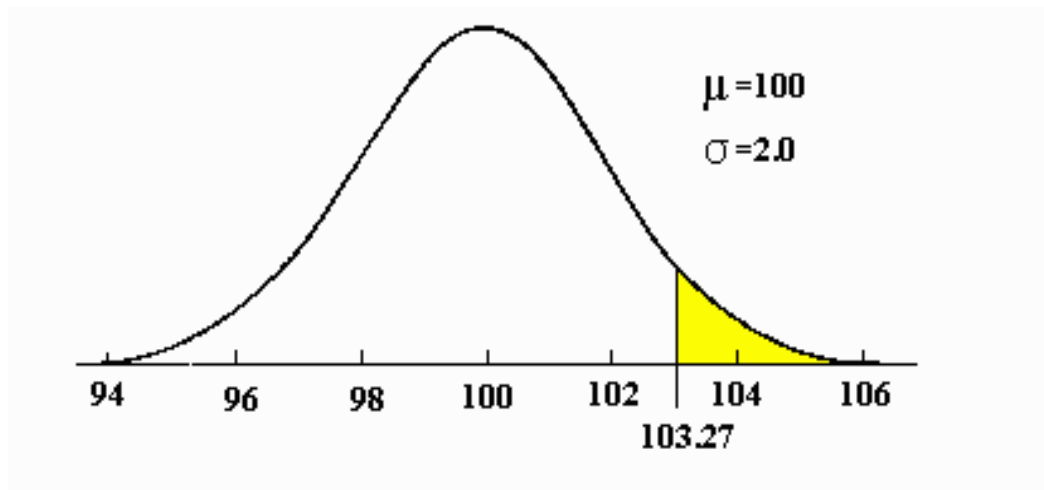


# ***CURSO DE ESTATÍSTICA APLICADA***



**Prof. Henrique Dantas Neder**  
**Instituto de Economia – Universidade Federal**  
**de Uberlândia.**

## **SUMÁRIO**

1. Introdução .....	4
2. Estatística Descritiva .....	8
2.1 Tipos de Variáveis .....	8
2.2 Tabelas e Distribuições de Frequência .....	10
2.3 Histogramas .....	12
2.4 Tabulação de Frequência e Histograma para Variáveis Contínuas .....	13
2.5 Medidas de Posição e de Dispersão .....	16
2.5.1 Uma Nota sobre Notação Estatística .....	17
2.5.2 A Média Aritmética Não Ponderada .....	18
2.5.3 A Média Aritmética Ponderada .....	19
2.5.4 Proporções como Médias .....	20
2.5.5 A Média Geométrica .....	21
2.5.6 A Média Harmônica .....	25
2.5.7 A Mediana .....	25
2.5.8 A Média para Dados Agrupados .....	27
2.5.9 A Mediana para dados Agrupados .....	28
2.5.10 A Moda para dados Agrupados .....	30
2.5.11 O Intervalo (ou amplitude) .....	37
2.5.13 Variância e Desvio Padrão .....	39
2.5.14 Variância e Desvio Padrão para Dados Agrupados .....	42
2.5.15 Interpretando e Aplicando o Desvio Padrão .....	43
2.5.16 Coeficiente de Variação .....	45
2.6 Medidas de Assimetria .....	46
2.7 Curtose: uma medida de achatamento .....	48
3. Probabilidade .....	50
3.1 Definição Clássica de Probabilidade .....	51
3.2 Conceito da Frequência Relativa .....	52
3.3 Probabilidade Subjetiva .....	53
3.4 Algumas Regras Básicas de Probabilidade .....	53
3.5 A Regra do Complemento .....	55
3.6 A Regra Geral da Adição .....	56
3.7 Regras de Multiplicação .....	58
3.8 Probabilidade Condicional .....	60
3.9 Diagramas em Árvore .....	62
3.10 Teorema de Bayes .....	64
Anexo 1 – Recordando Definições e Conceitos .....	65
Anexo 2 - Independência e Modelos de Árvore para Calcular Probabilidades .....	68
Anexo 3 - Probabilidade Condicional .....	74
Anexo 4 – Revisando os conceitos .....	77
Resumo do Cálculo de Probabilidades .....	96
Exercícios de Probabilidade .....	97
4. Variáveis Aleatórias Discretas .....	114

4.1 O Valor Esperado (média) de uma Distribuição de Probabilidade Discreta.....	118
4.2 A Variância e o Desvio Padrão de uma Distribuição de Probabilidade Discreta ..	119
4.3 A Distribuição de Probabilidade Binomial .....	121
4.4 A Média e Variância De Uma Distribuição Binomial .....	125
Apêndice 1 (Recordação).....	126
Apêndice 2 (Recordação).....	127
Apêndice 3 (Recordação).....	128
Apêndice 4 (Recordação) Valor Esperado e Variância de uma Variável Aleatória....	132
Variáveis Aleatórias Independentes.....	140
Apêndice 4 (recordação) .....	141
5. Variáveis Aleatórias Contínuas e Distribuição Normal .....	144
5.1 Variáveis Aleatórias Contínuas .....	144
5.2 Média e Variância de uma Variável Aleatória Contínua .....	146
5.3 Variável Aleatória Normal.....	165
5.4 Distribuição Normal Padrão.....	167
5.5 Áreas Abaixo da Curva Normal .....	168
6. Métodos de Amostragem e Distribuições Amostras .....	175
6.1 Amostragem Probabilística.....	180
6.2 Teorema do Limite Central.....	184
6.3 Estimativa de Ponto .....	186
6.4 Estimativa de Intervalo.....	186
6.5 Intervalo de Confiança para Uma Proporção Populacional .....	188
6.6 Fator de Correção de População Finita.....	189
6.7 Selecionando uma Amostra .....	190
6.8 Tamanho Amostral para Estimativa de Proporções.....	191
7. Teste de Hipóteses – Amostras Grandes .....	192
7.1 Testes de Significância Unicaudais.....	194
7.2 Testes de Significância Bicaudais.....	194
7.3 P-value de um Teste de Hipótese.....	196
7.4 Cálculo do p-value.....	196
7.5 Teste de Hipóteses: Duas Médias Populacionais .....	198
7.6 Testes Referentes à Proporção.....	201
EXERCÍCIOS :	204

