q_1)}{D_9 - D_1}$$

Por meio do coeficiente de curtose, classificamos diferentes graus de achatamento em três categorias: *leptocúrtica*, *platicúrtica* e *mesocúrtica* (ver figura, a seguir). Uma distribuição leptocúrtica (curva a) tem a maior parte de suas observações concentradas no centro. Consequentemente, a diferença entre as duas distâncias, (Q_3 - Q_1) e (D_9 - D_1) tende a ser muito pequena. Para um dado grau de dispersão, quanto menor for o achatamento da distribuição, menor será diferença entre estas duas distâncias. Desde que ½ (Q_3 - Q_1) < (D_9 - D_1) para uma distribuição com forma muito pontiaguda, K aproxima-se de 0,5 no limite, quando Q_3 - Q_1 = D_9 - D_1. Ao contrário, quanto mais platicúrtica é a distribuição (curva b), mais o intervalo entre os decis 9 e 1 tende a exceder o intervalo interquartílico. Portanto, quando o intervalo de uma variável tende ao infinito e para uma curva completamente achatada, K tende a zero. Em vista destas considerações, parece razoável estabelecer valores próximos de 0,25 para representar distribuições mesocúrticas (curva c). Esta escolha é reforçada pelo facto de que para a variável normal padronizada, k = 0,2630.

Fig 2.17



Na figura acima compara-se a curtose de duas distribuições com a curtose de uma distribuição mesocúrtica (em linha tracejada). Na figura da esquerda temos uma distribuição platicúrtica (linha cheia) e na figura da direita temos uma distribuição leptocúrtica (linha cheia).

Após o cálculo dos quartis e decis a partir dos dados agrupados para a distribuição de gastos com alimentação, temos que:

$K = \frac{\frac{1}{2}(Q_3 - Q_1)}{D_9 - D_1} = \frac{(1/2)(188.39 - 154{,}83)}{209{,}78 - 146{,}58} = 0{,}2655$ Este resultado indica que a distribuição de gastos com alimentos é aproximadamente mesocúrtica, já que é muito próximo de 0,25.

Principais parâmetros e estatísticas: definição e operacionalização

Deve-se ter cuidado com a notação, uma vez que se pode estar trabalhando tanto com dados populacionais, quanto amostrais. Notação das principais estatísticas:

Tabela 2.34 – Parâmetros e estatísticas (estimadores)

	Parâmetro populacional	Estimador	Variável aleatória
Tamanho	Tamanho da população N	Tamanho da amostra n	
Média	Média populacional μ	Média amostral X	Esperança matemática E(X)
Proporção	Proporção populacional π ou p	Proporção amostral P	Esperança matemática E(X)
Variância	Variância populacional σ^2	Variância amostral S^2	Variância matemática V(X)
Desvio padrão	Desvio padrão populacional σ	Desvio padrão amostral S	
Coeficiente de correlação	Coef. correlação populac. ρ	Coef. correlação amostral r	

6. INDICADORES GENÉRICOS

6.1 Proporções

A proporção permite avaliar uma parte relativa a um todo, a qual ela é integrante.

Tomemos o exemplo de predominância da cólera em Maputo-Moçambique (variável X) e que esta distribuição é feita por 10 bairros (Hulene, Mafalala, Chamanculo, Xipamanine, Luís Cabral, Inhagóia, Benfica, Maxaquene, Aeroporto e Malanga). A distribuição é a seguinte:

Tabela 2.35 – Determinação de proporções

Bairros	Nº de Doentes (p_i^*)	Proporção (p_i)	Percentagem ($p_i^{\%}$) %
Hulene	410	$p_1 = \frac{p_1^*}{n} = \frac{410}{3700} = 0{,}111$	11,1
Mafalala	513	$p_2 = \frac{p_2^*}{n} = \frac{519}{3700} = 0{,}140$	14,0
Chamanculo	517	$p_3 = \frac{p_3^*}{n} = \frac{517}{3700} = 0{,}140$	14,0
Xipamanine	613	$p_4 = \frac{p_4^*}{n} = \frac{613}{3700} = 0{,}166$	16,6
Luís Cabral	12	$p_5 = \frac{p_5^*}{n} = \frac{12}{3700} = 0{,}003$	0,3
Inhagóia	100	$p_6 = \frac{p_6^*}{n} = \frac{100}{3700} = 0{,}027$	2,7
Benfica	409	$p_7 = \frac{p_7^*}{n} = \frac{409}{3700} = 0{,}110$	11,0

Maxaquene	407	$p_8 = \frac{p_8^*}{n} = \frac{407}{3700} = 0{,}110$	11,0
Aeroporto	215	$p_9 = \frac{p_9^*}{n} = \frac{215}{3700} = 0{,}058$	5,8
Malanga	498	$p_{10} = \frac{p_{10}^*}{n} = \frac{498}{3700} = 0{,}135$	13,5
Total	3700	1,000	100

Nota 4: Para as proporções usaremos sempre 3 casas decimais de modo que o seu valor percentual tenha uma casa decimal.

Tomando a tabela 2.35, é fácil aperceber-se de que: $p_i = \frac{p_i^*}{n}$ (1), o n é o número total de casos de cólera que é de 3700. Vejamos agora o cálculo das proporções e percentagens de doentes constantes da tabela acima.

Atente-se para o facto de que n é tamanho duma amostra, mas se for da população usaremos N.

Repare-se ainda que em todos pressupostos $\sum_{i=1}^{n} p_i^*$ (2) ou $\sum_{i=1}^{N} p_i^*$ (3), se for amostra ou população respectivamente. $\sum_{i=1}^{n} p_i = 1{,}00$ (4)

6.2 Percentagens

Na maior parte dos serviços, o indicador mais usado é a percentagem devido à sua maneira mais clara de ilustração das ocorrências. A percentagem é 100 vezes o valor da respectiva proporção, isto é, $p_i^{\%} = p_i \times 100$ (5), sendo que $\sum_{i=1}^{n} p_i^{\%} = 100$ (6). A fórmula (5) pode ser vista ainda como $p_i^{\%} = p_i \times 100 = \frac{p_i^*}{n} \times 100$ (7).

Para a tabela 2.35 as percentagens foram encontradas dos seguintes cálculos:

$$p_1^{\%} = \frac{p_1^*}{n} \times 100 = \frac{410}{3700} \times 100 = 0{,}111 \times 100 = 11{,}1$$

$$p_2^{\%} = \frac{p_2^*}{n} \times 100 = \frac{519}{3700} \times 100 = 0{,}140 \times 100 = 14{,}0$$

$$p_3^{\%} = \frac{p_3^*}{n} \times 100 = \frac{517}{3700} \times 100 = 0{,}140 \times 100 = 14{,}0$$

$$p_4^{\%} = \frac{p_4^*}{n} \times 100 = \frac{613}{3700} \times 100 = 0{,}166 \times 100 = 16{,}6$$

$$p_5^{\%} = \frac{p_5^*}{n} \times 100 = \frac{12}{3700} \times 100 = 0{,}003 \times 100 = 0{,}3$$

$$p_6^{\%} = \frac{p_6^*}{n} \times 100 = \frac{100}{3700} \times 100 = 0{,}027 \times 100 = 2{,}7$$

$$p_7^{\%} = \frac{p_7^*}{n} \times 100 = \frac{409}{3700} \times 100 = 0{,}110 \times 100 = 11{,}0$$

$$p_8^{\%} = \frac{p_8^*}{n} \times 100 = \frac{407}{3700} \times 100 = 0{,}110 \times 100 = 11{,}0$$

$$p_9^{\%} = \frac{p_9^*}{n} \times 100 = \frac{215}{3700} \times 100 = 0{,}058 \times 100 = 5{,}8$$

$$p_{10}^{\%} = \frac{p_{10}^*}{n} \times 100 = \frac{498}{3700} \times 100 = 0{,}135 \times 100 = 13{,}5$$

Então $\frac{p_1^*}{n} \times 100 + \frac{p_2^*}{n} \times 100 + \frac{p_3^*}{n} \times 100 + \ldots + \frac{p_n^*}{n} \times 100 = 100$ (8).

Voltando para a tabela 2.35, pode-se constatar que:

$$\frac{410}{3700} \times 100 + \frac{519}{3700} \times 100 + \frac{517}{3700} \times 100 + \frac{613}{3700} \times 100 + \frac{12}{3700} \times 100 + \frac{100}{3700} \times 100 +$$

$$+ \frac{409}{3700} \times 100 + \frac{407}{3700} \times 100 + \frac{215}{3700} \times 100 + \frac{498}{3700} \times 100 = 100$$

Assim podemos afirmar que as proporções por um lado e as percentagens por outro facilitam a leitura e comparações entre duas ou mais distribuições de frequências com diferentes dimensões amostrais ou populacionais, que se refiram à mesma variável ou mesma categoria.

Exemplo:
A Rádio Cidade da Beira colocou à disposição dos ouvintes para ligarem à rádio e dizer se preferiam música Pop ou Jazz durante os seus divertimentos, e os resultados foram os seguintes:

Tabela 2.36 - Número de ouvintes que preferem a música Pop

Idade	M	F	Total
<25	19	26	45
25-50	38	34	72
>50	48	60	108
Total	105	120	225

Tabela 2.37 -Número de ouvintes que preferem a música Jazz

Idade	M	F	Total
<25	63	45	108
25-50	38	33	71
>50	44	52	96
Total	145	130	275

a) Calcule a Percentagem de Masculinos que preferem a música Pop.

Resposta: $\%_{M_{pop}} = \frac{105}{225} \times 100\% = 46{,}7\%$

b) Calcule a percentagem dos ouvintes que preferem a música Jazz.

Resposta: $\% = \frac{275}{275+225} \times 100\% = \frac{275}{500} \times 100\% = 55{,}0\%$

c) Calcule a percentagem das mulheres que preferem a música pop ou música Jazz.

Resposta: $\% = \frac{120+130}{225+275} \times 100\% = \frac{250}{500} \times 100\% = 50{,}0\%$

d) Calcule a percentagem de homens que preferem a música Pop ou mulheres que preferem a música Jazz.

Resposta: $\% = \frac{105+130}{225+275} \times 100\% = \frac{235}{500} \times 100\% = 47{,}0\%$

Exemplo: Os dados abaixo referem-se ao produto interno bruto (PIB) de Moçambique (óptica da despesa no que diz respeito ao comércio externo para o ano de 1991)

Tabela 2.38 – PIB, óptica da despesa

Descrição	Valor em 1996 (10^6Mt)
Exportações	*1,934,256.00*
Bens	1,548,003.00
Serviços	386,253.00
Importações	*-12,534,808.00*
Bens	-11,386,175.00
Serviços	-1,148,633.00
Total COMÉRCIO EXTERNO	-10,600,552.00

Fonte: www.ine.org.mz - Instituto Nacional de Estatística– extraído no dia 31/01/2004

a) Calcule a percentagem das importações.

Resposta: $\% = \frac{-12{,}534{,}808.00 \times 10^6}{-10{,}600{,}552.00 \times 10^6} \times 100\% = 118{,}25\%$

b) Calcule a percentagem das exportações.

Resposta: $\% = \frac{1{,}934{,}256.00 \times 10^6}{-10{,}600{,}552.00 \times 10^6} \times 100\% = -18{,}25\%$

Conclusões: O país exportou muito menos do que importou, causando um elevado défice, o que leva a crer que o país só poderá ter um crescimento positivo com base noutros meios, casos de maior arrecadação de receitas pelo combate à fuga aos impostos, fortificação do sector de serviços (pelo aumento de tecnologias de outsourcing) e combate ao despesismo, de modo a equilibrar a balança de pagamentos, o que passa pelo aumento de fiscalização e controle.

6.3 Taxas

As taxas são quocientes que exprimem o peso do valor registado para um dado fenómeno face ao seu valor potencial, expressos em percentagens.
Suponhamos que o Santos seja amigo do Miguel. O Santos solicita ao Miguel um empréstimo de 250.000.000,00Mt (Duzentos e cinquenta milhões de meticais). Admitamos que o Miguel concede ao Santos 200.000.000,00Mt (Duzentos milhões de Meticais). A partir daquí, podemos apurar duas taxas.
- A Taxa de Empréstimo, que se obtém dividindo o valor concedido pelo total solicitado e o resultado multiplicado em cem por cento.

$$Taxadeemprestimo = \frac{Concedido}{Solicitado} \times 100\% = \frac{200.000.000,00Mt}{250.000.000,00Mt} \times 100\% = 80\%$$, ou seja, o Santos foi emprestado 80% do solicitado pelo Miguel.
- A Taxa do que falta, a qual se obtém dividindo os 50.000.000,00Mt (Cinquenta milhões de Meticais que faltam) para completar o valor, pelo total solicitado

$$Taxade\,\mathrm{Re}\,manescente = \frac{\mathrm{Re}\,manescente}{Solicitado} \times 100\% = \frac{50.000.000,00Mt}{250.000.000,00Mt} \times 100\% = 20\%$$, ou seja, o Santos terá de encontrar outros meios para conseguir o restante montante avaliado em 20% do total necessário.

Há casos em que as taxas podem ter valor superior a 100%. São casos do número de doentes de cólera numa enfermaria da Beira, em relação à capacidade de camas instaladas. Neste caso tem-se vislumbrado que, na maior dos casos, o número de doentes tem sido superior ao número de camas da unidade sanitária. Pode-se considerar também as existências relativamente à capacidade de ocupação das penitenciárias em Moçambique.

6.4 Taxa de Variação

Traduz percentualmente o acréscimo ou decréscimo global entre os registos relativos a um determinado fenómeno, num período de tempo definido.
Tomemos os dados publicados pelo INE no que diz respeito à prevalência de Unidades Sanitárias na províncias da Zambézia, Gaza e Manica, respectivamente, nos anos 1996 e 1997.

Tabela 2.39 – Distribuição de Unidades Sanitárias entre 1996 e1997

Províncias	Anos	
	1996	1997
Zambézia	152	166
Gaza	76	85
Manica	77	75

Fonte: www.ine.org.mz - Instituto Nacional de Estatística– extraído no dia 31/01/2004

Calculemos a taxa de variação Δ, relativa aos dois anos, considerados para a província de Gaza

$$\Delta = \frac{T_1 - T_0}{T_0} \times 100\% \qquad \text{ou} \qquad TV = \Delta = \left(\frac{T_1}{T_o} - 1\right) \times 100\%$$

$$\Delta_{Gaza} = \frac{T_{1997} - T_{1996}}{T_{1996}} \times 100\% = \frac{85 - 76}{76} \times 100\% = 11{,}84\%$$

$$\Delta_{Zambézia} = \frac{T_{1997} - T_{1996}}{T_{1996}} \times 100\% = \frac{166 - 152}{152} \times 100\% = 9{,}21\%$$

$$\Delta_{Manica} = \frac{T_{1997} - T_{1996}}{T_{1996}} \times 100\% = \frac{75 - 77}{77} \times 100\% = -2{,}6\%$$

Ou

$$TV_{Gaza} = \Delta = \left(\frac{T_{1997}}{T_{1996}} - 1\right) \times 100\% = \left(\frac{85}{76} - 1\right) \times 100\% = 11{,}84\%$$

$$TV_{Zambézia} = \Delta = \left(\frac{T_{1997}}{T_{1996}} - 1\right) \times 100\% = \left(\frac{166}{152} - 1\right) \times 100\% = 9{,}21\%$$

$$TV_{Manica} = \Delta = \left(\frac{T_{1997}}{T_{1996}} - 1\right) \times 100\% = \left(\frac{75}{77} - 1\right) \times 100\% = -2{,}6\%$$

Assim foram apuradas as taxas percentuais das diferenças positivas da prevalência de unidades sanitárias para as três províncias no período referenciado na tabela.
Para os casos da Zambézia e Gaza traduziu-se em aumento de unidades sanitárias em 11,84% e 9,21% respectivamente e para a província de Manica um decréscimo na ordem de 2,6%.

6.5 Taxa de Variação Média ou Taxa de Crescimento Médio

É uma taxa usada para medir a variação dos registos observados num determinado fenómeno (em termos percentuais), verificada entre dois períodos de tempo, como se a mesma tivesse mantido igual ao longo dos subperíodos considerados (anos, semestres, trimestres, dias, semanas, etc.) ou seja, ela dá-nos um acréscimo ou decréscimo, como se ele traduzisse uma variação verificada sub-período a sub-período, de forma constante.

Suponhamos que a vendedeira "Mariamo" do mercado Brandão em Quelimane seja cliente do Novo Banco e pretenda um crédito de 100.000.000,00Mt (Cem milhões de Meticais) para a reconstrução da sua Banca. Suponhamos que o Novo Banco fixe para ela uma taxa anual de 4% sobre o montante aplicado, independentemente de eventuais flutuações a que as taxas de juros estão sujeitas. Se ela pretender a simulação do mapa de amortizações com base num plano, para saber que montante pagará depois de n anos, poderia proceder aos seguintes cálculos:

Tabela 2.40 – Tabela de amortizações a devolver ao Banco, incluindo juros

Anos	Valor em (1.000.000Mt) milhão	
0	100	V_o
1	$100+,04\times100=100\times(1+0,04)=100\times1,04=104$	$V_1=V_o+V_0i=V_o\times(1+i)$
2	$104\times1,04=100\times1,04\times1,04=100\times1,04^2$	$V_2=V_o\times(1+i)\times(1+i)=V_o\times(1+i)^2$
……	……	……
n	$100\times1,04^n$	$V_n=V_o\times(1+i)^n$

Neste caso,

V_n - É o valor final da variável

V_o - Valor inicial da variável

i - Taxa a aplicar no decorrer do tempo

n - Tempo correspondente ao valor final

$$V_n=V_o\times(1+i)^n\Rightarrow\frac{V_n}{V_o}=(1+i)^n\Leftrightarrow1+i=\sqrt[n]{\frac{V_n}{V_o}}\Leftrightarrow i=\left(\sqrt[n]{\frac{V_n}{V_o}}-1\right)\times100\%$$

Neste caso a taxa de variação ou crescimento médio pode ser calculada como sendo $i=\left(\sqrt[n]{\frac{V_n}{V_o}}-1\right)\times100\%$

Imaginemos que esse pagamento tenha de ser feito em dez anos, então o valor a retornar ao credor será de: $V_{10}=V_o\times(1+i)^{10}=100.000.000,00Mt\times(1+0,04)^{10}=148.024.428,00Mt$

Observação: Geralmente na Banca Comercial ou instituições de crédito, aplica-se o conceito de taxa de juro (Não a calculada sobre capital em dívida, mas sim sobre o capital concedido), como no exemplo atrás.

Juro- É uma compensação recebida pelo credor por não dispor do objecto da relação estabelecida com o devedor durante o tempo acordado. É importante que desde logo, se pondere da exequibilidade do devedor gerar uma mais valia, de forma a que, findo o prazo combinado, possa restituir ao credor, para além do juro, o inicialmente acordado.

Único: a taxa que caracteriza o juro denomina-se Taxa de Juros.

Exercício

Distribuição percentual da população de 15 anos por nível de ensino concluído segundo área de residência, idade e sexo, na Província de Inhambane, em 1997.

Tabela 2.41 Nível de alfabetização da população Moçambicana

Grupos de idade	N (1000)	Nível concluído								
		Total	Alfabet.	Primário	Secundário	Técnico	C.F.P.	Superior	Nenhum	Desconh.
Total										
Total	638.5	100.0	0.5	19.4	1.0	0.2	0.2	0.0	78.5	0.2
15 – 19	126.1	100.0	0.1	30.8	0.5	0.0	0.0	0.0	68.1	0.3
20 – 24	90.0	100.0	0.1	31.7	1.8	0.3	0.1	0.0	65.7	0.4
25 – 29	69.3	100.0	0.2	25.9	1.8	0.4	0.2	0.0	71.4	0.2
30 – 39	108.3	100.0	0.3	20.6	1.8	0.5	0.5	0.1	76.1	0.1
40 – 49	85.8	100.0	0.8	9.2	0.8	0.2	0.1	0.0	88.9	0.1
50 – 59	72.2	100.0	1.0	6.5	0.3	0.1	0.1	0.0	91.9	0.0
60 +	86.7	100.0	1.2	4.1	0.1	0.1	0.0	0.0	94.5	0.0
Homens	250.3	100.0	0.9	26.7	1.9	0.4	0.3	0.1	69.5	0.3
15 – 19	55.9	100.0	0.1	34.5	0.8	0.1	0.0	0.0	64.0	0.5
20 – 24	32.4	100.0	0.1	38.1	3.2	0.6	0.1	0.0	57.2	0.6
25 – 29	24.1	100.0	0.2	35.5	3.4	0.8	0.4	0.1	59.2	0.4
30 – 39	40.7	100.0	0.5	33.3	3.7	1.0	1.0	0.1	60.0	0.3
40 – 49	32.1	100.0	1.4	18.2	1.7	0.4	0.2	0.1	77.8	0.1
50 – 59	28.4	100.0	1.9	14.0	0.7	0.3	0.1	0.0	82.9	0.0
60 +	36.6	100.0	2.3	8.6	0.3	0.2	0.1	0.0	88.5	0.0
Mulheres	388.3	100.0	0.2	14.7	0.5	0.1	0.1	0.0	84.4	0.1
15 – 19	70.2	100.0	0.1	27.9	0.4	0.0	0.0	0.0	71.3	0.2
20 – 24	57.6	100.0	0.1	28.0	1.0	0.1	0.1	0.0	70.5	0.2
25 – 29	45.2	100.0	0.1	20.7	0.9	0.1	0.1	0.0	77.8	0.1
30 – 39	67.6	100.0	0.2	12.9	0.7	0.1	0.2	0.0	85.8	0.1
40 – 49	53.8	100.0	0.4	3.8	0.2	0.0	0.0	0.0	95.5	0.0
50 – 59	43.8	100.0	0.4	1.7	0.1	0.0	0.0	0.0	97.8	0.0
60 +	50.2	100.0	0.3	0.8	0.0	0.0	0.0	0.0	98.8	0.0

Fonte: www.ine.org.mz - Instituto Nacional de Estatística– extraído no dia 31/01/2004

a) Compare a estrutura percentual das mulheres com o nível secundário e com o nível técnico entre os 20-24 anos de idade. (Proposta: Repare nos dados e determine os rácios possíveis).

b) Compare a evolução do número de homens do escalão 15-19 anos com o do escalão 25-29 anos no que diz respeito ao nível desconhecido.

c) Determine a moda das mulheres.

d) Determine a mediana das mulheres e construa o respectivo polígono de frequências.

e) É possível com esses dados construir um diagrama de Box-Plot?

7. ECONOMIA-CONCEITO

Diversos conceitos já foram formulados e são bem aceites para a ciência económica. Todavia, o que será citado a seguir consegue sintetizar o tema de forma objectiva. Assim: "Economia é a ciência que estuda a produção, distribuição e consumo dos bens de serviços, com o objectivo de aproveitá-los plena e combinadamente".
Entenda-se de forma literal que, aproveitar os recursos plenamente, refere-se ao facto de evitar a toda prova a ociosidade dos recursos de produção tais como, terra, trabalho, capital e capacidade empresarial, sendo este último incluso e aceite pelos economistas clássicos.

Já aproveitar os recursos combinadamente, refere-se ao facto de optimizar os resultados da produção, ou seja, definir correctamente o que produzir, como produzir e para quem produzir, dentro de infinitas possibilidades. Neste caso, busca-se a produção ideal que atenda as necessidades dos consumidores, ao menor custo e com a melhor qualidade possível, satisfazendo também a maior remuneração do capital investido.

7.1 Problema central da economia

Constitui-se indubitavelmente o problema central de qualquer economia decidir: O que produzir?, Como produzir? e Para quem produzir? Sabe-se que escolher correctamente o investimento é tão importante quanto a torná-lo operacional. Definiremos o conceito Rácio muito usado como indicador social ou económico para a seguir tratarmos da Estatística Económica/Social, reparando em Índices (para determinar alguns indicadores sociais, nomeadamente "Índice de preço do consumidor"), Séries Temporais (para previsões económicas/Sociais) e dados bivariados (relação de duas ou mais variáveis e regressão).

7.2 Rácios

Se pretendermos comparar dois valores de uma mesma variável ou de variáveis diferentes, poder-se-á fazê-lo recorrendo a um rácio, o qual consiste na mera divisão dos referidos valores e sua interpretação.

Tomemos um conjunto de Notas do Exame de recorrência dos Estudantes aprovados da Faculdade de Engenharias da Universidade Eduardo Mondlane à cadeira de Probabilidades e Métodos Estatísticos no Semestre Julho-Dezembro 2003.

Tabela 2.42 - Resultados do Exame de Recorrência – Engª Mecânica

ENGENHARIA MECÂNICA

	Nomes	Resultado	Exame	Média
01.	ABDUL	13.0	19.0	16.0
02.	CUMBE	12.0	14.0	13.0
03.	DIOGO	10.0	15.0	13.0
04.	MADIJA	13.0	11.0	12.0
05.	MANJICHE	12.0	13.0	13.0
06.	MATEUS	10.0	17.0	14.0
07.	MAÚSSE	10.0	15.0	13.0
08.	MIAMBO	10.0	11.0	11.0
09.	TONDO	11.0	14.0	13.0
10.	TONELA	11.0	14.0	13.0
	Média	11,2	14,3	13,1

Tabela 2.43 - Resultados do Exame de Recorrência – Engª Civil

ENGENHARIA CIVIL

	Nomes	Resultado	Exame	Média
01	ABADA	12	10	11,0
02	ABDALA	10,0	11,0	11,0
03	CHENENE	10,0	12,0	11,0
04	CHIPANGA	11,0	11,0	11,0
05	COIMBRA	10,0	10,0	10,0
06	CRUZ	10,0	14,0	12,0
07	LEITE	11,0	10,0	11,0
08	MACITA	13,0	11,0	12,0
09	MIGUEL	10,0	12,0	11,0
10	MOISÉS	10,0	13,0	12,0
	Média	10,7	11,4	11,2

Comparando a média das médias da Engenharia Mecânica com a Engenharia Civil

$\frac{MédiaMecânica}{MédiaCivil} = \frac{13,1}{11,2} = 1,17$, que se pode lêr da seguinte forma:

- Na cadeira de Estatística por cada valor da média do estudante da Engª Mecânica , há aproximadamente um valor na média para o estudante da Engª Civil
- Na cadeira de Estatística, o rácio (ou a razão) entre a média dos estudantes da Engª Mecânica e a média dos Estudantes da Engª Civil é de aproximadamente um para um.

Tratando-se de pessoas (que não são fraccionáveis) e sempre que o valor do rácio seja inferior à unidade, é usual multiplicá-lo por dez, cem, mil a fim de facilitar a respectiva leitura.

O que significa que para o caso do exemplo, teríamos $\frac{MédiaMecânica}{MédiaCivil} = \frac{13,1}{11,2} = 1,17 \times 10 = 11,7$, ficando assim : Na cadeira de Estatística por cada dez valores da média do estudante da Engª Mecânica , há aproximadamente dezoito valores na média para o estudante da Engª Civil.

Observação: O bom senso reza que sempre se use a regra mais coerente para facilitação da leitura, mediante a convicção da equipa que pretenda essa informação.

8. CORRELAÇÃO E REGRESSÃO

Por vezes certos fenómenos em estudo não são descritos apenas através de uma variável, sendo necessária a observação de duas ou mais variáveis para termos uma visão global do problema. Quando tal ocorre, cada unidade estatística contribui com um conjunto de dois valores (ou duas variáveis) passando a trabalhar-se com dados bivariados (ao passo que os anteriormente estudados eram univariados).

Exemplos:

Se pretendessemos verificar se os pesos dos pais é um factor herdável pelos filhos mais velhos, recolherí informações sobre pesos dos pais e dos filhos mais velhos e mais tarde observaríamos se possuem alguma relação.

Pretendendo saber a existência de alguma relação entre o comportamento violento de jovens de um bairro que gostam muito de bebida tradicional, precisaríamos de entrevistar alguns jovens para verificar se são ou não violentos e saber deles se gostam ou não da bebida tradicional e no fim fazer a respectiva comparação.

Se pretendemos saber se o aumento de criminalidade é condicionado pelos aumentos mensais do número de desempregados, conduzimos um inquérito para se apurar o número de desempregados por mês e o respectivo número de crimes, para depois verificar a existência, ou não, de alguma relação.

Definição de correlação: Grau de associação entre variáveis quantitativas.

Um problema essencial com o qual nos deparamos na maior dos casos é se determinada característica de uma população está ou não relacionada com outra(s) e em que grau.

O nosso país é um dos mais abalados por anemia, devido às diversas condições a que está sujeita a maioria da população. Suponha que desejássemos realizar uma investigação sobre a ocorrência de anemia e infecção numa comunidade localizada algures na zona de Mavonde na província de Manica. Seria interessante poder estimar a concentração de hemoglobina e a contagem de eritrócitos e leucócitos no sangue pela medida do hematócrito.

Para verificar a possibilidade de se usar tal procedimento veja alguns resultados da rotina de um laboratório de hematologia dum hospital central.

Tabela 2.44 – Resultados (hipotéticos) de uma rotina de um laboratório de hematologia

Exame n^o	leucócitos (x 10^3/mm^3)	eritrócitos (x 10^6/mm^3)	hemoglobina (g/dl)	Hematócrito (%)
1	6.81	4.51	14.7	41
2	9.69	5.21	15.5	47
3	4.30	4.54	14.4	41
4	7.89	4.64	14.3	41
5	7.41	4.41	13.7	40
6	7.60	4.40	14.0	40
7	2.81	4.31	13.6	40
8	7.81	4.60	13.8	42
9	5.49	4.92	15.1	44
10	4.60	4.11	13.0	39
11	8.01	5.01	17.1	46
12	7.02	5.16	16.0	47
13	7.10	4.21	11.6	35
14	10.49	4.49	13.5	39
15	6.91	4.49	14.2	40
16	13.51	4.44	13.6	40
17	8.31	3.69	11.1	33
18	7.00	4.29	12.7	38
19	4.30	4.68	14.0	43
20	2.71	4.39	12.7	39
21	11.19	4.39	13.3	38
22	5.90	4.41	11.9	37
23	12.29	4.24	10.0	31

A interpretação dos dados acima fica bastante mais fácil sob forma gráfica.
Assim sendo, como estamos interessados na relação entre hematócrito e outras medidas hematológicas, observe os gráficos que se seguem:

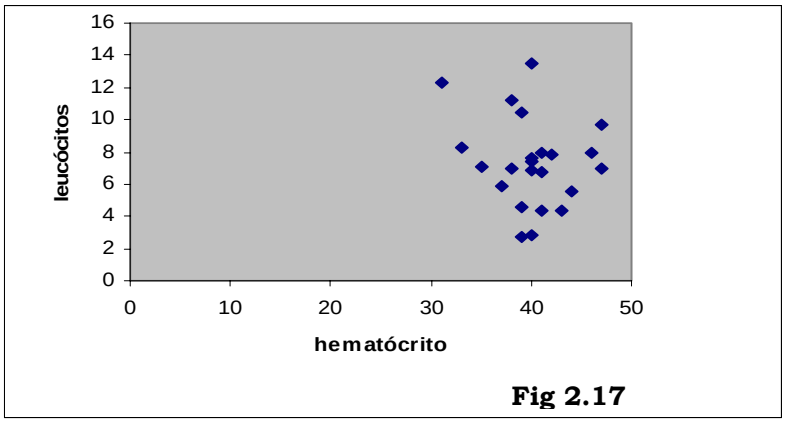


Fig 2.17

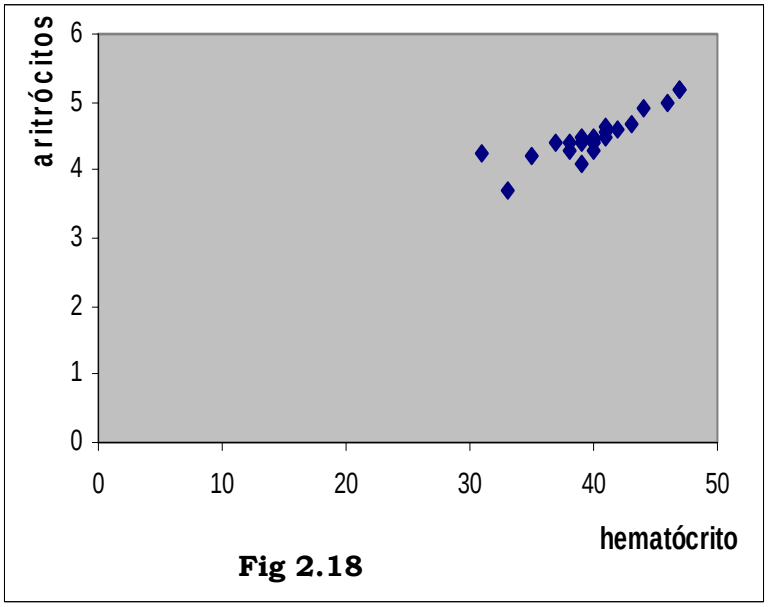


Fig 2.18

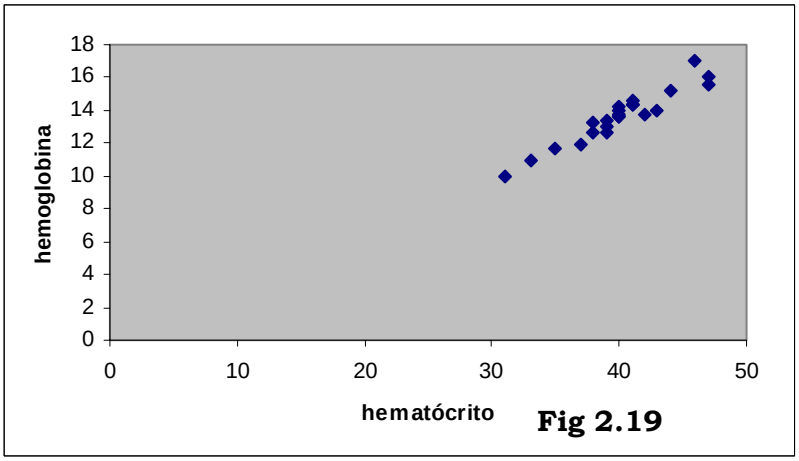


Fig 2.19

8.1 Diagrama de dispersão

É a representação dos pontos que compõem o conjunto dos pares na relação das variáveis. Cada ponto provém de uma das variáveis em análise. Se forem duas variáveis X e Y cada ponto será composto pelo valor de X e pelo valor de Y, como abaixo se apresenta (por exemplo, hematócrito e hemoglobina). Denominado também de *gráfico* XY.

Note que simplesmente se representam pontos e não se traça alguma linha.

8.2 Rectas de Regressão

É a recta que ajusta os dados representados no diagrama de dispersão. Observe que ajustar os dados não significa, necessariamente, passar por todos pontos. Só se passa por todos pontos naqueles casos em que por coincidência os pontos estão bem alinhados. Para obter a recta de regressão é necessário calcular o Coeficiente angular (Coeficiente de regressão) e o intercepto (ponto onde o gráfico corta o eixo y) da recta com o eixo das ordenadas. $y = (\overline{y} - b_{y,x}\overline{x}) + b_{y,x}x$

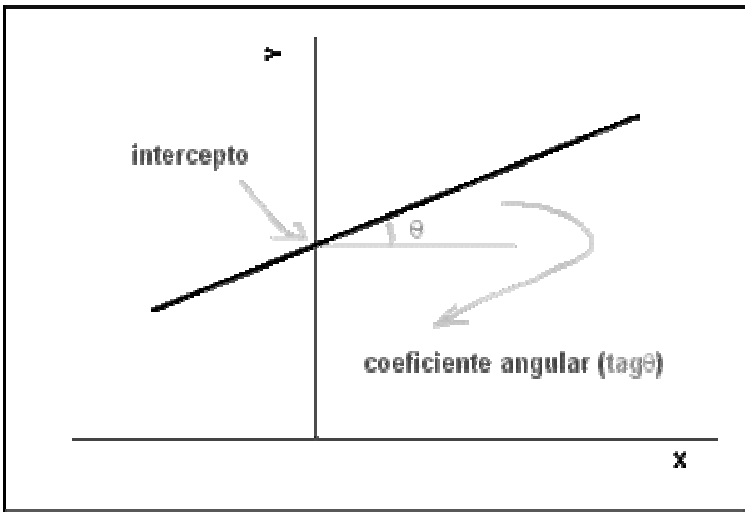


Fig 2.20

8.3 Técnica de ajuste – mínimos quadrados

Em geral quando se pretende ajustar os dados através de uma curva (recta) de regressão deve-se determinar o coeficiente angular, o intercepto, de modo que se possa fazer uma recta que esteja de acordo com os dados e não viciada. Para tal, se usa como uma das técnicas, a de *Mínimos quadrados,* para determinar a equação da recta de regressão.
Tomemos o seguinte exemplo: Numa análise feita a diversos pacientes, em consultas médicas, procurou-se verificar se havia alguma relação entre a pressão arterial diastólica e

o tempo de repouso por paciente. Para tal retiraram-se os seguintes dados referentes a 5 pacientes, para essa análise, segundo a tabela a seguir.

Tabela 2.45 - Pressão arterial diastólica e o tempo de repouso de 5 pacientes

Paciente	1	2	3	4	5
X_i	0	5	10	15	20
Y_i	72	66	70	64	66

Pelo gráfico a seguir é possível observar que há uma relação negativa e fraca. Fraca porque os pontos estão afastados da recta de ajustamento. Mas também o gráfico resulta de um ajuste manual em virtude de não resultar de uma equação de regressão, mas sim da análise feita sobre a disposição dos pontos.

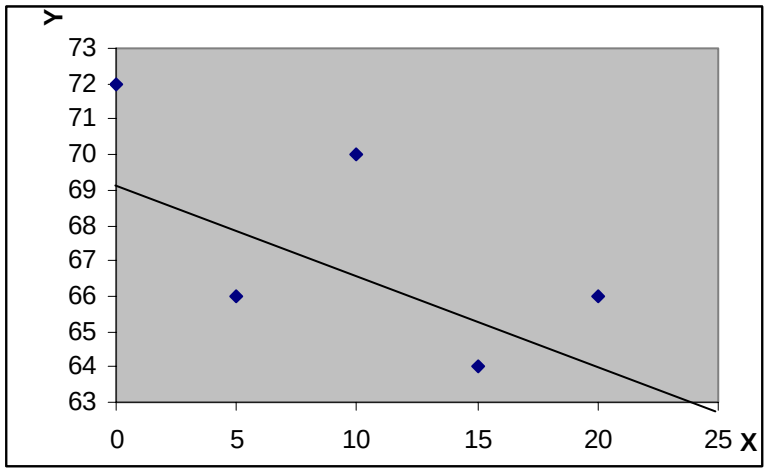


Fig 2.21

onde: Y = Pressão arterial diastólica, X = tempo (minutos) de repouso

Para fazer-se um esboço que não resulte de ajustes individuais é sempre aconselhável obter a respectiva equação de regressão antes. Vejamos a seguir:

$\sum X_i = 50 \quad \sum Y_i = 338 \quad \sum X_i Y_i = 3310 \quad \sum X_i^2 = 750 \quad \sum Y_i^2 = 22982$

$$b_{y,x} = \frac{3310 - \frac{50 \times 338}{5}}{750 - \frac{50}{5}} = -0,28 \qquad a = \frac{338}{5} - (-0,28 \times \frac{50}{5}) = 70,4$$

sendo assim, a equação será dada por $y = 70,4 - 0,28\,x$. Com esta equação já pode-se produzir a melhor recta que ajuste os dados

8.4 Coeficiente Angular (*by,x*)

Mede a variação que ocorre numa característica quando outra característica se modifica de uma unidade. É chamado também de Inclinação do gráfico e ainda coeficiente angular $(tg\theta)$. $b_{y,x} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-x)^2} = \frac{\sum xy - \left[(\sum x)(\sum y)/n\right]}{\sum x^2 - \left[(\sum x)^2\right]/n}$

Intercepto (a) - Ponto de intersecção da recta com a ordenada (eixo Y). Equivale ao valor de Y quando $X=0$. $a = \bar{y} - b_{y,x}\bar{x}$

Equação de Regressão - equação que define a linha recta que descreve a associação entre duas características e que permite estimar o valor de uma medida pela outra. $(y-\bar{y}) = b_{y,x}(x-\bar{x})$

Para se determinar a relação entre cada duas variáveis deve-se calcular o respectivo coeficiente de relacionamento, a que chamamos de coeficiente de correlação. Esse coeficiente de correlação varia entre -1 e 1.

8.5 Coeficiente de Correlação (r)

Karl Pearson, 27 de Março de 1857 - 27 de Abril de 1936 (Londres), foi a pessoa que determinou o primeiro coeficiente de correlação a que se atribui o seu nome, passando a se chamar de Coeficiente de Correlação de Pearson.

Um Coeficiente de Correlação (r)- é a medida que indica o grau de associação entre duas variáveis a partir de uma série de observações. $r = \frac{\text{cov}(x,y)}{S_x S_y} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$, onde $\text{cov}(x,y)$ indica a variância simultânea das duas variáveis, denominada covariância. S_x *e* S_y são os desvios amostrais de X e de Y, respectivamente. Recorde-se que na maioria dos casos determinamos este coeficiente a partir duma amostra.