# 1.Introdução

## **A Significância e a Abrangência da Estatística** **Porque a estatística é importante?**

Os métodos estatísticos são usados hoje em quase todos os campos de investigação científica, já que eles capacitam-nos a responder a um vasto número de questões, tais como as listadas abaixo:

- 1) Como os cientistas avaliam a validade de novas teorias?
- 2) Como os pesquisadores médicos testam a eficiência de novas drogas ?
- 3) Como os demógrafos prevêm o tamanho da população do mundo em qualquer tempo futuro?
- 4) Como pode um economista verificar se a mudança atual no Índice de Preços ao Consumidor é a continuação de uma tendência secular, ou simplesmente um desvio aleatório?
- 5) Como é possível para alguém prever o resultado de uma eleição entrevistando apenas algumas centenas de eleitores ?

Estes são poucos exemplos nos quais a aplicação da estatística é necessária. Podemos presumir que a matemática é uma das rainhas das ciências porque ela fornece a estrutura teórica para quase todas as outras ciências. Se você já fez um curso básico de física, já está familiarizado com algumas das leis matemáticas que governam temas tão diversificados como gravidade, energia, luz, eletricidade, etc. Mas também devemos considerar o fato de que as teorias matemáticas estão sendo desenvolvidas todos os dias em muitas áreas por estatísticos teóricos - pessoas treinadas em teoria estatística e probabilidade. Para citar alguns poucos casos ilustrativos elas são desenvolvidas para teoria dos vãos espaciais em física; para teorias do conhecimento do comportamento animal e humano em psicologia; para teorias da migração e dos diferenciais de raça em sociologia; para teorias de epidemias em saúde pública;...

De fato, a estatística tornou-se uma ferramenta cotidiana para todos os tipos de profissionais que entram em contato com dados quantitativos ou tiram conclusões a partir destes.

### **O que é Estatística?**

A noção de “Estatística” foi originalmente derivada da mesma raiz da palavra “Estado”, já que foi a função tradicional de governos centrais no sentido de armazenar registros da população, nascimentos e mortes, produção das lavouras, taxas e muitas outras espécies de informação e atividades. A contagem e mensuração dessas quantidades gera todos os tipos de dados numéricos que são úteis para o desenvolvimento de muitos tipos de funções governamentais e formulação de políticas públicas.

Dados numéricos são de fato uma parte da Estatística, mas são apenas a matéria-prima, que precisa ser transformada pelos “métodos estatísticos” para posterior análise. A Estatística, como um método científico, refere-se ao projeto de experimentos e a descrição e interpretação de observações que são feitas. De um ponto de vista moderno, a Estatística é freqüentemente definida como um método de tomada de decisão em face da aleatoriedade dos fenômenos. Em uma mais vasta perspectiva, o escopo da estatística pode ser pensado em termos de três áreas diferentes de estudos: (1) a Estatística Descritiva (2) A Estatística Indutiva e (3) A Teoria da Decisão Estatística.

### **Estatística Descritiva**

A estatística Descritiva refere-se ao corpo de métodos desenvolvidos para coletar, organizar, apresentar e descrever dados numéricos. Essa área da Estatística refere-se às seguintes tarefas:

- 1) Encontrar um método apropriado de coletar dados numéricos eficientemente e acuradamente para um dado problema.
- 2) Determinar um formato eficiente, tal como uma apresentação tabular, para a organização dos dados de uma forma sistemática e ordenada, de maneira que a

informação fornecida pelos dados possa ser observada com grande facilidade e precisão.

- 3) Apresentar dados numéricos, seja organizados ou não, de forma que as características e o comportamento dos dados são clara e facilmente revelados. Tais apresentações São feitas por meio de métodos gráficos.
- 4) Sumarizar ou descrever cada característica ou propriedade dos dados por um simples número, tal como uma média, uma porcentagem ou alguma outra medida apropriada, a qual é calculada a partir dos dados por meio de uma fórmula derivada a partir de algum princípio válido.

### **Estatística Indutiva**

A Estatística Indutiva, que é também freqüentemente chamada de inferência estatística ou estatística inferencial, em contraste com a estatística descritiva, é essencialmente analítica em sua natureza. Consiste de um conjunto de princípios ou teoremas que nos permitem generalizar acerca de alguma característica de uma “população” a partir das características observadas de uma “amostra”. Nessa definição, uma *população* é o conjunto de todos os itens, objetos, coisas ou pessoas a respeito das quais a informação é desejada para a solução de um problema. Uma *amostra* é um grupo de itens selecionados por um método cuidadosamente concebido e projetado a partir de uma população. Existem diferentes tipos de amostras, dependendo dos diferentes métodos de seleção disponíveis. Uma amostra aleatória simples, falando em termos simplificados, é aquela que é selecionada de tal forma que cada e todos os itens na população têm a mesma chance de serem incluídos na amostra.

Se uma medida descritiva é calculada a partir dos dados da população ela é chamada de *parâmetro populacional*, ou simplesmente *parâmetro*; se é calculada a partir dos dados da amostra ela é chamada de *estatística amostral*, ou simplesmente *estatística*. Considerando esses conceitos podemos definir *estatística indutiva* como o processo de generalizar acerca de do valor de um parâmetro a partir do valor de uma estatística. Existem dois procedimentos de inferência distintos mas relacionados: estimação e teste de hipóteses. *Estimação* é processo de usar o valor de uma estatística amostral para

estimar o valor de um parâmetro que é desconhecido, mas é uma constante. Como um exemplo, suponhamos que temos uma população de 100.000 bolas de gude em um saco, todas as quais são idênticas exceto pela cor, e que não podemos vê-las embora saibamos que uma parte delas são brancas e o restante são pretas. Suponha que desejamos ter uma idéia da proporção de, digamos, bolas brancas nessa população. Suponha que para conseguir isso selecionamos 1.000 bolas aleatoriamente do saco e verificamos que 350 são brancas. Isso significa que nossa proporção amostral de bolas brancas é 35 %. A partir disso concluímos que a proporção populacional de bolas brancas é também 35 %. Fazendo isso nós realizamos o que é chamado de *estatística pontual*.