O coeficiente de Correlação tem o seguinte comportamento:

$r = 1$ A relação é perfeita e positiva (há uma proporcionalidade directa)

$r = -1$ A relação é perfeita e negativa (há proporcionalidade inversa)

$-1 < r < -0{,}5$ A relação é negativa e forte

$-0{,}5 \le r < 0$ A relação é negativa e fraca

$r = 0$ Indica a ausência de relação

$0 < r \le 0{,}5$ A relação é positiva e fraca

$0{,}5 < r < 1$ A relação é positiva e forte

Precauções no uso e interpretação

- A relação deve ser representável por uma linha recta (curva de regressão)
- A recta não pode ser extendida além dos pontos medidos
- A associação não implica necessariamente uma relação casual
- Depende da variabilidade amostral

Exercícios Resolvidos

1- A tabela a seguir mostra os preços de gasolina que vigoraram em Moçambique entre os anos de 1999 e 2003

Preço da Gasolina

Mês/Ano	1999	2000	2001	2002	2003
Janeiro	6.190,00	6.620,00	9.150,00	9.010,00	10.650,00
Fevereiro	6.190,00	7.580,00	8.800,00	8.860,00	11.045,00
Março	6.190,00	8.040,00	8.800,00	9.365,00	12.366,00
Abril	6.190,00	8.040,00	9.330,00	9.870,00	13.074,00
Maio	6.190,00	9.320,00	10.910,00	10.638,00	13.029,00
Junho	6.190,00	9.320,00	11.320,00	11.310,00	13.913,00
Julho	6.190,00	9.850,00	11.320,00	11.310,00	13.790,00
Agosto	6.190,00	9.850,00	11.320,00	11.310,00	14.763,00
Setembro	6.310,00	9.850,00	10.289,00	11.471,00	15.343,00
Outubro	6.620,00	9.150,00	9.500,00	11.750,00	15.380,00
Novembro	6.620,00	9.150,00	9.330,00	11.794,00	15.380,00
Dezembro	6.620,00	9.150,00	9.160,00	10.650,00	15.380,00

a) Determine as medidas de Tendência Central, Posição e de Dispersão

Resolução:

Medidas de localização	1999	2000	2001	2002	2003
Média	6.307,50	8.826,67	9.935,75	10.611,50	13.676,08
Máximo	6.620,00	9.850,00	11.320,00	11.794,00	15.380,00
Minimo	6.190,00	6.620,00	8.800,00	8.860,00	10.650,00
Mediana	6.190,00	9.150,00	9.415,00	10.980,00	13.851,50
Moda	6.190,00	9.850,00	11.320,00	11.310,00	15.380,00

Medidas de posição	1999	2000	2001	2002	2003
Quartil 1	6.190,00	8.040,00	9.157,50	9.743,75	12.963,25
Quartil 2	6.190,00	9.150,00	9.150,00	10.980,00	13.851,50
Quartil 3	6.387,50	9.452,50	11.012,50	11.350,25	15.352,25
InterQuartil	197,50	1.412,50	1.855,00	1.606,50	2.489,00
Percentil 10	6.190,00	7.626,00	8.835,00	9.045,50	11.177,10
Percentil 25	6.190,00	8.040,00	9.157,50	9.743,75	12.863,25
Percentil 50	6.190,00	9.150,00	9.415,00	10.980,00	13.851,50
Percentil 75	6.387,50	9.452,50	11.012,50	11.350,25	15.352,25
Percentil 90	6.620,00	9.850,00	11.320,00	11.722,10	15.380,00
Percentil 95	6.620,00	9.850,00	11.320,00	11.769,80	15.380,00

MEDIDAS DE DISPERSÃO	1999	2000	2001	2002	2003
Desvio Médio	156,67	837,78	912,38	890,17	1369,40
Variância	33618,75	966938,89	962098,69	1052938,25	2609701,74
Desvio Padrão	183,35	983,33	980,87	1026,13	1615,46
Assimetria e sua classificação	1,92	-0,99	1,59	-1,08	-0,33
	Positiva	Negativa	Positiva	Negativa	Negativa
Curtose e sua classificação	0,23	0,32	0,37	0,30	0,30
	Leptocúrtica	Platicúrtica	Platicúrtica	Platicúrtica	Platicúrtica

2- Os visitantes do Parque Nacional de Gorongosa consideram um Leopardo com um dos filhotes, "um gato", uma atração que não pode ser perdida. A tabela de frequências a seguir resume uma amostra de tempos (em minutos) entre as presenças do dito Leopardo.

	Tempo	Frequência	F_i
	40-49	8	8
	50-59	44	52
	60-69	23	75
	70-79	6	81
Classe Mediana	80-89	107	188
	90-99	11	199
	100-109	1	200

a) Construa um polígono de frequências para a tabela de frequências dada.

Resolução:

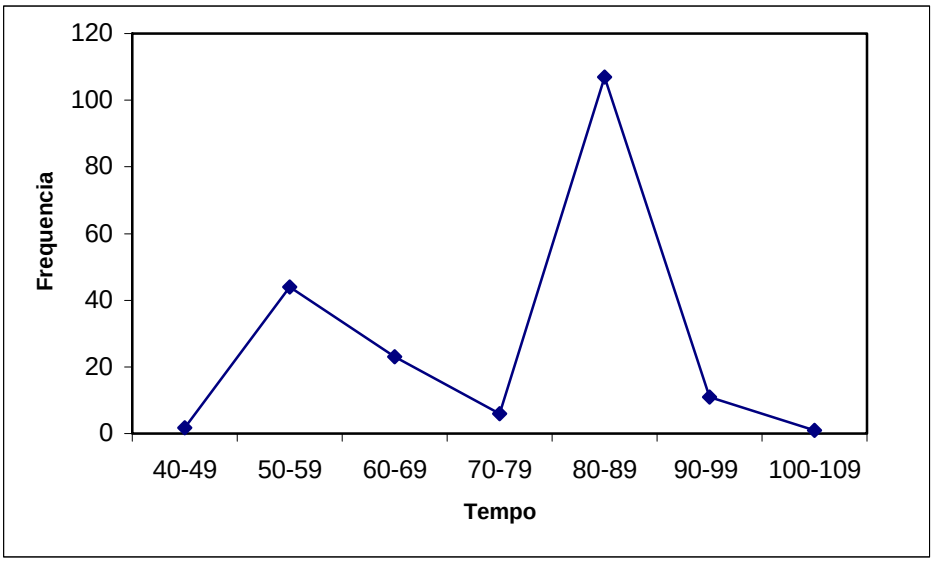


Fig 2.22

b) Se um guia turístico deseja garantir que seus turistas presenciem o facto, qual será o tempo mínimo que devem permanecer no parque?

Resposta:
O tempo mínimo a permanecer no parque será de 40 minutos conforme os dados da tabela acima.

c) Qual é a informação que lhe é sugerida pela mediana dos dados da tabela acima?

Resolução: Começa-se por calcular a mediana segundo a fórmula: $M_e = l_1 + \frac{\frac{n}{2} - \sum f_1}{f_{med}} \times c$, a classe mediana é calculada por $\frac{n}{2} = \frac{200}{2} = 100$. Procuramos na tabela onde aparece o valor 100 na tabela acima. É fácil ver que será na classe já marcada, que possui $l_1 = 80$,

$\sum f_1 = 8+44+23+6 = 81 \quad f_{med} = 107 \quad \lambda = 109-40 = 69 \quad k = 7 \quad c = \frac{\lambda}{k} = \frac{69}{7} = 9{,}9$ então:

$M_e = l_1 + \frac{\frac{n}{2} - \sum f_1}{f_{med}} \times c = 81{,}76$. Assim podemos afrmar que 81 presenças do Leopardo acontecem antes de 81,76 minutos e 12 presenças depois dos 81,76 minutos

3- Os dados a seguir dão as velocidades de carros cujos motoristas foram multados pela polícia duma cidade no percurso entre duas avenidas. Esses motoristas estavam dirigindo num trecho da zona de 30 Km/h.
a) Construa um histograma para esta tabela de frequências.

Velocidade	Frequência
[42-44[	14
[44-46[	11
[46-48[	8
[48-50[	6
[50-52[	4
[52-54[	3
[54-56[	1
[56-58[	2
[58-60[	1

Resolução:

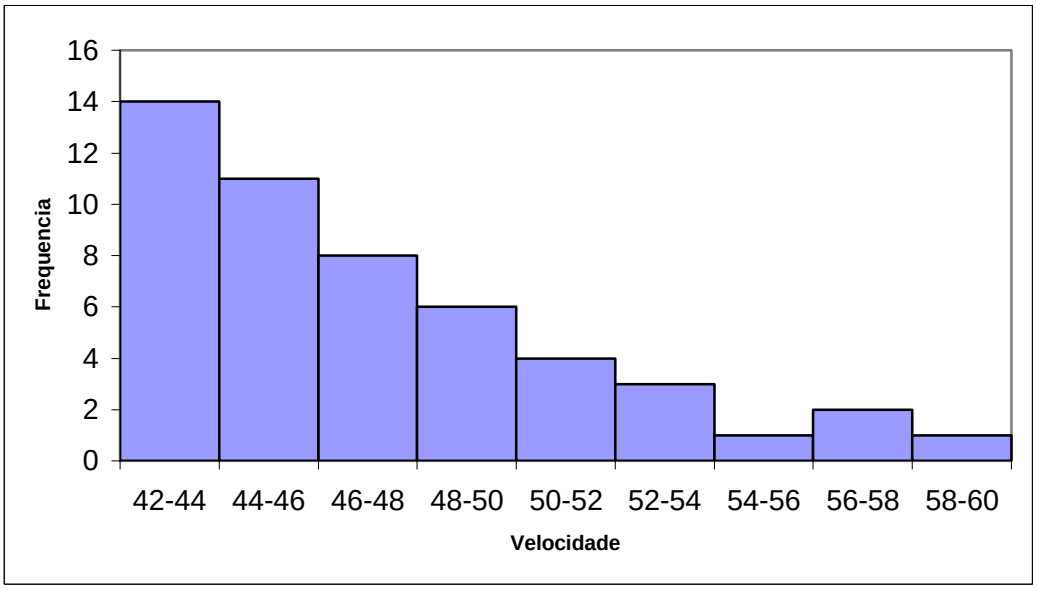


Fig 2.23

b) O que esta distribuição sugere sobre o limite fixado comparado com o limite de velocidade constatado?

Resposta:

A partida todos multados andavam a uma velocidade entre os 42Km/h e os 60Km/h. Sendo assim, pode-se dizer claramente que todos eles estavam num excesso de velocidade pois estavam sempre acima da velocidade normal de 30 Km/h.

4- Indique uma vantagem que existe de efectuar um diagrama de ramo-e-folhas em relação ao histograma

Resolução: Num diagrama de ramo-e-folhas não perdemos (ou perdemos pouca) informação sobre os dados do que num histograma, em virtude de que os dados para construir um histograma provém de classes cujos pontos médios são obtidos de resultados arredondados.

5- Dadas as tabelas a seguir:

Preços de Cebola por Kilo e por Mês entre Janeiro 2001 e Dezembro 2003

Mês	Preço	Mês	Preço	Mês	Preço
1	7729	13	11721	25	13019
2	7706	14	11212	26	13173
3	8106	15	11308	27	12308
4	9808	16	11462	28	13404
5	11019	17	11269	29	16954
6	10846	18	14500	30	17202
7	10529	19	14404	31	15433
8	10135	20	13635	32	14527
9	9750	21	14481	33	13590
10	10471	22	13725	34	13877
11	11000	23	12731	35	13748
12	12135	24	13981	36	15596

Preços de Tomate por Kilo e por Mês entre Janeiro 2001 e Dezembro 2003

Mês	Preço	Mês	Preço	Mês	Preço
1	8298	13	12140	25	9401
2	7648	14	10629	26	12148
3	7722	15	9805	27	16227
4	8021	16	9536	28	17242
5	8412	17	8651	29	15071
6	8156	18	7308	30	10321
7	8669	19	7170	31	9860
8	8214	20	6593	32	9725
9	5776	21	6092	33	10259
10	6732	22	6546	34	10095
11	9200	23	8978	35	10080
12	12560	24	12087	36	12201

a) Determine a equação de tendência para as duas tabelas usando o método de mínimos quadrados.
Resolução: Deve-se determinar a melhor curva que ajuste oa pontos de cada produto e, sendo assim terá: y=8850,67+191,98x para cebola e y=7481,78+120,91x para tomate

b) Calcule o coeficiente de determinação para o caso em a).
Resolução: r^2=0,739

6- Uma empresa financeira decidiu atribuir subsídio de isenção de horário a todos trabalhadores para melhor responder à demanda dos seus clientes. Após negociações entre a Administração e os sindicatos, fixaram-se os valores indicados na tabela seguinte:

Grau do Trabalhador	Subsídio em 1000,00Mt	Quantidade de Trabalhadores (y_i)
Servente	500 - 700	40
Caixas	700 - 1000	60
Técnicos	1000 - 1400	24
Consultores	1400 - 1900	70
Chefes de Serviço	1900 - 2500	17
Directores	2500 - 4000	8
Administradores	4000 - 11000	3

Determine a moda

Grau do Trabalhador	Subsídio em 1000,00Mt	Quantidade de Trabalhadores (y_i)	Ponto Médio (X_i)	Pesos de King
Servente	500 - 700	40	600	0,200000
Caixas	700 - 1000	60	850	0,200000
Técnicos	1000 - 1400	24	1200	0,060000
Consultores	1400 - 1900	70	1650	0,140000
Chefes de Serviço	1900 - 2500	17	2200	0,028333
Directores	2500 - 4000	8	3250	0,005333
Administradores	4000 - 11000	3	7500	0,000429

Resolução: O peso é dado por $\frac{f_i}{c_i}$. A classe com maior peso é denominada de classe modal.
Neste caso as classes dos Caixas e dos Serventes são as classes modais (isto é Bimodal). Usando a fórmula de King, porque os intervalos de classe não são iguais, teremos:

$$m_{Caixas} = l_i + \frac{\frac{f_{i+1}}{c_{i+1}}}{\frac{f_{i+1}}{c_{i+1}} + \frac{f_{i-1}}{c_{i-1}}} \times c_i = 700 + \frac{\frac{24}{400}}{\frac{24}{400} + \frac{40}{200}} \times 300 = 769{,}2\,.$$

$$m_{Serventes} = l_i + \frac{\frac{f_{i+1}}{c_{i+1}}}{\frac{f_{i+1}}{c_{i+1}} + \frac{f_{i-1}}{c_{i-1}}} \times c_i = 500 + \frac{\frac{60}{300}}{\frac{60}{300} + 0} \times 200 = 700$$

Se um trabalhador fôr elevado ao grau de Consultor, qual será o seu subsídio estimado assumindo que passarão a ser 71 consultores?

Resolução:

Grau do Trabalhador	Ponto Médio (X_i)-subsídio (em 1000,00Mt)	Quant (y_i)	Xy	y_i^2
Servente	600	40	24000	1600
Caixas	850	60	51000	3600
Técnicos	1200	24	28800	576
Consultores	1650	70	115500	4900
Chefes de Serviço	2200	17	37400	289
Directores	3250	8	26000	64
Administradores	7500	3	22500	9
$\sum$	17250	222	305200	11038

$$b_o = \frac{\sum X \sum y^2 - \sum y \sum Xy}{N\sum y^2 - \left(\sum y\right)^2} = \frac{17250 \times 11038 - 222 \times 305200}{7 \times 11038 - 222^2} = 4383{,}214$$

$$b_1 = \frac{N\sum Xy - \sum y \sum X}{N\sum y^2 - \left(\sum y\right)^2} = \frac{7 \times 305200 - 222 \times 17250}{7 \times 11038 - 222^2} = -60{,}5068$$

E. ntão, $X = 4383{,}214 - 60{,}507y$. *C*om $y = 71$ *vem* $X = 87{,}217$. O subsídio estimado será de 87.217,00Mt.

Exercícios Propostos

1- Com as tabelas tabelas dadas em 5) dos exercícios resolvidos, determine:
A sazonalidade.

a) Faça uma estimativa do custo da cebola para o mês de Dezembro de 2004.
b) Construa um índice para o preço de tomate, no príodo de 2001 e 2003, tendo como base Janeiro de 2002 e interprete o seu valor.

2- Com base em Dezembro de 1998 o Instituto Nacional de Estatística lançou na sua página (internet) os dados constantes na tabela, sobre o índice de preço ao consumidor na cidade de Maputo

Inflação Mensal (%) a/ */Monthly inflation (%) a/*

Ano/*Year*	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
1996	4.4	10.4	1.6	0.6	-3.5	0.2	0.4	0.1	0	0.3	1.6	0
1997	4.1	3.4	-0.7	-1.1	-1.7	-0.3	-0.4	-0.3	-0.6	0.4	1.3	1.7

a) Faça o diagrama de dispersão entre índices de 1996 e 1997.
b) Haverá alguma relação entre esses preços? Se sim, determine a equação da recta de regressão e desenhe a respectiva curva.

c) Indique os pontos que são resíduos.

3- Qual é o interesse na estimação da recta de regressão?

a) *Fazer previsões;*
b) *Saber se existe correlação entre as variáveis;*
c) *Determinar o coeficiente de correlação;*
d) *Determinar a média das variáveis.*

Assinale a opção correcta.

4- Taxas específicas de analfabetismo por sexo segundo área de residência e idade, Província de Nampula, 1997 (por cada 1000 pessoas entrevistadas)

Grupos de idade	Taxas de analfabetismo (%)		
	Total	Homens	Mulheres
Total	71.7	56.7	85.9
15-19	66.3	54.4	77.7
20-24	69.0	53.9	81.2
25-29	68.0	50.2	83.3
30-39	67.4	47.2	87.2
40-49	76.1	59.1	93.4
50-59	83.0	70.9	95.9
60 +	88.6	82.4	96.6

a) Determine graficamente o intervalo de idades com maior taxa de analfabetismo
b) Será que o analfabetismo reinante nos homens condiciona (tem alguma relação com) o analfabetismo das mulheres?
c) Faça o respectivo diagrama de dispersão entre homens e mulheres
d) Determine o coeficiente de correlação e se houver alguma relação, faça a respectiva recta de regressão.
e) Indique os pares que formam resíduos dessa distribuição

5- Os dados abaixo referem-se à dureza de 30 peças de alumínio

53,0	70,2	84,3	69,5	77,8	87,5	53,4	82,5	67,3	54,1
70,5	71,4	95,4	51,1	74,4	55,7	63,5	85,8	53,5	64,3
82,7	78,5	55,7	69,1	72,3	59,5	55,3	73,0	52,4	50,7

1- Faça o respectivo diagrama de ramo e folhas.

Dica: Opte por cortar valor, omitindo as décimas, de modo que 69,1 e 69,5, por exemplo, tornem-se 69 e 69 e aparecem como 9 na linha que corresponde ao ramo 6.

2- O Gráfico a seguir representa o aproveitamento pedagógico a Estatística, tendo em conta a seguinte situação, "se teve ou não matemática no ensino secundário geral"

a) Como se denomina este tipo de gráfico?
b) Indique a nota mínima e a máxima para cada grupo.
c) Descreva algumas conclusões acerca da dispersão.

No grupo dos que tiveram matemática existem pontos acima e abaixo, como se denominam? Que influência têm sobre a média.

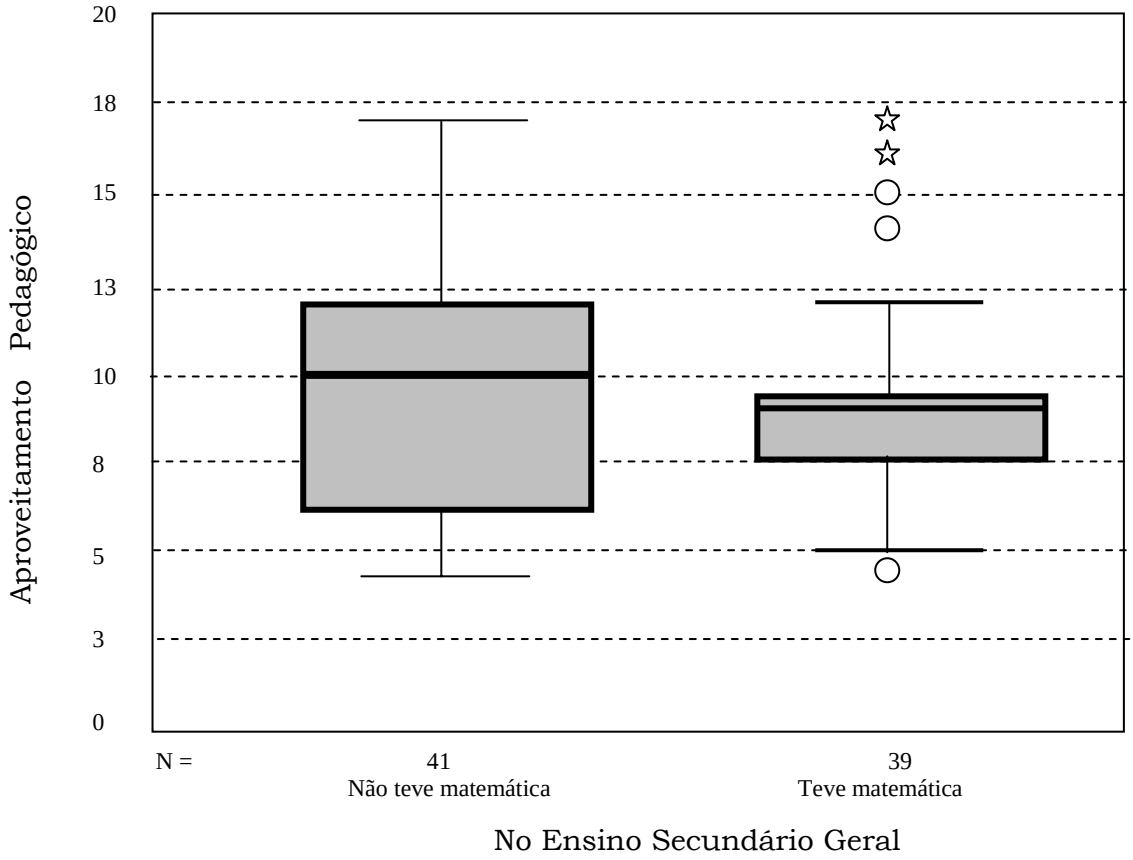


Fig 2.24

3- De um inquérito realizado a 20 famílias na pequena aldeia de Unango no Niassa, obtiveram-se os seguintes resultados, relativos ao rendimento médio mensal (em mil meticais)

125	106	98	107	162
113	125	110	140	145
100	91	130	128	126
96	98	145	130	160

a) Construa um quadro de distribuição de frequências, utilizando classes de amplitude constante, com base no mesmo calcule a média, mediana e moda e compare com os mesmos para cálculos sem agrupamento.
b) Faça o histograma e o polígono de frequência. Realize algumas leituras dos mesmos dados a partir do Histograma ou polígono de frequências sem reparar na fonte dos dados.
c) Determine a variabilidade dos dados relactivamente à média. O que lhe sugere a variância que acabou de calcular?
d) Determine o desvio padrão. Determine $X \pm 3S$ e $X \pm 2S$. Realize alguma leitura relativa a esses cálculos, interpretando-os relativamente aos dados. Desenhe o gráfico de frequências para $X \pm 2S$.

4- A antiguidade dos trabalhadores numa empresa constitui uma variável importante na análise e formulação de políticas de pessoal. Numa Instituição Bancária em 1996, a distribuição dos efectivos por sexo e por escalões de antiguidade era a seguinte:

Antiguidade (Anos)	Homens	Mulheres	Total
0 – 2	210	110	320
2 – 6	212	242	454
6 – 11	212	152	364
11 – 16	432	216	648
16 – 21	315	129	444
21 – 26	278	119	397
≥ 26	360	42	402
Total	2019	1010	3029

a) Determine a antiguidade média dos efectivos totais.
b) Determine a antiguidade mais frequente nas mulheres.
c) Elabore o histograma e respectivo polígono de frequências respeitante à distribuição das mulheres
d) Suponha agora que o banco decide admitir no seu quadro de pessoal um novo funcionário com uma carreira bancária de 12 anos. Determine em que percentil se localizaria este novo funcionário na presente estrutura de antiguidade, referente ao sector masculino, explicitando o respectivo significado.

5- O Transporte público e o transporte semicolectivo, vulgo "chapa 100" são dois meios que um professor pode usar para ir ao trabalho diariamente. Amostras de tempo para cada meio de transporte estão registradas a seguir. Os tempos estão em minutos.

Chapa 100 *28 29 32 37 33 25 29 32 41 34*
Transporte público *29 31 33 32 34 30 31 32 35 33*
Que meio de transporte deve ser preferido? Explique.

6- Os gráficos seguintes mostram a mesma informação. No entanto, eles apresentam uma imagem diferente. Um deles foi apresentado pela administração e outro pelo delegado sindical de uma empresa numa reunião de renovação do contrato salarial.

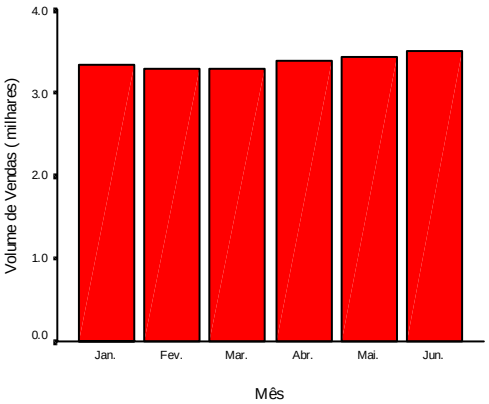


Fig 2.24

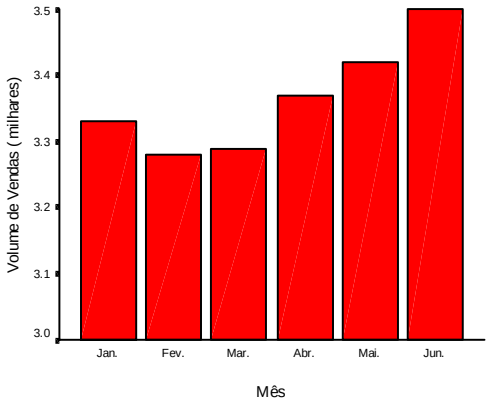


Fig 2.25

a) Diga qual dos gráficos teria apresentado o delegado sindical para pedir aumento salarial. Justifique.
b) Qual foi a estratégia usada para dar a mesma informação de uma forma aparentemente tão diferente?

7- Contou-se o número de carros que passam pela Avenida OUA perto do Hospital José Macamo em Maputo, na hora de ponta durante 50 horas consecutivas, obtendo-se os resultados abaixo:

9	8	13	9	13	11	15	10	10	9
12	10	7	10	8	14	17	15	12	14
13	12	6	14	10	17	16	14	16	13
7	11	8	13	11	15	14	17	17	7
11	14	10	12	11	13	12	13	9	8

a) Represente os dados num diagrama de pontos
b) Faça um histograma e um diagrama ramo-e-folhas

8- Dadas alturas (em cm) de estudantes duma turma, construir um diagrama em caixa e analisar os resultados: 170, 182, 152, 162, 178, 192, 154, 156, 158, 176, 174, 172, 179, 174, 173, 184, 176, 168, 197, 173, 185, 149, 157, 167, 180, 180, 170, 167, 191.

CAPÍTULO 3. PROBABILIDADES

Objectivos do capítulo:

- Definir probabilidade.
- Descrever as abordagens clássica, da frequência relativa e subjectiva da probabilidade.
- Explicar os termos experimento, espaço amostral e evento.
- Definir os termos probabilidade condicional e probabilidade conjunta
- Calcular probabilidades aplicando as regras da adição e da multiplicação.
- Determinar o número de possíveis permutações e combinações.
- Calcular uma probabilidade usando o Teorema de Bayes.

1. HISTÓRIA E SURGIMENTO DA PROBABILIDADE

A Teoria de probabilidade é uma invenção surgida a partir dos jogos de azar no Mónaco. Foi a necessidade de tentar prever alguma tentativa em que o apostador poderia ter prémio. Niccollo Fontana (1500-1557), Matemático Italiano, iniciou o estudo no séc XVI. Teve seguidores casos dos Franceses Pierre Fermat (1601-1665) e Blaise Pascal (1623-1662).
A Teoria de Probabilidades é um ramo da Matemática extremamente útil para o estudo e investigação das regularidades dos chamados fenómenos dum mero acaso, denominados *Aleatórios*.
Atendamos a seguinte curiosidade para poder-se entender melhor o fenómeno de uma experiência aleatória.

Curiosidade sobre teoria de probabilidade

John Kerrich, um matemático sul africano estava visitando Copenhague quando a Segunda Guerra Mundial começou. Dois dias antes de seu voo marcado para a Inglaterra, os alemães invadiram a Dinamarca. Kerrich passou o resto da guerra internado num acampamento em Jutland e para passar o tempo ele levou a cabo uma série de experimentos em teoria da probabilidade. Num desses experimentos, lançou uma moeda 10.000 vezes. Seus resultados são mostrados no gráfico a seguir.

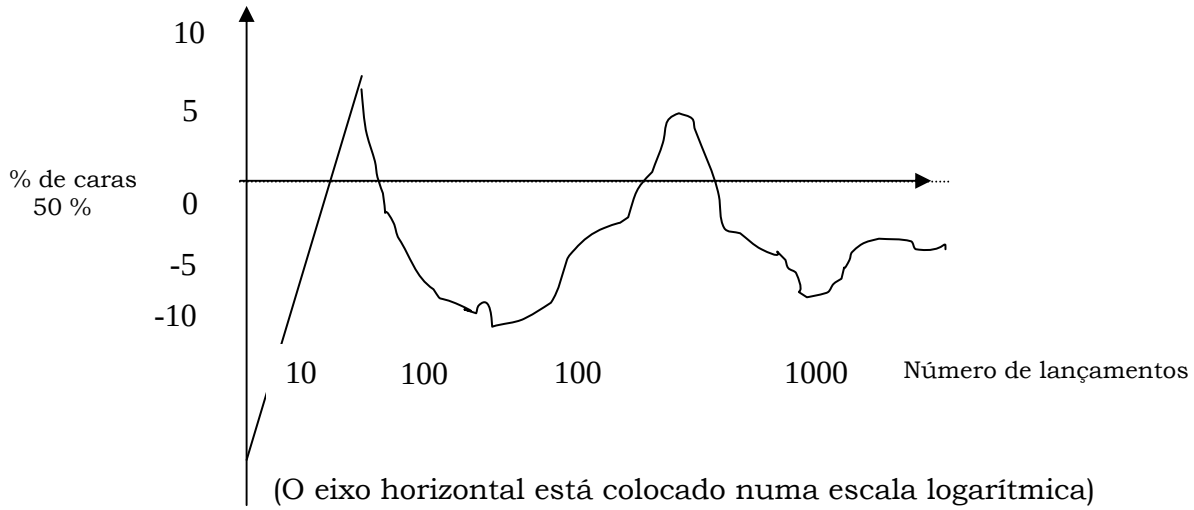


Fig 3.1

O lançamento de uma moeda 10 vezes é um exemplo de um experimento aleatório. A maioria dos experimentos está sujeito a Variação Aleatória. A Teoria de probabilidade é a aproximação matemática que busca quantificar em temos de modelos o que ocorre com estes experimentos.

Definições:

Probabilidade: é uma medida de possibilidade de ocorrência de um determinado evento; ela pode assumir um valor entre 0 e 1.

Evento: É uma colecção de um ou mais resultados de um experimento

Experimento é uma experiência cujos resultados são imprevisíveis.

Exemplo dum Experimento: lançar uma moeda duas vezes, podendo ter como resultados possíveis (espaço amostral) { KK, KC, CK, CC }, onde C – Saída de Cara e K-Saída de Coroa

Evento 2: No mínimo uma cara = {CC, CK, KC}

Exemplo: Um jogador de futebol quando bate uma penalidade, os resultados possíveis são "marcar" ou "não marcar". Em cada tentativa não é possível prever o resultado que se irá conseguir, embora ele seja determinado por causas perfeitamente bem definidas. Entre as diversas causas se destacam o poderio do remate, o tipo de guarda redes, a simulação do marcador, etc. Pode-se ver que isto resulta numa diversidade de parâmetros que podemos controlar mas que leva-nos ao resultado imprevisível, *Se vai ser golo ou não*.

O exemplo descrito faz parte de diversas experiências que podem ser feitas, tendo o imprevisivel resultado dentro dos casos possíveis.

Experiência Processo ou conjunto de causas e processamentos que podem produzir resultados observáveis. Se essa experiência está sujeita a várias situações e que os resultados são um mero acaso, diz que ela é Aleatória.

Voltando ao exemplo atrás, aparentemente pode parecer que nele existe alguma regularidade. Mas se o número de observações for elevado, alguma regularidade surge. O número de penaltes marcados é aproximadamente cerca de 95% do que os falhados (defendidos, que tocaram na barra ou que foram ao lado da baliza) 5%.

Podemos afirmar que as experiências aleatórias caracterizam-se por:

- Poderem repetir-se um grande número de vezes nas mesmas condições ou em condições muito semelhantes;
- Cada vez que a experiência é executada obtém-se um resultado, mas que não é possível de ser previsto antes.
- Os resultados dessas experiências individuais são irregulares, mas que analisados na globalidade ao longo de muito tempo (muitas experiências), parecem ter uma certa regularidade estatística quando tomados em conjunto.

Cumprindo-se *os três pressupostos* dizemos que estamos perante um **Experimento**.

Se está a trabalhar num computador e lhe dizem que dentro de dez minutos não haverá energia eléctrica, já pode prever o resultado de que o computador irá desligar.
A teoria de probabilidades opera sobre os fenómenos aleatórios e na maior parte dos casos sobre experimentos.

Será que o jogar citado no exemplo acima, ao chutar o penalte, vai mesmo marcar?
Neste caso recorremos à teoria de probabilidade, pois ela tenta dar significado a experimentos tais que o resultado não pode ser completamente pré-determinado. *Atente para o facto de que pré determinado não significa pré definido.*

Calcular a probabilidade é medir a incerteza ou associar um grau de confiança aos resultados possíveis. Por exemplo, escolha uma carta qualquer num baralho depois de ter sido bem embaralhado, para garantir que todas as cartas terão a mesma possibilidade de serem seleccionadas. O que é mais provável, sair uma figura (Q, K, J) ou sair o dois de espadas?

Entendamos evento como um acontecimento.
As probabilidades associam às possíveis combinações dos resultados, que chamamos de *eventos*, um valor entre 0 e 1. Quanto maior o valor, maior a certeza de sua possibilidade de ocorrência.

2. DEFINIÇÃO AXIOMÁTICA DE PROBABILIDADE

Se A_i for um evento qualquer e S for espaço de resultados (conjunto universal), então:

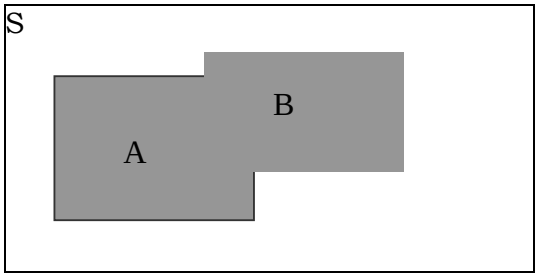
a) $P(A_i) \geq 0$

b) $P(S) = 1$

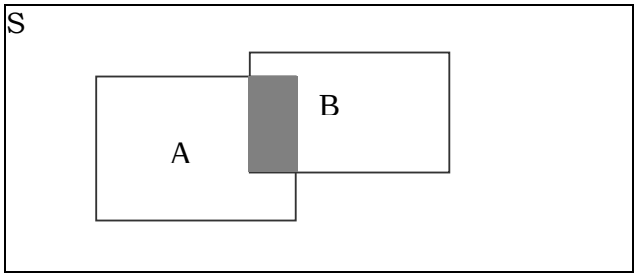
c) $\sum_{i=1}^{n} P(A_i) = P(\bigcup_i Ai),\ \cap Ai = \emptyset$

Operações com eventos: Sejam A e B dois eventos associados a um espaço amostral S.

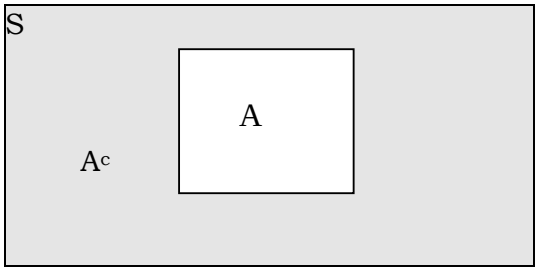
União $A \cup B$: só é verdadeiro se pelo menos um dos eventos ocorre



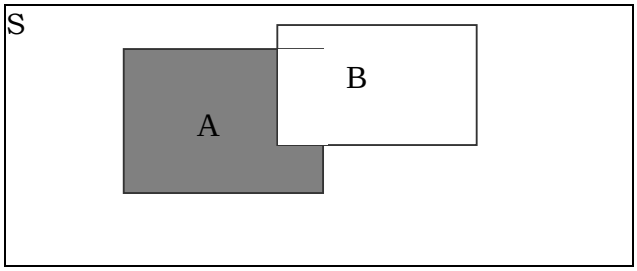
Interseção $A \cap B$: quando os dois Eventos coexistem



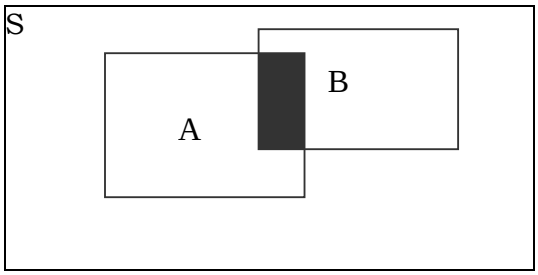
Complemento de A (A^c ou $\overline{A}$): ocorre quando não ocorre A



Diferença A-B: quando ocorre A mas não ocorre B



Diferença simétrica A ∴ B: ocorre apenas um dois evento



Eventos mutuamente exclusivos: quando a interseção deles é o evento impossível

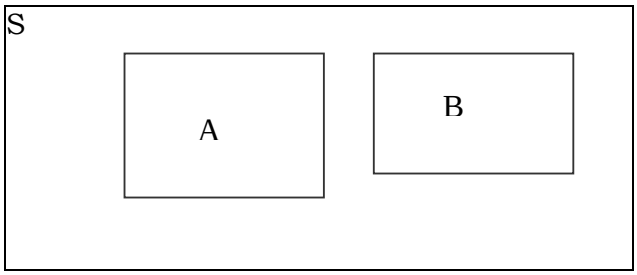


Fig 3.2

Exemplo: Consideremos o lançamento dum dado. Um dado possui seis faces numeradas de 1 a 6 nesse lançamento pode sair uma face par. Acontecimento é qualquer um dos subconjuntos do conjunto que contém todos resultados possíveis da experiência aleatória
É evento, subconjunto do *conjunto que contém todos os resultados possíveis do experimento aleatório.*

Denotemos por S tal conjunto, a que chamamos de espaço amostral.

Considere o experimento aleatório que consiste no lançamento de um dado. Um dado contém seis faces numeradas de 1 a 6. Então S pode ser escrito como S = { 1, 2, 3, 4, 5, 6 } e alguns possíveis eventos poderiam ser:

A = { face par } = { 2,4 ,6 } - caso em que num lançamento saia uma face par.

B = { face maior ou igual a 5 } = { 5, 6 } - caso em que num lançamento saia uma face maior ou igual a 5.

C = { face maior ou igual a 7 } = $\emptyset$ = conjunto vazio.

Definições: Tomemos o lançamento dum dado. Um dado contém seis faces numeradas de 1 a 6. Neste caso o espaço de resultados é S={1,2,3,4,5,6}. Vejamos as definições que se seguem:

Dois eventos (acontecimentos) quaisquer **dizem-se mutuamente exclusivos** se possuem intersecção vazia, são incompatíveis, isto é, eles não podem ocorrer simultaneamente.
Exemplo:

A = { face par } = { 2,4 ,6 }
D = { face impar} = { 1,3,5} $A \cap D = \emptyset$

Os subconjuntos de S designam-se por acontecimentos; Os subconjuntos formados por um único elemento chamam-se acontecimentos elementares.

Exemplos: $E = \{1\}$; $F = \{2\}$; $G = \{3\}$; $H = \{4\}$; $I = \{5\}$; $J = \{6\}$

O acontecimento que contém nenhum elemento de S, chama-se de acontecimento impossível
Exemplo: N = { 7,8 }

A reunião de dois acontecimentos A e B é o acontecimento que se realiza se, e somente se, A ou B se realizam. O que significa que A se realiza ou B se realiza ou ainda ambos se realizam. Abrevia-se por $A \cup B$ ou $A + B$ e é formado pelos elementos que vem de A e os que vem de B. Alerte-se para o facto de que um elemento que coexiste nos dois conjuntos, deve ser escrito (participar) uma única vez.

Exemplo:
A = { face par } = { 2,4 ,6 }
B = { face maior ou igual a 5 } = { 5, 6 } $A \cup B = \{2,4,5,6\}$

A intersecção dos acontecimentos A e B é o acontecimento que se realiza se, e somente se, A e B se realizam conjuntamente. Representa-se por $A \cap B$ ou AB, e é formado por elementos que comumente existem em A e em B.