Mas afirmar que a proporção de bolas brancas em toda a população é exatamente igual a proporção daquela amostra particular é como dar um tiro no escuro: o valor da proporção amostral é um resultado aleatório e depende de cada amostra de 1.000 bolas escolhida da população. Pode ser que por uma enorme casualidade o resultado daquela amostra que escolhemos coincida exatamente com o valor da proporção de bolas brancas em toda a população. Mas as chances de que isso não ocorra são muito grandes. Uma forma de contornarmos esse problema é afirmarmos que as chances são de 95 em 100 (ou de 95 %) de que o intervalo formado pela proporção amostral acrescida e diminuída de 3 pontos percentuais contenha o verdadeiro valor da proporção populacional desconhecido. Ou seja, construímos um intervalo com limites  $35 + 0,03 \times 35 = 36,05$  e  $35 - 0,03 \times 35 = 33,95$  e afirmamos (com base em algum princípio obtido a partir da teoria estatística) que as chances são de 95 em 100 de que o verdadeiro valor da proporção populacional esteja localizado dentro desse intervalo. Quando uma afirmativa dessa natureza é feita estamos realizando o que se chama de *estimativa por intervalo*.

Quanto ao segundo procedimento da estatística inferencial deixaremos para comentá-lo quando for abordado em sua íntegra. E o terceiro campo de estudos da Estatística, a Teoria da Decisão Estatística não será discutido nessa apresentação.

## 2. Estatística Descritiva

### 2.1 Tipos de Variáveis

Existem diversos tipos de variáveis que serão utilizadas em um estudo estatístico. É importante compreender o conceito matemático de variável. Variável é uma abstração que se refere a um determinado aspecto do fenômeno que está sendo estudado. Podemos afirmar que a quantidade colhida da safra anual de soja é uma variável. Representemos essa variável pela letra  $X$ . Essa variável pode assumir diversos valores específicos, dependendo do anos de safra, por exemplo,  $X_{1986}$ ,  $X_{1990}$  e  $X_{1992}$ . Esses valores que a variável assume em determinados anos não são a própria variável, mas valores assumidos ela para determinados objetos ou pessoas da amostra ou da população. Se uma amostra tiver 50 indivíduos podemos referimo-nos a  $X$  como sendo a variável nota de estatística e a  $X_{30}$  como a nota de um indivíduo particular, no caso o trigésimo. É freqüente também na literatura utilizar-se letras maiúsculas para a notação de variáveis e as correspondentes letras minúsculas para referência aos valores particulares assumidos por essa variável mas nesse resumo procuraremos evitar essa forma de notação.

**Variáveis quantitativas** - referem-se a quantidades e podem ser medidas em uma escala numérica. Exemplos: idade de pessoas, preço de produtos, peso de recém nascidos.

As variáveis quantitativas subdividem-se em dois grupos: variáveis quantitativas discretas e variáveis quantitativas contínuas. Variáveis discretas são aquelas que assumem apenas determinados valores tais como 0,1,2,3,4,5,6 dando saltos de descontinuidade entre seus valores. Normalmente referem-se a contagens. Por exemplo: número de vendas diárias em uma empresa, número de pessoas por família, quantidade de doentes por hospital.<sup>1</sup> As

---

<sup>1</sup> Uma variável quantitativa discreta não precisa assumir necessariamente apenas valores de contagem, ou seja números inteiros ou números naturais em seqüência. Um exemplo de variável quantitativa discreta seria, por exemplo, uma que assumisse apenas os seguintes valores : { 1; 3,5 ;



variáveis quantitativas contínuas são aquelas cujos valores assumem uma faixa contínua e não apresentam saltos de descontinuidade. Exemplos dessas variáveis são o peso de pessoas, a renda familiar, o consumo mensal de energia elétrica, o preço de um produto agrícola.<sup>2</sup> As variáveis quantitativas contínuas referem-se ao conjunto dos números reais ou a um de seus subconjuntos contínuos.

**Variáveis Qualitativas** - referem-se a dados não numéricos.<sup>3</sup> Exemplos dessas variáveis são o sexo das pessoas, a cor, o grau de instrução.

As variáveis qualitativas subdividem-se também em dois grupos: as variáveis qualitativas ordinais e as variáveis qualitativas nominais. As variáveis qualitativas ordinais são aquelas que definem um ordenamento ou uma hierarquia. Exemplos são o grau de instrução, a classificação de um estudante no curso de estatística, as posições das 100 empresas mais lucrativas, etc. As variáveis qualitativas nominais por sua vez não definem qualquer ordenamento ou hierarquia. São exemplos destas a cor, o sexo, o local de nascimento, etc.<sup>4</sup>

Dependendo da situação uma variável qualitativa pode ser representada (codificada) através de emprego de números (por exemplo: em sexo representamos homens como sendo “0” e mulheres como sendo “1”). Mas no tratamento estatístico

---

5,75 ; 10 }. Apesar dessa variável abranger valores não inteiros ela apresenta saltos de descontinuidade: nesse exemplo ela não pode assumir nenhum valor intermediário entre 1 e 3,5 ou entre 5,75 e 10.

<sup>2</sup> Seria impossível obter na prática uma variável perfeitamente contínua já que os instrumentos de medida não tem precisão infinita. Por exemplo., o peso de pessoas é medido com uma balança com precisão, digamos, de décimos de gramas. Então jamais conseguiremos obter um valor para essa variável que se localize entre 50.000,1 e 50.000,2 gramas, por exemplo, 50.000,15 gramas. Ocorre portanto um salto de descontinuidade entre os dois valores possíveis de serem medidos e a variável, do ponto de vista teórico, não pode ser considerada como variável quantitativa contínua, mas variável quantitativa discreta. Mas do ponto de vista prático, acabamos freqüentemente por considerá-la e tratá-la como sendo uma variável quantitativa contínua, apesar dessa falta de precisão absoluta. O mesmo podemos dizer para o caso da renda ou qualquer outra variável econômica medida em unidades monetária: não existe uma renda de por exemplo R\$ 200,345 já que o centavo é a menor divisão do sistema monetário. Mas de qualquer forma, costuma-se tratar a renda como variável quantitativa contínua e não discreta.

<sup>3</sup> É muito comum considerar-se que a estatística apenas abrange os estudos que utilizam as variáveis quantitativas. Nada mais equivocado. Existe um vasto campo de aplicações estatísticas em que são empregadas as variáveis qualitativas, tanto isoladamente como em conjunto com variáveis quantitativas.

<sup>4</sup> Não podemos dizer que a cor X é superior a cor Y mas podemos afirmar que o terceiro ano do segundo grau é superior hierarquicamente ao primeiro ano do primeiro grau.

dessa variável codificada não podemos considerá-la como sendo quantitativa. Ela continua sendo uma variável qualitativa (pois o é em sua essência e natureza) apesar de sua codificação numérica que tem como finalidade uma maior facilidade de tabulação de resultados.

Não podemos dizer que para qualquer uma destas categorias qualquer método estatístico pode ser adequadamente aplicado. As variáveis quantitativas contínuas são aquelas que permitem a utilização de um conjunto maior e superior de métodos estatísticos e são, sem dúvida, as variáveis mais passíveis de um rico tratamento estatístico. Em seguida vêm, nessa ordem, as variáveis quantitativas discretas, as variáveis qualitativas ordinais e por último, as variáveis qualitativas nominais. Essas últimas são as que permitem a utilização de um menor e menos poderoso arsenal de instrumentos estatísticos de análise.

## **2.2 Tabelas e Distribuições de Frequência**

A análise estatística se inicia quando um conjunto 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 contem 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. Os dados precisam então ser organizados e apresentados em uma forma sistemática e seqüencial 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.

Suponhamos o seguinte conjunto de dados:

141213111213

161414151714

111314151312

141314131516

1212

Para montarmos uma distribuição de freqüências desses dados verificamos quais são os valores não repetidos que existem e em uma primeira coluna de uma tabela colocamos esses valores e na segunda coluna colocamos o número de repetições de cada um desses valores. Para o exemplo acima, a distribuição de freqüências será:

Variável	freqüência
11	2
12	5
13	6
14	7
15	3
16	2
17	1

A freqüência de uma observação é o número de repetições dessa observação no conjunto de observações. A distribuição de freqüência é uma função formada por pares de valores sendo que o primeiro é o valor da observação (ou valor da variável) e o segundo é o número de repetições desse valor.

### **Freqüências Relativas e Acumuladas**

Para o exemplo acima também podemos calcular a freqüência relativa referente a cada valor observado da variável. A freqüência relativa é o valor da freqüência absoluta dividido pelo número total de observações.

Variável	freqüência absoluta	freqüência relativa
11	2	$2/26 = 0,0769$
12	5	$5/26 = 0,1923$
13	6	$6/26 = 0,2308$
14	7	$7/26 = 0,2692$
15	3	$3/26 = 0,1154$

16	2	$2/26 = 0,0769$
17	1	$1/26 = 0,0385$
TOTAL	26	1,0000

Podemos também calcular as frequências acumuladas. Nesse caso existem as frequências absolutas acumuladas e as frequências relativas acumuladas.<sup>5</sup>

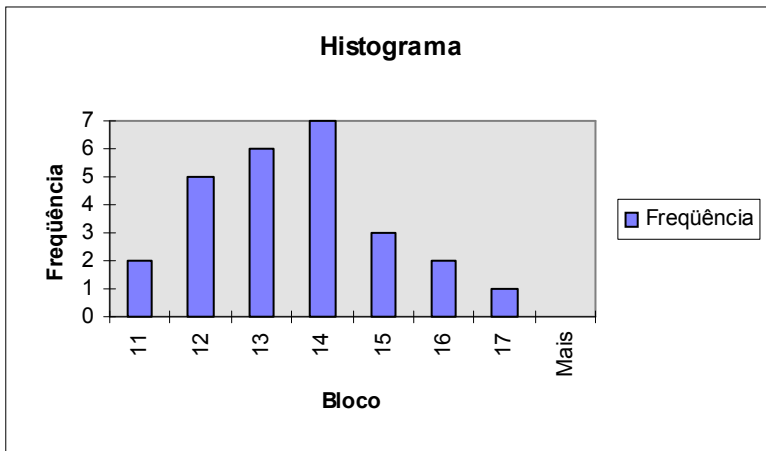
Variável	frequência absoluta	frequência relativa	frequência absoluta acumulada	frequência relativa acumulada
11	2	$2/26 = 0,0769$	2	$2/26 = 0,0769$
12	5	$5/26 = 0,1923$	7	$7/26 = 0,2692$
13	6	$6/26 = 0,2308$	13	$13/26 = 0,5000$
14	7	$7/26 = 0,2692$	20	$20/26 = 0,7692$
15	3	$3/26 = 0,1154$	23	$23/26 = 0,8846$
16	2	$2/26 = 0,0769$	25	$25/26 = 0,9615$
17	1	$1/26 = 0,0385$	26	$26/26 = 1,0000$
TOTAL	26	1,0000		