Exemplo:

A = { face par } = { 2,4 ,6 }
B = { face maior ou igual a 5 } = { 5, 6 } $A \cap B = \{6\}$

Diz-se que Q é um acontecimento complementar de A se é formado por todos elementos do espaço amostral excepto os existentes em A. Escreve-se $Q = S \setminus A$. Para o nosso exemplo o $Q = \{1,3,5\} \equiv D$. Também pode-se designar de $\overline{A}$ ou A^c ao complementar de A.

Na lógica matemática, segundo as **leis de Morgan**, definem-se seguintes axiomas para conjuntos:
Se A é um conjunto qualquer e $\overline{A}$ *ou* A^c definido como sendo complementar do conjunto A

1. $\overline{\overline{A}} = A$
2. $(A \cup B)^C = A^c \cap B^c$
3. $(A \cap B)^C = A^c \cup B^c$

Uma das principais preocupações da estatística é inferir os parâmetros populacionais, baseando-se nos resultados observados/calculados duma amostra. Quando uma amostra é seleccionada aleatoriamente não podemos determinar, ou prever a priori, os resultados (experimento aleatório), como atrás fora descrito. Contudo, podemos construir modelos probabilísticos que permitem calcular as chances de ocorrência dos possíveis resultados, através da teoria de probabilidades.

Suponhamos que esteja interessado em conhecer a possibilidade relativa teórica de sair cara no experimento lançar "n" vezes uma moeda não viciada. Existem duas formas de abordar o problema, uma através da experimentação e a outra através de um modelo probabilístico.

Através duma experiência, observamos a frequência relativa com que cara aparece nos "n" lançamentos. Se repetirmos o experimento teremos outra frequência relativa observada, que não é necessariamente igual à anterior, mas esperamos que esteja muito próximo dela. Assim, se repetirmos várias vezes os "n" lançamentos, esperamos que as frequências observadas convirjam para um número chamado probabilidade. Buffon e Pearson (um dos grandes estudiosos da correlação entre duas variáveis) realizaram esse tipo de experimento com os seguintes resultados:

Estimativa da probabilidade através das frequências observadas

Tabela 3.1- Frequências observadas de Buffon e Pearson

	Buffon		Pearson	
Possíveis resultados	Frequência Absoluta	Frequência Relativa	Frequência Absoluta	Frequência Relativa
Cara	2048	0,5069	12012	0,5005
Coroa	1992	0,4931	11988	0,4995
Total	4040	1,0000	24000	1,0000

Outra forma de se chegar à frequência relativa teórica igual da tabela atrás, é através da construção de um modelo probabilístico teórico sob certas suposições adequadas. Assim, no exemplo, sabemos que existem somente dois possíveis resultados: cara ou coroa, sendo que

as duas faces tem as mesmas chances de ocorrer. Então, a frequência relativa teórica para a ocorrência de cada resultado é ½ .

Tabela 3.2- Resumo das frequências observadas de Buffon e Pearson

Possíveis resultados	cara	coroa	total
Frequência teórica	½	½	1

Este modelo representa de forma adequada o resultado do experimento e, quando falamos de probabilidades da ocorrência dos possíveis resultados do experimento, referimo-nos às chances teóricas deles acontecerem.

Afinal como é que uma probabilidade é expressa?
Uma probabilidade é expressa como um número decimal, tal como 0,70 ; 0,27 ; ou 0,50. Entretanto ela pode ser representada como uma percentagem tal com 70 %, 27 % ou 50 %. O valor de uma probabilidade está localizado no intervalo de número reais que vai de 0 a 1, inclusive as extremidades deste intervalo.

Quando uma probabilidade é 0, o evento a ela associado é improvável de ocorrer o que se denomina por **Evento Impossível**.

Quando uma probabilidade é 1, o evento a ela associado é mais provável de ocorrer e é denominado de **Evento Exacto**.

Definição: Dado um evento A associado a um experimento aleatório com espaço amostral S a probabilidade do evento A, denotada por P (A) é o quociente :

$$P(A) = \frac{n}{N} = \frac{n^o \, de \; resultados \;\; favoráveis}{n^o \, de \; resultados \;\; possíveis} \; .$$

Voltemos para o caso do lançamento de dado e que nos interessemos somente pelo aparecimento da face par. O evento A = { face par } = { 2,4 ,6 }, terá como probabilidade $P(A) = \frac{n(A)}{N} = \frac{3}{6} = \frac{1}{2}$, Isto é, número de resultados favoráveis (ao evento A) dividido pelo número de resultados possíveis no conjunto S.

Definições

1) Pela noção Frequencista, a probabilidade de um acontecimento é o valor para o qual tende a frequência relativa do acontecimento, quando o número de repetições da experiência aumenta.

3. ANÁLISE COMBINATÓRIA

3.1 Factorial de um número

Seja n um número inteiro não negativo. Define-se factorial de n (indicado pelo símbolo n!) como sendo: $n! = n \times (n-1) \times (n-2) \times (n-3) \times ... \times 3 \times 2 \times 1$ e lê-se n factorial

Casos Especiais
0! = 1 1! = 1

Exemplos:
a) $6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$, lê-se seis factorial
b) $4! = 4 \times 3 \times 2 \times 1 = 24$, lê-se quatro factorial
c) Repare ainda que $6! = 6 \times 5 \times 4!$, lê-se seis factorial igual a seis vezes cinco vezes quatro Factorial.

3.2 Princípio fundamental da contagem

Se uma tarefa pode ser realizada em n etapas diferentes, e se a primeira etapa pode ocorrer de k_1 maneiras diferentes, a segunda de k_2 maneiras diferentes, e assim sucessivamente , então o número total T de maneiras de realização da tarefa é dado por: $T=k_1.k_2.k_3...k_n$.

Exemplo:
O INAV (Instituto Nacional de Viação) tem como norma mandar colocar placas de matrículas dos veículos automóveis usando-se 3 letras do alfabeto (sendo a primeira letra M) e 4 algarismos. Qual o número máximo de veículos que poderão ser licenciados?

Resolução:

Usando o Princípio fundamental da contagem, podemos imaginar uma placa do tipo MLN 10 15. Como o alfabeto possui 26 letras e nosso sistema numérico possui 10 algarismos (de 0 a 9), podemos concluir que: para a 1ª posição, temos 1 (uma) alternativa que é a letra M, como pode haver repetição, para a 2ª, e 3ª também, estas duas terão 26 alternativas cada. Com relação aos algarismos, concluímos facilmente que temos 10 alternativas para cada um dos 4 lugares. Podemos então afirmar que o número total de veículos que podem ser licenciados será igual a: 1.26.26.10.10.10.10 que resulta em 6760000. Observe que se no país existissem 6760001 veículos, todos com matricula nacional, o sistema de códigos de emplacamento teria que ser modificado, já que não existiriam números suficientes para codificar todos os veículos.

3.3 Permutações simples

Permutações simples de n elementos distintos são agrupamentos formados por todos os n elementos tomados n a n e que diferem na ordem de sua colocação. Neste caso o número total de permutações simples de n elementos distintos é dado por n!, e lê-se, n factorial.

Exemplo 1: Com os elementos a, b, c de um conjunto qualquer, são possíveis as seguintes permutações: abc, acb, bac, bca, cab e cba.

Exemplo 2: De quantas maneiras 5 pessoas podem se sentar num banco rectangular. $P_5 = 5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$

Permutações com elementos repetidos

Se entre os n elementos de um agrupamento, existem k elementos repetidos com uma característica, t elementos repetidos com outra característica e s elementos repetidos com outra característica e assim sucessivamente, o número total de permutações que podemos formar é dado por:

$$P_n^{(k,t,s,\ldots)} = \frac{n!}{k!t!s!\ldots}$$

Exemplo:
Imagine que as marcas de carros viessem da combinação de um conjunto de letras de uma palavra pré definida. Quantas marcas de viaturas podíamos ter a partir da palavra correspondente ?

Solução: Temos 14 elementos, com repetição. Observe que a letra o aparece duas vezes, a letra r duas, a letra e três vezes e a letra n duas vezes. Na fórmula anterior, teremos: n=14, o=2, r=2, e=3 e n=2: a resposta seria $\frac{14!}{2!2!3!2!} = 1816214400$

Resposta: Podemos ter 1816214400 marcas de carros diferentes.

3.4 Arranjos simples

Dado um conjunto com n elementos, chama-se arranjo simples tomados k a k, a todo agrupamento de k elementos diferentes dispostos numa certa ordem. Especial atenção deve existir para evitar perigo de confusão, porque dois arranjos diferem entre sí, pela ordem de colocação dos elementos. Assim, no conjunto E = {a,b,c}, teremos:

a) Arranjos de três elementos tomados dois a dois: ab, ac, bc, ba, ca, cb.
b) Arranjos de três elementos tomados três a três: abc, acb, bac, bca, cab, cba.

Arranjos simples são representados pela fórmula: $A_k^n = \frac{n!}{(n-K)!}$. Lê-se: Arranjos de n, k a k.

É fácil perceber que $A_n^n = \frac{n!}{(n-n)!} = n! = P_n$

Exemplo:
Um cartão multibanco possui um número de 16 dígitos. O segredo (PIN) do cartão é marcado por uma sequência de 4 dígitos distintos. Se uma pessoa tentar usar o cartão, quantas tentativas deverá fazer (no máximo) para conseguir fazer levantamento, se a norma fosse de cada pessoa escolher 4 dígitos para usar como PIN dentre os constantes como

número do cartão?

Resolução:

As sequências serão do tipo xypz. Para a primeira posição teremos 16 alternativas, para a segunda 15, para a terceira 14 e para a última 13. Podemos aplicar a fórmula de arranjos, mas pelo princípio fundamental de contagem, chegaremos ao mesmo resultado: $16\times15\times14\times13=43680$. $A_4^{16}=\frac{16!}{(16-4)!}=\frac{16\times15\times14\times13\times12!}{12!}=16\times15\times14\times13=43680$

3.5 Combinações simples

Denominamos combinações simples de n elementos distintos tomados k a k aos subconjuntos formados por k elementos distintos escolhidos entre os n elementos dados. A diferença entre arranjos e combinação reside no facto de nos arranjos a ordem é importante enquanto que nas combinações a ordem não é importante.

Exemplo

Dado um conjunto A={a,b.c,d} podemos considerar: a) combinações dos 4 elementos tomados 2 a 2, o seguinte ab, ac, ad, bc, bd, cd. b) combinações dos quatro elementos tomados 3 a 3, os seguintes: abc, abd, acd, bcd. c) combinações dos 4 elementos tomados 4 a 4: abcd.

Analíticamente as combinações são representadas por: $C_k^n=\frac{n!}{k!(n-k)!}$, que se lê, combinação de n, k a k. A fórmula de combinações é mais conhecida como sendo Número binomial e indicado por: $\binom{n}{k}=\frac{n!}{k!(n-k)!}$ e que será usado nas Distribuições de Probabilidade teóricas discretas (Binomial e Hipergeométrica)

Resumo:

Numa permutação, todos os elementos do conjunto devem fazer parte. Num arranjo toma-se uma parte dos elementos do conjunto para formar agrupamentos; quando essa parte dos elementos é o todo (o universo), o arranjo se transforma em permutação. Num arranjo a ordem dos elementos interessa, sendo que se no primeiro agrupamento um elemento estiver na primeira posição e no segundo agrupamento o mesmo elemento passar para uma outra posição diferente da primeira teremos um outro arranjo diferente do primeiro. Arranjos são seqüências, sendo por isso diferentes, desde que se altere a ordem. Uma combinação é um agrupamento onde a ordem não interessa, interessando somente os elementos envolvidos. Combinações são subconjuntos.

4. Algumas considerações sobre leis, axiomas e teoremas:

Se A e B, são dois eventos quaisquer de S, teremos:

1) $P(A_i)\geq 0$

2) $0 \le P(A) \le 1$

3) $\sum_{i=1}^{n} P(A_i) = 1$

4) $P(A \cup B) = P(A) + P(B)$, *se* $A \cap B = \emptyset$, *i.e, A e B são mutuamente exclusivos*

5) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, *se* $A \cap B \neq \emptyset$, *i.e, A e B não são mutuamente exclusivos*

6) $P(A \cap B) = P(A)P(B) \Leftrightarrow$ *A e B são independentes*

7) $P(A \cap B) = P(A)P(B/A)$ ou $P(A \cap B) = P(B)P(A/B) \Leftrightarrow$ *A e B são dependentes*

8) $P(A/B) = \dfrac{P(A \cap B)}{P(B)} \Leftrightarrow$ *A ocorre depois de B ter ocorrido*, define Probabilidade condicional

9) $P(S) = 1$

10) $P(\emptyset) = 0$

11) $P(\overline{A}) = 1 - P(A)$

Exemplos

Na zona do Bairro Canongole em Tete, existem duas padarias, produzindo pães de forma e arofadas, respectivamente. Nesta zona, que se dedica essencialmente à venda de cabritos, 9,8% dos residentes compram sempre pão forma, 22,9% compram sempre arofadas e 5,1% compram sempre ambos tipos de pães. Determine a pobabilidade de:

a) Uma pessoa comprar pelo menos um dos dois tipos de pães.
b) Uma pessoa comprar somente pão forma.
c) Uma pessoa não comprar nem pão forma nem arofadas.

Resolução

Designemos por A o acontecimento c*omprar pão forma* e B c*omprar arofadas*. A probabilidade de uma pessoa comprar pelo menos um dos dois tipos de pães será $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0{,}098 + 0{,}229 - 0{,}051 = 0{,}276$, visto que A e B não são mutuamnete exclusivos. A probabilidade de uma pessoa comprar somente pão forma é a probabilidade da pessoa comprar pão forma e não arofadas, ou seja $P(A \cap \overline{B}) = P(A) - P(A \cap B) = 0{,}098 - 0{,}051 = 0{,}047$. A probabilidade de uma pessoa não comprar nem pão forma nem arofadas, corresponde a não comprar pão forma nem comprar arofadas, logo $P(\overline{A} \cap \overline{B}) = P((A \cup B)^c) = 1 - P(A \cup B) = 1 - 0{,}276 = 0{,}724$.

5. Definição para acontecimentos independentes

A regra especial de multiplicação requer que dois eventos A, B, ... sejam <u>independentes</u>.
Já tinhamos visto atrás, onde figuram as diversas propriedades, para casos de dois eventos. E definimos que, dois eventos A e B dizem-se <u>independentes</u> se a ocorrência de um não tem efeito sobre a probabilidade de ocorrência do outro, isto é $P(A \cap B) = P(A) \times P(B)$.

Para três eventos independentes A,B e C, a regra especial da multiplicação usada para determinar a probabilidade de que todos os eventos ocorram é:
$P(A \cap B \cap C) = P(A) \times P(B) \times P(C)$

6. PROBABILIDADE CONDICIONAL

Exemplo: Um grupo de pessoas é classificado de acordo com o seu gosto pela bebida e a incidência de qualidades de violência que possa praticar. As proporções das diversas categorias, aparecem na tabela seguinte:

Tabela 3.3 Proporções de gosto pela bebida e a incidência de violência

	Bébedo	Bêbado	Não Bebe	Total
Violento	0,10	0,08	0,02	0,20
Não Violento	0,15	0,45	0,20	0,80
Total	0,25	0,53	0,22	1,00

a) Qual é a probabilidade de uma pessoa escolhida ao acaso seja violenta?
b) Qual é a probabilidade de uma pessoa bébeda ser violenta?

A **resposta** para a primeira questão é imediata. Basta ver a proporção de violentos dentro da população que é de 0,20.
Para responder à segunda questão há que tomar em atenção que o que se pretende é a proporção de violentos dentre a população de bébedos, i.e, $\frac{0,10}{0,25} = 0,4$. Por outras palavras quer-se calcular a probabilidade do acontecimento "ser violento", sabendo que ocorreu o acontecimento "ser bébedo". Repare-se que este quociente resulta da divisão entre a probabilidade de uma pessoa ser violenta e bébeda e a probabilidade de uma pessoa ser bébeda. Chamemos por B o acontecimento "ser bébedo" e por V o acontecimento "ser violento" podemos escrever simbolicamente que a probabilidade pretendida é dada por $P(V/O) = \frac{P(H \cap O)}{P(O)}$, o que equivale à definição de probabilidade condicional.

Definição: A probabilidade condicional de um evento A dado o evento B é dada por $P(A/B) = \frac{P(A \cap B)}{P(B)}$, o que equivale a $P(A \cap B) = P(B) \times P(A/B)$

Definição: A e B dizem-se independentes se, e somente se, $P(A \cap B) = P(A) \times P(B)$, o que equivale a dizer $P(A/B) = P(A)$ e $P(B/A) = P(B)$.

Regra Geral da Multiplicação numa Probabilidade Condicional

A Regra Geral da Multiplicação é usada para encontrar a probabilidade conjunta de que dois eventos ocorram.

Ela estabelece que para dois eventos A e B, a probabilidade conjunta de que os dois eventos ocorram é obtida pela multiplicação da probabilidade de que o evento A ocorra pela probabilidade condicional de B, dado que A ocorreu. Pela notação simbólica fica assim: $P(A \cap B) = P(B) \times P(A/B)$. Alternativamente, podemos também escrever:

$$P(A \cap B) = P(A) \times P(B/A)$$

7. PROBABILIDADE TOTAL

Suponhamos que um acontecimento C possa ocorrer se ocorrer um dos acontecimentos que são mutuamente exclusivos A_1, A_2,...,A_n, os quais formam um conjunto. Suponhamos que sejam conhecidas as probabilidades destes acontecimentos e as probabilidades condicionais $P(C/A_1), P(C/A_2),..., P(C/A_n)$, do acontecimento C. Pretendemos calcular a probabilidade de que C tenha ocorrido.

Exemplo: A Administração de um fundo de investimentos em acções pretende divulgar após o encerramento do pregão, a probabilidade de queda de um índice da bolsa no dia seguinte, baseando-se nas informações disponíveis até aquele momento. Suponha que a previsão inicial seja de 0,10. Após encerrado o pregão, nova informação sugere uma alta do dólar em relação ao metical. A experiência passada indica que quando houve queda da bolsa no dia seguinte 20% das vezes foram precedidas por esse tipo de notícias, enquanto nos dias em que a bolsa esteve em alta, apenas em 5% das vezes houve esse tipo de notícia no dia anterior. Determine a probabilidade de que a bolsa caia pelo simples facto de ter caído o dólar. Determine a probabilidade de que haja alta do dólar.

Resolução:

Seja E- Queda da bolsa. $P(E) = 0{,}10 \Rightarrow P(\overline{E}) = 0{,}90$. Se B- é Alta do dólar, então:

$P(B/E) = 0{,}20$ - probabilidade de queda da bolsa devido a alta do dólar

$P(B/\overline{E}) = 0{,}05$

$P(B) = P(E)P(B/E) + P(\overline{E})P(B/\overline{E}) = 0{,}10 \times 0{,}20 + 0{,}90 \times 0{,}05 = 0{,}065$ Portanto a informação aumenta a probabilidade de alta do dólar em 6,5%.

Teorema da Probabilidade Total:

A probabilidade de um acontecimento C, que pode ocorrer apenas sob a condição de que ocorra um dos acontecimentos que excluem mutuamente A_1, A_2,...,A_n e que formam o conjunto A, é igual a soma do produto das probabilidades de cada um desses acontecimentos, pela correspondente probabilidade condicional do acontecimento C, isto é: $P(C) = P(A_1)P(C/A_1) + P(A_2)P(C/A_2) + ... + P(A_n)P(C/A_n)$.

8. TEOREMA DE BAYES

Suponhamos que um acontecimento C pode ocorrer se ocorre um dos acontecimentos que se excluem mutuamente A_1, A_2,...,A_n e que formam um grupo completo. Visto que não se sabe qual desses acontecimentos ocorrerá, eles ainda estão sob a forma de hipóteses. A probabilidade de que C ocorra, determina-se pelo teorema de probabilidade total. Como

iríamos determinar a probabilidade de que um dos eventos que é partição de A ocorra dado que C ocorreu?

Exemplo: Os trabalhadores de uma fábrica são classificados segundo o nível de instrução escolar. A experiência mostra que 30% de indivíduos são de nível superior, destes 80% são considerados conhecedores das tarefas que desempenham. Dos 70% com nível inferior a superior, 50% são considerados conhecedores das tarefas que desempenham.

Dados

$P(S) = 0{,}30 \qquad P(C \mid S) = 0{,}80 \qquad P(\overline{S}) = 0{,}70 \qquad P(C \mid \overline{S}) = 0{,}50$

a) Qual é a probabilidade de que um trabalhador da fábrica encontrado no refeitório seja considerado conhecedor das tarefas que desempenha na fábrica? (2,0)

Resolução

$$P(C) = P(S) \times P(C \mid S) + P(\overline{S}) \times P(C \mid \overline{S}) = 0{,}30 \times 0{,}80 + 0{,}70 \times 0{,}50 = 0{,}59$$

Resposta:

A probabilidade de que um trabalhador da fábrica encontrado no refeitório seja considerado conhecedor das tarefas que desempenha na fábrica é de 0,59

b) Qual é a probabilidade de que um trabalhador conhecedor da matéria tenha nível superior? (2,0)

Resolução

$$P(S \mid C) = \frac{P(S) \times P(C \mid S)}{P(S) \times P(C \mid S) + P(\overline{S}) \times P(C \mid \overline{S})} = \frac{0{,}30 \times 0{,}80}{0{,}30 \times 0{,}80 + 0{,}70 \times 0{,}50} = 0{,}4067797$$

Resposta:

A probabilidade de que um trabalhador conhecedor da matéria tenha nível superior é de 0,4067797

$$P(A_i / B) = \frac{P(A_i)P(B/A_i)}{P(A_1)P(B/A_1) + P(A_2)P(B/A_2) + ... + P(A_n)P(B/A_n)} = \frac{P(A_i)P(B/A_i)}{\sum_{i=1}^{n} P(A_i)P(B/A_i)}$$

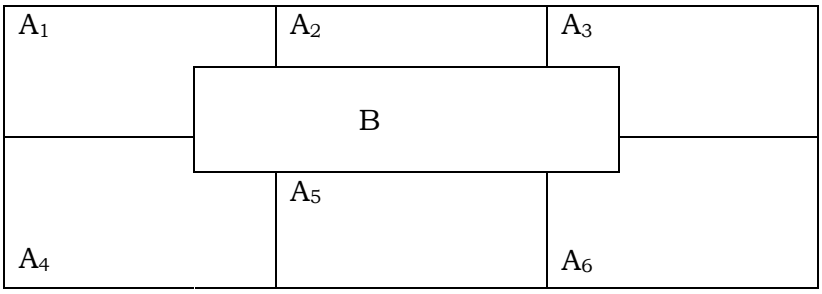


Fig 3.3

Assim temos $A_1 \cap B$; $A_2 \cap B$; $A_3 \cap B$; $A_4 \cap B$; $A_5 \cap B$; $A_6 \cap B$;

Em que A_1, A_2, A_3, A_4, A_5 e A_6 são subconjuntos (subeventos) de A e C é um evento qualquer. A_1, A_2, A_3, A_4, A_5 e A_6 devem ser mutuamente exclusivos e C, realiza intersecção com todos subeventos de A.

Regra Geral da Multiplicação numa Probabilidade Condicional

Exemplo
Uma Faculdade colectou a seguinte informação sobre seus estudantes de Licenciatura:

Tabela 3.4- Número de Estudantes de licenciatura duma Faculdade por cursos e sexos

Curso	Homens	Mulheres	Total
Engª Civil	120	80	200
Finanças	110	70	180
Administração	70	50	120
Gestão	110	100	210
Estatística	50	10	60
Informática	140	90	230
Total	600	400	1000

a) Um estudante é seleccionado ao acaso. Qual é a probabilidade de que o(a) estudante seja mulher e que esteja cursando Engª Civil?

Seja A o evento: o(a) estudante é de Engª Civil e F o evento: o(a) estudante é mulher. P(A e F) = 80 / 1000

b) Qual é a probabilidade de seleccionar uma mulher ? P(F) = 400/1000

c) Dado que o(a) estudante é mulher, qual é a probabilidade de que seja de Engª Civil ? Precisamos calcular P(A/F).

P(A/F) = P(A e F)/P(F) = [80/1000]/[400/1000]=0,20

Exemplo: Teorema de Bayes

A Companhia de bebidas Muzondwa Lda, produz vinho de litchi ou toranja. Recentemente, recebeu diversas reclamações de que suas garrafas estão sendo preenchidas com conteúdo abaixo do especificado. Uma reclamação foi recebida hoje mas o administrador da produção não é capaz de identificar qual das duas plantas (A ou B) preencheu a garrafa. Qual é a probabilidade de que a garrafa com pouco preenchimento provenha da planta A?

Resolução: Seja S o evento - A garrafa foi preenchida com conteúdo abaixo do especificado.

Tabela 3.5 – Distribuição de produção do vinho

	% da Produção Total	% de garrafas com pouco preenchimento
Vinho de tipo A (Litchi)	55	3
Vinho de tipo B (toranja)	45	4

$$P(A/S) = \frac{0,55 \times 0,03}{0,55 \times 0,03 + 0,45 \times 0,04} = 0,4783$$

Exercícios resolvidos

1 – Na maior parte das empresas as reuniões começam se o número acima de 50% estiver presente para a tomada de decisões. Uma empresa tem a Directoria com seis membros. Quantas comissões de quatro membros da Directoria podem ser formadas, com a condição de que em cada comissão figurem sempre o Presidente e o Vice-Presidente?

Resolução:
Os agrupamentos são do tipo combinações, já que a ordem dos elementos não muda o agrupamento. O número procurado é igual a: $C_{4-2}^{6-2} = C_2^4 = \frac{4!}{2!2!} = \frac{4.3.2.1}{2.1.2.1} = 6$. Observe que ao raciocinar na formação das comissões de 2 membros escolhidos entre 4, é preciso ter em conta que há já duas posições fixas na comissão: a do Presidente e do Vice.

2 - Numa Assembleia de quarenta cientistas, oito são estaticistas. Quantas comissões de cinco membros podem ser formadas incluindo no mínimo um estaticista?

Resolução:
A expressão "no mínimo um Estaticista" significa a presença de 1, 2, 3, 4 ou 5 estaticistas nas comissões. Podemos raciocinar da seguinte forma: em quantas comissões não possuem Estaticistas e subtrair este número do total de agrupamentos possíveis. Ora, existem C_5^{40} comissões possíveis de 5 membros escolhidos entre 40 e, existem $C_5^{40-8} = C_5^{32}$ comissões nas quais não aparecem estaticistas. Assim, teremos: $C_5^{40} - C_5^{32} = 656948$ comissões.

3 - Ordenando de modo crescente as permutações dos algarismos 2, 5, 6, 7 e 8, qual o lugar que ocupará a permutação 68275?

Resolução:
O número 68275 será precedido pelos números das formas:
a) 2xxxx, 5xxxx que dão um total de 4! + 4! = 48 permutações
b) 62xxx, 65xxx, 67xxx que dão um total de 3.3! = 18 permutações
c) 6825x que dá um total de 1! = 1 permutação.
Logo o número 68275 será precedido por 48+18+1 = 67 números. Logo, sua posição será a de número 68.

4- A Assembleia da República de Moçambique, para o ano de 2004 tem 250 deputados, sendo 132 do Partido Frelimo e 118 da Renamo União Eleitoral. Quantas comissões de sete deputados podem ser formadas com quatro membros da Frelimo e três da Renamo?

Resolução:

Quantas são as combinações dos 132 deputados da Frelimo tomados 4 a 4, isto é: C_4^{132}. Podemos escolher 3 da Oposição, entre os 118 existentes, de C_3^{118} maneiras distintas; portanto o número total de comissões é igual a $C_3^{118} \times C_4^{132}$, isto é, $C_4^{132} \times C_3^{118} = \frac{132!}{4!128!} \times \frac{118!}{3!115!} = \frac{132.131.130.129}{4.3.2.1} \times \frac{118.117.116}{3.2.1} = 3225088600000$ comissões distintas!.

5- Quantos números são ímpares se de 2,3,5,6,7 e 9 formamos números de três dígitos, sem que sejam permitidas repetições?
Resposta: O lugar da direita pode ser preenchido somente de 4 maneiras, por 3,5,7,9, já que os números devem ser ímpares; o da esquerda pode ser preenchido de 5 maneiras, e finalmente o do meio pode ser preenchido de 4 maneiras. Assim, existem 5.4.4=80 números.

6- Seja um dado viciado de modo que a probabilidade de aparecer um número seja proporcional ao número dado. Encontre a probabilidade de cada ponto amostral
Resolução: Seja P(1)=p, então P(2)=2p; P(3)=3p; P(4)=4p; P(5)=5p; P(6)=6p, como a soma de probabilidade deve ser 1, obtemos $p = \frac{1}{21}$. Assim $P(1) = \frac{1}{21}$; $P(2) = \frac{2}{21}$; $P(3) = \frac{3}{21}$; $P(4) = \frac{4}{21}$; $P(5) = \frac{5}{21}$; $P(6) = \frac{2}{7}$.

7- Determinar o erro da resolução do problema: Foram atirados dois dados. Determinar a probabilidade de que a soma dos pontos que saíram seja igual a 4 (acontecimento A).
Resolução:
Ao todo, são possíveis 2 casos da prova: a soma dos pontos que saíram é igual a 4, e a soma dos pontos que saíram não é igual a 4. Um caso favorece o acontecimento A; o número total de casos é igual a dois. Logo, a probabilidade procurada é: P(A)=1/2. o erro desta resolução consiste em que os casos examinados não são igualmente prováveis.
A solução correcta: o número de casos igualmente prováveis é 6.6=36; entre os casos que favorecem o acontecimento A há apenas (1,3), (3,1), (2,2). Logo a probabilidade procurada é P(A)=3/36=1/12.

8- Uma determinada peça é manufacturada por três fábricas. Sabendo que a fábrica 1, produz o dobro de peças que 2, e 2 e 3 produziram o mesmo número de peças (durante um período de produção especificado). Sabe-se também que dois por cento das peças produzidas por 1 e por 2 são defeituosas, enquanto quatro por cento daquelas produzidas por 4 são defeituosas. Todas peças produzidas são colocadas num depósito e depois uma peça é extraída ao acaso. Qual é a probabilidade de que essa peça seja defeituosa?

Resolução:
Seja:
A – peça é defeituosa B_1 – a peça provém da fábrica 1 B_2 – a peça provém da fábrica 2
B_3 – a peça provém da fábrica 3. pede-se $P(A)=P(A/B_1)P(B_1)+P(A/B_2)P(B_2)+P(A/B_3)P(B_3)$
$P(A/B_1)= P(A/B_2)= 0,02$ $P(B_1)=1/2$ $P(B_2)=(B_3)=1/4$ $P(A/B_3)=0,04$
$P(A)=P(A/B_1)P(B_1)+P(A/B_2)P(B_2)+P(A/B_3)P(B_3)=0,02\times1/2+0,02\times1/4+0,04\times1/4=0,025$

9- Uma cerâmica produz manilhas com dois tipos de defeitos: peças trincadas e peças tortas. A probabilidade de ocorrência do primeiro defeito é de 0,15 e do segundo de 0,25. Qual é a probabilidade de que 5 manilhas escolhidas ao acaso sejam todas defeituosas?

Resolução:
A probabilidade duma manilha defeituosa é
$P(A)+P(B)-P(A\cap B)=0{,}15+0{,}25-0{,}15\times 0{,}25=0{,}3625$

$P(X=5)=0{,}3625^5=0{,}00626$

10- Imagine que tenha sido contratado por uma empresa de pesquisas (Sondage Lda), sediada em Maputo, para fazer uma pesquisa em cada um dos 11 circulos eleitorais. Pretende-se determinar o número de caminhos distintos possíveis a usar. Quantos caminhos serão?
Resolução: $11!=11\times 10\times 9\times 8\times 7\times 6\times 5\times 4\times 3\times 2\times 1=39916800$ caminhos distintos

11-Um investidor possui duas acções. Uma numa companhia de processamento de castanha de Cajú e a outra é de uma cadeia de supermercados, de forma que podemos assumir que suas cotações são independentes. A probabilidade de que a acção da companhia de processamento de castanha de Cajú suba no próximo ano é 0,50. A probabilidade de que a cotação da cadeia de supermercados aumente em valor no próximo ano é 0,70.

a) Qual é a probabilidade de que ambas as acções cresçam em valor no próximo ano?
Resolução: Seja A o evento: a cotação da companhia de processamento de castanha de Cajú cresce no próximo ano e seja B o evento: a cotação da cadeia de supermercados cresce no próximo ano.
P(A e B) = $(0{,}50)\times(0{,}70)$ = 0,35

b) Qual é a probabilidade de que <u>ao menos</u> uma destas acções aumentem em valor no próximo ano?
Resolução: Isto implica que tanto uma pode aumentar (sem que a outra aumente) assim como ambas. Portanto, P(no mínimo uma) = $(0{,}50)\times(0{,}30)+(0{,}50)\times(0{,}70)+(0{,}70)\times(0{,}50)$ = 0,85

12- Um estudo recente constatou que 60 % das mães com crianças de idades de 0 até 10 anos empregam-se em tempo integral. Três mães são seleccionadas aos acaso. Assumiremos que as mães são empregadas de forma independente umas das outras.
a) Qual é a probabilidade de que <u>todas</u> sejam empregadas em período integral?
Resolução: P(todas as três empregadas em período integral) = $(0{,}60)\times(0{,}60)\times(0{,}60)=0{,}216$

b) Qual é a probabilidade de que no mínimo umas das mães seja empregada em período integral?

Resolução: P(no mínimo uma) = 1 – P(nenhuma empregada em período integral) =
$1-[(0{,}40)\times(0{,}40)\times(0{,}40)]$ = 0,936

13- Um exame consta de 15 questões das quais o aluno deve resolver 10. De quantas formas ele poderá escolher as 10 questões?

Resolução:
Observe que a ordem das questões não muda o teste. Logo, podemos concluir que se trata de um problema de combinação de 15 elementos dez a dez.
Aplicando simplesmente a fórmula chegaremos a:

$$C_{10}^{15} = \frac{15!}{10!(15-10)!} = \frac{15!}{10!5!} = \frac{15\times14\times13\times12\times11\times10!}{10!120} = 3003$$

14- Um Estádio tem 6 portões de acesso (entrada). De quantas formas distintas esse estádio pode estar aberto?

Resolução:
Para o primeiro portão temos duas opções: aberto ou fechado
Para o segundo portão temos também, duas opções, e assim sucessivamente. Para os seis portões, teremos então, pelo Princípio Fundamental da Contagem:
$N = 2\times2\times2\times2\times2\times2-1 = 64-1 = 63$
Resposta: o Estádio pode estar aberto de 63 modos possíveis.

Exercícios Propostos

1 – Com seis homens e quatro mulheres, quantas comissões de quatro pessoas podemos formar?
Resposta: 210

2 – Com seis homens e quatro mulheres, quantas comissões de cinco pessoas podemos formar, constituídas por dois homens e três mulheres?
Resposta: 60

3) Um coquetel é preparado com duas ou mais bebidas distintas. Se existem 7 bebidas distintas, quantos coquetéis diferentes podem ser preparados?
Resp: 120

4) Uma família com 5 pessoas possui um automóvel de 5 lugares. Sabendo que somente 2 pessoas sabem conduzir, de quantos modos poderão se acomodar para uma viagem?
Resp: 48

5) Numa sala de oito técnicos do nível superior, 3 são sociólogos e os restantes são das diferentes áreas. Pretende-se formar uma comissão de três técnicos para participarem num determinado workshop (a escolha será feita de forma aleatória).
Determine a probabilidade de dois serem sociólogos.

6) Num determinado grupo de jovens, apenas cinco têm tuberculose. Das pessoas que têm tuberculose 70% reagem positivamente a um determinado teste, enquanto 20% dos que não têm tuberculose reagem negativamente. Uma pessoa da população é escolhida ao acaso. Qual é a probabilidade de que essa pessoa tenha reagido positivamente ao teste?

7) Os pesquisadores estão preocupados com o declíneo do nível de cooperação por parte dos entrevistados em pesquisas, dado que muitas empresas de sondagem estão a encerrar portas por sempre estarem a publicar resultados não fiáveis. Um pesquisador aborda 100 pessoas na faixa etária 18-25 e constata que 73 respondem, enquanto 23 recusam responder. Quando são abordadas 300 pessoas na faixa etária 25-35, 250 respondem e 50 recusam responder. Suponha que um dos 450 indivíduos seja escolhido aleatoriamente. Determine a probabilidade de obter alguém na faixa etária 18-25 ou alguém que recuse responder.

8) Os problemas de assédio sexual tem recebido muita atenção nos últimos anos. Basta ouvir entrevistas dadas pela Associação de Secretárias ou mesmo conversando com diversas individualidades ligadas aos recursos humanos. Numa pesquisa, 420 trabalhadores (200 dos quais homens) consideram uma simples batida no ombro como uma forma de assédio sexual, enquanto que 580 trabalhadores (380 dos quais homens) não consideram isso como assédio. Escolhido aleatoriamente um dos trabalhadores pesquisados, determine a probabilidade de obter alguém que não considere um simples tapa no ombro como uma forma de assédio sexual.

9) O rendimento anual das famílias de uma certa cidade pode ser expresso através de uma variável contínua. Sabe-se que a mediana do rendimento é de 60000000,00 MT e que 40% das famílias da cidade tem um rendimento superior a 72000000,00MT.
Escolhida uma família aleatoriamente, qual é a probabilidade desta ter um rendimento entre 60000000,00MT e 72000000,00MT?
b) Sem mais informação adicional, o que pode dizer sobre a probabilidade de uma família ter um rendimento menor que 65000000,00Mt?

10) De acordo com uma pesquisa, a Coca-cola e a Pepsi se posicionaram como número um e dois, respectivamente em vendas no ano de 1996. Suponha que de um grupo de 10 indivíduos, seis prefiram Coca-cola e quatro prefiram a Pepsi. Uma amostra de três indivíduos é seleccionada.
Qual é a probabilidade de que exactamente dois prefiram a Coca-cola?
Qual é a probabilidade de que a maioria prefira a Pepsi?

CAPÍTULO 4 VARIÁVEIS ALEATÓRIA, FUNÇÕES DE DISTRIBUIÇÃO E DISTRIBUIÇÕES TEÓRICAS DE PROBABILIDADE

Objetivos do Capítulo:

- Distinguir uma distribuição de probabilidade discreta da contínua.
- Calcular a média, a variância e o desvio padrão de uma distribuição de probabilidade discreta.
- Definir os termos Distribuição de Probabilidade e Variável Aleatória.
- Definir uma distribuição de probabilidade discreta.
- Calcular a média, a variância e o desvio padrão de uma distribuição de probabilidade discreta.
- Definir os termos Distribuição de Probabilidade e Variável Aleatória.
- Descrever as características das distribuições Binomial, Geométrica, Multinomial, Hipergeométrica e de Poisson.
- Definir uma distribuição de probabilidade contínua.
- Calcular a média, a variância e o desvio padrão de uma distribuição de probabilidade contínua.
- Definir os termos Distribuição de Probabilidade e Variável Aleatória contínua.
- Descrever as características das distribuições Normal

Definição 1: Uma variável aleatória (v.a.) é um conjunto de valores numéricos que são associados a eventos de um experimento. Em geral as variáveis aleatórias são designadas por letras maiúsculas, X, Y, Z, etc e podem ser discretas ou contínuas.

Definição 2: Uma variável aleatória X num espaço amostral S é uma função de S no conjunto $\Re$ dos números reais tal que a imagem inversa de cada intervalo de $\Re$ seja um evento de S (Lipschutz, 1993).
Exemplos:

1- Seja a V.A. X, o número de sementes de milho que germinam em 100 plantios. Possíveis valores para X são 0,1,2,100, (discreta), pois nós contamos planta a planta, pelo que estamos em presença de uma Variável Aleatória Discreta (VAD)

2- Seja X a variável aleatória que indica a temperatura máxima diária em Tete. Possíveis valores são 0 - 45 ºC, por exemplo 36.1573 (contínua), pois resulta de uma medição, pelo que estamos em presença de uma Variável Aleatória Contínua (VAC).

3- Seja X a resposta a uma questão se é ou não Moçambicano, cuja resposta é 'Sim', 'Não', 'Não Responde'. Repare que para este caso o X não é uma v.a (por ser não numérica). O 'Sim', 'Não', 'Não Responde' são resultados qualitativos. Se tivéssemos que contar a quantidade de pessoas que responderam 'Sim', 'Não', 'Não Responde', geraria variáveis discretas. Repare-se para o facto de que este exemplo não leva a variável contínua porque as respostas dadas não são medíveis no sistema básico internacional de medidas.

4- Considere um experimento aleatório no qual uma moeda é lançada 3 vezes. Sejam X o número de caras, C o resultado cara e K o resultado de coroa.
O espaço amostral para este experimento será:
S={KKK, KKC, KCK, CKK, KCC, CKC, CCK, CCC}

Assim, os possíveis valores de X (número de caras) serão: X = {0, 1, 2, 3}.

Da definição de uma variável aleatória X, neste experimento, é uma variável aleatória discreta. Seus valores são resultados de uma contagem.