## 2.3 Histogramas

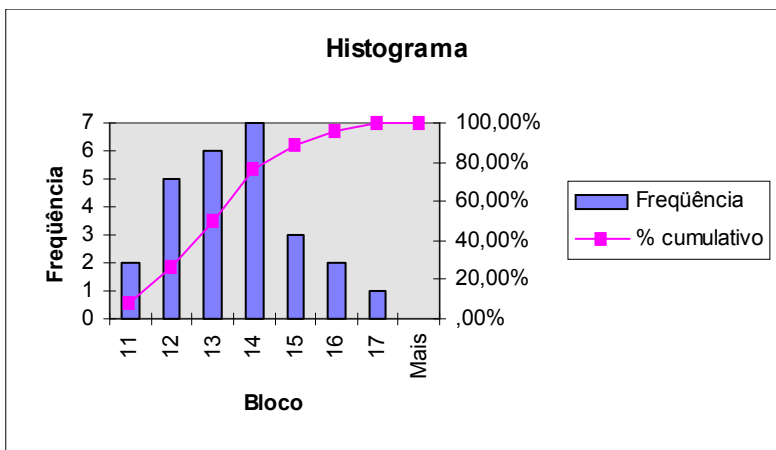
Histograma é uma representação gráfica de uma tabela de distribuição de frequências. Desenhamos um par de eixos cartesianos e no eixo horizontal (abscissas) colocamos os valores da variável em estudo e no eixo vertical (ordenadas) colocamos os valores das frequências. O histograma tanto pode ser representado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

---

<sup>5</sup> Observe que os valores da última coluna (frequência relativa acumulada) podem ser calculados de duas maneiras. Na primeira, tal como é feito na tabela a seguir, dividimos 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 acúmulos de erros, de forma que o último valor de frequência relativa acumulado se distancie consideravelmente de 1.



histograma de frequência acumulada (ou ogiva) é a representação gráfica do comportamento da frequência acumulada. Na figura abaixo a ogiva é mostrada em sobreposição ao histograma.



## 2.4 Tabulação de Frequência e Histograma para Variáveis Contínuas

Até agora vimos como são calculadas as frequências (relativas e acumuladas) para variáveis quantitativas discretas. Nesse caso a tabulação dos resultados é mais simples. Mas quando tratamos de variáveis quantitativas contínuas os valores observados devem ser tabulados em intervalos de 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. Suponhamos que as safras agrícolas de um determinado produto, em uma determinada região, sejam dadas pela tabela a seguir:

Ano	Safra (1000 t)	Ano	Safra (1000 t)
1	280	10	365
2	305	11	280
3	320	12	375
4	330	13	380
5	310	14	400
6	340	15	371
7	310	16	390
8	340	17	400
9	369	18	370

Devem ser seguidos alguns passos para a tabulação de frequências de dados que se referem a uma variável quantitativa contínua, como é o caso de nosso exemplo.

- 1. Definir o número de classes.** O número de classes não deve ser muito baixo nem muito alto. Um número de classes pequeno gera amplitudes de classes grandes o que pode causar distorções na visualização do histograma. Um número de classes grande gera amplitude de classes muito reduzidas. Foram definidas regras práticas para a determinação do número de classes, sendo que este deve variar entre 5 e 20 (5 para um número muito reduzido de observações e 20 para um número muito elevado). Se **n** representa o número de observações (na amostra ou na população, conforme for o caso) o número aproximado de classes pode ser calculado por Número de Classes =  $\sqrt{n}$  arredondando os resultados. No caso do exemplo anterior temos  $n = 18$  e  $\sqrt{18} = 4,24$  e podemos adotar um número de 5 classes, que será razoável.
- 2. Calcular a amplitude das classes.** Essa será obtida conhecendo-se o número de classes e amplitude total dos dados. A amplitude total dos dados é o resultado da subtração valor máximo - valor mínimo da série de dados. A amplitude de classe será:

$$\text{Amplitude de classe} = \frac{\text{Valor Máximo} - \text{Valor Mínimo}}{\text{número de classes}}$$

Em geral, o valor do resultado é também arredondado para um número inteiro mais adequado. No nosso exemplo temos:

$$\text{Amplitude de Classe} = \frac{430 - 280}{5} = 30$$

**3. Preparar a tabela de seleção com os limites de cada classe.** Na tabela abaixo apresentamos para os dados do nosso exemplo os limites inferior e superior de cada uma das 5 classes de frequência.

Classe	Limite inferior	Limite Superior
1	280	310
2	310	340
3	340	370
4	370	400
5	400	430

Observa-se na tabela acima que o limite superior de cada classe coincide com o limite inferior da classe seguinte. Prevendo-se que pode ocorrer que o valor de uma observação seja exatamente igual ao valor do limite de classe deve-se estabelecer um critério de inclusão. Para evitar esse tipo de dificuldade normalmente se estabelece que o limite superior de cada classe é aberto (e conseqüentemente, o limite inferior de cada classe é fechado), ou seja, cada intervalo de classe não inclui o valor de seu limite superior, com exceção da última classe.

**4. Tabular os dados por classe de frequência.** A partir da listagem de dados seleciona-se para cada um deles qual é a sua classe de frequência e acumula-se o total de frequência de cada classe. De acordo com nosso exemplo, teremos:

<b>Classe</b>	<b>Frequência Absoluta Simples</b>	<b>Frequência Relativa Simples</b>
280 - 310	3	0,12 (12 %)
310 - 340	4	0,16 (16 %)
340 - 370	6	0,24 (24 %)
370 - 400	7	0,28 (28 %)
400 - 430	5	0,20 (20%)
Total	25	1,00 (100 %)

Veremos adiante, quando discutirmos as medidas de posição e de dispersão, que quando agrupamos dados numéricos em intervalos de classe ocorre perda de informação o que leva a resultados não tão precisos do que aqueles que seriam obtidos a partir dos dados originais sem agrupamento.

## **2.5 Medidas de Posição e de Dispersão**

Podemos considerar que a Estatística Descritiva subdivide-se em duas partes. Na primeira, abordada anteriormente, são estudadas as formas de apresentação dos dados para que fiquem salientadas as suas características principais. Na segunda, que começaremos a tratar agora, abrange as medidas descritivas na forma de simples números que representam de forma sintética essas características da distribuição estatística dos dados. Estudaremos, a rigor, quatro tipos de medidas:

1. Medidas de Tendência Central (ou medidas de posição). Essa propriedade dos dados refere-se a localização do centro de uma distribuição. Elas nos indicam qual é a localização dos dados ( no eixo que representa o conjunto dos números inteiros se estivermos tratando de uma variável quantitativa contínua).
2. Medidas de Dispersão. Essa propriedade revela o grau de variação dos valores individuais em torno do ponto central.



3. Assimetria. É a propriedade que indica a tendência de maior concentração dos dados em relação ao ponto central.
4. Curtose. É a característica que se refere ao grau de achatamento, ou a taxa na qual a distribuição cresce ou cai da direita para a esquerda.

### 2.5.1 Uma Nota sobre Notação Estatística

Utilizaremos as letras maiúsculas para representar as variáveis, como por exemplo a variável **X**. Os valores individuais que uma variável pode assumir são representados pelas correspondentes letras minúsculas. Por exemplo, se **X** é usado para designar o peso de uma amostra de 50 pessoas, então **x** é o valor numérico do peso de uma dessas 50 pessoas. Diferentes valores de uma variável são identificados por subscritos. Assim, os pesos de 50 pessoas em uma amostra podem ser denotados por **x<sub>1</sub>, x<sub>2</sub>, ..., x<sub>50</sub>**.

- número total de observações em uma população finita é designado por **N** e na amostra é representado por **n**. A distinção entre medidas descritivas para populações e amostras é muito importante. Denotaremos os parâmetros (medidas referentes a população) por letras gregas ou letras minúsculas em português. As estatísticas amostrais serão representadas por letras maiúsculas em português e os valores observados de uma estatística amostral pela correspondente letra minúscula em português. Por exemplo, as medidas descritivas a serem introduzidas nessa seção serão denotadas como segue:

Nome da Medida	Parâmetro	Notação da Estatística	Valor observado
média aritmética	$\mu$	$\bar{X}$	$\bar{x}$
proporção	$\pi$	$\bar{P}$	$\bar{p}$
média geométrica	$\tilde{g}$	G	g
média harmônica	$\tilde{h}$	H	h
mediana	$\tilde{x}_{.5}$	$X_{.5}$	$x_{.5}$
moda	$\tilde{x}_m$	$X_m$	$x_m$

### 2.5.2 A Média Aritmética Não Ponderada

A média é definida como a soma das observações dividida pelo número de observações. Se tivermos, por exemplo,  $n$  valores, temos:

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

Propriedades da média aritmética não ponderada:

1. A média é um valor típico, ou seja, ela é o centro de gravidade da distribuição, um ponto de equilíbrio. Seu valor pode ser substituído pelo valor de cada item na série de dados sem mudar o total. Simbolicamente temos:

$$n(\bar{X}) = \sum x \quad (6)$$

2. A soma dos desvios das observações em relação a média é igual a zero.

$$\sum (x - \bar{X}) = 0$$

3. A soma dos desvios elevados ao quadrado das observações em relação a média é menor que qualquer soma de quadrados de desvios em relação a qualquer outro número. Em outras palavras,

$$\sum (x - \bar{X})^2 = \text{é um mínimo.}$$

A idéia básica de selecionar um número tal que a soma dos quadrados dos desvios em relação a este número é minimizada tem grande importância na teoria estatística. Ela chega a ter um nome especial: o “*princípio dos mínimos quadrados*”. Ela é, por exemplo, a base racional do *método dos mínimos quadrados* que é usado para ajustar a melhor

---

<sup>6</sup> - Utilizaremos muito freqüentemente a notação  $\sum x$  simplificadamente para representar  $\sum_{i=1}^n x_i$ .

curva através de um conjunto de pontos em um sistema de eixos cartesianos, como veremos adiante. Esta propriedade é também a base para o cálculo de uma importante medida de dispersão, que veremos logo a seguir.

A validade dessas três propriedades pode ser facilmente demonstrada por um exemplo numérico simples, mostrado na tabela a seguir. Nesta tabela, a coluna (1) contém o conjunto de dados cuja soma é 9 e cuja média é 3. A coluna (2) demonstra a primeira propriedade da média, ou seja, se cada uma das observações individuais dos dados é substituída pela média, a soma permanece igual a 9. A coluna (3) verifica que de fato  $\sum (x - \bar{X}) = 0$ . Finalmente, as colunas (4), (5) e (6) demonstram que  $\sum (x - \bar{X})^2 = 14$ , que é menor que somas quando os desvios individuais são tomados a partir do número 2 e do número 5, respectivamente.

(1) x	(2) $\bar{x}$	(3) $(x - \bar{x})$	(4) $(x - \bar{x})^2$	(5) $(x - 2)^2$	(6) $(x - 5)^2$
1	3	-2	4	1	16
2	3	-1	1	0	9
6	3	+3	9	16	1
Soma 9	9	0	14	17	26

### 2.5.3 A Média Aritmética Ponderada

No cálculo da média aritmética não ponderada todos os valores observados foram somados atribuindo-se o mesmo peso a todas as observações. Agora veremos uma nova forma de calcular a média. Consideremos um exemplo familiar de cálculo da média de notas de estudantes, quando o exame final vale duas vezes mais do que as duas provas comuns realizadas no decorrer do semestre. Se um determinado aluno obtiver as notas 7, 5 e 8 a sua média ponderada final será:

$$\frac{1 \times (7) + 1 \times (5) + 2 \times 8}{1 + 1 + 2} = 7$$

Em termos gerais, a fórmula para a média aritmética ponderada é:

$$\bar{X}_w = \sum_{i=1}^n w_i \times x_i = \sum wx$$

onde  $w_i$  é o peso da observação  $i$

e  $n$  é o número de observações.