Nota: A variável aleatória X é uma associação de pontos no espaço amostral com pontos na recta dos números reais (0,1, 2,3). Na realidade, uma variável aleatória é definida através de uma função em que o domínio é o conjunto de todos os resultados possíveis do experimento e a imagem é o conjunto de todos os valores assumidos pela variável aleatória. Note que a variável aleatória não é resultado do experimento, mas sim um valor associado a este.

1 FUNÇÃO DE DISTRIBUIÇÃO

Exemplo: Consideremos sucessivos lançamentos dum dado. Cada vez que o dado é lançado seis possíveis resultados podem ocorrer 1, 2, 3, 4, 5 ou 6. Isto significa que a nossa variável aleatória pode tomar um dos seis resultados, em que todos possuem a mesma probabilidade $P(X = x_i) = \frac{1}{6}$. Não há hipóteses de que $x_i < 1$, o que significa que é um acontecimento impossível, logo $P(X < 1) = 0$

Consideremos uma variável aleatória X, um intervalo real $T_x =]-\infty, x]$ e a respectiva imagem inversa $X^{-1}(T_x)$. Com $P(X \leq x)$ sendo que $P(X \leq x)$ depende de x, a igualdade $F(x) = P(X \leq x)$ define uma função real de variável real. Essa função real é designada por Função de Distribuição da variável Aleatória X.

Uma distribuição de probabilidades discreta, é uma lista dos possíveis valores da variável aleatória e as probabilidades correspondentes (que tem que somar 1). As probabilidades podem ser escritas:

$P(X = x_i) = p_i$ para $\mathrm{i} = 1,2,...,\mathrm{k}\ \mathrm{e}\ 0 \leq \mathrm{p_i} \leq 1$ $\sum_{i=1}^{k} p_i = 1$

Definição: Uma Distribuição de Probabilidade é uma lista de todos os resultados de um experimento e suas probabilidades associadas. De forma mais rigorosa, é uma função matemática em que o domínio são os valores possíveis de uma variável aleatória e a imagem são as suas probabilidades associadas.
A distribuição de probabilidade de uma variável aleatória X (número de caras) nas três jogadas de uma moeda é mostrada a seguir.

Tabela 4.1- Distribuição de Probabilidade para três lançamentos duma Moeda

Número de Caras	Probabilidade
0	1/8 = 0,125
1	3/8 = 0,375
2	3/8 = 0,375
3	1/8 = 0,125
Total	8/8 = 1

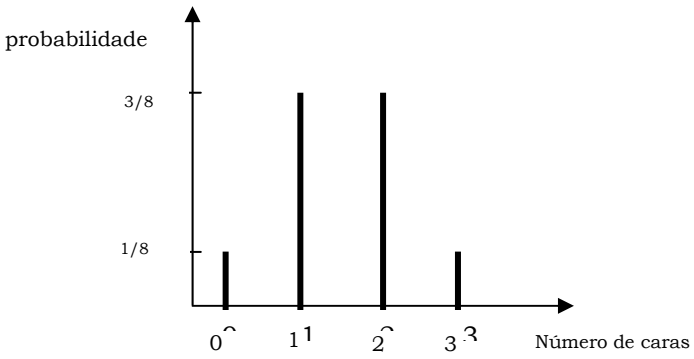


Fig 4.1

Características de uma distribuição de probabilidade

- A probabilidade de um resultado deve estar sempre situada entre 0 e 1.

Exemplo: P(0 caras) = 0,125, P(1 cara) = 0,375 , etc. no experimento de jogar 3 moedas.

- A soma das probabilidades de todos os resultados mutuamente exclusivos é sempre 1, veja a tabela 12.1 de distribuição de probabilidade acima.

2 VARIÁVEL ALEATÓRIA DISCRETA (VAD)

Definição: Uma variável aleatória discreta é uma variável que pode assumir somente certos valores claramente separados (em descontinuidade) resultantes, por exemplo, de uma contagem de algum item.

Exemplo: Seja X o número de caras quando uma moeda é jogada 3 vezes. Aqui os valores de X são 0,1,2 ou 3.

Nota: uma variável aleatória discreta não precisa necessariamente assumir apenas valores inteiros. Poderia, por exemplo, ser uma variável que apresentasse os seguintes valores: 0, 23/7 , 72/25, etc. A condição que deve ser cumprida é seus valores sejam descontínuos.

2.1 Valor Esperado (média) de uma Distribuição de Probabilidade Discreta

A média refere-se a localização central_ de um conjunto de dados. Ela pode ser considerada como um valor de "longo prazo" de uma variável aleatória e é também chamada de valor esperado (ou esperança matemática), E(X).

A média de uma distribuição de probabilidade discreta é determinada pela fórmula:

$\mu = E(X) = \sum [X.P(X)]$

onde μ (letra grega, míu) representa a média (ou valor esperado) e P(X) é a probabilidade dos vários resultados de X.

2.2 Valor Esperado do Produto de Variáveis Aleatórias Independentes

Se as variáveis aleatórias X e Y são independentes, então $E[X \times Y] = E[X] \times E[Y]$
O inverso (recíproca) não é verdadeiro em geral: $E[X \times Y] = E[X] \times E[Y]$ não implica que X e Y sejam independentes.

2.3 Variância e o Desvio Padrão de uma Distribuição de Probabilidade Discreta
A variância mede a quantidade de dispersão ou variabilidade de uma distribuição. Ela é denotada pela letra grega σ^2 (sigma ao quadrado).
O desvio padrão é obtido através da raiz quadrada de σ^2.
A variância de uma distribuição de probabilidade discreta é calculada através da fórmula:

$\sigma^2 = \sum[(X-\mu)^2 P(X)]$ O desvio padrão é: $\sigma = \sqrt{\sigma^2}$

Exemplo
Uma empresa especializa-se no aluguer de carros para famílias que necessitam de um carro adicional por um período curto de tempo. O presidente da empresa tem estudado seus registos para as últimas 20 semanas e apresentou os seguintes números de carros alugados por semana.

Tabela 4.2- Número de carros alugados por semana

Número de Carros alugados	Semanas
10 11 12 13	5 6 7 2

Os dados acima podem ser considerados como uma distribuição de probabilidade?

Convertendo o número de carros alugados por semana numa distribuição de probabilidade, teremos:

Tabela 4.2a)- Distribuição de Probabilidades do número de carros alugados por semana

Número de carros alugados	Probabilidade P(X)
10 11 12 13	0,25 0,30 0,35 0,10
Total	1,00

Como se pode ver, trata-se de uma distribuição de probabilidade, já que todos pressupostos são satisfeitos.

a) Calcule o número médio de carros alugados por semana.

A média

$$\mu = E(X) = \sum[X.P(X)] = (10)\times(0{,}25) + (11)\times(0{,}30) + (12)\times(0{,}35) + (13)\times(0{,}10) = 11{,}3$$

b) Calcule a variância do número de carros alugados por semana.

A variância

$$\sigma^2 = \sum[(X-\mu)^2.P(X)] = (10-11{,}3)^2\times 0{,}25 + (11-11{,}3)^2\times 0{,}30 + \ldots + (13-11{,}3)^2\times 0{,}10 = 0{,}91$$

$$\sigma = \sqrt{0{,}9135} = 0{,}9558$$

3. DISTRIBUIÇÕES DE PROBABILIDADE TEÓRICAS DISCRETAS

3.1 A Distribuição Binomial

Exemplo

O Departamento de Estatística do Ministério de Trabalho de Moçambique estimou que 20 % da força de trabalho está desempregada. Uma amostra de 14 trabalhadores é obtida do município de Maputo. Calcule as seguintes probabilidades:

Três trabalhadores estão desempregados. Neste caso teremos n = 14 e p = 0,2 o que leva com que $q = 1-p = 1-0{,}2 = 0{,}8$

$$P(X=3) = \frac{14!}{3!(14-3)!}0{,}2^3 0{,}8^{14-3} = 0{,}250$$

No mínimo um dos trabalhadores da amostra esteja desempregados.

$$P(X \geq 1) = 1 - P(X=0) = 1 - \frac{14!}{0!(14-0)!}0{,}2^0 0{,}8^{14-0} = 0{,}956$$

No máximo dois dos trabalhadores estejam desempregados.

$$P(X \leq 2) = 0{,}044 + 0{,}154 + 0{,}250 = 0{,}448$$

A **Distribuição Binomial tem as seguintes características**:

Consideremos um experimento que apresenta somente dois resultados possíveis que são categorias mutuamente exclusivas: sucesso e fracasso (ocorre ou não ocorre). Referindo-se a uma dicotomia

a) O experimento é repetido várias vezes.
b) A probabilidade de sucesso permanece constante para cada tentativa (consequentemente, a probabilidade de falha também permanece constante).
c) As tentativas são independentes, significando que o resultado de uma tentativa não afecta o resultado de qualquer outra tentativa. Se for o caso de tirar algo de uma urna, a independência é garantida por retirada com reposição.

Para construir uma distribuição binomial, é assumir que:

n é o número de tentativas
r é o número de sucessos observados
p é a probabilidade de sucesso em cada tentativa
q é a probabilidade de (insucesso) falha em cada tentativa, em que $q = 1 - p$

A distribuição de probabilidade para uma distribuição discreta binomial é dada por

$$P(X = r) = C_n^r \times p^r \times q^{n-r} = \frac{n!}{r! \times (n - r)!} \times p^r \times q^{n-r}$$

A Média e Variância de uma Distribuição Binomial
A média é dada por: $\mu = np$
A variância é dada por: $\sigma^2 = np(1-p)$

Para o exemplo anterior:
p = 0,05 e n = 6

$\mu = np = 6 \times 0{,}05 = 0{,}3$
$\sigma^2 = np(1-p) = 6 \times 0{,}05 \times 0{,}95 = 0{,}285$

3.2. Distribuição HipergeométrIca

Exemplo:
De uma urna que contém 30 bolas diferentes apenas no tamanho das quais 254 são azuis e 30-25 são amarelas, tiram-se ao acaso 9 bolas. Qual é a probabilidade de obter 5 bolas amarelas?
Aqui estamos perante uma experiências em que:

a) uma amostra aleatória de n elementos é seleccionada a partir de uma população com N elementos;
b) k dos N elementos da população são designados por sucessos e N-K por insucessos.

Neste caso teremos: $\binom{25}{5} \times \binom{30-25}{9-5} = \frac{25!}{20!5!} \times \frac{5!}{1!4!} = 265650$ casos favoráveis e $\binom{30}{9} = \frac{30!}{21!9!} = 14307149$ casos possíveis, donde a respectiva probabilidade será $\frac{265650}{14307149} = 0{,}0186$.

Considere-se a população finita constituída por N elementos de dois tipos (ou seja com dois atributos principais e mutuamente exclusivos). Seja $M \leq N$ o número de elementos de um dos referidos tipos. Admita-se que, nesta população se retiram sucessivamente, sem reposição (ou, o que é equivalente; de uma só vez, em blocos), m elementos. Se X representar o número de elementos de um dos tipos que figuram entre os n que foram retirados da população, então, X segue uma distribuição hipergeométrica.

Tomemos um caso concreto. Achemos a probabilidade de que X=m, ou seja, que entre as n peças extraídas, exactamente m sejam standarizadas. Empreguemos para isso uma definição clássica de probabilidade. O número total de casos elementares possíveis da prova é igual ao número de maneiras com as quais se podem extrair n peças de N, ou seja, igual ao número de combinações C_N^n. Achemos o número de casos favoráveis ao acontecimento X=m (entre as n peças extraídas há exactamente m standarizadas); as m peças standarizadas podem ser extraídas das M peças standarizadas de C_M^m maneiras; neste caso, as n-m peças restantes devem ser não standarizadas; igualmente, pode-se extrair n-m não standarizadas das N-m peças não standarizadas de C_{N-M}^{n-m} maneiras. Consequentemente, o número de casos favoráveis é igual a $C_M^N C_{N-M}^{n-m}$ (de acordo com a regra de produto para obtenção de casos favoráveis).

A probabilidade procurada para X=k casos favoráveis será $P(X=k)=\frac{C_M^k C_{N-M}^{n-k}}{C_N^n}$, a sua esperança matemática ou média populacional será $\mu=E(X)=\frac{kn}{N}$, a variância é dada pela fórmula seguinte $\sigma^2=\left(\frac{nk}{N}\right)\left(\frac{1-k}{N}\right)\left(\frac{N-n}{N-1}\right)$

3.3 DistribuiÇÃo Multinomial ou Polinomial

Exemplo:
Uma roleta de casino é composta por 16 números de cor vermelha, 16 números de cor preta, e 1 número de cor verde. Após 10 partidas consecutivas, o vermelho saiu 7 vezes, o preto 2 vezes, o verde 1 vez. Calcular a probabilidade desse acontecimento

Resolução:

Em cada partida, a probabilidade de saída de um número vermelho é: $P(R)=\frac{16}{33}$. Em cada partida, a probabilidade de um número preto é: $P(N)=\frac{16}{33}$. Em cada partida, a probabilidade de saída do número verde é $P(V)=\frac{1}{33}$.

A probabilidade de ocorrerem 7 números vermelhos, 2 pretos e 1 verde é dad pela aplicação da lei multinomial: $P((7R)\cap(2N)\cap(1V))=\frac{10!}{7!2!1!}(P(R))^7\times(P(N))^2\times(P(V))^1=0{,}15$.

A distribuição multinomial é a generalização da distribuição binomial. Suponha que o espaço amostral S de um experimento E seja repartido em K eventos mutuamente exclusivos A_1, A_2, A_3, .., A_k com probabilidades P_1, P_2, P_3, ..., P_k, respectivamente. Onde $P_1+P_2+P_3+..+P_k=1$, então em n ensaios repetidos, a probabilidade de que A_1 ocorra K_1, A_2

ocorra K_2, A_3 ocorra K_3, … , e A_k ocorra K_k vezes e igual a: $P(X_1 = k_1, X_2 = k_2, X_3 = k_3, ..., X_k = k_k) = \frac{n!}{k_1 \times k_2 \times k_3 \times ... \times k_k} \times p_1^{k_1} \times p_2^{k_2} \times p_3^{k_3} \times ... \times p_k^{k_k}$, onde $k_1 + k_2 + k_3 + ... + k_k = n$ $E(X_i) = np_i$ $Var(X) = n \times p_i \times q_i$

3.4 Distribuição Geométrica

Exemplo: Num processo de fabricação 2% de peças produzidas são defeituosas. A probabilidade de que numa amostra de X peças não haja nenhuma defeituosa é de $0,98^X$. Se de 200 peças nenhuma foi encontrada com defeito, qual é a probabilidade de que a peça 201ª seja defeituosa?

Resolução:
Trata-se de uma distribuição geométrica.
$P(X = x) = pq^{x-1}$, $p = 0,02$, $q = 0,98$, $P(X = 201) = 0,02 \times 0,98^{200} = 0,020$

Suponha-se que realizemos um experimento ε e que estejamos interessados apenas na ocorrência ou não de algum evento A. Admita-se, tal como na distribuição binomial, que realizemos ε repetidamente, que as repetições sejam independentes, e que em cada repetição P(A)=p e permaneça constante. Consideremos que repetimos o experimento até que A ocorra pela primeira vez. Define-se a variável aleatória X como o número de repetições necessárias para obter a primeira ocorrência de A; assim, X toma os valores possíveis 1, 2, … com X= k se, e somente se, as primeiras K-1 repetições de ε derem o resultado A, enquanto K-ésima repetição dê o resultado A, teremos: $P(X = k) = q^{k-1} \times p$ $E(X) = \mu = \frac{1}{p}$

$Var(X) = \frac{q}{p^2}$

Exemplo:
Num processo de fabricação, 2% de peças produzidas são defeituosas. A probabilidade de que numa amostra de X peças não haja nenhuma defeituosa é de $0,98^X$. Se de 200 peças nenhuma foi encontrada com defeito, qual é a probabilidade de que a peça 201ª seja defeituosa?

3.5 Distribuição de pascal

Exemplo:
Seja uma urna coontendo uma população p de bolas brancas e $q = 1 - p$ de bolas pretas. Efectuam-se extracções com reposição, até a obtenção de r bolas brancas.
Calcular a probabilidade de k extracções necessárias.

Resolução:

$P(X = k) = C_{k-1}^{r-1} p^{k-r} q^r$

Uma generalização da distribuição geométrica é a distribuição de Pascal. Suponhamos que um experimento seja continuado até que particular evento A ocorra na r-ésima vez. Se P(A)=p e q=1–p(A), em cada repetição, definiremos a variável aleatória X como segue: X é o número de repetições necessárias a fim de que A possa ocorrer exactamente r vezes. X=k se, e somente se, A ocorrer na k-ésima repetição e A tiver ocorrido exactamente r-1 vezes nas $k-1$ repetições anteriores. A probabilidade deste evento é $P(X = k) = C_{r-1}^{k-1} \times p^r \times q^{k-r}$, onde k = r, r+1, ... $E(X) = \frac{r}{q}$ $Var(X) = \frac{rq}{p^2}$

3.6. Distribuição de Poisson

Exemplo:
Admitamos que o número de defeitos de um sapato obedece a lei de poisson de parâmetro $\lambda = 2$. Calcular a probabilidade de que não haja nenhum defeito no sapato.

Resolução:

O número de defeitos X no sapato é uma variável aleatória que obedece uma lei de Poisson com $\lambda = 2$. $P(X = k) = e^{-2} \times \frac{2^k}{k!} = e^{-2} \times \frac{2^0}{0!} \approx 0{,}135$ aasumindo que k=0.

Seja X uma variavel aleatória discreta, tomando seguintes valores: o, 1, ..., n, ..., se $P(X) = \frac{e^{-\lambda} \times \lambda^k}{k!}$, onde k = o, 1, 2, 3, ..., n, ..., diremos que X tem uma distribuição de Poisson com parâmetro $\lambda \succ 0$, simbolicamente $X \sim P_o(\lambda)$ $E(X) = \lambda$ $Var(X) = \lambda$

Distribuições Discretas no controlo de qualidade

Iniciamos esta matéria ainda neste ponto em virtude da necessidade premente de ser tão importante e necessário executar o control de qualidade em diversas áreas de aplicação estatística.
Se um lote de produto precisa de ser vendido, seria interessante verificar a sua fiabilidade se pode ou não ser comercializável. É nesta perspectiva que podemos assumir que todo o produto precisa de passar por algum control de qualidade antes de chegar aos terminais consumidores.

Exemplo
Um Sistema de Control de Qualidade duma companhia vendedora de leite requer que cada manhã, antes da saída dos carros à rua, se teste 10 pacotes para se ter a certeza de que o produto esteja dentro dos patamares da empresa. Se dois ou mais pacotes de leite do total dos tais 10, mostrarem que possuem algum problema de lactose, a amostra é retirada e consequentemente a produção do dia é considerada perdida. A probabilidade de um pacote defeituoso é de 0,05 e mantém-se constante.
(i) qual é a probabilidade de existirem dois 2 pacotes na amostra com problemas de lactose?
(ii) Qual é a probabilidade do lote ser rejeitado?

Resolução:
Seja X a V.A. = número de defeituosos na amostra de n = 10 itens. Portanto,
X~Bin (10; 0,05)

$$P(X=2)=\binom{10}{2}(0{,}05)^2(0{,}95)^8=0{,}0746$$

P(rejeitar o lote) = $P(X\geq 2)=\sum_{x=2}^{10}\binom{10}{x}(0{,}05)^x(0{,}95)^{10-x}$ o que é muito trabalhoso de calcular.

Mas:

$$P(X\geq 2)=1-P(X<2)=1-P(X=0 \text{ ou } X=1)$$
$$=1-[P(X=0)+P(X=1)] \text{ mutuamente exclusivos}$$
$$=1-[\binom{10}{0}(0{,}05)^0(0{,}95)^{10}+\binom{10}{1}(0{,}05)^1(0{,}95)^9]$$
$$=0{,}0862$$

Exemplo - exploração de petróleo
Uma companhia de exploração de petróleo tem um arrendamento para o qual precisa decidir se:
(i) vende agora,
(ii) segura durante um ano e então vende, ou
(iii) perfura agora.
O custo de perfurar é $200,000
Perfurando conduzirá a um dos resultados seguintes

Tabela 4.3 Resultado de Exploração de Petróleo

Resultado	Probabilidade	Receita
Poço Seco	0.5	$0
Poço com pouco petróleo	0.4	$400000
Poço com jorro	0.1	$1500000

Se vende agora, a companhia pode adquirir $125000.

Se aguentar durante um ano e os preços do petróleo sobem (probabilidade =0.6) pode vender por $300000 ou se os preços do petróleo caem (probabilidade = 0.4) pode adquirir $100000. O que deveria fazer? A melhor decisão é segurar durante um ano e então vender. Assim está de novo ilustrado o conceito do valor esperado de uma variável aleatória .

Se a distribuição de probabilidade de uma variável aleatória X é

Tabela 4.3. a) Forma de Distribuição de probabilidades

Valores de X	X_1	X_2	...	X_k
Probabilidades	P_1	P_2	...	P_k

seu valor esperado é

$$E(X) = x_1 p_1 + x_2 p_2 + ... + x_K p_K = \sum_{i=1}^{K} x_i p_i$$

Exemplo da Perfuração do Petróleo
Seja X a variável aleatória lucro financeiro, segundo tabela 13.1
= Receita – custo de perfuração
= Receita - $200000
A distribuição de probabilidade para X é

Tabela 4.4 Distribuição de Probabilidades do resultado de Exploração de Petróleo

x	-200	200	1300
P(X=x)	0.5	0.4	0.1

Portanto, o valor esperado (média) de X é

E(X) = -200 x 0.5 + 200 x 0.4 + 1300 x 0.1 = $110K

Isto é directamente análogo à média amostral. E(X) pode ser considerada como uma idealização do teórico para, a média da amostra. E(X) é denotado frequentemente pela letra grega μ (pronuncia-se miu).

Algumas propriedades do valor esperado e variância para uma função de variáveis aleatórias

Se Y = aX + b onde X é uma variável aleatória e a e b são valores constantes conhecidos, então, E(Y) = a E(X) + b

$\text{var}(Y) = a^2 \text{var}(X)$ Portanto, $\sigma_Y = \sqrt{a^2 \text{var}(X)} = \sqrt{a^2 \sigma_x^2} = a\sigma_X$ e semelhantemente se

T = a X + b Y + c onde X e Y são variáveis aleatórias e a, b e c são constantes conhecidas, então,

E(T) = a E (X) + b E (Y) + c. e $Var(T) = a^2 \text{var}(X) + b^2 \text{var}(Y) + 2ab\,\text{cov}(X,Y)$

Em particular, se X e Y são independentes então a covariância cov(X,Y) é zero. Portanto $Var(T) = a^2 \text{var}(X) + b^2 \text{var}(Y)$. Prova: Segue das definições de E(X) e var(X).

Exemplo - **Lucro previsto estimado**
Uma companhia faz produtos para mercados locais e de exportacção.
O número de vendas do próximo ano não pode ser predito exactamente mas estimativas podem ser feitas como a seguir

Tabela 4.5 Distribuição de Probabilidades de vendas num mercado local

unidades de X,local	1,000	3,000	5,000	10,000
Probabilidade	0.1	0.3	0.4	0.2

Tabela 4.6 Distribuição de Probabilidades de vendas num mercado externo

unidades Y, export.	300	500	700
Probabilidade	0.4	0.5	0.1

Consequentemente E(X) = 1000 x 0.1 + 3000 x 0.3 + 5000 x 0.4 + 10000 x 0.2
= 5000 (= esperou vendas locais)
E(Y) = 300 x 0.4 + 500 x 0.5 + 700 x 0.1
= 440 (= vendas de exportação esperadas)
A companhia lucra $2000 em cada unidade vendida no mercado local e $3500 em cada unidade exportada.
Consequentemente o lucro total é T = 2000 X + 3500 Y. Usando a fórmula acima
E(T) = 2000 E(X) + 3500 E(Y)
= 2000 x 5000 + 3500 x 440
= $11,540,000 - este é o lucro estimado (previsto) durante o próximo ano.

Exemplo:
Um componente é feito cortando um pedaço de metal de comprimento X e reduzindo este valor da quantidade Y. Ambos processos são um pouco imprecisos. O comprimento líquido é então: T = X - Y. Isto pode ser escrito na forma T = a X + b Y com a = 1 e b = -1
assim E(T) = a E (X) + b E (Y) = 1 E(X) + (-1)E(Y) = E(X) - E(Y)

$Var(T) = a^2 \text{ var}(X) + b^2 \text{ var}(Y)$ portanto $\text{var(T)} = 1^2 \text{ var}(X) + (-1)^2 \text{ var}(Y) = \text{var(X)} + \text{var(Y)}$
ou seja, var(T) é maior tanto que var(X) ou var(Y), embora T = X - Y, porque X e Y contribuem à variabilidade em T.

3.7 Variáveis Aleatórias Independentes

Lembremos que dois eventos A e B são independentes se e somente se $P(A \cap B) = P(A)P(B)$, se a probabilidade da interseção de A e B é o produto das probabilidades de A e de B. Podemos relacionar variáveis aleatórias a eventos, ou seja, podemos definir eventos em termos de valor(es) que uma variável aleatória assume. Por exemplo, o evento $A = \{a < X \leq b\}$ ocorre se X é maior do que a e menor do que b. Duas variáveis aleatórias, X e Y, são independentes se e somente se todo evento da forma $\{a < X \leq b\}$ é independente de todo evento da forma $\{c < Y \leq d\}$. Duas variáveis aleatórias são independentes se conhecendo o valor de uma não ajuda a predizer o valor da outra.

Exemplos: Considere a jogada de uma moeda 10 vezes.

Seja X o número de caras nas primeiras 6 jogadas e seja Y o número de caras nas últimas 4 jogadas. Portanto X e Y são independentes. Conhecer o valor de X não ajuda a predizer o valor de Y e vice-versa.

Seja X o número de caras nas primeiras 6 jogadas e seja Y o número de caras nas últimas 5 jogadas. Então X e Y são dependentes porque, por exemplo, o evento $\{5 < X \leq 6)$ e o evento $\{-1 < Y \leq 0\}$ são dependentes (e mutuamente exclusivos).

Seja X o número de caras nas primeiras 6 jogadas e seja Y o número de coroas nas primeiras 2 jogadas. Então X e Y são dependentes porque, por exemplo, o evento $\{5 < X \leq 6\}$ e o evento $\{2 < Y \leq 3\}$ são dependentes (e mutuamente exclusivos).

Que espécies de experimentos conduzem a variáveis aleatórias independentes? Somas e médias de sequências que não se sobrepõem seja de jogadas de moedas, de jogadas de dados são alguns exemplos. Os segundo e terceiro exemplos acima mostram porque existe a necessidade das sequências serem não sobrepostas (ou seja, não tenham intersecção).

Tabela 4.7- Resumo das Estatísticas e respectivas probabilidades

EMPÍRICO (baseado nos dados) QUANTIDADE	TEÓRICO (MATEMÁTICO) QUANTIDADE	
Frequência relativa $x_i = \frac{f_i}{n}$	PROB[X = x_i] = p_i	$\frac{f_i}{n} \to \infty$ *quando* $n \to 0$
$(b)\sum_i \frac{f_i}{n} = 1$	$\sum_{i=1}^{n} p_i = 1$	
(c) média $\bar{x} = \frac{1}{n}\sum_i x_i f_i$	ESPERANÇA, $\mu =$ $E(X) = \sum_i p_i x_i$	$\bar{x} \to E(X)$ *quando* $n \to \infty$
(d) VARIÂNCIA $S^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2 f_i}{n-1}$	$VAR(X) = \sum_{ii=1}^{n} (x_i - x)^2 p_i$	$S^2 \to VAR(X)$ *quando* $n \to \infty$

4. VARIÁVEIS ALEATÓRIAS CONTÍNUAS E DISTRIBUIÇÕES TEÓRICAS CONTÍNUAS

4.1 Variável Aleatória Contínua (VAC)

Definição: Uma variável aleatória contínua é uma variável que pode assumir um número infinitamente grande de valores (com certas limitações práticas).
Tomando X como sendo a variável e F(x), como sendo a respectiva função de distribuição, então $D = \{a : P(X = a) > \} = \emptyset$. Resulta então que F(x) não apresentará descontinuidade. Se além disso existe uma função não negativa $f(x) \geq 0$, tal que para qualquer número real x se verifica a relação $F(x) = \int_{-\infty}^{x} f(m)dm$, então dizemos que a variável aleatória X é contínua. Repare-se que uma variável para ser integrável deve ser diferenciável e para ser diferenciável deve ser contínua.

Exemplo: (a) Peso de um estudante
(b) comprimento de um carro

O Valor Esperado (média) de uma Distribuição de Probabilidade Contínua

Tal como na Distribuição Discreta, a média refere-se a localização central de um conjunto de dados. Ela pode ser considerada como um valor de "longo prazo" de uma variável aleatória e é também chamada de valor esperado (ou esperança matemática), E(X).

$$E(x) = \mu(x) = \int_{-\infty}^{+\infty} x^k f(x)dx$$

A Variância e o Desvio Padrão de uma Distribuição de Probabilidade Contínua

A variância é dada por $\sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx$ e desvio $\sigma = \sqrt{\sigma^2}$

Diferente de uma VAD (Variável Aleatória Discreta), uma variável aleatória contínua é definida a partir de uma função, dado que as variáveis contínuas resultam de medição. Por ser contínua e por ser expressa ou definida através de uma função, denominamos essa função por função densidade de probabilidade ou simplesmente densidade de probabilidade (f.d.p.). No entanto, podemos definir o que se chama de uma *função densidade de probabilidade* para as variáveis aleatórias contínuas. Por exemplo, suponhamos uma distribuição uniforme do tipo:

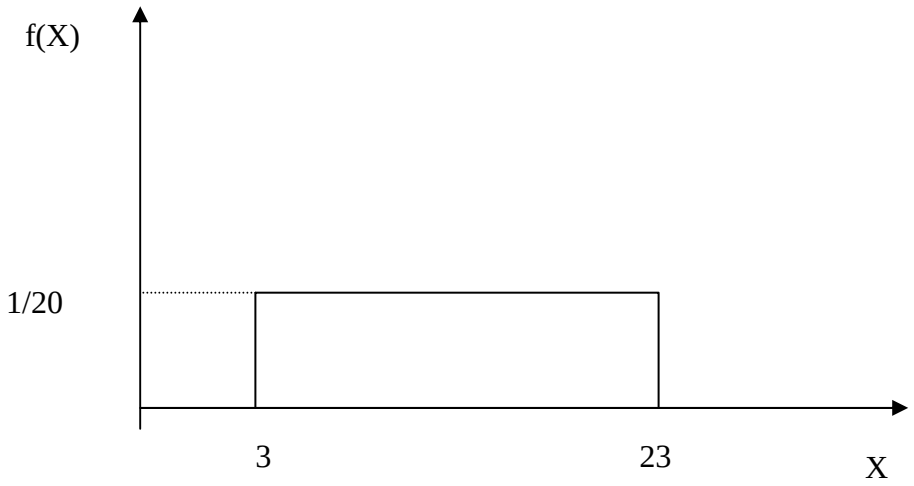


Fig 4.2

A f(X) é uma função constante cujo valor é $1/20$, tal que $x \in [3,23]$. Essa função goza das seguintes propriedades:

1) É sempre positiva ou nula. Isto é, $f(X) \geq 0$, para $\forall x$
2) Integrando-a no intervalo $3 \leq X \leq 23$ o valor desta integral definida será igual a 1. Ou seja, $\int_3^{23} f(X)dx = \int_3^{23} (1/20)dx = [x/20]_3^{23} = \frac{23}{20} - \frac{3}{20} = 1$

Toda função que satisfizer essas duas propriedades **é denominada função densidade de probabilidade**. Essa função é usada no lugar da função de distribuição de probabilidade quando falamos de variáveis aleatórias discretas. Suponhamos que queiramos calcular a probabilidade da variável aleatória contínua X estar contida no intervalo $15 \leq X \leq 20$ teremos:

$$P(15 \leq X \leq 20) = \int_{15}^{20} f(X)dx = \int_{15}^{20} (1/20)dx = [x/20]_{15}^{20} = \frac{20}{20} - \frac{15}{20} = \frac{5}{20} = 0{,}4$$

Resumindo: $P(a \leq X \leq b) = \int_a^b f(X)dx$

Média e Variância de uma Variável Aleatória Contínua

A média (ou valor esperado) de uma variável aleatória contínua é dada pela expressão:

$E[X] = \int_{-\infty}^{+\infty} Xf(X)dx$ sendo que a sua variância é $V[X] = \int_{-\infty}^{+\infty} (X - E[X])^2 f(X)dx$

4.2 Variável Aleatória Normal

Na maior parte dos casos trabalhamos com situações em que as variáveis tem um comportamento normal, mensurável a partir de alguma medida. A maior parte dos estudos são conduzidos a partir de estudos em que a respectiva população tem comportamento normal. Por isso tomamos uma especial atenção para este tipo de distribuição por ser a mais usada nos estudos que necessitem de análises estatísticas minuciosas. Por isso, a distribuição normal :
É a mais importante (e mais utilizada na prática) porque muitas variáveis da natureza comportam-se normalmente. Exemplo, peso dum adulto (são poucos com peso muito baixo, poucos com peso exagerado, etc.)
Tem uma função densidade de probabilidade (chamada de curva normal) que apresenta a forma de um sino e é unimodal no centro da distribuição, onde também se localizam a média e mediana.
Tem a mediada coincidente com a moda e média, garante que metade da área sob a curva esteja acima do ponto central e a outra metade abaixo dele.
É simétrica em relacção a sua média e assintóptica, aproximando-se cada vez mais do eixo X mas nunca o toca.

Características de uma Função Densidade de Probabilidade Normal (Distribuição Normal)
Para distribuições normais pode existir casos em que haja diferença de variâncias, mas com igual média, havendo dois gráficos em que uma é mesocúrtica e outra platicúrtica (veja nas medidas de curtose). Na realidade distribuição normal é um nome que define uma família de infinitas distribuições normais particulares, cada uma com os seus valores específicos de média e desvio padrão. O que caracteriza, portanto, e diferencia uma distribuição normal de outra são os valores destes dois parâmetros: a sua média e a variância. A função densidade de probabilidade de uma variável aleatória normal é dada por: $f(X)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-(X-\mu)^2}{2\sigma^2}}$.

Repare que a função densidade mostra exactamente que a média e variância continuam a ser os únicos que podem diferenciar duas distribuições normais.

A sua média e variância são dadas por $E[X]=\int_{-\infty}^{+\infty}Xf(X)dx=\int_{-\infty}^{+\infty}X\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-(X-\mu)^2}{2\sigma^2}}dx=\mu$

e $V[X]=\int_{-\infty}^{+\infty}(X-E[X])^2 f(X)dx=\int_{-\infty}^{+\infty}(X-\mu)^2\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-(X-\mu)^2}{2\sigma^2}}dx=\sigma^2$, respectivamente.

4.3 Distribuição Normal Padrão

Repare-se que a todo o momento que fosse necessário calcular a probabilidade seria necessário integrar a função densidade. O cálculo do integral requer métodos mais complexos, daí o uso da forma padronizada. O padrão criado é nas condições em que a média é zero e a variância 1, representado por $Z\sim N(0,1)$, que se lê: Z segue uma distribuição normal com média 0 e variância 1.

Se X é uma variável aleatória normal com média μ diferente de zero e desvio padrão σ diferente de 1 podemos "converter" essa distribuição em uma distribuição normal padrão através da transformação linear: $Z = \frac{X - \mu}{\sigma}$, repare que se média é zero e desvio 1, teremos $Z = \frac{X - 0}{1} \Leftrightarrow Z = X$.

Exemplo: Os salários mensais de licenciados, no aparelho do estado em Moçambique são em média \$300 com um desvio padrão de \$83. Foi entrevistado um professor que falou do seu salário de cerca de \$260. Qual é o valor de Z

Para X = 260 ➔ $Z = \frac{X - \mu}{\sigma} = \frac{260 - 300}{83} = -0{,}48$

Um valor de Z = -0,48 indica que o valor de \$260 está localizado 0,48 desvio padrão à esquerda da média de \$300.

Áreas a esquerda da Curva Normal (Controle de qualidade)

Para uma variável normalmente distribuída, observa-se que:

1) Cerca de 68 % da área sob a curva normal está entre menos um e mais um desvio padrão da média ($\mu \pm 1\sigma$)

2) Cerca de 95 % da área sob a curva normal está entre menos dois e mais dois desvios padrões da média ($\mu \pm 2\sigma$)

3) Praticamente toda (99,74 %) a área sob a curva normal está entre menos três e mais três desvios padrões da média ($\mu \pm 3\sigma$)

Exemplo:

O uso diário de água por pessoa numa determinada cidade é normalmente distribuído com média μ igual a 20 litros e desvio padrão σ igual a 5 litros. Entre que valores caem cerca de 68 % das pessoas que fazem uso da água?

$\mu \pm 1\sigma = 20 \pm 1(5)$. Ou seja, cerca de 68 % das pessoas usam de 15 a 25 litros de água por dia.

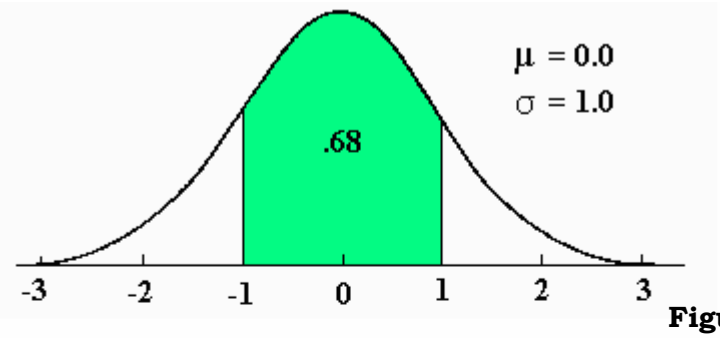


Figura 4.3 A

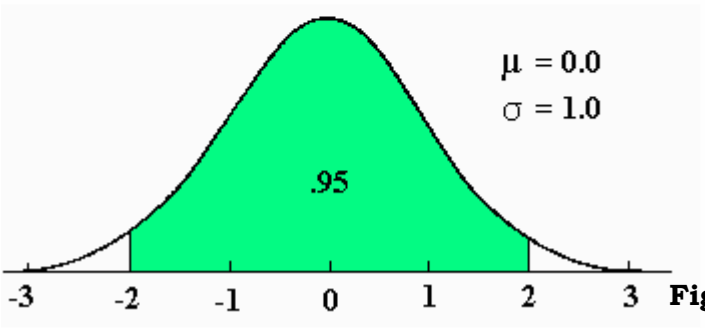


Figura 4.3 B

Similarmente, para 95 % e 99 %, os intervalos serão de 10 a 30 litros e 5 a 35 litros.

Qual é a probabilidade de que uma pessoa seleccionada ao acaso use menos do que 20 litros por dia ?

O valor de Z é Z = (20 – 20) / 5 = 0. Portanto P(X < 20) = $P(\frac{X-\mu}{\sigma} < \frac{20-20}{5})$ =P(Z < 0) = $\phi(0)$ =0,5. Repare que o valor de $\phi(0)$ é um valor tabelado, na tabela da distribuição normal reduzida.

Teorema do Limite Central

Para uma População com média μ e uma variância σ^2, a distribuição amostral das médias das possíveis amostras de tamanho n, geradas a partir da População, será normalmente distribuída – com a média da distribuição amostral igual μ e variância igual σ^2/n - assumindo que o tamanho amostral é suficientemente grande, ou seja, $n > 30$.

Noutras palavras, se a População tem qualquer distribuição *não precisa ser necessariamente normal* com média igual a μ e variância igual a σ^2, então a distribuição amostral dos valores médios amostrais é normalmente distribuída com a média das médias

($\mu_{\overline{X}}$) igual a média da População e o erro padrão das médias amostrais igual a $\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}$, desde que $n > 30$.

Note que o erro padrão das médias amostrais mostra quão próximo da média da População a média amostral tende a ser. Se σ não é conhecido e n ≥ 30 (considerada uma amostra grande), o desvio padrão da amostra, designado por s, é usado para aproximar o desvio padrão da População, σ. A fórmula para o erro padrão torna-se:

$$s_{\overline{X}} = \frac{s}{\sqrt{n}} \qquad \text{onde} \qquad s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

Exercícios Resolvidos

1- Alguns casais preferem ter filhos do sexo femenino, porque as mães são portadoras de um distúrbio recessivo que é herdado por 50 % dos filhos, mas por nenhuma das filhas. O método Ericsson se selecção de sexo tem uma taxa admitida de 75% de sucesso. Suponha que 100 casais utilizam o método Ericsson com o resultado de que, dentre 100 recém nascidos, há 75 meninas.

a) Supondo que o método Ericsson não produza efeito e admitindo que menino e menina sejam igualmente prováveis, determine a média e o desvio padrão do número de meninas num grupo de 100 crianças.

Resolução:

X-número de meninas em 100 nascimentos
Supondo que o método não produza efeito e que as meninas e meninos sejam igualmente prováveis, tendo $n = 100;\ \ p = 0{,}5 \quad q = 0{,}5 \Rightarrow \mu = np = 50 \quad \sigma = \sqrt{npq} = 5$
Resposta: Para grupos de 100 casais com um filho cada, o número médio de meninas é 50 com um desvio de 5.

b) Interprete os resultados de a) para determinar se o resultado de 75 meninas em 100 bebés confirma a alegação de eficiência do método.

Resolução:
$mínimo \approx \overline{X} - 2\sigma = 50 - 2\times 5 = 40 \quad máximo \approx \overline{X} + 2\sigma = 50 + 2\times 5 = 60$. Neste caso os resultados típicos estão entre 40 e 60.
Resposta: 75 meninas não parece um resultado devido unicamente ao acaso ou $z = \frac{X - \mu}{\sigma} = \frac{75 - 50}{5} = 5$, logo o resultado de 75 meninas não é usual. Pode-se concluir que o método Ericsson é eficiente.

2- As maternidades do hospital Central A e do Hospital B, estão interessadas em controlar o absentismo ao serviço das senhoras grávidas. Abriu-se um ficheiro que continha vários dados referentes a dias de abstenção ao serviço para consultas com o ginecologista, o que produziu os dados constantes nas tabelas abaixo.

Hospital A

25	28	63	70	42	33	46	79	85
14	89	98	56	75	28	56	48	91
85	76	42	81	93	45	41	86	98
78	95	84	20	53	40	80	60	70
75	43	51	61	81	72	94	83	54
86	75	43	61	68	59	72	47	41
78	45	76	84	73	51	48	92	85
76	87	92	43	57	86	75	84	91
73	84	91	42	86	75	10	14	11

Hospital B

15	18	43	92	74	73	61	63	84
75	76	94	82	76	84	83	91	87
87	79	78	96	94	92	83	84	54
56	58	59	52	57	53	55	84	61
60	63	68	69	62	67	64	85	71
73	74	77	76	84	79	81	80	82
91	90	80	70	75	76	84	73	94
84	71	84	71	24	28	27	39	37
29	34	81	46	48	47	45	41	90

Para as alíneas c) e d) devia-se calcular antes, as seguintes medidas:

	Hospital	
	A	B
N	81	81
$\sum X_i$	5228	5527
$\sum (X_i - \overline{X})^2$	42772,10	31238,54
μ	64.54	68.23
σ^2	534.65	390.48
σ	23.12	19.76

a) Retire das duas populações amostras aleatórias de tamanho 14, usando a tabela de números aleatórios.

Resolução:
Hospital A: 92; 45; 86; 25; 75; 73; 28; 89; 98; 84; 84; 91; 86; 86.
Hospital B: 84; 92; 28; 15; 47; 75; 84; 76; 94; 70; 34; 81; 73; 84.

Diga qual é o hospital com maior variância em faltas?
Resposta: É o hospital B.

$$S^2_{HCM} = \frac{\sum (X_i - \overline{X})}{n-1} = \frac{7403,43}{13} = 569,49 \qquad S^2_{MACAMO} = \frac{\sum (X_i - \overline{X})}{n-1} = \frac{8320,93}{13} = 640,07$$

b) Determine a média de cada amostra
Resolução:

$$\overline{X}_A = \frac{\sum X_i}{n} = \frac{1042}{14} = 74{,}4 \qquad \overline{X}_B = \frac{\sum X_i}{n} = \frac{937}{14} = 66{,}9$$

c) Qual é a probabilidade de que uma senhora escolhida ao acaso tenha faltado mais de 61 dias relativamente aos dados do B
Resolução:

$$P(X > 61) = P(Z > \frac{61-\mu}{\sigma}) = P(Z > \frac{61-68{,}23}{19{,}76}) = P(Z > -0{,}366) = \emptyset(0{,}366) = 0{,}6428$$

d) Qual é a probabilidade de que o número de faltas esteja ente 25 e 50 para uma doente do Hospital A
Resposta:

$$P(25 < X < 50) = P(\frac{25-64{,}54}{23{,}12} < Z < \frac{50-64{,}54}{23{,}12}) = P(-1{,}71 < Z < -0{,}629) = \emptyset(-0{,}629) - \emptyset(-1{,}71)$$
$$= \emptyset(1{,}71) - \emptyset(0{,}629) = 0{,}9564 - 0{,}7353 = 0{,}2211$$

Exercícios Propostos

1- Dada a distribuição binomial na forma b(r,n,p), encontre (i) $b\left(1; 5, \frac{1}{3}\right)$, (ii) $b\left(2; 7, \frac{1}{2}\right)$, (iii) $b\left(2; 4, \frac{1}{4}\right)$. **[2]

2- Três cartas são selecionadas, com reposição, de um baralho comum de 52 cartas. Encontre a probalidade de (i) duas copas serem retiradas, (ii) três serem retiradas, (iii) pelo menos uma copa ser retirada. **

3- A média de rebatidas de um jogador de beisebol é de 0,300. Se ele bate 4 vezes, qual é a probabilidade dele obter (i) dois acertos, (ii) pelo menos um acerto? **

4- Uma caixa contém 3 bolas vermelhas e 2 branca. Três bolas são retiradas, com reposição, da caixa. Encontre a probabilida de de (i) 1 bola vermelha ser retirada, (ii) 2 bolas vermelhas serem retiradas, (iii) pelo menos uma bola vermelha ser retirada. **

5- A equipa A tem $\frac{2}{5}$ de probabilidade de vitória sempre que joga. Se disputa 4 partidas, encontre a probabilidade de vencer (i) 2 partidas, (ii) pelo menos uma partida, (iii) mais que a metade das partidas. **

[2] Os exercícios com ** foram retirados do Seymour, L. (1993). Probabilidade.

6- Uma carta retirada e recolocada em um baralho comum. Quantas vezes devemos repitir esse procedimento, de modo que (i) a probabilidade de retirar uma copa maior que $\frac{1}{2}$, (ii) a probabilidade de retirar uma copa seja maior que $\frac{3}{4}$? **

7- A probabilidade de um homem acertar o alvo é $\frac{1}{3}$. (i) Se ele atira 5 vezes, qual é a probabilidade de acertar o alvo pelo menos duas vezes? (ii) Quantas vezes deve ele atirar?, de modo que a probabilidade de acertar , ao menos uma vez, seja maior de 90%? **

8- O departamento de Matemática tem 8 professores graduados ocupando o mesmo gabinete. Cada um tanto estuda em casa como no gabinete. Quantas escrivaninhas deve haver no gabinete, de modo que cada um tenha pelo menos 90% do tempo? **

9- Dos parafusos produzidos por uma fábrica, 2% são defeituosos. Em um depósito de 3.600 parafusos de fábrica, encontre o nº esperado de parafusos com defeitos e desvio-padrão. **

10- Um dado não-viciado é atirado 1.620 vezes. Encontre o número esperado de vezes em que o número 6 ocorre e o desvio-padrão. **

11- Seja X uma variável aleatória com a distribuição binoinal com $E(X)=2$ e Var $(X)=\frac{4}{3}$. Encontre a distribuição de X. **

12- Considere a distribuição binominal $P(K)=b(K;n,p)$. Mostre que: **

(i) $$\frac{P(K)}{P(K-1}=\frac{b(K;n,p)}{b(K-1;n,p)}=\frac{(n-K+1)p}{Kq}$$

(ii) $$P(K)\rangle P(K-1) \text{ para } K\langle (n+1)p \quad e$$
$$P(K)\langle P(K-1) \text{ para } K\rangle (n+1)p$$

13- Admita-se que o atraso nas chegadas a estação de uma cidade dos comboios directos provenientes de outra segue uma distribuição normal T(0,12). Calcular a probabilidade de ocorrer um atraso compreendido entre os 5 e os 1o minutos. Determine o valor esperado e a variância da variável atraso na chegada.

14- A função densidade de probabilidade da variável X, que representa o tempo de funcionamento sem avarias, expresso em dias, de um determinado equipamento tem o parâmetro igual a o,5. Calcule a probabilidade de o equipamento funcionar sem avarias durante um período compreendido entre 1 e 3 dias. Determine o valor esperado e a variância.

15- A variação relativa diária da cotação de fecho de um determinado fundo transacionado numa bolsa de valores pode ser razoavelmente aproximada por uma distribuição normal com valor esperado de 0,2% e desvio padrao 1,6%. Pretende-se calcular:

a) A probabilidade de a proxima variação do preço de fecho ultrapassar 1%.
b) Aprobabilidade de a proxima variação do preço de preço de fecho se situar entre 1% e 1,16%.

16- tempo de funcionamento sem avarias de uma determinada máquina de produção continua segue uma lei exponencial negativa com valor esperado igual a 4.5 horas. Imagine a máquina a ser colocada em funcionamento no instante t=o horas.
a) Qual é a probabilidade de não ocorrer qualquer avaria antes do instante t=6 horas?
b) Admitindo que a máquina se encontrava ainda em funcionamento no instante t=4 horas, qual a probabilidade de não ocorrer qualquer avaria antes do instante t=6 horas?
c) Qual e a probabilidade de se verificarem duas avarias durante as primeiras 6 horas de funcionamento da máquina?

17- A altura dos cidadãos adultos de um determinado país segue uma distribuição normal com valor esperado igual a 1,70 m e com desvio padrão igual a 0,05m.
a) Qual é a probabilidade da altura de um cidadão ser de 1,80 m?
b) Qual é a probabilidade de a altura de um cidadao ultrapassar 1,80 m?
Sabe-se que um determinado cidadão tem uma altura superior a 1,75m. Qual é a probabilidade de ter uma altura superior a 1,80 m?

18- As durações da gravidez têm distribuição normal com média de 268 dias de desvio padrão de 15 dias.
a) Selecciona-se aleatoriamente uma mulher grávida; determine a probabilidade de que a duração da gravidez não seja inferior a 265 dias e inferior a média

b) Se 25 mulheres escolhidas aleatoriamente são submetidas a uma dieta especial a partir do dia em se engravidam, determine a probabilidade dos prazos de duração da sua gravidez terem média não inferior a 265 dias e inferior a média (admitindo que a dieta não produz algum efeito).

19- A admissão a uma empresa de segurança privada é limit ada a mulheres e homens. A exigênia da altura mínima para as mulheres é de 1,70ms. Asalturas das mulheres tem média de 1,63m e desvio pdrão de 0,25m. Ache o valor de Z, correpondente a uma mulher com altura de 1,70m e determine se se trata de uma altura fora do comum
20- Em determinada empresa a utilização da matéria prima F é uma variável aleatória com distribuição normal de parâmetros μ = 600 kg e σ = 40 kg. No início de determinada semana, a empresa tem em stock 634 kg de matéria prima, não sendo viável no decurso dessa semana realizar mais aprovisionamentos.
Determine a probabilidade de ruptura de stock da matéria prima.
Qual deveria ser o stock de modo a que fosse de 0,01 a probabilidade de ruptura ?

21- O rendimento anual das famílias de uma certa cidade pode ser expresso através de uma variável contínua. Sabe-se que a mediana do rendimento é de 60.000.000,00 MT e que 40% das famílias da cidade tem um rendimento superior a 72.000.000,00MT.

Escolhida uma família aleatoriamente, qual é a probabilidade desta ter um rendimento entre 60.000.000,00Mt e 72.000.000,00Mt?
Sem mais informação adicional, o que pode dizer sobre a probabilidade de uma
família ter um rendimento menor que 65000000,00Mt?

22- Suponha que o tempo que um operador leva para executar uma certa actividade seja normalmente distribuído, com o tempo médio de 12 minutos e desvio de padrão de 1,5 minutos. Se o operador está realizando esta actividade repetidamente, qual é probabilidade de que, em certo momento, ele leve entre 9 e 15 minutos para executar uma operação deste tipo? ***

23- Suponha que uma fábrica tenha estabelecido que a vida média dos pneus para automóveis, de sua fabricação, é de 35.000 Km rodados, com um desvio padrão de 3.000 Km. Suponha ainda que o tempo de duração dos pneus seja uma variável aleatória normalmente distribuída. ***

Se a fábrica fornecer uma garantia de 30.000 Km, em condições normais de uso do veículo, qual a probabilidade de que um pneu vendido tenha de ser substituído?

Nas condições do item (a), qual a percentagem de pneus que terão de ser substituidos?

Qual quilometragem a fábrica deve oferecer como garantia, para que nenhum pneu vendido tenha de ser substituido?

A fábrica está preocupada em melhorar a qualidade dos pneus e, para isso, está sendo estudada a possibilidade de se aumentar a duração média dos pneus. Desta forma, qual deveria ser a duração média para que, com uma garantia de 30.000 Km, somente 1% dos pneus vendidos tenham de ser trocados?

24- Um fabricante de refrigerantes vende um dos seus produtos engarrafados em vasilhames de 1 litro. Para engarrafar este produto é utilizado uma máquina, que, calibrada, permite obter o volume desejado, segundo uma normal, com um desvio de 30 ml ***[3]

Se o órgão fiscalizador do governo (OFG) faz a exigência de que não mais de 8% de garrafas tenham um volume menor do que o nominal, em quanto deve ser regulada a máquina para que o fabricante não seja autuado?

b) Se a máquina for calibrada para colocar 1.035 ml de líquido no vasilhame, qual a porcentagem de vasilhames que não estarão atendendo às especificações do OFG?

c) Para qual valor deve ser ajustada a precisão da máquina, para que, estando calibrada em 1.350 ml, as especificaçãoes do OFG seja atendidas?

25- Uma amostra aleatória de 10 pacotes de café foi selecionada do estoque de um grande supermercado. Observou-se os seguintes pesos (em g): 497,5; 499,2; 500,3; 491,8; 502,7; 493,9; 497,4; 509,8; 503,2. Encontre estimativas para o peso médio e a variância. Considerando o peso dos pacotes como uma variável normalmente distribuida, obtenha também intervalos com 95% de confiança para os mesmos parâmetros estimados. ***

26- Foram observados os tempos de duração do intervalo para o "cafezinho", para uma amostra de 20 empregados de uma empresa, obtendo-se os seguintes resultados, em minutos: ***

15,79 15,75 18,11 14,54 10,06 17,32 18,52 16,11 13,59 18,63 16,27 13,75 15,16 14,75 13,03 18,47 12,14 14,67 16,52 12,47

Encontre a média e a variabilidade estimadas do tempo de duração do intervalo para o "cafezinho" dos funcionários da empresa. Encontr, ainda, intervalos de 90% de confiança para a média e a variância, supondo a variável tempo distribuida segundo uma Norma.

[3] Todos exercícios com *** foram retirados do livro: Exercícios retirados do Reginaldo, C., at al (1999). ***Análise de Modelos de Regressão Linear***. pp. 22-23

27- Suponha que a pressão sanguinea sistólica seja uma variável distribuida segundo uma norma. Foi observada a pressão de um grupo 16 pacientes de uma clínica, obtendo-se os seguintes resultados, em mm de Hg: ***
121,3 118,8 127,9 132,5 146,3 110,7 152,3 126,7
120,9 110,8 142,3 135,7 140,8 137,6 128,3 113,9
Estime e encontre um intervalo de 99,5% de confiança para a pressõ sistólica média dos pacientes desta clínica.
28- Para o estudo do consumo médio de combustível para uma determinada marca de automóvel, foi observado o consumo de uma amostra de 20 destes veículos, obtendo-se uma média de 16,7 KM/l e um desvio padrão de 2,3Km/l. Construa um intervalo de 95% de confiança para o consumo médio de combustível, para este tipo de veículo. Suponha o consumo de combustível aproximadamente normal. ***

29- Seja ϕ a distribuição normal padrão **
Encontre $\phi\left(\frac{1}{4}\right), \quad \phi\left(\frac{1}{2}\right) e\, \phi\left(-\frac{3}{4}\right).$
Encontre τ de modo que a) $\phi(\tau) = 0{,}100$, b) $\phi(\tau) = 0{,}2500$, c) $\phi(\tau) = 0{,}4500.$

30- Seja X uma variável aleatória com distribuição normal padrão ϕ. **[4]
Encontre
(i)
$P(-0{,}81 \le X \le 1{,}13),$

(ii) $P(0{,}53 \le X \le 2{,}03),$
(iii) $P(X \le 0{,}73),$

(iv) $P\left(|X| \le \frac{1}{4}\right).$

[4] Exercícios com ** foram retirados do Seymour, L. (1993). Probabilidade.

CAPÍTULO 5 INFERÊNCIA ESTATÍSTICA

Objectivos do Capítulo:

- Explicar as razões da necessidade de recolha de amostras para diversos estudos (antropológicos, sociais, económicos, políticos, etc).
- Indicar as principais técnicas a usar para recolha duma amostra.
- Diferenciar amostragem probabilística de amostragem não probabilística.
- Saber retirar amostras para realizar inferências estatísticas.
- Definir estimador e estimativa.
- Descrever a estimação por ponto e por intervalo.
- Descrever intervalo de confiança da média, variância, proporção e desvio padrão de uma amostragem.
- Definir hipóteses e Testes de Hipóteses.
- Descrever os 5 passos do procedimento de Teste de Hipóteses.
- Distinguir entre Teste de Hipóteses Unicaudal e Bicaudal.
- Realizar um teste para a média populacional e dar exemplo para proporções e variâncias.
- Realizar um teste para a diferença entre duas médias ou proporções populacionais.
- Descrever os erros estatísticos associados aos testes de hipóteses.

1. SONDAGEM E TÉCNICAS DE AMOSTRAGEM

1.1 Introdução

Na perspectiva etimológica, sondagem tem origem na palavra francesa *sondage,* (Costa e Melo, 1976), que surgiu provavelmente no séc. XIV para expressar o acto de, com recurso a uma sonda, investigar a profundidade da água e a natureza do fundo de um rio ou mar. No séc. XIX, Balzac utiliza-o para expressar a ideia de uma pesquisa ou investigação rápida (Droesbeke et al., 1987). A associação do termo sondagem ao domínio marítimo ainda hoje permanece, mas coexiste já com a aplicação a outras áreas, como sejam a geologia, a medicina ou a estatística. A língua portuguesa não apresenta distinção vocabular para os diversos domínios, mas por exemplo a língua inglesa diferencia todas estas formas de

sondagem. Sounding, boring, probing, designam respectivamente a sondagem marítima, a geológica e a médica (Hornby, 1980). No domínio estatístico diferencia a sondagem de opinião – *poll*- dos outros tipos de sondagem que designam de *survey sampling*.

Em geral, quando se pretende fazer o estudo de uma população com muitos indivíduos/características somos obrigados a retirar uma amostra. A preferência na amostra é importante devido a vários constrangimentos operacionais, económicos, temporais, demográficos, políticos e/ou históricos.

Quando nos propomos a obter uma amostra porque a população tem um tamanho grande[5] devemos elaborar um inquérito/entrevista[6], a primeira questão a colocar, depois de definido o problema e equacionadas as hipóteses é "a quem questionar?". Esta questão remete-nos a duas etapas seguintes:

a. Qual é a população que é objecto de estudo?

b. Como escolher, nessa população, as unidades a questionar efectivamente, dado que, na maior parte dos casos, se exclui a hipótese, de as interrogar todas juntas?

Como a primeira questão não é sempre explicitada (por exemplo o conjunto de bovinos numa província), então, centramo-nos na segunda, que cobre os problemas dos métodos de amostragem e da dimensão da amostra, tendo em vista o melhoramento do questionário e/ou diminuir-lhe os custos, escolhendo de forma adequada as populações a tomar em consideração.

A falta de coerência nas sondagens de opinião (resultantes de erros na previsão de vitórias em casos de eleições), ensinou ao grande público que é possível obter uma informação digna de confiança, sobre uma população de várias dezenas de milhões, interrogando apenas alguns milhares. Para tal, o recurso às técnicas de amostragem não é exclusivo das sondagens de opinião. Podem ser utilizadas nos mais variados fins, como por exemplo um determinado nível de ensino analisar uma amostra representativa das milhares de classificações ao longo de um determinado ano lectivo, para obter informações relativas à totalidade das classificações. Como população podemos entender como sendo conjunto de pessoas, objectos, coisas, animais de qualquer natureza, que tenham uma determinada característica em comum. Assim, também definimos técnica de pesquisa por amostragem em interrogar um subconjunto da totalidade da população que interessa aos objectivos do questionário (o inverso seria chamado recenseamento, ou censo).

Esse subconjunto populacional (a amostra), deverá apresentar as características da totalidade da população para que, depois de retiradas as conclusões sobre a amostra, seja válido alargá-las a toda a população (processo designado de inferência), sendo a amostragem a parte da estatística que estuda os processos de selecção de amostra.

Porquê recolher amostra numa População?

a) Natureza destrutiva de certos testes.

Exemplo: Imagine que pretenda ter a certeza de que os palitos de fósforo que uma fábrica produz acendem ou não, não será necessário experimentar todos palitos fabricados para a venda, sob pena de não ter algum no armazém para venda.

[5] Refere-se ao número de habitantes dum país, da população bovina dum país, de estudantes dum nível de ensino, de doentes contaminados pelo HIV, etc

[6] Vamos usar o termo questionário por uma questão de simplificação na linguagem

b) A impossibilidade física de obter todos os itens na População

Exemplos: Se pretender estudar a razão das pessoas gostarem de bebida tradicional, seria difícil entrevistar toda a população do país por motivos de falta de acessibilidades, problemas de conflito armado ou porque uma parte dos nacionais emigraram do país.

c) O custo de estudar todos os itens numa População é frequentemente proibitivo

Exemplo: Para o caso de Moçambique, se quiséssemos inquirir a quem as pessoas votariam em eleições presidenciais se as eleições ocorressem naquela semana, precisaríamos de muito dinheiro para comprar muitos carros a tracção de modo que se conseguisse cobrir todo o país.

d) Muitas vezes as estimativas baseadas numa amostra são mais precisas do que os resultados obtidos através de um levantamento censitário.

Exemplo: Ao se realizar o recenseamento geral da população podem nalguns casos certos pais mentirem que não tem filhos ou diminuir o número de filhos por várias situações. Para o caso de Moçambique às vezes os pais podem pensar que talvez queira-se saber o número de filhos para os levar ao serviço militar obrigatório que seria penoso para aquela família, dependendo dos seus valores sociais.

e) Tempo elevado para apurar resultados em censos.
Exemplo: Em Moçambique os resultados do censo geral da população levaram mais de um ano para serem divulgados, o que as vezes faz com que diversas políticas falhem por se trabalhar com previsões cesgadas.

Razões de uso da técnica de Amostragem:

1- A população é infinita - quando a população é grande, como é o caso dos infectados pelo vírus de HIV, para fins práticos, podemos admitir como infinita, em virtude do número de infectados no mundo subir de minuto a minuto;

2- Economia – uma amostra é menos dispendiosa, pois observam-se menos unidades;

3- Rapidez – como se constata facilmente. Geralmente para se garantir que o estudo seja fiável, deve levar pouco tempo;

4- Maior precisão – é possível examinar uma amostra com mais cuidado do que a população;

5- Problemas de acessos - casos de dificuldades de acesso porque a rede viária (vias de acesso) não está em condições, as populações estão em zonas de difícil acesso;

6- Instabilidade Política – referindo-se a guerras, tumultos e outros;

7- Problemas Demográficos – Algumas zonas com maior extenção territorial podem apresentar pouca densidade populacional, etc.

8- Outra razão para recorrer a amostras é que, por vezes, os outros **processos que manipulam a população são destrutivos**. Imagine-se que um fabricante de fósforo pretenda estudar se a concentração do fósforo no palito em contacto com a caixa acende ou não. Para isso, é conveniente utilizar utilizar apenas uma amostra de fósforo.

1.2 A Técnica da Amostragem

Como foi referido atrás, na impossibilidade de questionar toda a população, deve-se recorrer a amostra. Efectivamente pelo facto desta ser demasiado grande ou infinita, a maior parte das vezes torna-se difícil ou mesmo inviável recolher as opiniões da totalidade da população a inquirir. Assim teremos de usar um processo que nos permite tirar conclusões válidas, auscultando apenas uma parte ou um subgrupo do universo. A esta técnica, que consiste em questionar um subconjunto, considerando as suas opiniões representativas de todo o universo, é costume chamar-se técnica de amostragem. Ao subconjunto questionado, tirado do universo, designamos por amostra ou parte representativa do universo conceptual (população a estudar).

Fixada a população em estudo, a operação seguinte a levar a cabo é a da recolha da amostra. O problema que se põem nesta operação é o da sua representatividade e o critério a usar para a sua obtenção, que se traduz no facto da estrutura da amostra coincidir com a estrutura da população. Atente-se para o facto de que a maior amostra é a própria população. Aconselha-se que se deva evitar ter mais dados na amostra que se aproxime ao tamanho da população A validade da generalização das conclusões tiradas, depende dessa correspondência estrutural. Na posse da amostra, a operação seguinte, consiste na recolha da informação que vai ser tratada, de forma a poderem ser tiradas as respectivas conclusões, resultantes do cálculo dos estimadores (medidas estatísticas da amostra).

As conclusões a que o encarregado por estudar a amostra chega, apenas, e por enquanto, são válidas para a amostra. De seguida terá de proceder à sua generalização à população, em linguagem técnica, este processo de generalizar denomina-se inferência ou diz-se que se está a inferir.

Quando se fazem inferências, cometem-se, em geral, dois tipos de erros que podem de certa forma afectar os resultados:

1- Erros de observação – resultantes, por exemplo, da ausência ou imprecisão das respostas a certas questões. Imagine casos em que o inquirido não responde enquanto queríamos que a pergunta fosse respondida, o outro caso é do inquirido mentir.

2- Erros de amostragem (ou aleatórios) – que são devidos ao facto de se extrapolar os dados da amostra para o universo, já que é pouco provável que um parâmetro da amostra seja igual ao seu equivalente da população.
Num recenseamento (ou censo), não existem erros aleatórios, mas existem os erros de observação.

1.3 Fases para construção de uma Amostra

Existem três fases fundamentais para construção duma amostra. Eis a descrição de cada fase a seguir:

a) **Determinar o tipo de amostragem** que o inquérito requer mediante a extensão e conhecimento da população, território do inquérito, acessibilidades, duração, custos...;

b) **Calcular ou determinar o tamanho da amostra**, tomando em consideração as principais características da população que constitui a unidade de análise. Algumas Técnicas de amostragem, que será tratado ainda neste capítulo, ajuda a partir do tamanho da amostra, determinar quanto a tirar por estrato;

c) **Construir a amostra** (através da listagem de pessoas ou dos elementos da população a interrogar, procurando recolher as suas informações de forma mais precisa...).

1.4 Amostra Representativa

Uma amostra para que seja considerada num estudo sério, fiável e de responsabilidade, deve ser uma amostra representativa. Afinal de contas o que é uma amostra representativa? Uma população apresenta um certo número de características. O que interessa ao inquiridor/entrevistador é obter uma fracção desta população na qual alguma das características estaria distribuída do mesmo modo que na população, ou aproximadamente. Que características? Aquelas que têm uma relação directa com o questionário que vamos realizar. Um estudo sobre as opiniões de professores numa escola não tem, evidentemente, nada a ver com a cor dos cabelos, ao passo que um estudo sobre as modas capilares se pode interessar por essas características.

A escolha das características a reter remete ao investigador, que tome geralmente algumas em consideração, quatro ou cinco no máximo. Uma amostra nunca é representativa senão em função das características retidas, com exclusão de todas as outras (estas características são também chamadas variáveis quando são de natureza numérica ou de natureza qualitativa: a idade, o rendimento, cor dos olhos, atitudes, etc.). Esta precisão é muito importante para compreender a filosofia da amostragem frequentemente posta em causa pelos desencontrados com esta matéria.

Calculando a dimensão da amostra, o organizador do questionário aceita um certo grau de erro. É o preço que é preciso pagar para ser dispensado de interrogar a população inteira. Assim, pode decidir trabalhar com 5% ou 10% de erro: nestes casos há 5 ou 10 possibilidades em 100 para que a amostra, depois de feitas a triagem e a verificação da distribuição das características retidas, não seja representativa. Ao mesmo tempo ele aceita um outro erro que diz respeito ao grau de precisão das distribuições da características da amostra. Assim, se uma das variáveis retidas é a idade, e se a média de idades da população é de 50 anos, pode aceitar que a média de idade da amostra seja 50 anos, mais ou menos 2 anos, isto é que ela esteja compreendida entre 48 e 52 anos. Esta margem é deixada à sua apreciação. No que diz respeito à idade, por exemplo, é evidente que um comportamento varia pouco conforme se tem 60 ou 62 anos, mas, ao contrário, varia fortemente segundo se tem 12 ou 14 anos. O realizador do questionário deve avaliar os erros aceites em função do papel que apresenta a característica dos fenómenos que quer estudar. A combinação das duas latitudes, a da probabilidade de que amostra seja inválida,

e a que se refere ao grau de precisão das características retidas, determina a dimensão da amostra.

Uma amostra é representativa se as unidades que a constituem forem escolhidas por um processo tal que todos os membros da população tenham a mesma probabilidade de fazer parte dela. Se não for esse o caso, diremos que a amostra é enviesada ou viciada, visto que certos indivíduos tiveram mais hipóteses do que outros de serem escolhidos e as categorias a que pertencem ocuparam mais espaço na amostra do que deveriam: as características da amostra serão então sistematicamente diferentes das da população. Por exemplo, se quiséssemos estudar toda a população de professores de uma escola, não poderíamos ir para essa escola num determinado dia e interrogar um determinado número de professores (amostra) num determinado local, a uma determinada hora. Estaríamos perante uma amostra enviesada, uma vez que nem todos os professores vão à escola todos os dias, nem todos frequentam a sala de professores. O ideal seria dispor de uma lista de todos os professores da escola (base de sondagem), tirando à sorte um número igual (superior, na prática, tendo em conta as possíveis recusas e outras) à dimensão da amostra desejada. Neste caso estaríamos perante uma amostra representativa: todos os professores da escola teriam a mesma probabilidade de serem escolhidos.

Colocar o problema da representatividade por si só, e querer a qualquer preço uma amostra representativa, é impor uma condição difícil de satisfazer e, muitas vezes inútil. É necessário substituir a noção de representatividade por uma noção mais ampla, a de adequação da amostra aos objectivos estabelecidos, sabendo-se que um questionário, visa em geral, diversos objectivos (na prática, isso significa que estão previstos diversos tipos de análise) e que não é necessariamente a mesma amostra que, inicialmente, seria considerada óptima para cada um deles. Certos compromissos são então necessários. Quando o objectivo de um questionário é fazer a estimativa das grandezas, a representatividade exacta da amostra é uma condição necessária para a validade do resultado. Em contrapartida, a condição da representatividade é muito menos rigorosa quando tentamos verificar hipóteses sobre relações.

1.5 Sondagem

Numa sondagem a ideia fundamental é que se tenha uma amostra interrogando um número reduzido de pessoas, pertencendo a um ou vários grupos definidos, que são considerados representativos do ponto de vista estatístico. Como já foi referido, é raro, que se possa interrogar todos os membros de um grupo. Passa-se quase sempre por uma amostra dos membros do grupo, que podemos determinar por métodos estatísticos apropriados e que representa bem esse grupo. É necessário, pois dispor de uma amostra representativa. O questionário, apoiando-se numa amostra desse tipo chama-se sondagem.

Exemplo Prático de Uma Amostra Representativa

Queremos saber a opinião dos alunos sobre a qualidade de determinado serviço prestado pelas escolas de uma Cidade (EP2 + secundário), por exemplo, a biblioteca. Para tal, teremos, primeiramente, de identificar os indivíduos que podem responder validamente à questão. Com isto queremos dizer que as pessoas a inquirir deverão poder dar uma resposta que interesse aos objectivos da sondagem. Ora, neste caso, o universo, isto é o conjunto de

indivíduos que interessaria questionar, de acordo com os objectivos da pesquisa em curso, é formado por todos os alunos da escola, visto que todos terão, em princípio, uma opinião sobre a biblioteca escolar.

Identificada a população e reconhecendo a impossibilidade de interrogar a sua totalidade, há que recorrer à técnica da amostragem. Teremos, então, de definir uma amostra que seja representativa do todo o universo. Para tal, a amostra deverá respeitar as características do universo consideradas relevantes para o questionário que estamos a realizar. Neste exemplo, será importante considerarmos, o grau e ensino. No ensino secundário os alunos deverão, no sentido de aprofundarem e diversificarem os seus conhecimentos, recorrer a bibliografia complementar. Poderíamos ainda distinguir outros níveis de ensino, com diferentes necessidades em relação aos recursos da biblioteca. Por uma questão de simplificação vamos considerar os alunos do ensino secundário e os alunos do ensino básico. Ao que acabámos de fazer damos o nome técnico de *break-down* do universo (resultando em amostras estratificadas através de estratos homogéneos), ou seja, dividir o universo em subconjuntos, de acordo com critérios definidos, que nos possam revelar opiniões eventualmente diferentes.

No caso desta sondagem a estratificação do universo seria, por exemplo:
Ensino básico 1200 alunos, ensino secundário 800 alunos, dando um total de 2000 alunos
Em seguida teríamos um problema grave a resolver, que seria determinar o tamanho da amostra, isto é, o número de alunos a interrogar. Este problema é apresentado teoricamente adiante. Mas para já, vamos aceitar como representativa dos alunos da escola, uma amostra de 100 alunos. Aceitando o tamanho da amostra, haveria então a necessidade de calcular, que percentagem representam os 100 alunos em relação à totalidade dos 2000 alunos que a escola tem e que constitui a nossa população:

$$p = \frac{tamanhoamostra}{tamanhoPopulação} \times 100\% = 5\%$$

A percentagem encontrada será respeitada quando determinarmos a estratificação da amostra, ou seja, o número de alunos que vamos interrogar do ensino básico e do ensino secundário. Nesse sentido, deverá inquirir-se 5% do número total de alunos de cada um dos níveis considerados. Assim teremos:

Ensino básico 1200 x 5% = 60 alunos Ensino secundário 800 x 5% = 40 alunos
Poderemos, agora interrogar, os 60 alunos do ensino básico e os 40 alunos do ensino secundário, sabendo que o grupo assim constituído, a amostra, pode representar a opinião de todos os alunos da escola "a população", já que aceitámos serem os 100 alunos suficientemente representativos.

1.6 Tamanho duma amostra

O que é tamanho da amostra? A qualidade e a validade dos resultados de um questionário dependem da dimensão da amostra, ou por outras palavras, o número de pessoas/itens a interrogar depende da precisão desejada. Num grande número de casos, mais do que aumentar simplesmente a dimensão da amostra, o que pode levar a aumentar o número de pessoas/itens pertencentes a categorias já suficientes, há vantagem em construir uma amostra experimental, concebida de forma exacta em função das análises previstas, por forma a evitar amostras inutilmente grandes.

Atrás foi referido que deverá haver uma combinação de duas latitudes, a da probabilidade de que amostra seja inválida, e a que se refere ao grau de precisão das características retidas, é que determina a dimensão da amostra. Sem querermos aprofundar os cálculos matemáticos que a este nível seria preciso efectuar, remetemos apenas para as ideias principais que lhes estão subjacentes. Trata-se então, no essencial, de retirar a uma população determinada fracção na qual os diferentes caracteres possuam a frequência semelhante à da população inicial. Se, por exemplo, no âmbito do conjunto da população da escola (no nosso exemplo anterior), existe uma percentagem elevada de alunos que não frequenta a biblioteca, há que tentar transpor esse índice de frequência para a amostra. Logo, segundo a "lei das probabilidades", é necessário que a frequência relativa de um dado carácter (atitude do aluno na escola), se aproxime na amostra o mais possível da sua probabilidade de ocorrência, fornecida pela sua frequência relativa ao conjunto da população.

Por outro lado, a "lei dos grandes números"[7], identifica as condições nas quais uma exigência pode ser satisfeita: os acontecimentos cuja probabilidade é fraca, raramente podem ocorrer; a probabilidade de que a frequência relativa na amostra não se afaste mais do que um dado valor, é tanto maior, quanto maior for o número de observações registadas.

1.7 Tipos de Amostragem e Métodos de Amostragem

A capacidade que um investigador deve revelar na obtenção duma amostra deve ser aliada ao facto de ter uma margem para quantificar, a prior, a margem de erro que possa ser cometida durante o estudo. Isso facilita e reduz o nível de enviesamento, o que determina sobremaneira rigor nos resultados e sobretudo uma maior fieldade.

Ao dizermos que a margem de erro que nos levou a conclusões foi com a probabilidade de 2%, 5% ou 10%, estamos a dar uma informação sobre a credibilidade das conclusões generalizadas, garantindo uma confiança de 98%, 95% ou 90% respectivamente.

O controlo da probabilidade de erro na generalização está directamente relacionada com a forma como a amostra foi recolhida. A primeira e a última operação da técnica da amostragem estão directamente relacionadas. Só se pode quantificar o erro se a amostra for recolhida de uma forma aleatória, isto é, se a probabilidade de todos os elementos da população fazerem parte da amostra for conhecida e diferente de zero (e só nesta situação a amostra é representativa). Assim podemos definir dois tipos de amostragem:

A. Amostragem probabilística ou aleatória;
B. Amostragem não probabilística ou não aleatória.

A. Amostras aleatórias ou casuais

Obtém-se por um sorteio que respeite a condição de definição das amostras representativas: deve-se garantir que cada elemento (ou dado) da população tenha a mesma probabilidade

[7] formulada por C. Javeau

de fazer parte da amostra. Para esse efeito, a situação ideal é aquela em que dispomos de uma lista exaustiva da população, ou seja, de uma base de sondagem. É conhecida e diferente de zero a probabilidade dos elementos da população fazerem parte da amostra; a possibilidade de erro na generalização é quantificável.

Nas amostras não probabilística ou não aleatórias é igual a zero ou desconhecida a probabilidade de alguns membros da população fazerem parte da amostra, o que torna impossível a quantificação da possibilidade de erro na generalização.

A.1 Amostra Aleatória Simples

É aleatória porque a escolha é meramente por meio dum acaso; É simples porque todos elementos componentes da população tem igual probabilidade de serem escolhidos para amostra.

Para este tipo de amostragem tem sido vulgar enumerar todos elementos da população e colocá-los numa urna, o que geralmente acontece em sorteios de bilhetes nas lojas comerciais para um determinado concurso. Também é usual para casos em que após atribuição de números aos elementos da população e sem os colocar numa urna se proceda ao sorteio extra para mais tarde procurar-se pelo vencedor (casos de lotaria). Ainda se pode considerar outro caso em que se possa recorrer a tabela de números aleatórios.

Passos a considerar para seleccionar uma amostra aleatória:

a) Definir o número de elementos que devem fazer parte da amostrra (tamanho da amostra n);

b) Numerar sucessivamente os elementos da população de 1 a N;

c) Escolher os n elementos da amostra usando um procedimento aleatório, como tabela de números aleatórios ou outros. Deve-se garantir que o número seleccionado da população para amostra não seja re-seleccionado (garantia de não repetição na escolha);

d) Estabelecer a correspondência entre o número aleatório com a numeração dos elementos da população. Para casos em que haja coincidência (repetição do número aleatório) esse elemento da população é considerado seleccionado para fazer parte da amostra, não podendo ser repetido (escolha sem reposição), mas se a coincidência for movida pela repetição do elemento da população e não do número aleatório, então será considerada mera coincidência sendo que o elemento deverá ser seleccionado. Este procedimento deve ser repetido até atingir n elementos requeridos para a amostra.

Nem sempre é fácil executar esse tipo de amostragem, em virtude de ser uma tarefa penosa, "na maior parte dos casos, porque a população é infinita", tornando difícil enumerar todos elementos da população. Por outro lado, esta amostragem pode fazer com que elementos da amostra estejam dispersos geograficamente (se os elementos da população estiverem em regiões geográficas dispersas).

A precisão dos estimadores segundo um esquema aleatório simples pode ser avaliada a partir da sua variabilidade. Tomando como exemplo o caso da média amostral como estimador da média da população, vem que: $V(\overline{X}) = \frac{\sigma^2}{n}$, ignorando o factor de correcção de populações finitas[8]

A.2 Amostragem Sistemática

Se tomarmos em linha de conta que o tamanho da população é N e o tamanho da amostra é n, onde a razão k é $k = \frac{N}{n}$. De entre as primeiras k unidades selecciona-se uma (por exemplo t, onde $1 \le t \le k$) e, obtermos a lista de elementos a seleccionar da população nas posições t, $t+k$, $t+2k$, ..., $t+(n-1)k$, então estaremos na presença duma amostragem aleatória sistemática.

Passos para recolha duma amostragem sistemática:

a) Determinar k. $k = \frac{N}{n}$. Quando se trata de medidas expressas duma forma discreta é aconselhável que k seja inteiro, o que quer dizer que é susceptível de arredondamento;

b) Escolher um valor da posição t dentre as primeiras k unidades;

c) Partindo de t, seleccionar respectivamente os elementos das posições $t+k$, $t+2k$, ..., $t+(n-1)k$

Repare-se que a maior garantia da aleatoriedade é feita por escolha do primeiro elemento. Os elementos seguintes são obtidos e dependentes do primeiro.