A soma dos pesos não pode ser igual a zero. Fora disto, não existe restrição para os valores dos pesos. Se todos os pesos forem iguais a 1, a média ponderada recai em seu caso particular, a média aritmética não ponderada. O mesmo ocorre se todos os pesos forem iguais a uma constante  $c$ . Portanto, a média aritmética não ponderada na realidade é uma média aritmética ponderada com pesos iguais.

#### 2.5.4 Proporções como Médias

Freqüentemente encontramos populações cujas unidades elementares podem ser classificadas em duas categorias: uma que tem certo atributo e outra que não tem esse atributo. Nesse caso, estamos interessados na proporção de casos que possuem esse atributo. Uma proporção comumente é pensada como uma fração ou porcentagem, mas também pode ser pensada como um caso especial de média.

Suponha que queremos determinar a proporção de votantes entre os cidadãos brasileiros. Devemos primeiro designar um valor 1 para cada pessoa qualificada como eleitor e um valor 0 para cada pessoa não qualificada como eleitor. Então, a soma dos 1's seria  $\sum x$  e a média seria a média seria obtida pela divisão da soma pelo número  $N$  total de pessoas no Brasil.

A média da variável  $x$  é  $\mu = \sum x / N$ . No entanto essa média é também uma proporção, a proporção de eleitores na população brasileira.

### 2.5.5 A Média Geométrica

A média geométrica de uma amostra é definida como a raiz enésima do produto nos **n** valores amostrais.

$$G = \sqrt[n]{(x_1)(x_2)\dots(x_n)}$$

Por exemplo, a média geométrica de 5, 9 e 13 é:

$$G = \sqrt[3]{(5)(9)(13)} = 8,36$$

Para a mesma série de dados a média é 9. É sempre verdade que a média aritmética é maior do que a média geométrica para qualquer série de valores positivos, com exceção do caso em que os valores da série são todos iguais, quando as duas médias coincidem.

- cálculo da média geométrica é muito simples. Mas a sua interpretação e as suas propriedades tornam-se mais evidentes quando reduzimos a fórmula a sua forma logarítmica. Tomando logaritmos de ambos os lados da equação anterior teremos:

$$\log G = \log(\sqrt[n]{(x_1)(x_2)\dots(x_n)}) = \frac{\log x_1 + \log x_2 + \dots + \log x_n}{n} = \frac{\sum \log x}{n}$$

A conclusão que chegamos é que o logaritmo da média geométrica é igual à média aritmética dos logaritmos dos valores da série. Verifica-se que a média geométrica somente tem significado quando todos os valores da série são todos positivos.

Suponhamos como exemplo de aplicação de cálculo da média geométrica os dados da tabela seguinte que mostram as mudanças de preços de duas mercadorias, **A** e **B**, de 1980 a 1985. Durante esse período o preço de **A** subiu 100 % e o preço de **B** decresceu 50 %. Qual foi a mudança média relativa de preços? Em outras palavras, qual foi o percentual médio de mudança de preços?

Preços das Mercadorias **A** e **B** em 1980 e 1985

Preço		Relativo de Preços				
		1980 = 100		1985 = 100		
Mercadoria	1980	1985	1980	1985	1980	1985
<b>A</b>	R\$ 50	R\$ 100	100	200	50	100
<b>B</b>	R\$ 20	R\$ 10	100	50	200	100
Média Aritmética			100	125	125	100
Média Geométrica			100	100	100	100

A média aritmética fornece uma resposta incorreta para essa questão. Como indicado pelos cálculos da tabela acima leva a duas conclusões opostas. Se 1980 é tomado como base para o relativo de preços, os preços são em média 25 % maiores em 1985 do que em 1980. Se 1985 é tomado como base, os preços de 1980 são 25 % maiores do que os preços de 1985. Portanto, a média aritmética dos relativos de preços conduz a resultados inconsistentes.

No entanto, um resultado consistente é obtido quando a média geométrica é aplicada:

1. Se 1980 é escolhido como a base, os preços de 1985 são 100 % dos preços de 1980, ou seja:

$$g = \sqrt{200 \times 50} = \sqrt{10.000} = 100$$

2. Se 1985 é escolhido como a base, os preços de 1980 serão também 100 % dos preços em 1985, ou seja:

$$g = \sqrt{50 \times 200} = 100$$

A mais importante aplicação da média geométrica refere-se talvez ao cálculo de taxas de crescimento médias, desde que essas podem ser corretamente medidas somente por esse método. Para exemplificar, no campo da economia, esse ponto, suponha que a produção anual de um setor industrial cresceu de 10.000 para 17.280 unidades durante o período 1985-1988 como mostrado na tabela a seguir; qual é a taxa média de crescimento anual? A taxa média anual de crescimento pode ser calculada a partir dos valores em

porcentagem da produção em relação aos anos anteriores. Se calcularmos a média aritmética desses valores teríamos:

$$\bar{x} = (60 + 96 + 300) / 3 = 152$$

implicando uma taxa de crescimento média de  $152 - 100 = 52 \%$ . Se a produção cresce  $52 \%$  ao ano, começando da produção de 1985 de 10.000 unidades, então a produção de 1986 seria de

$$23.0 + 0,52(10.000) = 15.200;$$

a produção de 1987 seria de

$$15.200 + 0,52(15.200) = 23.104;$$

a produção de 1988 seria de

$$23.104 + 0,52(23.104) = 35.118,08$$

Ano	1985	1986	1987	1988
Produção	10.000	6.000	5.760	17.280
Porcentagem do ano anterior		60	96	300

Observe-se que este último valor é quase  $200 \%$  do valor efetivamente observado em 1988, de 17.200.