A sua variância é dada por $V(X) = \frac{\sigma^2}{n}\left[1+\rho(n-1)\right]$. Esta amostragem pode ser mais eficiente que a aleatória simples, dependendo do valor de ρ a partir da seguinte equação

$$\frac{V(\overline{X}_s)}{V(\overline{X})} = \frac{\frac{\sigma^2}{n}\left[1+\rho(n-1)\right]}{\frac{\sigma^2}{n}} = 1+\rho(n-1)$$, veja[9]

[8] Amostragem como Factor decisivo na qualidade- Paula Vicente, Elizabeth Reis e Fátima Ferrão, 2ª edição, Edições Sílabo, pp 53

[9] Amostragem como Factor decisivo na qualidade- Paula Vicente, Elizabeth Reis e Fátima Ferrão, 2ª edição, Edições Sílabo, pp 56

Docente: Rodrigues Zicai Fazenda z_fazenda@yahoo.com.br ou rzfazenda@ustm.ac.mz 84-4934100

Reiterando: As amostras sistemáticas podem ser aleatórias ou não, dependendo da forma como o primeiro elemento foi escolhido. Se o primeiro elemento localizado dentre as primeiras t unidades for escolhido usando aleatoriedade, estaremos sob presença duma amostragem aleatória sistemática, caso contrário, diremos somente que estamos sob presença duma amostragem sistemática.

A precisão de uma amostragem sistemática é, geralmente, superior à de uma amostragem aleatória simples da mesma dimensão. Mais exactamente:

- Se a ordem das unidades no ficheiro que serviu de base de sondagem pode ser considerada como aleatória, os dois tipos de sondagem serão equivalentes;

- Se a ordem dos indivíduos tiverem uma determinada característica, isto é, se os indivíduos a quem foram atribuídos números semelhantes tiverem características semelhantes, a precisão obtida através da amostra sistemática será maior do que aquela que obteríamos com uma amostra aleatória simples.

Observação: Uma amostragem sistemática é uma amostragem que deve ser usada no campo de trabalho, e é melhor que a aleatória para uma população que se encontre geograficamente dispersa.

A.3 Amostra Estratificada

As duas amostragem anteriores, tomam a população como um todo. Esta amostragem é diferente e que é mais usada para casos em que haja certeza de que a população é heterogénea em relação à(s) característica(s) a estudar. Neste caso, deve proceder-se a uma prévia decomposição da população em estratos homogéneos, isto é, a população deve ser subdividida em estratos. Os elementos que compõem o estrato deverão ser homogéneos. A obtenção dos estratos pode basear-se num único critério (por exemplo a raça, o que permitirá estabelecer quatro estratos: branca, negra, amarela e caneca) ou na combinação de dois ou mais critérios (por exemplo a raça e o estatuto social, obtendo-se deste modo pelos menos oito estratos). A amostra total será constituída pelos conjunto de subamostras referentes a cada um dos estratos e obtidas através de um método aplicável no da amostra simples. Reduz-se assim, a margem de erro e os custos da operação. Além destas vantagens, pode-se salientar a possibilidade de realização de análises mais profundas de cada estrato separadamente e permite-nos ainda uma melhor estimativa de certas grandezas.

Importa observar que, na hipótese dos estratos terem uma dimensão suficiente, as dimensões da amostra em cada estrato podem ser idênticas. Contudo, torna-se preferível fazer a dimensão da cada amostra do grau de homogeneidade do estrato. Quanto menor for, mais ampla deverá ser a amostra para que possa ser compensado o fenómeno da dispersão. Podemos dividir a amostra estratificada em representativa ou proporcional, quando as taxas de amostragem são iguais em todos os estratos, e estratificada óptima (no sentido de Neyman), quando as dimensões dos estratos forem escolhidos de modo a minimizar a variância da média.

Há sempre vantagem em estratificar. No caso de não se conhecer em cada estrato o desvio padrão da variável utilizada como critério de estratificação, não se pode calcular a repartição óptima da amostra. No entanto uma estratificação com taxa de amostragem

uniforme (amostra estratificada proporcional) é preferível a ausência de estratificação. A eficácia da estratificação depende da homogeneidade dos estratos. Os estratos devem ser o mais homogéneos possível e heterogéneos entre si.

A seguir os passos para esta técnica de amostragem:

1. Definir o estrato- Pode ser a partir de estudo anteriores cujos resultados são conhecidos ou foram divulgados, trabalhos pilotos, conhecimento prévio da situação, etc. Em geral usam-se variáveis geográficas, demográficas, económicas ou outras que facilitem a estratificação. São exemplos de localidades, bairros, quarteirões, idade, sexo, nível salarial, etc. A ideia fundamental é que os elementos constituintes do estrato sejam mais homogéneos. Quanto mais estratos existirem, facilita a análise e aumenta a variabilidade;

2. Organizar as bases de amostragem- Cada estrato é tratado como uma população independente no estudo, levando com que se defina diversas formas para analisar diferentes estratos;

3. Seleccionar os elementos de cada estrato usando amostragem aleatória simples ou aleatória sistemática. Deve-se garantir que $\frac{n_1}{N_1} = \frac{n_2}{N_2} = ... = \frac{n_i}{N_i} = \frac{n}{N}$, isso é possível se o tamanho de cada estrato for dado por $n_i = N_i \frac{n}{N}$.

A.4 Amostragem por Conglomerado

As amostragens aleatória, sistemática e estratificada, requerem uma identificação individual dos elementos da população. Geralmente é difícil ter a listagem de todos elementos da população para se realizar a escolha dos constituintes da amostra, mas que sempre é mais fácil encontrar conglomerados. Neste caso a única exigência é que se disponha de uma lista completa dos grupos da população, mais conhecido por Unidades Primárias.

De realçar que os conglomerados devem ser grupos mutuamente exclusivos de modo a garantir que cada elemento seleccionado não faça parte da intersecção entre conglomerados.

Vejamos alguns exemplos deste tipo de amostragem:

Tabela 5.1 – Escolha de unidades elementares em conglomerados

Unidade Primária (conglomerado)	Unidade Elementar	Exemplo
Mercados Informais	Vendedor Informal	Conhecer a opinião dos vendedores dum mercado de Maputo acerca das taxas de lixo
Hospitais	Doentes	Estimar o tempo médio de espera para atendimento matinal na Enfermaria
Banco	Trabalhador	Conhecer a opinião do trabalhador dum Banco qualquer a respeito da nova taxa de juros para consumo dirigido

Passos para recolha de uma amostragem por conglomerados:

a) Especificar os conglomerados tomando-os como unidades primárias. Os elementos dos conglomerados estão geralmente próximos, na maior parte dos casos eles devem apresentar características comuns. Há dois cenários, ou define conglomerados grandes com maior variabilidade ou conglomerados pequenos em casos de garantia de não existência de variabilidade. A ideia fundamental é que a variabilidade dentro do conglomerado deve ser próxima à da população de modo que se garanta respostas do inquérito mais fiáveis e com diversidade de opiniões. Os conglomerados com maior homogeneidade geram muita informação redundante o que pode causar inconsistência na análise que se pretenda realizar;

b) Seleccionar aleatoriamente uma amostra de unidades primárias e incluir na amostra todos os elementos resultantes da totalidade dos conglomerados seleccionados.

Esta amostragem diminui custos, com a desvantagem de possuir um desvio padrão maior entre os conglomerados, causado pela maior homogeneidade dos elementos dentro do conglomerado.

A.5 Amostragem em duas ou mais etapas

Quando a população a estudar é muito vasta e dispersa (por exemplo, a população dos funcionários do Ministério da Educação), é muito difícil construir uma base de sondagem (lista exaustiva e não repetitiva de todas as unidades estatísticas que compõem a população). Tal operação seria muito demorada e provavelmente o seu custo proibitivo. Daí que se recorra à amostragem em duas (ou mais) etapas.

Suponhamos que se pretende fazer um estudo sobre o aproveitamento escolar dos alunos do EP1. Como proceder para obter uma amostra usando este método?
Em primeiro lugar teríamos que dividir a população num certo número de unidades primárias (conglomerados), de modo que cada unidade estatística seja afecta sem ambiguidade a uma unidade primária bem determinada;

De seguida a selecção ao acaso é feita em duas etapas: na primeira etapa, selecciona-se ao acaso uma amostra de unidades primárias (conglomerados), na segunda etapa, em cada unidade primária (conglomerados) seleccionada(o), tira-se ao acaso uma amostra de elementos ou unidades secundárias.

Tem a vantagem de precisar apenas a lista das unidades primárias seleccionadas dispensando a lista de todos os indivíduos do universo. Permite-nos reduzir as despesas de deslocação, pois observa-se uma menor dispersão geográfica das unidades estatísticas. Também o custo é sempre menor do que aquele em que a amostra é seleccionada numa só etapa.

No entanto a precisão das estimativas será sempre menor numa amostra seleccionada em duas etapas. Isto resulta do facto da amostra ser menos dispersa geograficamente e da diferença entre unidade secundárias de uma unidade primária ser menor que em unidades secundárias pertencentes a unidades primárias diferentes. Podemos aumentar a precisão das estimativas aumentando a dimensão da amostra, sem que se verifique um grande acréscimo no custo do inquérito.

Exemplo:

Tabela 5.2 – Escolha de unidades secundárias e terciárias

Unidade amostral primária	Unidade amostral secundária	Unidade amostral terciária
Cidade	Bairro	Quarteirão
Universidade	Faculdade	Departamento
Alunos da Escola	Alunos duma turma	Um grupo da turma

A variância na unidade é dada por $V(\overline{X}) = \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n} + \frac{\sigma_3^2}{p}$, em que σ_1^2, σ_2^2 e σ_3^2 são as variâncias em cada unidade (primária, secundária e terciária) e m, n e p os respectivos tamanhos.

A.6 Amostragem Multi – Fases

Não entenda a amostragem multi etapa como multi fase, em virtude delas serem completamente diferentes. A amostragem multi fases considera que em cada fase de amostragem está sempre em causa o mesmo tipo de unidade amostral.

São os seguintes os passos desta amostragem:

a) Listar os elementos da população e seleccionar uma amostra aleatória. Esta amostra servirá de população para a fase seguinte;

b) Seleccionar uma segunda amostra cujos elementos serão inquiridos com maior profundidade. Deve-se questionar se na primeira fase não se teria recolhido toda informação suficiente para dispensar a fase seguinte. Caso a resposta seja negativa repetimos este passo sucintamente até que obtenhamos uma amostra de uma fase mais avançada que possamos questionar melhor.

B. Amostragens não aleatórias

B.1 Amostragem Intencional

Nesta amostragem os constituintes da amostra são seleccionados deliberadamente pelo investigador, geralmente porque este considera serem os mais fundamentais para facilitar o seu estudo ou para emitir algum parecer. Não se baseia na imposição de que deva ser exactamente selecção de elementos. Pode também ser um exemplo da imaginação do investigador.

Este é o caso que é mais usado pelas empresas de sondagem de opiniões pré-eleições quando acham ser necessário fazer um marketing político usando dados concretos e não rejeitáveis. Entrevistam na zona popular do candidato, filmam e publicam os resultados. A ideia é que eles tenham um retorno de informação pré-programada pelo candidato e não a realidade social.

Vejamos alguns casos da sua aplicabilidade:

Obtenção de amostra de dimensão reduzida: é aconselhável usá-la neste tipo de caso do que obter uma amostra aleatória porque os custos são menores para este tipo de amostragem do que na aleatória;

Casos em que não se consiga uma amostra aleatória: casos de vendedores ambulantes da cidade de Maputo. É difícil obter uma amostra aleatória destes em virtude de nos apercebermos que na maior parte dos casos exercem essa actividade duma forma ilícita;

Conseguir deliberadamente uma amostra viciada: é o caso de um fabricante que quer mostrar um impacto positivo de um produto que acaba de alinhar na nova produção. Se ninguém das autoridades faz um control efectivo, a preferência do industrial é viciar uma amostra e publicitar os resultados.

B.2 Amostragem Snowball

Consiste em localizar indivíduos com características desejadas ou próximas das requeridas no estudo que se pretende fazer. A amostra vai aumentando na medida em que um elemento da amostra da característica desejada encontre mais um com as mesmas características. Geralmente é usado este tipo de amostragem quando se pretende estudar uma característica específica, com a certeza de difícil localização dos possíveis constituintes da amostra.

A maior desvantagem reside no facto de que os amigos podem se indicar, resultando numa amostra em que todos elementos recolhidos tenham um pensamento ou respostas similares.

Exemplo: Esta amostragem é mais usada pela polícia quando pretende localizar mais companheiros de uma quadrilha de assaltantes depois de capturar o primeiro
.

B.3 Amostragem por Conveniência

É uma amostragem de uma pura coincidência, porque os elementos que possam ser constituintes da amostra se localizam na zona onde o inquérito está a decorrer passando a fazer parte dela por conveniência.

Esta é uma das amostragens que resulta num grande enviesamento (viciação). É muito importante ser usada quando pretendemos captar ideias gerais ou identificar aspectos críticos. Neste caso deve-se ter o cuidado de não se assumir qualquer tipo de objectividade científica.

Usualmente é denominada de amostragem de pré teste do questionário.

B.4 Amostragem usando Método das Escolhas Relacionadas ou "das Quotas"

Este método visa constituir um modelo reduzido da população a estudar. O que importa é realizar uma estrutura idêntica à do conjunto afim, a partir de um pequeno número de critérios considerados essenciais na definição da população em causa. Consiste então, em obter uma representatividade suficiente tentando reproduzir, na amostra, as distribuições de certas variáveis, tal como existem na população a estudar. Por exemplo se esta comporta tantos homens como mulheres proceder-se-á de forma a que o mesmo aconteça na amostra (50% de homens e 50% de mulheres).

A escolha dos indivíduos é deixada ao critério dos inquiridores, dentro dos limites impostos pelas quotas. Existe, a este nível, um risco de ver a amostra enviesada, pois o inquiridor tenderá a introduzir outros factores além do acaso na escolha dos indivíduos., interrogando as pessoas que conhece, que estão mais próximas, que são mais fáceis de contactar, mais disponíveis,...

Para assegurar uma melhor representatividade, devemos associar este método a um método aleatório, como por exemplo a sondagem aureolar, que consiste em tirar à sorte certas zonas para onde enviar os inquiridores.

Este método tem a vantagem de poder ser aplicado a qualquer população que se pretenda estudar ou a amostra em estudo seja apenas representativa, estratificada ou concebida segundo um plano experimental mais ou menos complexo. É um método que dá resultados muito satisfatórios e é utilizado com maior frequência.

Convém referir que, enquanto nas amostras por quotas, a escolha dos indivíduos a serem interrogados, fica ao arbítrio do inquiridor, nas amostras aleatórias estratificadas, isto é, aquelas que partem da divisão do universo em estratos, a escolha dos questionados é feita ao acaso por métodos probabilísticos, como a própria designação da amostra o indica.

Poderíamos aplicar este método no nosso exemplo sobre a opinião dos alunos em relação à biblioteca, reduzindo consideravelmente a amostra, se considerássemos somente os alunos que frequentam com mais regularidade a biblioteca.

B.5 Amostragem pelo Método Aureolar

É um processo de sondagem que permite evitar os inconvenientes de uma base de sondagem incompleta ou caduca. Consiste numa determinação de subconjuntos de população por uma área geográfica. Define-se num mapa um certo número de áreas geográficas, que podem ser constituídas por Municípios, Bairros, Quarteirões, grupos de casas. Procede-se em seguida a uma tiragem à sorte e exploram-se por fim sistematicamente as unidades de sondagem assim apuradas.

B.6 Amostragem por Cachos

Inscreve-se na mesma concepção geral. A tiragem à sorte do tipo de amostragem probabilística é estendida a diferentes tipos de conjunto de indivíduos (cachos), tais como os

alunos de uma determinada escola, os professores de uma escola, os professores de um determinado grupo de disciplina, os funcionários auxiliares de uma escola, uma família, etc.

Os membros de um mesmo cacho são, muitas vezes, mais parecidos entre eles do que, em comparação com o resto da população. Se por exemplo, estudarmos opiniões, há muitas hipóteses de que as dos diferentes membros do cacho sejam relativamente semelhantes. No plano dos comportamentos, poderemos igualmente encontrar semelhanças entre as actividades de lazer ou de viagem realizadas pelos membros de uma família. Os professores do mesmo departamento numa escola, conhecem-se, trabalham nas mesmas condições, são confrontados com problemas comuns e isso exprime-se, evidentemente, nas suas respostas.

O método de Kish é um processo aleatório de designação da pessoa a inquirir que evita enviesamentos na sua escolha aplicado ao cacho familiar. Poderemos, contudo, "inventar" métodos que evitem enviesamentos, como por exemplo, interrogar aquele que fez anos há menos tempo num determinado cacho.

B.7 Amostragem no Local

Quando nos interessamos por uma população restrita, para a qual não existe uma base de sondagem específica, como por exemplo nos casos de uma determinada categoria profissional, podemos construir uma amostra do conjunto da população, por sorteio ou por quota, e conservar apenas aqueles que pertencem à categoria visada. Podemos ainda apoiarmo-nos no facto de que certas pessoas se encontram em lugares particulares: os funcionários auxiliares de uma escola, nos corredores dessa escola ou na sua sala específica, os professores nas salas de aulas ou na sala dos professores, etc. Quando uma amostra de uma sub-população é por si só suficiente, não sendo necessário um grupo de comparação, é possível construir uma amostra correcta indo a determinados lugares e procedendo, no local, a um sorteio entre as pessoas presentes.

É necessário juntar a esta amostragem geral espacial, acrescentar uma amostragem temporal. Com efeito, os professores de uma escola não se encontram todos na sala dos professores num determinado dia. A importância destas precauções dependerá da natureza do problema estudado. A amostragem espacial e temporal permite eliminar um certo número de enviesamentos, mas não assegura necessariamente uma amostragem representativa.

B.8 Amostragem Random Route

Passos a considerar para obter uma Amostra Random Route

a) Selecção de um ponto de partida através de uma listagem, mapa ou registo de endereço ou ponto de referência da zona onde irá decorrer o estudo;

b) Definição de regras de orientação para o entrevistador. Assumimos que o inquiridor tem uma circunscrição ou um itinerário aleatório na escolha de unidades a inquirir. Imaginemos que queira entrevistar residentes de um certo bairro. Seria mais fácil perguntar onde fica a igreja X. Daí pode seguir a rua em frente, virando á esquerda ou direita. Se a data do seu nascimento for 24, a soma de 2 e 4 dá 6; pode-se

procurar entrevistar naquela rua todas casas que tenham a soma dos algarismos de seus números de casa 6, casos de 6, 15, 51, 24, 42, 33. dentro de cada casa interessa em saber que serão os indivíduos a entrevistar, processo que pode ser fácil usando a tabela de números aleatórios ou usando amostragem sistemática

1.8 Questionário

Quando pretendemos recolher uma informação com base em sondagens, nalgumas vezes recorre-se a um questionário.

Pressupostos para elaboração dum questionário
Verificar:

a) Se para o estudo que se pretende realizar já existe algum estudo preliminar (piloto);

b) Se para o estudo que se pretende realizar existe algum estudo anterior;

c) Se temos conhecimento íntegro sobre a população que pretendemos inquirir;

d) Se pretendemos testar o actual questionário (que se asume comopreliminar), para elaborar um questionário definitivo.

Num questionário deve-se tomar em consideração diversos **tipos de perguntas**:

Abertas: Em que o inquirido pode tecer comentários à volta da questão que é colocada
Exemplo: Porquê se considera a sua comunidade muito forte?

Fechadas: Em que as respostas são táxicas.
Exemplo: Acha que o candidato X ganha eleições?
Aí a resposta é táxica, pois terá como opções: Sim, Não, Não Sabe, Não Responde e Talvez.

Mistas: Em que se faz a mistura de aberta e fechada
Exemplo: Acha que o candidato X ganha? Porquê?
O inquirido dará resposta como no caso b) e fará comentários do porquê com em caso a).

Observação: Um questionário é usado em estrevistas semi estruturadas, em estudos de casos, em sondagens, nos censos, em pesquisas e muito mais. Ao tocar no tema questionário, o autor pretende convidar o estimado leitor a ler mais sobre o assunto, porque será deveras importante no futuro, principalmente quando o indivíduo atinge um nível académico em que queira dedicar-se à investigação.

Nota importante para tamanho duma amostra

Há 3 factores que determinam o tamanho duma amostra, nenhum dos quais tendo uma relação directa com o tamanho da população. Eles são:

1- O grau de confiança adoptado-com o aumento do grau de confiança aumenta também o valor de Z, t ou χ^2, o que faz com que o tamanho da amostra também cresça.

2- O máximo erro permissível- quanto mais aumenta o erro padrão de estimativa, diminui o tamanho da amostra.

3- A variabilidade da população- com o aumento da variabilidade, isto é, aumento do desvio ou da variância, também aumenta o tamanho da amostra.

2. INTERVALOS DE CONFIANÇA

Observação: Vimos modelos que procuram medir a variabilidade de fenómenos casuais de acordo com suas ocorrências (probabilidades). Nunca chegamos a ter a certeza sobre os parâmetros que especificamos. O propósito do pesquisador seria descobrir os parâmetros da distribuição. Raramente é possível obter a distribuição exacta de alguma variável, porque é despendioso, demorado ou porque é um processo destrutivo. Daí leva-nos a recolher amostras e inferir sobre os parâmetros populacionais. Se tivéssemos informação completa sobre a função de probabilidade (caso discreto) ou afunção densidade de probabilidade (caso contínuo) da variável em questão, não haveria necessidade de seleccionar uma amostra. Teríamos toda a informação pela distribuição de probabilidade. Isso acontece porque:
Não temos informação em relação à variável ou
Desconhecemos a curva (gráfico)

2.1 Estimadores

Chama-se estatística a uma variável aleatória que seja apenas função de uma amostra aleatória, que não contenha parâmetros desconhecidos. Em geral os parâmetros de uma população designam-se por letras gregas μ, σ, etc. As respectivas estatísticas são calculadas a partir das amostras. Alguns exemplos apresentam-se a seguir:

Tabela 5.3 – Estatísticas e Parâmetros

Estatísticas ou Estimadores	Parâmetros ou Estimativas
$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	$\mu = E(x)$
$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$	$\sigma^2 = E\left[(X-\mu)^2\right]$
$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$	$\sigma = \sqrt{\sigma^2} = \sqrt{E\left[(X-\mu)^2\right]}$

Definições

1- **Estimador**- É qualquer estatística usada para estimar o valor de um parâmetro.

2- **Um estimador diz-se convergente** ou consistente- se o limite da esperança do estimador é igual ao parâmetro e limite da variância é também igual ao parâmetro.

3- É denominado de **desvio do estimador** a diferença entre a esperança do estimador e o parâmetro.

4- Um **estimador, diz-se centrado** ou não enviesado ou não viciado quando o seu desvio é nulo, isto é, se a esperança do estimador é igual à do parâmetro.

5- **Estimativa de um parâmetro** de uma população- é qualquer valor específico de uma estatística desse parâmetro

6- **Estimação**- É todo o processo que se baseia em utilizar um estimador para produzir uma estimativa de um parâmetro.

Para encontrar as estimativas dum parâmetro podemos usar duas alternativas. Estimativa por ponto ou estimativa por intervalo.

Estimativa por Ponto

Estimativa por ponto é valor único obtido no cálculo da estatística ou parâmetro (para o caso de proporções) que é usado para estimar um parâmetro populacional.

Exemplos de estimativas por ponto são a média amostral, o desvio padrão amostral, a variância amostral, a proporção populacional, etc.

Exemplo: Uma turma de Engª qumíca é composta por 20 estudantes. Qual é a média de notas de 5 estudantes que tiveram 15, 19, 8, 12, 13, respectivamente? A média dos 5 estudantes será $\overline{X} = \frac{\sum X_i}{n} = \frac{15+19+8+12+13}{5} = 13{,}4$. Assim a estimativa de ponto para a média dos 5 estudantes é 13,4.

2.2 Estimativa Por Intervalo

O segundo tipo de estimação sobre o qual nos vamos debruçar mais, denomina-se estimação por intervalo. Ela estabelece um intervalo de valor e dentro do qual um parâmetro populacional provavelmente caia a um determinado nível de confiança. O intervalo de confiança, é o intervalo dentro do qual um parâmetro populacional é esperado ocorrer. Os intervalos de confiança que são em geral usados são os de 95 %, 98% e 99 %. Um intervalo de confiança de 98 % significa que cerca de 98 % dos intervalos construídos similarmente conterão o parâmetro que está sendo estimado. Se tomarmos 95%, poder-se-á dizer que 95 % das médias amostrais para um tamanho de amostra especificado cairão a uma distância máxima de 1,96 desvios padrões da média populacional.

I- Intervalo de confiança para a média

Para casos referentes a média é necessário destacar três casos:
Intervalo de confiança para a média, se o desvio é conhecido, o que é tratado pela distribuição normal;
Intervalo de confiança para a média, se o desvio não é conhecido e a amostra é grande, o que é tratado pela distribuição normal com desvio amostral no lugar do desvio populacional e
Intervalo de confiança para a média, se o desvio não é conhecido e tamanho da amostra pequeno, o que é tratado pela distribuição t-student com $n-1$ graus de liberdade

Erro padrão da média amostral

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$ usado para situações em que a população é finita, isto é $\frac{n}{N} > 0{,}05$. Para os casos em que a população é infinita, isto é, $\frac{n}{N} \leq 0{,}05$ aí usa-se $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

a) Intervalo de confiança para a média com desvio conhecido e $\forall n$.
Estamos sob presença de uma distribuição normal

Determinemos a respectiva probabilidade que é dada por $P(\bar{x} - z\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z\frac{\sigma}{\sqrt{n}}) = 1-\alpha$ e que o respectivo intervalo de confiança será $\bar{x} - z\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z\frac{\sigma}{\sqrt{n}}$. O valor de Z que é tabelado, deverá ser consultado para $Z_{1-\frac{\alpha}{2}}$. $1-\alpha$ é o nível de confiança. O respectivo erro padrão de estimativa será $\varepsilon = z\frac{\sigma}{\sqrt{n}}$, a amplitude $A = 2z\frac{\sigma}{\sqrt{n}}$ e o tamanho da amostra $n = \left(\frac{z\sigma}{\bar{X} - \mu}\right)^2$

Exemplo : Quando certos dados foram submetidos a análise por uma equipa que se dedica a revisão curricular de certa faculdade descobriu-se que todos tinham o mesmo erro de estimativa (no valor de 3,52) numa amplitude de 6. Qual deveria ter sido o tamanho da amostra, sabendo que tomaram $\sigma = 6$ a um nível de significância de 95%.
Solução:

$z\frac{\sigma}{\sqrt{n}} = 3{,}52 \quad A = 6 \quad \sigma = 6 \;\; 1-\alpha = 0{,}95 \qquad Z_{0{,}975} = 1{,}96$

$A = 2Z\frac{\sigma}{\sqrt{n}} \Leftrightarrow 6 = 2\times 1{,}96 \times \frac{6}{\sqrt{n}} \Leftrightarrow \left(\sqrt{n}\right)^2 = (3{,}92)^2 \Leftrightarrow n \approx 15$ ou

$A = 2Z\frac{\sigma}{\sqrt{n}} \Leftrightarrow \sqrt{n} = \frac{2Z\sigma}{A} \Leftrightarrow \left(\sqrt{n}\right)^2 = \left(\frac{2Z\sigma}{A}\right)^2 \Leftrightarrow \left(\sqrt{n}\right)^2 = (3{,}92)^2 \Leftrightarrow n \approx 15$

b) Intervalo de confiança para a média com desvio desconhecido e amostra grande $n > 30$ (Distribuição Normal)

Determinemos a respectiva probabilidade que é dada por $P(\bar{x} - z\frac{s}{\sqrt{n}} < \mu < \bar{x} + z\frac{s}{\sqrt{n}}) = 1-\alpha$ e que o respectivo intervalo de confiança será $\bar{x} - z\frac{s}{\sqrt{n}} < \mu < \bar{x} + z\frac{s}{\sqrt{n}}$. O valor de Z que é tabelado, deverá ser consultado para $Z_{1-\frac{\alpha}{2}}$. $1-\alpha$ é o nível de confiança. O respectivo erro padrão de estimativa será $\varepsilon = z\frac{s}{\sqrt{n}}$, a amplitude $A = 2z\frac{s}{\sqrt{n}}$ e o tamanho da amostra $n = \left(\frac{zs}{\bar{X} - \mu}\right)^2$. Neste caso deve obter-se o s a partir de cálculo a ser feito com base nos dados recolhidos e que compõem a amostra.

Exemplo

Uma universidade quer estimar o número médio de horas trabalhadas por semana por seus estudantes. Uma amostra de 49 estudantes mostrou uma média de 24 horas com um desvio padrão de 4 horas. A estimativa por ponto do número médio de horas trabalhadas por semana é 24 horas (média amostral). Qual é o intervalo de confiança de 95 % para o número médio de horas trabalhadas por semana ?

Resposta: Usando a fórmula anterior ($\bar{X} \pm 1{,}96\frac{s}{\sqrt{n}}$) temos $24 \pm 1{,}96\frac{4}{\sqrt{49}}$ ou 22,88 a 25,12. O limite de confiança inferior é 22,88. O limite superior de confiança é 25,12. O grau de confiança (nível de confiança) utilizado é 0,95.

Interpretando os resultados

Se nós tivéssemos tempo para seleccionar aleatoriamente 100 amostras de tamanho 49 da população de alunos do campus universitário e calcular as médias amostrais, os intervalos de confiança para cada uma destas 100 amostras, a média populacional (parâmetro) do número de horas trabalhadas estariam contidos em cerca de 95 dos 100 intervalos de confiança. Cerca de 5 dos 100 intervalos de confiança não conteriam a média populacional.

c) Intervalo de confiança para a média com desvio desconhecido e amostra pequena $n \leq 30$ (Distribuição t-student)

Determinemos a respectiva probabilidade que é dada por $P(\bar{x} - t\frac{s}{\sqrt{n-1}} < \mu < \bar{x} + t\frac{s}{\sqrt{n-1}}) = 1-\alpha$ e que o respectivo intervalo de confiança será $\bar{x} - t\frac{s}{\sqrt{n-1}} < \mu < \bar{x} + t\frac{s}{\sqrt{n-1}}$. O valor de t que é tabelado, deverá ser consultado para $t_{(1-\frac{\alpha}{2};n-1)}$. $1-\alpha$ é o nível de confiança e n-1 graus de liberdade. O respectivo erro padrão de estimativa será $\varepsilon = t\frac{s}{\sqrt{n-1}}$, a amplitude $A = 2t\frac{s}{\sqrt{n-1}}$ e o tamanho da amostra $n-1 = \left(\frac{ts}{\bar{X} - \mu}\right)^2$. Neste caso deve obter-se o s a partir de cálculo a ser feito com base nos dados recolhidos e que compõem a amostra.

Exemplo: O tempo que uma máquina leva a executar determinada operação numa peça está sujeito a variações. Para verificar se as condições de funcionamento da máquina estão dentro das normas, registou-se 12 vezes o referido tempo. Os resultados (em segundos) foram os seguintes: 29 33 36 35 36 40 32 37 31 35 30 36. Construa um intervalo de confiança a 95% para o tempo médio de execução da tarefa pela máquina em análise, sabendo que esta segue uma distribuição aproximadamente normal.

Resolução: Podemos definir a nossa variável X como o "tempo, em segundos, que uma máquina leva a executar uma tarefa". Sabemos que X segue uma distribuição normal. Como desconhecemos os parâmetros da distribuição e n é pequeno, vamos usar distribuição t-student. $\overline{X} = \frac{1}{12}\sum x_i = 34{,}17$ $S^2 = \frac{1}{11}\sum (x_i - \overline{X})^2 = 10{,}08 \Rightarrow S = 3{,}18$ $1 - \frac{\alpha}{2} = 0{,}975$ $t_{0,975}(11) = 2{,}201$. Repare que é distribuição t-student com n-1=12-1=11 graus de liberdade. Consultamos na tabela o em anexo a linha 11, coluna $t_{0,975}$ e encontramos o valor de 2,201. Substituíndo na fórmula teremos a probabilidade $P(\overline{x} - 2{,}201\frac{s}{\sqrt{12}} < \mu < \overline{x} + 2{,}201\frac{s}{\sqrt{12}}) = 0{,}95$ e o intervalo será $]32{,}15;36{,}19[$.

II Intervalo de confiança para uma proporção populacional

$$P(\overline{p} - \mathrm{Z}\sqrt{\frac{p(1-p)}{n}} < p < \overline{p} + \mathrm{Z}\sqrt{\frac{p(1-p)}{n}}) = 1-\alpha$$

Um intervalo de confiança para uma proporção populacional é dado por:
$\overline{p} \pm \mathrm{Z}\sigma_{\overline{p}}$ onde: $\overline{p}$ é a proporção amostral $\sigma_{\overline{p}}$ é o erro padrão da proporção amostral e é dado por: $\sigma_{\overline{p}} = \sqrt{\frac{p(1-p)}{n}}$. O respectivo intervalo de confiança é dado por

$\overline{p} - \mathrm{Z}\sqrt{\frac{p(1-p)}{n}} < p < \overline{p} + \mathrm{Z}\sqrt{\frac{p(1-p)}{n}}$ onde: $\overline{p}$ é a proporção amostral Z é o valor da variável normal padrão para o grau de confiança $1-\alpha$. n é o tamanho da amostra. O valor de Z que é tabelado, deverá ser consultado para $Z_{1-\frac{\alpha}{2}}$. O respectivo erro padrão de estimativa será $\varepsilon = \mathrm{Z}\sqrt{\frac{p(1-p)}{n}}$, a amplitude $A = 2\mathrm{Z}\sqrt{\frac{p(1-p)}{n}}$ e o tamanho da amostra $n = \left(\frac{Zp(1-p)}{\overline{p} - p}\right)^2$.
Neste caso deve obter-se o s a partir de cálculo a ser feito com base nos dados recolhidos e que compõem a amostra.

Factor de Correcção de População Finita

Uma População denomina-se finita quando $n/N > 0{,}05$ (ou seja, quando a fracção amostral é maior do que 5 %).

Erro padrão da proporção amostral $\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}$ usado para situações em que a população é finita, isto é $\frac{n}{N} > 0{,}05$. Para os casos em que a população é infinita, isto é, $\frac{n}{N} \leq 0{,}05$ aí usa-se $\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$.

Exemplo

Uma amostra aleatória de 100 eleitores do Município de Maputo dá 55% como favoráveis a um certo candidato. Determine os limites de confiança para a proporção global de eleitores favoráveis ao candidato assumindo 95% de confiança.

Resolução:

$\bar{p} = 0{,}55 \Rightarrow q = 1 - p = 1 - 0{,}55 = 0{,}45$

$1-\alpha = 0{,}95 \Rightarrow \alpha = 0{,}05 \Rightarrow \frac{\alpha}{2} = 0{,}025 \Rightarrow 1-\frac{\alpha}{2} = 0{,}975 \qquad Z_{1-\frac{\alpha}{2}} = Z_{0{,}975} = 1{,}96$

$P(\bar{p} - Z\sqrt{\frac{p(1-p)}{n}} < p < \bar{p} + Z\sqrt{\frac{p(1-p)}{n}}) = 1-\alpha$

$\Rightarrow P(0{,}55 - 1{,}96\sqrt{\frac{0{,}25 \times 0{,}75}{100}} < p < 0{,}55 - 1{,}96\sqrt{\frac{0{,}25 \times 0{,}75}{100}}) = 0{,}95 \Rightarrow$

$0{,}55 - 1{,}96\sqrt{\frac{0{,}25 \times 0{,}75}{100}} < p < 0{,}55 - 1{,}96\sqrt{\frac{0{,}25 \times 0{,}75}{100}} \Leftrightarrow 0{,}57 < p < 0{,}53$

III Intervalo de confiança para uma variância

1- Intervalo de confiança para uma variância se a média é conhecida

Para o caso da variância, se conhecemos a média, podemos determinar a probabilidade como sendo $P(\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\chi^2_{1-\frac{\alpha}{2}}} < \sigma^2 < \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\chi^2_{\frac{\alpha}{2}}}) = 1-\alpha$ onde $1-\alpha$ é o nível de confiança. O respectivo intervalo de confiança será $\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\chi^2_{1-\frac{\alpha}{2}}} < \sigma^2 < \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\chi^2_{\frac{\alpha}{2}}}$ em que o $\chi^2_{\frac{\alpha}{2}}$ tem n-1 graus de liberdade.

Docente: Rodrigues Zicai Fazenda z_fazenda@yahoo.com.br ou rzfazenda@ustm.ac.mz 84-4934100

2- Intervalo de confiança para uma variância se a média é desconhecida

Para o caso da variância, se não conhecemos a média, podemos determinar a probabilidade a partir da seguinte expressão: $P(\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}}) = 1-\alpha$ onde $1-\alpha$ é o nível de confiança. O respectivo intervalo de confiança será $\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}}$ em que o $\chi^2_{\frac{\alpha}{2}}$ tem n-1 graus de liberdade.

E que $X^2 = (n-1)\frac{s^2}{\sigma^2} \sim \chi^2_{n-1}$, o que equivale a dizer que estamos em presença da distribuição qui-quadrático.

Exemplo:

Suponha-se em presença de uma população normal, com parâmetros desconhecidos. Com base numa amostra casual, com 16 observações, foi construído o seguinte intervalo de confiança para a média da população: $]7{,}398;12{,}602[$. Sabendo que, com a informação da amostra, se obteve $s = 4$. Com base na mesma amostra construa um intervalo de confiança a 95% para a variância da população

Resolução

$$1-\alpha = 0{,}95 \Rightarrow 1-\frac{\alpha}{2} = 0{,}975 \Rightarrow \chi^2_{(15;0{,}975)} = 27{,}5 \rightarrow \chi^2_{(15;0{,}025)} = 6{,}26$$

$$\chi^2 \in \left]\frac{15\times 16}{27{,}5}; \frac{15\times 16}{6{,}26}\right[\Leftrightarrow \chi^2 \in]8{,}7273;38{,}3387[$$

IV Intervalo de confiança para a diferença de médias

1- Intervalo de confiança para a diferença de médias se os dois desvios são conhecidos

Para o caso em que todos parâmetros são dados excepto a diferença de médias que se pretende estimar, a respectiva probabilidade é dada por:

$P(\left(\overline{X}_A - \overline{X}_B\right) - Z_{cal}\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}} < \mu_A - \mu_B < \left(\overline{X}_A - \overline{X}_B\right) + Z_{cal}\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}) = 1-\alpha$, onde $1-\alpha$ é o grau de confiança, e o respectivo intervalo de confiança é daddo por

$$\left(\overline{X}_A - \overline{X}_B\right) - Z_{cal}\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}} < \mu_A - \mu_B < \left(\overline{X}_A - \overline{X}_B\right) + Z_{cal}\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$$

Exemplo: Uma amostra de 150 lâmpadas eléctricas da marca A apresenta uma vida média de 1400 hr e um desvio padrão de 120 hr. Uma amostra de 200 lâmpadas da marca B apresenta média 1200 hr e desvio padrão de 80 hr. Determine os limites de confiança a 95 % para a diferença das vidas médias das lâmpadas das duas marcas.

$$P(\left(\overline{X}_A-\overline{X}_B\right)-Z_{cal}\sqrt{\frac{\sigma_A^2}{n_A}+\frac{\sigma_B^2}{n_B}}<\mu_A-\mu_B<\left(\overline{X}_A-\overline{X}_B\right)+Z_{cal}\sqrt{\frac{\sigma_A^2}{n_A}+\frac{\sigma_B^2}{n_B}})=1-\alpha \Leftrightarrow$$

$$P((1400-1200)-Z_{cal}\sqrt{\frac{120^2}{150}+\frac{80^2}{200}}<\mu_A-\mu_B<(1400-1200)-Z_{cal}\sqrt{\frac{120^2}{150}+\frac{80^2}{200}})=0{,}95 \Leftrightarrow$$

$$(1400-1200)-1{,}645\sqrt{\frac{120^2}{150}+\frac{80^2}{200}}<\mu_A-\mu_B<(1400-1200)-1{,}645\sqrt{\frac{120^2}{150}+\frac{80^2}{200}}$$

$$181{,}39<\mu_A-\mu_B<218{,}61$$

2- Intervalo de confiança para a diferença de médias se os dois desvios são desconhecidos e as amostras são grandes

A sua probabilidade é dada por

$$P\left(\overline{X}_1-\overline{X}_2-Z\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\langle\mu_1-\mu_2\langle\overline{X}_1-\overline{X}_2+Z\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\right)=1-\alpha\,,$$ o respectivo intervalo de confiança será: $\overline{X}_1-\overline{X}_2-Z\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\langle\mu_1-\mu_2\langle\overline{X}_1-\overline{X}_2+Z\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$, em que $1-\alpha$ é o nível de confiança, n_1 e n_2 são tamanhos da primeira e segunda amostra, respectivamente. $\overline{X}_1-\overline{X}_2$ (diferença de médias amostrais).

Exemplo: Foi realizado um estudo para determinar se um certo tratamento tinha efeito corrosivo sobre determinado metal. Uma amostra de 100 peças foi imersa num banho durante 24 horas com o tratamento, tendo sido removido uma média de 12.2 mm de metal com um desvio padrão de 1.1mm. Uma segunda amostra de 200 peças foi também imersa durante 24 horas mas sem tratamento, sendo a média do metal removido de 9.1mm, com um desvio padrão de 0.9mm. Determine um intervalo de confiança a 98% para a diferença entre as médias das populações, retirando conclusões quanto ao efeito do tratamento.

Resolução: Como n_1 e n_2 são grandes vamos utilizar $1-\alpha=0{,}98\Rightarrow\frac{\alpha}{2}=0{,}01$ e

$$Z=\frac{\left(\overline{X}_1-\overline{X}_2\right)-\left(\mu_1-\mu_2\right)}{\sqrt{\frac{S_1^2}{100}+\frac{S_2^2}{200}}}\rightarrow N(0;1)\qquad 1-\alpha=0{,}98\Rightarrow 1-\frac{\alpha}{2}=0{,}99\text{, logo}$$

$$P(Z\langle Z_{0{,}99})=0{,}99\Rightarrow Z_{0{,}99}=2{,}326\qquad P(-2{,}33\langle Z\langle 2{,}33)=0{,}98\ \Leftrightarrow$$

$$\Leftrightarrow\qquad P\left(\overline{X}_1-\overline{X}_2-2{,}326\sqrt{\frac{S_1^2}{100}+\frac{S_2^2}{200}}\langle\mu_1-\mu_2\langle\overline{X}_1-\overline{X}_2+2{,}326\sqrt{\frac{S_1^2}{100}+\frac{S_2^2}{200}}\right)=0{,}98$$

Então, para $S_1^2 = 1{,}1^2$ e $S_2^2 = 0{,}9^2$ o intervalo de confiança para $\mu_1 - \mu_2$ a 98% de confiança (ou com 2% de risco de erro) vai ser: $(12{,}2 - 9{,}1) - 2{,}326\sqrt{\frac{1{,}1^2}{100} + \frac{0{,}9^2}{200}} < \mu_1 - \mu_2 < (12{,}2 - 9{,}1) + 2{,}326\sqrt{\frac{1{,}1^2}{100} + \frac{0{,}9^2}{200}}$

$2{,}804 < \mu_1 - \mu_2 < 3{,}396$. Como $\mu_1 - \mu_2 > 0$, Concluiu-se que o tratamento tem efeito corrosivo no metal

3- Intervalo de confiança para a diferença de médias se os dois desvios são desconhecidos e tamanhos da amostra pequenos

A sua probabilidade é dada por

$$P\left(\overline{X}_1 - \overline{X}_2 - t\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}} < \mu_1 - \mu_2 < \overline{X}_1 - \overline{X}_2 + t\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\right) = 1-\alpha$$

o respectivo intervalo de confiança será:

$$\overline{X}_1 - \overline{X}_2 - t\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}} < \mu_1 - \mu_2 < \overline{X}_1 - \overline{X}_2 + t\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}},$$

em que $1-\alpha$ é o nível de confiança, n_1 e n_2 são tamanhos da primeira e segunda amostra, respectivamente. $\overline{X}_1 - \overline{X}_2$. Estamos na presença de uma distribuição t-student com $n_1 + n_2 - 2$ graus de liberdade.

Exemplo: Foi realizado um estudo para determinar se um certo tratamento tinha efeito corrosivo sobre determinado metal. Uma amostra de 16 peças foi imersa num banho durante 24 horas com o tratamento, tendo sido removido uma média de 12.2 mm de metal com um desvio padrão de 1.1mm. Uma segunda amostra de 25 peças foi também imersa durante 24 horas mas sem tratamento, sendo a média do metal removido de 9.1mm, com um desvio padrão de 0.9mm. Determine um intervalo de confiança a 98% para a diferença entre as médias das populações, retirando conclusões quanto ao efeito do tratamento.

Resolução: Como n_1 e n_2 são grandes vamos utilizar $1-\alpha = 0{,}98 \Rightarrow \frac{\alpha}{2} = 0{,}01$ e

$1-\alpha = 0{,}98 \Rightarrow 1 - \frac{\alpha}{2} = 0{,}99$, logo

$$P\left(\overline{X}_1 - \overline{X}_2 - t\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}} < \mu_1 - \mu_2 < \overline{X}_1 - \overline{X}_2 + t\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\right) = 1-\alpha$$

$$P\left(12{,}2-9{,}1-t\sqrt{(\frac{1}{16} + \frac{1}{25})\frac{(16-1)1{,}1^2 + (25-1)0{,}9^2}{16+25-2}} < \mu_1 - \mu_2 < 12{,}2-9{,}1+t\sqrt{(\frac{1}{16} + \frac{1}{25})\frac{(16-1)1{,}1^2 + (25-1)0{,}9^2}{16+25-2}}\right) = 0{,}98$$

Então, para $S_1^2 = 1{,}1^2$ e $S_2^2 = 0{,}9^2$ o intervalo de confiança para $\mu_1 - \mu_2$ a 98% de confiança (ou com 2% de risco de erro) vai ser: $12{,}2 - 9{,}1 - 2{,}42\sqrt{(\frac{1}{16}+\frac{1}{25})\frac{(16-1)1{,}1^2+(25-1)0{,}9^2}{16+25-2}} \langle \mu_1 - \mu_2 \langle 12{,}2 - 9{,}1 + 2{,}42\sqrt{(\frac{1}{16}+\frac{1}{25})\frac{(16-1)1{,}1^2+(25-1)0{,}9^2}{16+25-2}}$

Se o resultado for positivo, então o tratamento tem efeito corrosivo.

V Intervalo de confiança para a diferença de proporções

$$P(\overline{p}_1 - \overline{p}_2 - Z\sqrt{\frac{\overline{p}_1(1-\overline{p}_1)}{n_1}+\frac{\overline{p}_2(1-\overline{p}_2)}{n_2}} < p_1 - p_2 < \overline{p}_1 - \overline{p}_2 - Z\sqrt{\frac{\overline{p}_1(1-\overline{p}_1)}{n_1}+\frac{\overline{p}_2(1-\overline{p}_2)}{n_2}}) = 1-\alpha$$

Um intervalo de confiança para uma proporção populacional é dado por:

$$\overline{p}_1 - \overline{p}_2 - Z\sqrt{\frac{\overline{p}_1(1-\overline{p}_1)}{n_1}+\frac{\overline{p}_2(1-\overline{p}_2)}{n_2}} < p_1 - p_2 < \overline{p}_1 - \overline{p}_2 - Z\sqrt{\frac{\overline{p}_1(1-\overline{p}_1)}{n_1}+\frac{\overline{p}_2(1-\overline{p}_2)}{n_2}}$$

onde: $\overline{p}_1 - \overline{p}_2$ é a diferença de proporção amostral Z é o valor da variável normal padrão para o grau de confiança $1-\alpha$. n_1 e n_2 são os tamanhos das respectivas amostras. O valor de Z que é tabelado, deverá ser consultado para $Z_{1-\frac{\alpha}{2}}$.

Exemplo

Uma amostra aleatória de 100 eleitores do Município de Maputo, para certos bairros, dá 55% como favoráveis a um certo candidato, e outra de 100 eleitores para outro bairro, dá 52% ao mesmo candidato. Determine os limites de confiança para a diferença de proporções de eleitores favoráveis ao candidato assumindo 95% de confiança.

Resolução:

$\overline{p}_1 = 0{,}55 \Rightarrow q_1 = 1 - p_1 = 1 - 0{,}55 = 0{,}45 \qquad \overline{p}_2 = 0{,}52 \Rightarrow q_2 = 1 - p_2 = 1 - 0{,}52 = 0{,}48$

$1-\alpha = 0{,}95 \Rightarrow \alpha = 0{,}05 \Rightarrow \frac{\alpha}{2} = 0{,}025 \Rightarrow 1-\frac{\alpha}{2} = 0{,}975 \qquad Z_{1-\frac{\alpha}{2}} = Z_{0{,}975} = 1{,}96$

$$P(0{,}55 - 0{,}52 - 1{,}96\sqrt{\frac{0{,}55\times 0{,}45}{100}+\frac{0{,}52\times 0{,}48}{100}} < p_1 - p_2 < 0{,}55 - 0{,}52 + 1{,}96\sqrt{\frac{0{,}55\times 0{,}45}{100}+\frac{0{,}52\times 0{,}48}{100}}) = 0{,}95$$

$$0{,}55 - 0{,}52 - 1{,}96\sqrt{\frac{0{,}55\times 0{,}45}{100}+\frac{0{,}52\times 0{,}48}{100}} < p_1 - p_2 < 0{,}55 - 0{,}52 + 1{,}96\sqrt{\frac{0{,}55\times 0{,}45}{100}+\frac{0{,}52\times 0{,}48}{100}}$$

3. TESTE DE HIPÓTESES

A. Teste de Hipóteses para variáveis Quantitativas

- Se nada é conhecido acerca da População, a <u>estimação</u> é usada para fornecer uma estimativa de ponto e de intervalo acerca da População.
- Se alguma informação acerca da População é proposta ou suspeitada, o <u>Teste de Hipóteses</u> é usado para determinar a plausibilidade desta informação.

Neste capítulo mostraremos a maneira de tratar a apresentação de um problema ligado a mais um dos casos de inferência. Continuaremos a associar as distribuições de probabilidade como foi feito no tema anterior (ligado a intervalos de confiança). Em vez de procurarmos a estimativa do parâmetro, frequentemente parecerá conveniente admitir um certo valor hipotético para depois recolher uma amostra e procurar provar se estaria na situação ideal de ser aceite ou rejeitado.

Exemplo 1: Se o tratamento do Sida usando uma droga tradicional traria efeitos secundários, seria melhor analisar o caso assumindo que não traz efeitos secundários e, depois com base em cálculos rejeitar ou não a essa afirmação.

Exemplo 2: Se a câmara de comércio de Maputo fixa o peso de um pacote de leite, independentemente do fabricante, na formulação da hipótese admitiríamos que para dois produtores quaisquer, os seus pacotes tivessem diferenças no peso.

Formular uma hipótese dessa forma levaria a tanta preocupação ao investigador.

Definição:

Hipótese: É uma sentença sobre o valor de um parâmetro populacional desenvolvida para o propósito de teste. Em geral as hipóteses resultam de questionamento de um valor achado hipotético, com o objectivo de conhecer as razões essenciais de se rejeitar enquanto está correcto (erro de tipo I- α) ou de não rejeitar enquanto estiver errado (erro de tipo II - β).

Exemplos de hipóteses, ou sentenças, feitas acerca de um parâmetro populacional são:

1) O Rendimento médio mensal proveniente de todas as vendas de Mapatana em 5 lojas é de 300.000.000,00Mt. Na formulação de hipótese admitiríamos que as 5 lojas não tivessem tido exactamente 300.000.000,00Mt de rendimento mensal

2) 10 % da produção do fósforo numa certa região é viciada.
Nós iríamos testar a hipótese de que a produção do fósforo naquela região não é viciada.

i. Hipóteses Estatísticas

Denomina-se hipótese nula, aquela hipótese que se pretende testar e abreviadamente escreve-se H_o. A hipótese contrária à hipótese nula denomina-se hipótese alternativa, abreviada por H_1.

ii. Definição

Teste de Hipóteses: é um procedimento, baseado na evidência amostral e na teoria da probabilidade, usado para determinar se a hipótese é uma afirmação razoável e não seria rejeitada, ou é não razoável e seria rejeitada.

Os 5 (cinco) passos essenciais para um teste de hipóteses:

Passo 1: Estabeleça a Hipótese Nula (H_o) e a Hipótese Alternativa (H_1)
Passo 2: Selecione um nível de significância (α)
Passo 3: Identifique a Estatística de teste ($\overline{X}; S; S^2; \overline{p}$)
Passo 4: Formule uma regra de decisão
Passo 5: Tome uma amostra e obtenha uma decisão: (Não rejeitar H_0) ou (rejeitar H_0 e admitir H_1)

Hipótese Nula H_0: Uma afirmação (sentença) sobre o valor de um parâmetro populacional. Revela aquilo que pretendemos testar.

Hipótese Alternativa H_1: Uma afirmação (sentença) que é aceite se os dados amostrais fornecem evidência de que a hipótese nula é falsa e pode ser rejeitada.

Nível de Significância: A probabilidade de rejeitar a hipótese nula quando ela é efectivamente verdadeira, ou seja, valor de α (alfa).

Erro Tipo I: Rejeitar a Hipótese Nula, H_0, quando ela é efectivamente verdadeira. A probabilidade do erro tipo I é igual ao nível de significância, α (alfa).

Erro Tipo II: Aceitar a Hipótese Nula, H_0, quando é efectivamente falsa. A probabilidade do erro tipo II é igual a β (beta)

Região Aceitável (RA) – é o conjunto de valores que não rejeitam H_o

Região Crítica (RC) – é o conjunto de valores que rejeitam H_o

Exemplo: Mrs. Llair é uma conhecida figura da sociedade que é célebre por pretender que é capaz de provar um chá e dizer com 65 % de segurança se foi adoçado antes ou depois do leite ter sido acrescentado. Uma senhora pouco delicada (talvez uma estrangeira....) resolveu pôr em dúvida as fenomenais capacidades gostativas de Mrs. Llair e propôs ingenuamente que lhe fossem dadas as provas, por uma ordem escolhida ao acaso, 10 chávenas de chá de preparação conhecida dos organizadores. O número de respostas erradas será a variável aleatória X.
a) Explique os conceitos de hipótese nula, hipótese alternativa, região crítica, erro de tipo I e erro de tipo II, utilizando esta situação concreta.

b) Calcule o número de chávenas de chá que Mrs. Llair teria de provar para que, simultaneamente, não pudesse falhar o teste com mais de 1% de probabilidade, caso a sua reindivicação fosse verdadeira, e não pudesse passar com mais de 1% de probabilidade, caso as suas respostas fossem fruto do acaso. Pode usar a aproximação por uma distribuição normal.

Resolução

a) Hipótese nula H_o**:** Mrs. Llair falha com probabilidade $p_o = 0{,}35$ a previsão da ordem de adição do açúcar e do leite;
Hipótese alternativa H_1**:** Mrs. Llair responde ao acaso e "prevê" com probabilidade $p_1 = 0{,}5$ a ordem de adição do açúcar e do leite;

Região crítica : escolhendo como estatística do teste o número de respostas erradas X para um dado número n de chávenas provadas, será o intervalo $R_C = [X_c, n]$ tal que se X $\in R_C$, H_o é rejeitada;

O erro de tipo I consiste em rejeitar erradamente H_o, porque o número de falhanços X caiu dentro da região crítica (X$\in R_C$).
O erro de tipo II consiste em aceitar erradamente H_o, porque o número de falhanços X não caiu dentro da região crítica (X$\notin R_C$), apesar de ser verdadeira H_1.

iii. Tipos de Erros

Tabela 5.4 – Tabela de Decisão em Relação H_o

Situação de H_o	Tipos de Decisão	
	Aceita H_0	Rejeita H_0
H_0 é verdadeira	Decisão Correcta	Erro Tipo I - α
H_0 é falsa	Erro Tipo II - β	Decisão Correcta

Estatística de Teste (ou z efetivo, valor de t ou χ^2): É um valor, determinado a partir da informação amostral, usado para determinar se devemos ou não rejeitar a hipótese nula.
Valor Crítico (ou z crítico, valor de t ou χ^2): O ponto divisor entre a região onde a hipótese nula é rejeitada (RC) e a região onde ela não é rejeitada (RA). Este valor é obtido a partir da tabela de z (normal padrão), da tabela de t (t de Student) ou da tabela de χ^2 (qui-quadrado).

Testes de significância unicaudais
Um teste é unicaudal quando a hipótese alternativa, H_1, estabelece uma direcção que nos leve ao uso de digualdade > - maior, < - menor, ≤ - menor ou igual ou ≥ - maior ou igual.

Testes de significância bicaudais

Um teste é bicaudal quando não existe uma direcção especificada para a hipótese alternativa H_1, representado por $\neq$ (diferença)

I- Teste da média populacional

a) Teste da média populacional- amostra grande, desvio padrão da população σ é conhecido.

Teremos $z = \dfrac{\overline{X} - \mu}{\sigma / \sqrt{n}}$

Exemplo:

Para $X \sim \mathrm{N}(\mu;100)$, $n = 25$, $\overline{x} = 1020$ e $\alpha = 0{,}05$, determine a RC, erros da 2ª espécie e a função potência (supondo $\mu = 1010$, 1030, 990 e 980) para:

$$H_o : \mu_o = 1000$$
$$H_1 : \mu_1 > 1000$$

Resolução:

$$P\left(\overline{X} \geq k\right) = 0{,}05 \Leftrightarrow P\left(Z \geq \frac{k - 1000}{\frac{100}{\sqrt{25}}}\right) = 0{,}05 \Leftrightarrow P\left(Z < \frac{k - 1000}{20}\right) = 0{,}95$$

$$\Rightarrow \frac{k - 1000}{20} = 1{,}645 \Leftrightarrow k = 1032{,}9 \qquad \mathrm{RC} \in [1032{,}9;+\infty[\quad 1020<1032{,}9 \text{ não se rejeita } H_o$$

P-value de um Teste de Hipótese

P-value: Esta é a probabilidade (considerando que a hipótese nula é verdadeira) de ter um valor para a estatística de teste no mínimo tão extremo como o valor calculado (efectivo) para o teste.
Se o p-value é menor que o nível de significância (α), H_0 é rejeitada.
Se o p-value é maior que o nível de siginificância (α), H_0 não é rejeitada.

Cálculo do P-value

Teste Unicaudal (para a direita ou cauda superior):
p-value = P{z ≥ valor da estatística de teste calculada}

Teste Unicaudal (para a esquerda ou cauda inferior):
p-value = P{z ≤ valor da estatística de teste calculada}

Teste Estatístico Bicaudal

p-value = 2P{z≥ valor absoluto do valor da estatística de teste calculado}

Docente: Rodrigues Zicai Fazenda z_fazenda@yahoo.com.br ou rzfazenda@ustm.ac.mz 84-4934100

b) Teste para a média populacional: grandes amostras, desvio padrão populacional σ desconhecido

Quando σ é desconhecido, estimamos com o desvio padrão amostral s.

2-a) Quanto maior for o tamanho amostral, n>30, o z efectivo pode ser aproximado com $z=\dfrac{\overline{X}-\mu}{\dfrac{s}{\sqrt{n}}}$. s é obtido apartir da amostra.

2-b) Quanto menor for o tamanho amostral, $n\leq 30$, z efetivo pode ser aproximado com $T=\dfrac{\overline{X}-\mu}{\dfrac{s}{\sqrt{n}}}\sim t_{(n-1)}$ com n-1 graus de liberdade.

Exemplo: De um universo normal de média e variância desconhecidas, foi retirada uma amostra aleatória de 9 observações, cujos resultados foram $\sum x_i=36$, $\sum x_i^2=162$. Proceda ao seguinte ensaio de hipóteses $H_o:\mu=5$ e $H_1:\mu=6$ para um nível de significância de 5%.

Resolução:

É possível observar que para H_1 a média é maior relativamente a da H_o. Também pode-se ver que o desvio não é conhecido e o tamanho da amostra é n=9, que significa ser amostra pequena. Sendo assim deve-se usar distribuição t-student com n-1 graus de liberdade.

$\overline{X}=\dfrac{1}{n}\sum x_i=\dfrac{1}{9}\times 36=4$ $\qquad$ $S^2=\dfrac{1}{n-1}\sum x_i^2-\dfrac{n}{n-1}\overline{X}^2=\dfrac{1}{9}\times 162-\dfrac{9}{8}\times 4^2=2{,}25\Rightarrow S=1{,}5$, logo partindo de $t_{(\alpha;n-1)}=t_{(0{,}05;8)}=1{,}86$, teremos

$$P\left(\overline{X}\geq k/\mu=5\right)=0{,}05\Leftrightarrow P\left(T\geq\frac{k-5}{\frac{1{,}5}{3}}\right)=0{,}05\Leftrightarrow P\left(T<\frac{k-5}{\frac{1}{2}}\right)=0{,}95$$

$\Rightarrow 2\times(k-5)=1{,}86\Leftrightarrow k=5{,}93$ $\qquad$ RC $\in[5{,}93;+\infty[$ $\qquad$ com $\overline{X}$=4<5,93 está na região de aceitação, logo não se rejeita H_o.

II Teste de hipóteses: duas médias populacionais

Assumamos que os parâmetros para duas Populações são: μ_1,μ_2,σ_1 e σ_2.

Caso I: Quando σ_1,σ_2 são conhecidos, o valor de Z será dado por:

$$z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Exemplo:

A altura média de 50 estudantes do sexo masculino que tiveram participação acima da média em competições desportivas da Univerdidade Eduardo Mondlane, é de 68,2 polegadas, com desvio padrão de 2,5 polegadas, enquanto que um grupo de 50 estudantes que não demonstraram interesse em tais competições acusa altura média de 67,5 polegadas, com desvio padrão de 2,8 polegadas. Teste a hipótese de que os estudantes que participam activamente nas competições são mais altos do que os que não se interessam por elas.

Resolução:

Devemos decidir entre as duas hipóteses

$H_o : \mu_1 = \mu_2$, não há diferença significativa entre as alturas médias

$H_1 : \mu_1 > \mu_2$, Altura média do 1º grupo é superior a altura média do 2º grupo

Pela hipótese H_o $\quad \mu_{\bar{x}_1 - \bar{x}_2} = 0 \quad \sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(2{,}5)^2}{50} + \frac{(2{,}8)^2}{50}} = 0{,}53$

Onde utilizamos o desvio padrão amostral como estimativa de σ_1 e σ_2

Então $Z = \frac{\overline{X}_1 - \overline{X}_2}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{68{,}2 - 67{,}5}{0{,}53} = 1{,}32$. Com base num teste unilateral ao nível de 0,05, rejeitaríamos a hipótese H_o se Z fosse superior a 1,645 assim não podemos rejeitar a hipótese ao nível de 0,05.

Caso II: Quando σ_1, σ_2 não são conhecidos mas os tamanhos amostrais n_1 e n_2 são maiores 30, a estatística de teste (Z efetivo) é:

$$z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Caso III: Quando σ_1, σ_2 não são conhecidos mas os tamanhos amostrais n_1 e n_2 são menores ou iguais a 30:

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Exemplos

1- O Instituto Agrário do Chimoio deseja testar um determinado Nitrato, para os campos de produção de trigo no Búzi. Para isso escolheram 24 espaços de terra da mesma área; em metade desses espaços usou-se o fertilizante e na outra metade não. Quanto ao mais, as condições são idênticas para os dois grupos de espaços de terra. A safra média do grupo de terra não tratada com o fertilizante foi de 4,8 celeiros com desvio padrão de 0,4 celeiros, enquqnto que a safra do grupo onde usou-se o fertilizante foi de 5,1 celeiros com desvio padrão de 0,36 celeiros. Podemos concluir que haja melhoria significativa na produção de trigo devida ao emprego do fertilizante adoptando o nível de significância de 5%?

Resolução:

Seja μ_1 e μ_2 médias das populações de terras tratadas e não tratadas com o fertilizante respectivamente. Temos que decidir entre as duas hipóteses a seguir

$$\begin{cases} H_o : \mu_1 = \mu_2 & \text{Não há diferença e se existir é devida ao acaso} \\ H_1 : \mu_1 > \mu_2 & \text{Há diferença devida ao emprego do fertilzante} \end{cases}$$

Como $n_1 < 30$ e $n_2 < 30$, usemos a distribuição t-student, com $n_1 + n_2$ -2 graus de liberdade.

$$T_{cal} = \frac{\overline{X}_1 - \overline{X}_2}{\sigma \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{5{,}1 - 4{,}8}{0{,}397\sqrt{\frac{1}{12} + \frac{1}{12}}} = 1{,}85 \qquad \sigma = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{12(0{,}4)^2 + 12(0{,}36)^2}{12 + 12 - 2}} = 0{,}397$$

$t_{(22;0,95)} = 1{,}82$ como $T_{cal} > t_{(22;0,95)}$, então rejeita-se H_o, o que quer dizer que a diferença é devida ao uso do fertilizante. Atente para o facto de que T_{cal} ser o valor calculado.

2- Uma exploração agrícola pretende testar o efeito de um novo estrume natural sobre a produção de batata. Para tal, escolheram-se 24 hectares de terra, metade dessa área foi tratada com o novo estrume, e a outra metade sem ele. As restantes condições são idênticas e a produção tem um comportamento normal. A colheita média por hectare tratado com o estrume foi 1,5t de batatas com um desvio padrão de 0,36t. A safra média por hectare restante foi 4,8t com um desvio padrão de 0,4t. Podemos concluir que há melhoria significativa na produção de batata devido ao emprego de novo estrume natural a um nível de significância de 0,05?

Resolução:

Dados $\alpha = 0{,}05 \quad n = 24 \quad \overline{X}_A = 1{,}5t \quad \overline{X}_B = 4{,}8t \quad S_A = 0{,}36t \quad S_A = 0{,}4t$

$H_o : \mu_A = \mu_B$ Não há nenhuma melhoria na produção da batata pelo emprego de novo estrume

$H_1 : \mu_A \langle \mu_B$ Há melhoria significativa na produção da Batata pelo emprego de novo estrume

$$P\left(\overline{X}_B \rangle \overline{X}_A\right) = P\left(\overline{X}_B - \overline{X}_A \rangle 0\right) = 0{,}05$$

$$P\left(t_{(n_1+n_2-2)} \langle \frac{K-(\mu_B-\mu_A)}{\sqrt{\left(\frac{1}{n_B}+\frac{1}{n_A}\right)\frac{(n_B-1)S_B^2+(n_A-1)S_A^2}{n_B+n_A-2}}} \right) = 0{,}05$$

$$2{,}82 \langle \frac{K-33}{\sqrt{\frac{1}{6}\times\frac{11\times 0{,}4^2+11\times 0{,}36^2}{22}}} \Leftrightarrow K \rangle 3{,}36 \Rightarrow RC = [3{,}36;+\infty[$$

$\overline{X}_B - \overline{X}_A = 4{,}8t - 1{,}5t = 3{,}3t$ Não rejeitamos H_o e pode-se afirmar que não há melhoria na produção

II Testes referentes à proporção

Proporção: Uma fracção ou percentagem que indica uma parte da População ou amostra que tem um particular traço de interesse. A proporção amostral é denotada por $\overline{p}$ onde:

$$\overline{p} = \frac{\textit{número} \text{ de sucessos na amostra}}{\textit{tamanho} \text{ da amostra}}$$

O valor de teste é $z = \dfrac{\overline{p}-p}{\sqrt{\dfrac{p(1-p)}{n}}}$ p≡ proporção populacional e $\overline{p} \equiv$ proporção amostral

Exemplo

3- Um fabricante de um determinado sal de cozinha alega que o mesmo acusou 90% de eficiência em aliviar o problema do bóssio. Pegou-se numa amostra de 200 indivíduos que sofriam de bóssio; o sal deu resultado positivo em 160. Determine se a alegação do fabricante é legítma ou não.

Resolução: Com se vê, teremos que testar se a alegação constitui uma realidade ou não

$\begin{cases} H_o : p = 0{,}9 & \text{A alegação é correcta} \\ H_1 : p < 0{,}9 & \text{A alegação é falsa} \end{cases}$

Escolhemos um teste unilateral á esquerda, pois o que queremos é determinar se a proporção de indivíduos beneficiados é muito baixa $1-\alpha = 0{,}90 \Rightarrow \alpha = 0{,}01$ $Z_\alpha = -2{,}33$. A alegaão será legítima caso $Z_{cal} > Z_\alpha$. Então vejamos: Tomando H_o-Verdadeiro $\mu = np = 200\times 0{,}9 = 180$

$\sigma = \sqrt{npq} = \sqrt{200 \times 0{,}9 \times 0{,}1} = 4{,}23$ $\qquad Z_{cal} = \dfrac{160 - 180}{4{,}23} = -4{,}73.$ Como $Z_{cal} < Z_{\alpha}$, rejeitamos H_o, pelo que a alegação é falsa

p

III- Teste de diferença entre duas proporções populacionais

n_1 é o tamanho da amostra da População 1.
n_2 é o tamanho da amostra da População 2.

$\overline{p}_c$ é a média ponderada das duas proporções amostrais, calculada por:

$$\overline{p}_c = \frac{\text{número total de sucessos}}{\text{tamanho total das duas amostras}} = \frac{X_1 + X_2}{n_1 + n_2}$$

X_1 é o número de sucessos em n_1.
X_2 é o número de sucessos em n_2.

Exemplo

4- Dois grupos A e B são constituídos cada um por 100 pessoas com a mesma doença. É ministrado um soro ao grupo A, mas não ao B (grupo de controlo). Verificou-se que 75 e 65 pessoas dos grupos A e B, respectivamente, se curaram da doença. Teste a hipótese do soro auxiliar a cura da doença para o níveil de significância de 1 %;

Resolução:

Sejam $P(A) = \dfrac{75}{100} = 0{,}75$ e $P(B) = \dfrac{65}{100} = 0{,}65$ as probabilidades de cura nos grupos A e B respectivamente, conduzindo as hipóteses:

$H_o : p_A = p_B$ - Não se nota nenhum efeito de aplicação do soro

$H_1 : p_A > p_B$ - A aplicação do soro auxilia na cura da doença

$Z_{0,99} = 2{,}326$ $\qquad Z_{cal} = \dfrac{\left(p_A^* - p_B^*\right) - \left(p_A - p_B\right)}{\sqrt{\dfrac{p_A^* q_A^*}{n_A} + \dfrac{p_B^* q_B^*}{n_B}}} = \dfrac{0{,}75 - 0{,}65}{\sqrt{\dfrac{0{,}75 \times 0{,}25}{100} + \dfrac{0{,}65 \times 0{,}35}{100}}} = \dfrac{0{,}10}{0{,}064} = 1{,}5625$

como $Z_{0,99} < Z_{cal}$ Não se rejeita H_o

IV Teste de hipóteses para variância

a) Teste de hipóteses para variância se a média μ não é conhecida

$\chi^2_{n-1} = \frac{(n-1)s^2}{\sigma^2}$ É uma distribuição qui-quadrático com n-1 graus de liberdade, deveremos comparar com o valor tabelado $\chi^2_{(n-1;1-\alpha)}$

Exemplo:

5- Um amostra de 10 elementos extraída duma população normal forneceu variância igual a 12,4. Esse resultado é suficiente para se concluir a 5% de significância, que a variância é inferior a 25?
Resolução:

$$\begin{cases} H_o : \sigma^2 = 25 \\ H_1 : \sigma^2 < 25 \end{cases} \qquad \chi^2_{cal} = \chi^2_9 = \frac{(n-1)s^2}{\sigma^2} = \frac{9 \times 12{,}4}{25} = 4{,}464$$

o crítico será $\chi^2_{9;95\%} = 3{,}325$, como $\chi^2_{cal} > \chi^2_{tabela}$, não se rejeita H_o

b) Teste de hipóteses para razão de variâncias

Para este caso usamos a distribuição de **F-Fisher**. Este tipo de teste é muito usado em laboratórios de Medicina com objectivo de apurar a razão de variabilidades entre a mostra antes do ensaio e a amostra depois do ensaio.

$F_{cal} = \frac{s_1^2}{s_2^2}$ F_{cal} - é o valor calculado $F_{tabela} = F_{(n_1-1,n_2-1,\alpha)}$

Exemplo:

6- Uma amostra de 10 elementos extraída duma população forneceu uma variância igual a 12,4. Esse resultado é suficiente para construir a 5% de significância, que a variância dessa população é inferior a 25.
Resolução:

$$\begin{cases} H_o : \sigma^2 = 25 \\ H_1 : \sigma^2 < 25 \end{cases}$$ Como F só traz valores críticos á direita, então, adaptamos o teste para:

$$\begin{cases} H_o : 25 = \sigma^2 \\ H_1 : 25 > \sigma^2 \end{cases}$$ O valor σ_o^2 é considerado como uma estimativa de variância com $\upsilon = \infty$ $F_{\infty,9,5\%} = \frac{s_1^2}{s_2^2} = \frac{25}{12{,}4} \approx 2{,}016$ $F_{tabela} = F_{\infty,9,5\%} = 2{,}71$ logo, não se pode rejeitar H_o

Erro tipo II ou da 2ª espécie (probabilidade de não rejeitar H_o dado que H_o é falso) ou simplesmente erro β

Este tipo de erro é útil para determinar o poder do teste que é $\pi = 1 - \beta$. Neste caso é dado por

β =P(Não rejeitar H_o/H_o é falso)= P(Não rejeitar H_o/H_1)= P($\vartheta^* \in RA$ /H_1) em que ϑ^*-Estimador e RA – Região Aceitável

Exemplo:

7.Para $X \sim N(\mu;100)$, $n = 25$, $\bar{x} = 1020$ e $\alpha = 0{,}05$, calcule a RC, erros da 2ª espécie e a função potência (supondo $\mu = 1010$, 1030, 990 e 980) para:

$$H_o : \mu_1 = 1000$$

$$H_1 : \mu_1 \neq 1000$$

Resolução:

$$P(\bar{X} \leq k_1) = \frac{0{,}05}{2} \Leftrightarrow P\left(Z \geq \frac{k_1 - 1000}{\frac{100}{\sqrt{25}}} \right) = 0{,}025 \Rightarrow \frac{k_1 - 1000}{20} = -1{,}96 \Leftrightarrow k_1 = 960{,}8$$

$$P(\bar{X} \geq k_2) = \frac{0{,}05}{2} \Leftrightarrow P\left(Z \geq \frac{k_2 - 1000}{\frac{100}{\sqrt{25}}} \right) = 0{,}025 \Rightarrow \frac{k_2 - 1000}{20} = 1{,}96 \Leftrightarrow k_2 = 1039{,}2$$

$RC \in]-\infty;960{,}8] \cup [1039{,}2;+\infty[$

Tabela 5.5 – Erros de 2ª espécie e potência do teste

μ	$\beta(\mu)$	$\pi(\mu)$
1010	0,9279	0,0721
1030	0,6772	0,3228
990	0,9279	0,0721
980	0,83	0,17

B. Teste de hipóteses para variáveis qualitativas

Recordemos que variáveis qualitativas são aquelas que podem ser representadas por letras porque exprimem uma qualidade. São casos de nacionalidade (Moçambicana, Zimbabweana, Malawiana, etc), côr dos cabelos (pretos, castanhos, etc), etc. Neste subcapítulo só se trata de realizar teste de hipóteses entre duas variáveis qualitativas, em que as respectivas observações podem ser classificadas em diversas categorias mutuamente exclusivas.

O problema de mensuração do grau de associação entre dois conjuntos de scores é de carácter bem diferente do teste da simples existência de uma associação numa determinada população, tomando uma ou mais variáveis quantitativas. Naturalmente, há interesse em avaliar o grau de associação entre dois conjuntos de scores referentes a um grupo de indivíduos, o que pode ser satisfeito fazendo uma avaliação dum teste qualitativo.

i. Tabelas de contingência

São tabelas usadas para testar a existência de relações entre duas variáveis. Em português não tem o mesmo significado. Segundo vários dicionários, caso do dicioário mais gramáticada Moçambique Editora, 3ª edição de 2002, só para citar, define "contingência" como sendo s.f. eventualidade, possibilidade imprevisível. Em estatística, a palavra "contingência" está mais próxima dos significados atribuídos em Inglês.

O teste chi-square ou chi-quadrado (χ^2) é o mais usado para avaliar a relação entre duas variáveis qualitativas. Este teste é não paramétrico, diferentemente do teste quantitativo que tem parâmetros $\mu, \sigma \, ou \, \sigma^2$ em que a variável suporta-se pela distribuição normal, o que é mais útil, já que não precisa da suposição de normalidade das variáveis para analisar o grau de associação entre elas. Porém, este teste é menos poderoso que o teste paramétrico.

O teste não paramétrico distingue-se em dois: o de independência e o de homogeneidade.

ii. Teste de independência e teste de homogeneidade

Suponha que a polícia da República esteja interessada no desempenho dos seus Agentes em actividades de segurança e na participação activa dos seus chefes. Suponha que ela categorize o desempenho dos Agentes em três grupos: baixo, médio, alto e, do mesmo modo, categoriza a participação dos chefes em dois grupos: activa e fraca. Suponha ainda que esta actividade é desenvolvida entre 300 Agentes.

Neste caso a polícia pode delinear sua pesquisa de duas formas:

Caso 1. Selecionar uma amostra de agentes aleatoriamente e examinar em que célula cada uma está alocada, logo o único valor fixo será o total geral que será de 300. Mas os totais de

colunas e de linhas serão frutos da pesquisa, portanto, aleatórios, neste caso estamos frente a um teste de independência de variáveis.
E a tabela de contingência será:

Tabela 5.5 – Distribuição aleatória de 300 Agentes por participação dos chefes e desempenho

Participação dos chefes	Desempenho dos Agentes			
	Baixo	Médio	Alto	Total
Activa				Aleatório
Fraca				aleatório
Total	Aleatório	Aleatório	Aleatório	300

⇨ Teste de independência

Atente para o facto de que poderá-se fixar o número de Agentes de acordo com seu desempenho.

Caso 2. Realizar uma amostra aleatória de 100 de cada grupo de agentes, logo os totais das colunas serão fixos, mas os totais das linhas serão aleatórios e assim estaremos frente a um teste de homogeneidade:

E a tabela de contingência será:

Tabela 5.5 a) – Distribuição de desempenho de 300 Agentes com participação aleatória dos Chefes Desempenho

Participação dos chefes	Desempenho dos Agentes			
	Baixo	Médio	Alto	Total
Ativa				Aleatório
Fraca				aleatório
Total	100	100	100	300

⇨ Teste de homogeneidade

Os totais das colunas e das linhas, são denominados totais marginais. Quando os totais marginais variam livremente, o teste de associação é chamado de *independência*, e quando um dos conjuntos, linha ou coluna é fixado pelo pesquisador então é chamado de teste de *homogeneidade*. Isso vai depender do pesquisador. No exemplo da polícia, observemos que para ela é muito mais fácil fixar o número de Agentes segundo seu desempenho, do que fixar pela participação dos chefes, que, apriori será difícil.

I. Teste de Independência

Repare na lógica do teste com base noutro exemplo bastante simples. Suponha que 125 jovens foram expostos a três tipos de aprendizagem, sobre química. Após a aprendizagem foi solicitado a cada jovem para indicar qual dos métodos mais gostou. O que se deseja saber é

se a escolha do método de ensino está relacionado ao género da pessoa: pois suspeita-se de que o género pode estar influenciando no desempenho. Veja os dados na tabela a seguir:

Tabela 5.6 – Distribuição de Tipo de aprendizagem por sexo

Género	Tipo de Aprendizagem			
	A	B	C	Total
Meninos	30	29	16	75
Meninas	12	33	5	50
Total	42	62	21	125

Analisando atentamente a Tabela, composta por valores absolutos, percebemos que:

A amostra está composta por mais meninos do que meninas,
Nas aprendizagens A e C o número de meninos é maior do que meninas, e
Na aprendizagem B essa relação se inverte, isto é, mais meninas que meninos.

Contudo, essa análise fica prejudicada pela composição da amostra, que tem mais meninos do que meninas. Portanto, a primeira coisa a fazer é analisar as estruturas percentuais, mostradas na Tabela a seguir, ou seja retirar a influência da amostragem.

Percentagens de jovens por

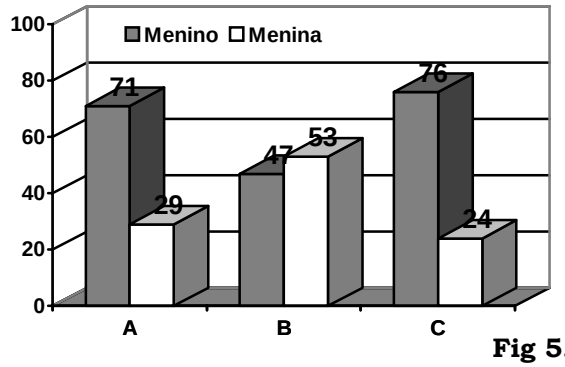


Fig 5.1

Tipo de Aprendizagem

Tabela 5.6 a) – Percentagens de jovens por tipo de aprendizagem escolhida

Género	Tipo de Aprendizagem			
	A	B	C	Total
Meninos	71%	47%	76%	60%
Meninas	29%	53%	24%	40%
Total	100%	100%	100%	100%

Observe cuidadosamente a tabela anterior, onde 60% da amostra é composta por meninos. Se a preferência dos jovens pelas aprendizagens não dependesse do género, esperaríamos

que a estrutura percentual para cada aprendizagem ficasse em torno de 60% para os meninos e 40% para as meninas, desvios grandes destes percentuais estariam mostrando evidências de que existe alguma relação entre essas variáveis. Essa inspecção intuitiva também, pode ser feita analisando a estrutura dentro de cada género como a seguir se ilustra.

Percentagem de jovens por género e tipo de aprendizagem escolhido

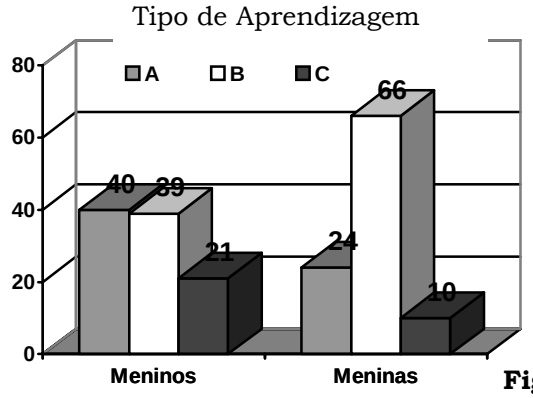


Fig 5.2

Tabela 5.7. Percentagem de jovens por género e tipo de aprendizagem escolhido

	Tipo de Aprendizagem			
Género	A	B	C	Total
Meninos	40%	39%	21%	100%
Meninas	24%	66%	10%	100%
Total	33%	50%	17%	100%

Analisando a Tabela atrás, observamos que as meninas tem uma forte preferência pela aprendizagem B, enquanto que os meninos se dividem entre a aprendizagem A e B.

Assim, intuitivamente percebemos que existe interferência do género na preferência, agora precisamos saber até que ponto essas diferenças se devem ao acaso, ou a existência de associação entre as duas variáveis:

X: preferência pela aprendizagem (A, B e C) ➔ qualitativa
Y: género (meninos, meninas) ➔ qualitativa

Hipótese nula: A preferência pela Aprendizagem não depende do género da criança
Hipótese alternativa: A preferência pela aprendizagem depende do género da criança (ou, o género interfere na preferência pela aprendizagem)

ou
H_o: independência de variáveis
H_1: dependência de variáveis

Como deveriam ser os valores a serem observados se as variáveis fossem não dependentes?, ou dito de outra forma, sob a hipótese nula, de independência de variáveis, como deveriam ser os valores a serem observados? A lógica nos diz que esses valores devem estar muito próximos da estrutura percentual global. Esses valores são denominados de valores esperados.

Tabela 5.8 Tipo de Aprendizagem

Tabela 5.9 Valores esperados

Género	A	B	C	Total
Meninos	60%	60%	60%	60%
Meninas	40%	40%	40%	40%
Total	42	62	21	150

Género	A	B	C	Total
Meninos	25	37	13	75
Meninas	17	25	8	50
Total	42	62	21	150

Observe que cada valor esperado foi calculado supondo que a estrutura percentual global se mantém em cada coluna:

Calculando os valores esperados, sobre a suposição de independência, teremos:
Valor esperado menino, aprendizagem A: 60% de 42 = 25,2
Valor esperado menino, aprendizagem B: 60% de 62 = 37,2
Valor esperado menino, aprendizagem C: 60% de 21 = 12,6
Valor esperado menina, aprendizagem A: 40% de 42 = 16,8
Valor esperado menina, aprendizagem B: 40% de 62 = 24,8
Valor esperado menina, aprendizagem C: 40% de 21 = 8,4

O mesmo teria acontecido se fixarmos primeiro a aprendizagem:
Valor esperado aprendizagem A, menino: 33,7% de 75= 25,2
Valor esperado aprendizagem A, menina: 33,7% de 50= 16,8
Valor esperado aprendizagem B, menino: 49,6% de 75= 37,2
Valor esperado aprendizagem B, menina: 49,6% de 50= 24,8
Valor esperado aprendizagem C, menino: 16,8% de 75= 12,6
Valor esperado aprendizagem C, menina: 16,8% de 50= 8,4

Tanto faz fixar a linha ou a coluna pois:

$$esperado = \frac{total\_linha * total\_coluna}{total\_geral} = total\_linha * \frac{total\_coluna}{total\_geral} = total\_coluna * \frac{total\_linha}{total\_geral}$$

Por exemplo, calculemos o valor esperado da primeira linha e primeira coluna:

$$esperado = \frac{75 * 42}{125} = 75 * \frac{42}{125} = 42 * \frac{75}{125} = 25{,}2$$

Assim calculando os valores esperados para todas as células temos:

Tabela 5.10 - Valores esperados para Tipo de Aprendizagem

Valores esperados para Tipo de Aprendizagem

Género	A		B		C		Total
	observado	Esperado	observado	esperado	observado	esperado	
Menino	30	25,2	29	37,2	16	12,6	75
diferença	+ 4,8		- 8,2		+ 3,4		
Menina	12	16,8	33	24,8	5	8,4	50
diferença	- 4,8		+ 8,2		- 3,4		
Total	42		62		21		125

Dentro de cada célula, no canto superior esquerdo colocamos o valor observado, no canto superior direito o valor esperado (sob a hipótese de independência) e, na parte inferior, a distância entre o observado e o esperado. Logo, se as variáveis fossem não dependentes, as distâncias entre os valores observados e esperados deveriam ser muito pequenas, caso contrário haverá indícios de dependência. A pergunta agora é: quando é que distância é pequena ou grande? Para isto devemos calcular o valor chi-quadrado da amostra:

$$\chi^2_{amostra} = \sum \frac{(observado - esperado)^2}{esperado}$$

que terá uma distribuição chi-quadrado com v graus de liberdade

$$\chi^2_{amostra} \sim \chi^2_v \quad \Longrightarrow$$

v=graus de liberdade
v=(nº colunas –1)*(nº linha-1)

No nosso exemplo:

$$\chi^2_{amostra} = \frac{(+4,8)^2}{25,2} + \frac{(-8,2)^2}{37,2} + \frac{(+3,4)^2}{12,6} + \frac{(-4,8)^2}{16,8} + \frac{(+8,2)^2}{24,8} + \frac{(-3,4)^2}{8,4}$$

$\chi^2_{amostra}$= 0,914 + 1,808 + 0,917 + 1,371 + 2,711 + 1,376

$\chi^2_{amostra}$= 9,09818 $\Longrightarrow$ onde v = (2-1)*(3-1)=1*2=2

Para aceitar ou rejeitar a hipótese devemos procurar na tabela chi-quadrado, com dois graus de liberdade. Para α =5%, o valor crítico é 5,991, como o valor da amostra é maior que o valor crítico, logo rejeitamos a hipótese nula, concluindo que o género interfere na preferência pelas aprendizagens.

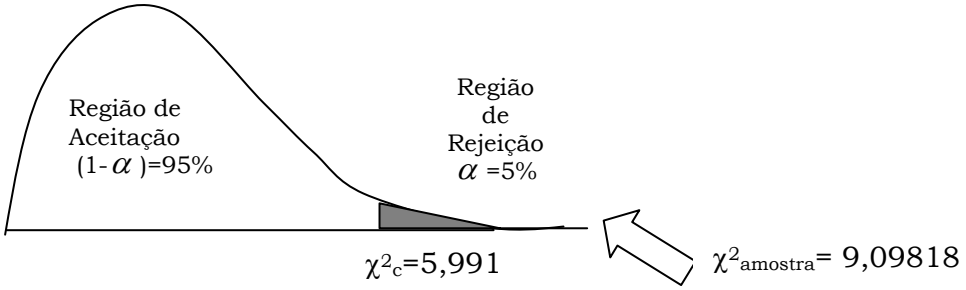


Fig 5.3

iii Limitações do teste χ^2:

Infelizmente, o teste chi-quadrado não permite concluir como se dá a relação, uma vez que ele testa apenas a hipótese geral de que as duas variáveis são não dependentes. Examinando a distância entre valor observado e esperado. Por exemplo, observamos que as meninas tem uma maior preferência pela aprendizagem B, porém não podemos concluir nada.

Uma outra limitação do teste chi-quadrado é que o valor esperado das células não deve ser menor ou igual a 5, pois isso torna vulnerável a estatística. Nesse caso, tem que se usar outra estratégia.

Teste de homogeneidade

Quando testamos independência de variáveis, o pesquisador só controla o tamanho total da amostra, mas os totais para cada coluna e linha são aleatórios. No caso do exemplo anterior, os pesquisadores selecionaram aleatoriamente 125 jovens, das quais 75 eram meninos e 50 meninas. Ele não fixou o número de meninos e o número de meninas.

Vejamos um exemplo de teste de homogeneidade. Retomemos o exemplo inicial e suponhamos que a polícia fixou o tamanho dentro de cada grupo de agentes e os resultados foram os seguintes:

Número de agentes segundo seu desempenho em Agente e
participação dos chefes nas actividades

Tabela 5.11 – Desempenho de Agentes

Participação dos chefes	Desempenho dos Agentes			
	Baixo	Médio	Alto	Total
Ativa	5	25	70	100
Fraca	95	75	30	200
Total	100	100	100	300

A hipótese nula está a testar que a proporção de agentes com baixo desempenho é igual à proporção de agentes médio e igual à proporção de agentes com desempenho alto, quando seus chefes participam activamente nas actividades, contra a hipótese alternativa que indica que existe pelo menos uma proporção diferente.

iv. O Coeficiente de contingência

O coeficiente de contingência é uma medida do alcance da associação ou relação entre dois conjuntos de atributos. Ele é calculado em função do valor calculado na tabela de contingência e não depende da ordenação das categorias das variáveis: $C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$

Onde N é o tamanho da amostra geral. No exemplo das aprendizagens de TV, o coeficiente de contingência será: $C = \sqrt{\frac{9{,}09818}{9{,}09818 + 125}} = 0{,}26047$

O Teste exacto de Fisher

A prova de Fisher é útil quando trabalhamos com variáveis categorizadas e quando o tamanho das amostras não dependentes é pequeno. É utilizado quando duas variáveis só podem ser catalogadas em duas possíveis categorias ou níveis, logo em tabelas de contingência 2x2.

Tabela 5.12 – Hipóses para teste de Fisher

Grupos não dependentes	Positivo	Negativo	Total
Grupo I	A	B	A+B (Fixo)
Grupo II	C	D	C+D (Fixo)
Total	A+C	B+D	N (Fixo)

H_0: independência
H_1: dependência

O método está baseado na distribuição hipergeométrica, calculando a probabilidade de observar um determinado conjunto de frequências numa tabela 2x2, quando se consideram fixos os totais marginais, sob a hipótese de nulidade, ou seja independência de variáveis.

$p = \frac{\binom{A+C}{A}\binom{B+D}{B}}{\binom{N}{A+B}} = \frac{(A+B)!*(C+D)!*(A+C)!*(B+D)!}{N!*A!*B!*C!*D!}$ Essa probabilidade, na realidade é o p-valor, ou seja a probabilidade de rejeitar a hipótese nula sob a suposição de independência; em outras palavras, é a probabilidade de afirmar que são dependentes quando na realidade as variáveis são não dependentes. Esse valor deve ser comparado com o nível de significância estipulado pelo pesquisador.

Exercícios Resolvidos

1- Uma população encontra-se dividida em três estratos com tamanhos respectivamente $N_1 = 80$, $N_2 = 120$, $N_3 = 60$. Ao realizar uma amostragem estratificada proporcional, 12 elementos da amostra foram retirados do 1º estrato. Qual é o número total dos elementos da amostra?

Resolução:

Calculemos a frequência para o estrato 1 que possui $n_1 = 12$. $f = \frac{n_1}{N_1} = \frac{12}{80} = 0{,}15$. Como é estratificada proporcional, significa que a frequência nos estratos é igual. Sendo assim, teremos:
$n_2 = N_2 \times f = 120 \times 0{,}15 = 18$ e $n_3 = N_3 \times f = 60 \times 0{,}15 = 9$. Portanto, o tamanho da amostra será: $n = n_1 + n_2 + n_3 = 12 + 18 + 9 = 39$

2- Realizou-se uma amostragem entre os moradores da cidade de Maputo da seguinte forma: em cada distrito sorteou-se um número de quarteirões proporcional à área do distrito. De cada quarteirão foram sorteados cinco residências cujos moradores são entevistados. De que amostragem se trata?

Resposta:

É estratificada pois possui unidade primária "distrito" e secundária "quarteirões"

3- Retire uma amostra aleatória de 32 elementos e determine a média e mediana, para os rendimentos de 220 moçambicanos residentes na cidade de Maputo. Os dados a seguir e estão expressos em milhões de meticais.

8,9	3,25	4,3	3,2	5,1	7,3	2,4	5,3	6,4	7,7
5,2	8,8	7,4	8,4	5,7	6,9	7,4	5,4	4,4	1,7
6,3	5,4	8,6	7,6	5,4	5,9	5,3	5,7	1,5	1,5
8,7	7,2	7,5	8,7	8,7	8,7	7,8	8,7	1,4	4,5
8,4	8,6	6,4	9,8	6,4	6,9	6,9	1,4	1,3	9,5
3,4	8,3	9,2	6,8	3,5	4,2	4,2	2,5	6,9	9,6
2,9	7,9	3,4	3,8	1,2	6,6	3,5	3,4	8,7	8,3
5,3	1,1	9,1	5,1	4,5	8,8	7,5	1,1	2,5	5,7
5,4	1,3	8,1	6,7	6,4	2,3	8,6	6,6	1,5	9,4
5,8	4,5	6,0	7,5	3,5	2,1	9,7	8,2	3,4	2,1
7,5	8,7	4,0	4,3	5,5	3,0	3,5	3,3	2,4	4,9
8,1	3,6	6,5,	6,1	6,6	5,0	4,6	2,2	1,4	8,5
8,3	3,5	8,7	7,2	2,3	8,6	2,8	2,4	5,6	6,7
9,1	4,6	6,8	4,6	1,1	4,4	3,9	7,9	3,6	8,4
4,5	5,3	9,4	5,8	7,2	3,9	3,7	4,6	2,4	6,7
5,3	4,6,	2,3	3,4	8,3	7,5	4,6	2,4	9,7	3,4
6,1	8,7	8,7	7,6	7,3	8,3	7,3	8,3	5,6	6,1
6,1	6,4	9,6	8,4	6,4	4,6	6,3	6,4	8,7	8,0
2,7	2,1	8,4	3,1	5,2	6,6	7,8	4,1	6,4	9,0
9,1	4,3	5,4	8,1	5,4	6,4	8,7	8,6	7,2	9,1
3,3	5,4	6,4	6,2	3,4	5,4	5,4	6,7	7,3	5,2
6,4	4,5	7,2	4,1	4,2	4,1	7,6	8,7	6,7	4,6

4- Considere uma variável aleatória normal de variância igual a 4, recolheu-se a seguinte amostra: 3, 7, 9, 10, 11, 12, 12, 14.

a) Determine um intervalo de confiança a 90% para a média.

b) Qual deveria ser o grau de confiança a utilizar para que a amplitude do intervalo fosse 2,77?

c) Indique a dimensão da amostra que consideraria para que o erro cometido fosse inferior a um, nas condições da alínea a)

d) Explique sucintamente o que aconteceria se aumentasse para 99% o grau de confiança, mantendo a amostra.

Resolução: Dados: $\sigma^2 = 4 \qquad \overline{X} = 9{,}75$

a) $1-\alpha = 0{,}90 \Rightarrow 1-\dfrac{\alpha}{2} = 0{,}95 \Rightarrow Z_{0{,}95} = 1{,}645$

$$P\left(\overline{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \Rightarrow \mu \in \left]\overline{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right[$$

$$\mu \in \left]9{,}75-1{,}645\frac{2}{\sqrt{8}};9{,}75+1{,}645\frac{2}{\sqrt{8}}\right[\Rightarrow \left]8{,}5868;10{,}9132\right[$$

b) $2Z_{1-\frac{\alpha}{2}}\frac{2}{\sqrt{8}}=2{,}77 \Rightarrow Z_{1-\frac{\alpha}{2}}=1{,}9587 \Rightarrow 1-\frac{\alpha}{2}=0{,}9747 \Rightarrow 1-\alpha \approx 0{,}95$

c) $2\times 1{,}645\times\frac{2}{\sqrt{n}}\leq 1 \Rightarrow n \geq (4\times 1{,}645)^2 \Rightarrow n \geq 44$

d) O intervalo fica com a maior amplitude. Mas que precisa achar esse intervalo como feito na alínea a)

5- Construiu-se uma máquina para produzir rolamentos com esferas de diâmetro médio de 0,500 cm e desvio padrão 0,08 cm a 99% de confiança. Para verificar o bom funcionamento da máquina é retirada uma amostra de 6 rolamentos de duas em duas horas e calcula-se o diâmetro médio de cada amostra, tendo-se observado uma média de 0,52. Será que o padrão ainda é cumprido?

Resolução:

A média dos diâmetros dos 6 rolamentos deverá estar compreendida no intervalo

$$P(\bar{x}-Z\frac{\sigma}{\sqrt{n}}<\mu<\bar{x}+Z\frac{\sigma}{\sqrt{n}})=1-\alpha \Leftrightarrow P(0{,}52-Z\frac{0{,}08}{\sqrt{6}}<\mu<0{,}52+Z\frac{0{,}08}{\sqrt{6}})=0{,}99$$

$0{,}52-2{,}575\frac{0{,}08}{\sqrt{6}}<\mu<0{,}52+2{,}575\frac{0{,}08}{\sqrt{6}} \Leftrightarrow 0{,}436<\mu<0{,}602$ Sim ainda é cumprido

Suponha-se em presença de uma população normal normal, com parâmetros desconhecidos. Com base numa amostra casual, com 16 observações, foi construído o seguinte intervalo de confiança para a média da população: $\left]7{,}398;12{,}602\right[$

Sabendo que, com a informação da amostra, se obteve $s=4$, qual o grau de confiança que pode atribuir ao intervalo atrás referido?

Com base na mesma amostra construa um intervalo de confiança a 95% para a variância da população.

Suponha que a verdadeira variância da população é 44. Se pretender construir um intervalo de confiança (a 95% para a média da população) cuja amplitude não exceda 6.5, qual deverá ser a dimensão da amostra a considerar?

Resolução:

a)

$$\bar{X}-t_{\left(15;1-\frac{\alpha}{2}\right)}\frac{4}{\sqrt{16}}=7{,}398$$

$$\bar{X}+t_{\left(15;1-\frac{\alpha}{2}\right)}\frac{4}{\sqrt{16}}=12{,}602$$

(-1)

$$\bar{X}-t_{\left(15;1-\frac{\alpha}{2}\right)}\frac{4}{\sqrt{16}}=7{,}398$$

$$\bar{X}-t_{\left(15;1-\frac{\alpha}{2}\right)}\frac{4}{\sqrt{16}}=7{,}398$$

$$2t_{15;1-\frac{\alpha}{2}}=5{,}204 \Rightarrow t_{15;1-\frac{\alpha}{2}}=2{,}602$$

$$\Rightarrow 1-\alpha=0{,}98$$

b) $1-\alpha = 0{,}95 \Rightarrow 1-\frac{\alpha}{2} = 0{,}975 \Rightarrow \chi^2_{(15;0{,}975)} = 27{,}5 \rightarrow \chi^2_{(15;0{,}025)} = 6{,}26$

$$\chi^2 \in \left]\frac{15\times 16}{27{,}5};\frac{15\times 16}{6{,}26}\right[\Leftrightarrow \chi^2 \in]8{,}7273;38{,}3387[$$

c) Com $s^2 = 44 \quad 1-\alpha = 0{,}975 \qquad \varepsilon = 6{,}5 \quad 2\times Z_{0{,}975}\times\frac{\sqrt{44}}{\sqrt{n}} \le 6{,}5 \Rightarrow n \ge \left(\frac{2\times 1{,}96\times\sqrt{44}}{6{,}5}\right)^2 \Rightarrow n \ge 16$

6- Para $X \sim N(\mu;100)$, $n = 25$, $\ddot{x} = 1020$ e $\alpha = 0{,}05$, calcule a RC, erros da 2ª espécie e a função potência (supondo $\mu = 1010$, 1030, 990 e 980) para:

$$\begin{cases} H_o : \mu_o = 1000 \\ H_1 : \mu_1 > 1000 \end{cases}$$

Resolução:

$$P(\overline{X} \ge k) = 0{,}05 \Leftrightarrow P\left(Z \ge \frac{k-1000}{\frac{100}{\sqrt{25}}}\right) = 0{,}05 \Leftrightarrow P\left(Z < \frac{k-1000}{20}\right) = 0{,}95$$

$$\Rightarrow \frac{k-1000}{20} = 1{,}645 \Leftrightarrow k = 1032{,}9 \qquad \text{RC} \in [1032{,}9;+\infty[\quad 1020<1032{,}9 \text{ não se rejeita } H_o$$

Tabela 17.3 – Erros de 2ª espécie
e potência do teste

μ	$\beta(\mu)$	$\pi(\mu)$
1010	0,8749	0,1251
1030	0,5596	0,4404

7- Dadas duas populações de notas (que variam de 00,00 á 200,00 Valores) de estudantes de Engª Química e Engª Electrotécnica com exame de recorrência, cujo tamanho da população para cada amostra é de 81.

25	28	63	70	42	33	46	79	85		15	18	43	92	74	73	61	63	84
14	89	98	56	75	28	56	48	91		75	76	94	82	76	84	83	91	87
85	76	42	81	93	45	41	86	98		87	79	78	96	94	92	83	84	54
78	95	84	20	53	40	80	60	70		56	58	59	52	57	53	55	84	61
75	43	51	61	81	72	94	83	54		60	63	68	69	62	67	64	85	71
86	75	43	61	68	59	72	47	41		73	74	77	76	84	79	81	80	82
78	45	76	84	73	51	48	92	85		91	90	80	70	75	76	84	73	94
76	87	92	43	57	86	75	84	91		84	71	84	71	24	28	27	39	37
73	84	91	42	86	75	10	14	11		29	34	81	46	48	47	45	41	90

a) Retire das duas populações amostras aleatórias de tamanho 14.

Resolução:
Neste caso deve-se usar a tabela de números aleatórios para conseguir retirar as amostras. As entradas na tabela devem ser consideradas por linhas e não colunas

Amostras
Turma 1: 92, 45, 86, 25, 75, 73, 28, 89, 98, 84, 84, 91, 86, 86 – Variável X
Turma 2: 84, 92, 28, 15, 47, 75, 84, 76, 94, 70, 34, 81, 73, 84 – Variável Y

X	Y	$x = X_i - \overline{X}$	$\left(X_i - \overline{X}\right)^2$	XY	$y = Y_i - \overline{Y}$	$\left(Y_i - \overline{Y}\right)^2$
92	84	17,57	308,70	299,92	17,07	291,39
45	92	-29,43	866,12	-737,81	25,07	628,50
86	28	11,57	133,86	-450,42	-38,93	1515,54
25	15	-49,43	2443,32	2566,90	-51,93	2696,72
75	47	0,57	0,32	-11,36	-19,93	397,20
73	75	-1,43	2,04	-11,54	8,07	65,12
28	84	-46,43	2155,74	-792,56	17,07	291,39
89	76	14,57	212,28	132,15	9,07	82,27
98	94	23,57	555,54	638,04	27,07	732,79
84	70	9,57	91,59	29,38	3,07	9,42
84	34	9,57	91,59	-315,14	-32,93	1084,39
91	81	16,57	274,56	233,14	14,07	197,97
86	73	11,57	133,86	70,23	6,07	36,85
86	84	11,57	133,86	197,50	17,07	291,39
1024		-0,02	7403,38	1848,43	-0,02	8320,94

b) Diga qual a turma que possui maior variância em notas.

$$\overline{X} = \frac{\sum\left(X_i\right)}{n} = \frac{1042}{14} = 74{,}43 \qquad S^2 = \frac{\sum\left(X_i - \overline{X}\right)^2}{n-1} = \frac{7403{,}38}{13} = 569{,}49$$

$$\overline{Y} = \frac{\sum\left(Y_i\right)}{n} = \frac{937}{14} = 66{,}93 \qquad S^2 = \frac{\sum\left(Y_i - \overline{Y}\right)^2}{n-1} = 640{,}07$$

É visível que as notas da turma 2 possuem maior variância

c) As estatísticas anteriores para os analistas levam com que estes afirmem que os estudantes da Engª Química são mais inteligentes que os da Electrotecnia. Será que é verdade, admitindo um erro de 5%?

Resolução:

Sabe-se que o σ é desconhecido e $n < 30$, então usaremos T-Student

$\alpha = 0{,}05$. Calculadas as medias populacionais teremos: $\mu_1 = 64{,}54 \quad \mu_2 = 67{,}20$

$H_o : \mu_1 <= \mu_2$ Estudantes da Turma 1 não são mais inteligentes que os da turma 2

$H_1 : \mu_1 > \mu_2$ Estudantes da Turma 1 são mais inteligentes que os da turma 2

$$T_{cal} = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}} = \frac{74{,}43 - 66{,}93 - 64{,}54 + 67{,}20}{\sqrt{\left(\frac{1}{14} + \frac{1}{14}\right)\frac{13 \times 569{,}49 + 13 \times 640{,}07}{26}}} = \frac{10{,}16}{9{,}295} = 1{,}093$$

$t_{(1-\alpha;n-1)} = 1{,}77$

$T_{cal} \langle t_{(n-1)}$, não se rejeita H_o, o que significa que os estudantes da Engª Química não são mais inteligentes que os da Engª Electrotécnica

d) Se tomarmos X como notas da Engª Química e Y como notas da Electrotecnia. Pode-se dizer que o comportamento dos estudantes cujas notas são X dependem ou relacionam-se com fieldade aos Y por estarem na mesma Faculdade?

Resolução:

$$r_{xy} = \frac{\sum xy}{\sqrt{\left(\sum x^2\right)\left(\sum y^2\right)}} = \frac{Cov(x, y)}{\sqrt{\left(\sum x^2\right)\left(\sum y^2\right)}} = \frac{\sum \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\left(\sum x^2\right)\left(\sum y^2\right)}} = \frac{1848{,}43}{\sqrt{6936967{,}06}} = \frac{1848{,}43}{2633{,}81} = 0{,}7$$

Eles se relacionam, o que significa que o comportamento desses alunos por estarem na mesma Faculdade é próxima.

e) Determine a equação da recta de regressão, considerando a variável Y como dependente.

Resolução:

$$y = a_o + a_1 y \quad \text{ou} \quad y = \bar{y} + \frac{\sum xy}{\sum x^2} y = 66{,}93 + \frac{71588}{7403{,}38} x = 66{,}93 + 9{,}67x$$

8- A despesa semanal em alimentação de um agregado familiar pertencente a certa classe de rendimentos tem desvio padrão $\sigma = 170$. Crê-se que a despesa semanal média é de 1500; sendo 1420 a hipótese alternativa, e fixado o nível de significância $\alpha = 0{,}05$, com base numa amostra aleatória de tamanho n, obteve-se a probabilidade do erro tipo II de 0,10 aproximadamente. Determine o tamanho da amostra.

Resolução

X- Despesa semanal em alimentação por agregado familiar pertencente a certa classe de rendimento com $\sigma = 170$ H_o: $\mu = 1500$ H_1: $\mu = 1420$ $\alpha = 0{,}05$ $\beta = 0{,}10$

$$P(\overline{X} \in RC / H_o = verdadeiro) = \alpha \Rightarrow P(\overline{X} \le k) = 0{,}05 \Leftrightarrow P(\frac{\overline{X} - \mu_o}{\frac{\sigma}{\sqrt{n}}} \le -1{,}645) \Leftrightarrow P(\overline{X} \le \mu_o - 1{,}645\frac{\sigma}{\sqrt{n}})$$

$$0{,}10 = \beta = P(\overline{X} \in RA / H_1) = \alpha \Rightarrow P(\overline{X} > k) = 0{,}05 \Leftrightarrow P(\frac{\overline{X} - \mu_A}{\frac{\sigma}{\sqrt{n}}} > -1{,}645) \Leftrightarrow P(\overline{X} > \mu_A - 1{,}645\frac{\sigma}{\sqrt{n}})$$

do erro tipo 1 temos $\overline{X} = \mu_o - 1{,}645\frac{\sigma}{\sqrt{n}}$ o valor crítico, substituindo na equação do erro tipo II, teremos: $P(Z > \frac{\left(\mu_o - 1{,}645\frac{\sigma}{\sqrt{n}}\right) - \mu_A}{\frac{\sigma}{\sqrt{n}}}) = P(Z > \frac{\left(1500 - 1{,}645\frac{170}{\sqrt{n}}\right) - 1420}{\frac{170}{\sqrt{n}}})$ para

$P(Z > z) = 0{,}10 \Rightarrow z = 1{,}282$ *dai que* $\frac{\left(1500 - 1{,}645\frac{170}{\sqrt{n}}\right) - 1420}{\frac{170}{\sqrt{n}}} = 1{,}282 \Rightarrow n = 39$

9- Onde é que usamos amostragem sistemática?

Resposta:
Quando estamos no campo de trabalho

Exercícios Propostos

1- Suponha que o tempo que um operador leva para executar uma certa actividade seja normalmente distribuído, com o tempo médio de 12 minutos e desvio de padrão de 1,5 minutos. Se o operador está realizando esta actividade repetidamente, qual é probabilidade de que, em certo momento, ele leve entre 9 e 15 minutos para executar uma operação deste tipo? ***

2- Suponha que uma fábrica tenha estabelecido que a vida média dos pneus para automóveis de sua fabricação, é de 35.000 Km rodados, com um desvio padrão de 3.000 Km. Suponha ainda que o tempo de duração dos pneus seja uma variável aleatória normalmente distribuída. ***

a) Se a fábrica fornecer uma garantia de 30.000 Km, em condições normais de uso do veículo, qual a probabilidade de que um pneu vendido tenha de ser substituído?

b) Nas condições do item (a), qual a percentagem de pneus que terão de ser substituidos?

c) Que quilometragem a fábrica deve oferecer como garantia, para que nenhum pneu vendido tenha de ser substituido?

d) A fábrica está preocupada em melhorar a qualidade dos pneus e, para isso, está sendo estudada a possibilidade de se aumentar a duração média dos pneus. Desta forma, qual deveria ser a duração média para que, com uma garantia de 30.000 Km, somente 1% dos pneus vendidos tenham de ser trocados?

3- Um fabricante de refrigerantes vende um dos seus produtos engarrafados em vasilhames de 1 litro. Para engarrafar este produto usa-se uma máquina, que, calibrada, permite obter o volume desejado, segundo uma curva normal, com um desvio de 30 ml ***[10]

a) Se o órgão fiscalizador do governo (OFG) faz a exigência de que não mais de 8% de garrafas tenham um volume menor do que o nominal, em quanto deve ser regulada a máquina para que o fabricante não seja autuado?

b) Se a máquina for calibrada para colocar 1.035 ml de líquido no vasilhame, qual a percentagem de vasilhames que não estarão atendendo às especificações do OFG?

c) Para qual valor deve ser ajustada a precisão da máquina, para que, estando calibrada em 1.350 ml, as especificaçãoes do OFG seja atendidas?

4- Uma amostra aleatória de 10 pacotes de café foi selecionada do stock de um grande supermercado. Observou-se os seguintes pesos (em g): 497,5; 499,2; 500,3; 491,8; 502,7; 493,9; 497,4; 509,8; 503,2. Encontre estimativas para o peso médio e a variância. Considerando o peso dos pacotes como uma variável normalmente distribuida, obtenha também intervalos com 95% de confiança para os mesmos parâmetros estimados. ***

5- Foram observados os tempos de duração do intervalo para o "cafezinho", para uma amostra de 20 empregados de uma empresa, obtendo-se os seguintes resultados, em minutos: ***
15,79 15,75 18,11 14,54 10,06 17,32 18,52 16,11 13,59 18,63 16,27 13,75 15,16 14,75 13,03 18,47 12,14 14,67 16,52 12,47

Encontre a média e a variabilidade estimadas do tempo de duração do intervalo para o "cafezinho" dos funcionários da empresa. Encontre, ainda, intervalos de 90% de confiança para a média e a variância, supondo a variável tempo distribuída segundo uma Norma.

6- Suponha que a pressão sanguínea sistólica seja uma variável distribuída segundo uma norma. Foi observada a pressão de um grupo 16 pacientes de uma clínica, obtendo-se os seguintes resultados, em mm de Hg: ***
121,3 118,8 127,9 132,5 146,3 110,7 152,3 126,7
120,9 110,8 142,3 135,7 140,8 137,6 128,3 113,9

Estime e encontre um intervalo de 99,5% de confiança para a pressão sistólica média dos pacientes desta clínica.

---

[10] Todos exercícios com *** foram retirados do livro: Exercícios retirados do Reginaldo, C., at al (1999). ***Análise de Modelos de Regressão Linear***. pp. 22-23

7- Para o estudo do consumo médio de combustível para uma determinada marca de automóvel, foi observado o consumo de uma amostra de 20 destes veículos, obtendo-se uma média de 16,7 KM/l e um desvio padrão de 2,3Km/l. Construa um intervalo de 95% de confiança para o consumo médio de combustível, para este tipo de veículo. Suponha o consumo de combustível aproximadamente normal. ***

Bibliografia

Barroso, M., Sampaio, E., & Ramos, M. (2003). *Exercícios de Estatística Descritiva para as Ciências Sociais* (1ª ed.). Lisboa, Portugal: Edições Silabo.

Bussab, W. O. (2002). *Estatística Básica* (5ª ed.). São Paulo, Brasil: Saraiva.

Dagnelie, P. (1973). *Estatística. Teoria e Métodos II* (2º Vol. ed. 6036/3666). Lisboa, Portugal: Publicações Europa-América.

Gmurman, V.E. (1977). *Teoria das Probabilidades e Estatística Matemática* (tradução para português em 1983). Moscovo, Rússia: Mir Moscou.

Hill, M. M., Hill, A. (2002). *Investigação por Questionário*. Lisboa, Portugal: Edições Silabo.

Labrousse, C. (2002). *Probabilidades. Resumos Teóricos. Exercícios Resolvidos*. Porto, Portugal: Rés.

Lipschutz, S. (1993). *Probabilidade* (4ª ed. revisada). São Paulo, Brasil: Makron Books.

Meyer, P. L.(1995). *Probabilidade. Aplicações à Estatística* (2ª ed.). Rio de Janeiro, Brasil: Livros Técnicos e Científicos.

Murteira, B. J. F., Muller, D. A., Turkman, K. F. (1993). *Análise de Sucessões Cronológicas*. Lisboa, Portugal: McGraw-Hill.

Pereira, A. (2003). *SPSS. Guia Prático de Utilização. Análise de Dados para Ciências Sociais e Psicologia* (4ª ed.- revista e aumentada). Lisboa, Portugal: Edições Silabo.

Pestana, M. H., & Gageiro, J. N. (2003). *Análise de Dados para Ciências Sociais. A Complementaridade do SPSS* (3ª ed., revista e aumentada). Lisboa, Portugal: Edições Silabo.

Reinaldo, C., Freire, C. A. L., Charnet, E. M. R., Bonvino, H. (1999). *Análise de Modelos de Regressão Linear com Aplicações*. São Paulo, Brasil: Editora da Unicamp.

Robalo, A. (2001).*Estatística. Exercícios. Distribuições e Inferência Estatística* (Vol. II, 5ª ed., 2ª reimpressão). Lisboa, Portugal: Edições Silabo.

Siegel, S. (1975). *Estatística não Paramétrica (para ciências do comportamento)*. São Paulo, Brasil: McGraw-Hill.

S.N. (1998). SPSS Base 8.0. User´s Guide. Chicago, USA: SPSS.

Tourinho, L. C. (1962). *Probabilidades. Cadernos de Estatística e Economia Nº. 1*. Curitiba, Brasil: Diretório Acadêmico de Engenharia do Paraná.

Triola, M. F. (1999). *Introdução à Estatística* (7ª ed.). Rio de janeiro, Brasil: LTC.

APÊNDICE

reas sob a curva normal padrão.

(Para os valores negativos de *z* as áreas são obtidas por simetria)

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2703	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4965	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4983	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3,1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993
3,2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4994	0,4995	0,4995	0,4995
3,3	0,4995	0,4995	0,4995	0,4996	0,4996	0,4996	0,4996	0,4996	0,4996	0,4997
3,49	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998
3,6	0,4998	0,4998	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,9	0,500									

Docente: Rodrigues Zicai Fazenda z_fazenda@yahoo.com.br ou rzfazenda@ustm.ac.mz 84-4934100

Distribuição de t (Student)

gl\P	0,55	0,60	0,65	0,70	0,75	0,80	0,85	0,90	0,95	0,975	0,99	0,995	0,999
01	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657	636,619
02	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,598
03	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,541	12,924
04	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
05	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,869
06	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
07	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,365	3,499	5,408
08	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
09	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,256	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,726
26	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,256	0,389	0,531	0,684	0,856	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,127	0,256	0,389	0,530	0,683	0,856	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,646
40	0,126	0,255	0,388	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	3,551
60	0,126	0,254	0,387	0,527	0,679	0,848	1,046	1,296	1,671	2,000	2,390	2,660	3,460
120	0,126	0,254	0,386	0,526	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617	3,373
∞	0,126	0,253	0,385	0,524	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,291

Tabela de Contigência da Distribuição Qui - Quadrado χ^2

GL\P	0,01	0,05	0,10	0,20	0,30	0,50	0,70	0,80	0,90	0,95	0,98	0,99	0,999
0l	,0002	0,004	0,0l6	0,064	0,l48	0,455	l,074	l,642	2,706	3,84l	5,4l2	6,635	l0,827
02	0,020	0 l03	0,2ll	0,446	0,7l3	1,386	2,408	3,219	4,605	5,991	7,824	9,210	13,815
03	0,115	0,352	0,584	1,005	1,424	2,366	3,665	4,642	6,251	7,815	9,837	11,345	16,266
04	0,297	0,711	1,064	1,649	2,195	3,357	4,878	5,989	7,779	9,488	11,668	13,277	18,467
05	0,554	1,145	1,610	2,343	3,000	4,351	6,064	7,289	9,236	11,070	13,388	15,080	20,515
06	0,872	1,635	2,204	3,070	3,828	5,348	7,231	8,558	10,645	12,592	15,033	16,812	22,457
07	1,239	2,167	2,833	3,822	4,671	6,346	8,383	9,803	12,017	14,067	16,622	18,475	24,322
08	1,646	2,733	3,490	4,594	5,527	7,344	9,524	11,030	13,362	15,507	18,168	20,090	26,125
09	2,088	3,325	4,168	5,380	6,393	8,343	10,656	12,242	14,684	16,919	19,679	21,666	27,877
10	2,558	3,940	4,865	6,179	7,267	9,342	11,781	13,442	15,987	18,307	21,161	23,209	29,588
11	3,053	4,575	5,578	6,989	8,l48	l0,34l	l2,899	l4,63l	l7,275	l9,675	22,6l8	24,725	3l,264
l2	3,57l	5,226	6,304	7,807	9,034	ll,340	l4,0ll	l5,8l2	l8,549	2l,026	24,054	26,2l7	32,909
13	4,l07	5,892	7,042	8,634	9,926	l2,340	l5,ll9	l6,985	l9,8l2	22,362	25,472	27,688	34,528
14	4,660	6,571	7,790	9,467	l0,821	l3,339	l6,222	l8,l5l	21,064	23,685	26,873	29,l4l	36,123
15	5,229	7,26l	8,547	l0,307	ll,721	l4,339	l7,322	l9,311	22,307	24,996	28,259	30,578	37,697
16	5,812	7,692	9,312	ll,l52	l2,624	l5,338	l8,4l8	20,465	23,542	26,296	29,633	32,000	39,252
17	6,408	8,672	l0,085	l2,002	l3,531	l6,338	l9,511	21,615	24,769	27,587	30 995	33,409	40,790
18	7,0l5	9,390	l0,865	l2,857	l4,440	l7,338	20,60l	22,760	25,989	28,869	32,346	34,805	42,312
19	7,633	l0,ll7	ll,651	l3,7l6	l5,532	l8,338	21,689	23,900	27,204	30,l44	33,687	36,l9l	43,820
20	8,260	l0,85l	l2,443	l4,572	l6,266	l9,337	22,775	25,038	28,4l2	31,4l0	35,020	37,566	45,315

Binômio de Newton

Denomina-se *Binómio de Newton*, a todo binómio da forma $(a + b)^n$, sendo n um número natural .

Exemplo:

B = $(3x - 2y)^4$ (onde a = 3x, b = -2y e n = 4 [grau do binômio]).

Isaac Newton - Físico e Matemático Inglês(1642 - 1727) deu diversas contribuições à Matemática, que estão reunidas na monumental obra Principia Mathematica, escrita em 1687. Tomemos alguns Exemplos de desenvolvimento de binómios de Newton:

a) $(a+b)^2=a^2 +2ab+b^2$

b) $(a+b)^3=a^3+3a^2b+3ab^2+b^3$

Observe que o expoente do primeiro e último termos são iguais ao expoente do binômio, ou seja, igual a 3. A partir do segundo termo, os coeficientes podem ser obtidos a partir da seguinte regra prática de fácil memorização: Multiplicamos o coeficiente de a pelo seu expoente e dividimos o resultado pela ordem do termo. O resultado será o coeficiente do próximo termo. Assim por exemplo, para obter o coeficiente do segundo termo do item (b) acima teríamos: 1.3 =3; agora dividimos 3 pela ordem do termo anterior (1 por se tratar do primeiro termo) 3:1=3 que é o coeficiente do segundo termo procurado.

Observe que os expoentes da variável a decrescem de n até 0 e os expoentes de b crescem de 0 até n.

Também os coeficientes binomiais para o binómio de Newton podem ser obtidos apartir do desenvolvimento da seguinte tabela:

Triângulo de Pascal

$(a+b)^1$					1		1			
$(a+b)^2$				1		2		1		
$(a+b)^3$			1		3		3		1	
$(a+b)^4$		1		4		6		4		1
$(a+b)^5$	1		5		10		10		5		1
$(a+b)^6$		1		6		15		20		15		6		1
$(a+b)^7$			1		7		21		35		35		21		7		1
$(a+b)^8$		1		8		28		56		70		56		28		8		1
$(a+b)^9$	1		9		36		84		126		126		84		36		9		1