A média geométrica, por sua vez, é:

$$g = \sqrt[3]{(60)(96)(300)} = 120$$

implicando uma taxa anual média de crescimento de  $120 - 100 = 20 \%$ . Verificando, teremos:

$$\text{no ano de 1986: } 10.000 + 0,20(10.000) = 12.000;$$

$$\text{no ano de 1987: } 12.000 + 0,20(12.000) = 14.400;$$

$$\text{no ano de 1988: } 12.000 + 0,20(14.400) = 17.280 \text{ que coincide com o valor observado efetivamente em 1988.}$$

Se o valor da média geométrica das porcentagens de crescimento for menor do que 100, implica em uma porcentagem média de crescimento negativa, o que indica uma taxa média de declínio ao invés de uma taxa média de crescimento.<sup>7</sup> Atente também para o fato de que as três porcentagens a partir das quais a média geométrica é calculada são *percentuais do ano anterior* ao invés de *mudança percentual do ano anterior*.<sup>8</sup>

- cálculo da taxa média de crescimento é baseado principalmente na hipótese de uma taxa constante de crescimento ou de que os valores individuais formam uma progressão geométrica. Quando o cálculo envolve um número considerável de períodos, utiliza-se com mais frequência uma fórmula que se relaciona com a média geométrica, que é:

$$R = \left( \sqrt[n]{\frac{x_f}{x_i}} \right) - 1$$

onde:

R = taxa de crescimento geométrica média,

n = número de períodos de tempo,

$x_f$  = valor no período final,

$x_i$  = valor no período inicial.

Para os dados da tabela anterior, teremos:

$$R = \left( \sqrt[3]{\frac{17.280}{10.000}} \right) - 1 = 0,20 \text{ ou } 20 \% , \text{ como obtido anteriormente.}^9 \text{ Note que } R = G - 1.$$

<sup>7</sup> Se, por exemplo, ao invés de 60, 96 e 300 %, como anteriormente, tivermos 60, 96 e 78 %, a taxa de crescimento geométrica média será de  $g = \sqrt[3]{(60)(96)(78)} = 76,59$ , o que indica um decréscimo médio de  $76,59 - 100 = -23,41$  %.

<sup>8</sup> Essas últimas porcentagens, referentes ao exemplo da tabela anterior, seriam  $(6.000 - 10.000)/10.000 = -0,40$ , ou seja - 40 %;  $(5.760 - 6.000)/6.000 = -0,04$ , ou seja, - 4 %; e  $(17.280 - 5.760)/5.760 = 2$ , ou seja + 200 %.

<sup>9</sup> - É interessante notar que pelo cálculo anterior empregam-se os valores dos anos intermediários, ao passo que nesse último, apenas empregam-se os valores do período inicial e final, não importando o que ocorreu nos períodos intermediários.



### 2.5.6 A Média Harmônica

A média harmônica é o inverso da média aritmética dos inversos dos valores observados. Simbolicamente, para uma amostra, temos:

$$H = \frac{\frac{1}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}}{n} = \frac{\frac{1}{\sum (1/x)}}{n} = \frac{n}{\sum (1/x)}$$

Para cálculos mais simples, a fórmula anterior pode ser reescrita como:

$$\frac{1}{H} = \frac{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}{n} = \frac{\sum (1/x)}{n}$$

A média harmônica dos três valores 4, 10 e 16 é:

$$\frac{1}{H} = \frac{\frac{1}{4} + \frac{1}{10} + \frac{1}{16}}{3} = 0,1375$$
$$H = 7,27$$

Para os mesmos dados a média aritmética é 10 e a média geométrica é 8,62. Para qualquer série de dados cujos valores não são todos os mesmos e que não incluem o zero, a média harmônica é sempre menor que tanto a média aritmética como a média geométrica.

### 2.5.7 A Mediana

A mediana é o valor do item central da série quando estes são arranjados em ordem de magnitude. Para a série R\$ 2, R\$ 4, R\$ 5, R\$ 7 e R\$ 8, a mediana é o valor do terceiro

item, R\$ 5. No caso do número de itens na série ser par, a mediana é a semi-soma dos dois valores mais centrais. Por exemplo, para a série 3, 5, 8, 10, 15 e 21 kg, a mediana é a média dos valores 8 e 10, ou seja 9.

A mediana pode ser formalmente definida como o valor que divide a série de tal forma que no mínimo 50 % dos itens são iguais ou menores do que ela, e no mínimo 50 % dos itens são iguais ou maiores do que ela. Mais rigorosamente, estabelecemos que:

$$X_{.5} = \text{o valor do } [(n+1)/2] \text{ -ésimo item}$$

Por exemplo, para uma série formada pelos valores 3, 5, 8, 10, 15 e 21 a mediana será o valor do  $[(6+1)/2] = 3,5$  ésimio item, ou seja, a semi soma do item de posto 3 e do item de posto 4, que são 8 e 10.

O valor da mediana não é influenciado pelos valores nas caudas de uma distribuição. Por exemplo, se temos a série de dados 1, 2, 3, 4, 5 a mediana é 3. Se substituirmos os valores das caudas dessa distribuição por quaisquer valores uma nova distribuição formada poderia ser formada pela série -1000, -100, 3, 500, 5000 e a mediana permanece sendo 3. Portanto, ela é uma medida de posição da distribuição bem adequada para distribuições assimétricas, tais como a distribuição de renda, já que não sabemos se a família mais rica ganha R\$7.000.000 ou R\$ 500.000.000. Veremos, mais a frente que ela possui vantagens em relação a média aritmética, como medida de posição (ou medida de tendência central) para dados agrupados em classes de frequência, quando a última classe tem limite superior indeterminado.

A mediana também tem a interessante propriedade de que a soma dos desvios absolutos das observações em relação a mediana é menor do que a soma dos desvios absolutos a partir de qualquer outro ponto na distribuição. Simbolicamente:

$$\sum |x - X_{.5}| = \text{um mínimo}$$

### 2.5.8 A Média para Dados Agrupados

Quando estamos tratando de amostras ou populações muito grandes é conveniente calcular as medidas descritivas a partir das distribuições de freqüência. A média não pode ser determinada exatamente a partir de distribuições de freqüência, mas uma boa aproximação pode ser obtida pela hipótese do ponto médio. A aproximação é quase sempre muito satisfatória se a distribuição é bem construída.<sup>10</sup> A hipótese do ponto médio refere-se a considerar-se de que todas as observações de uma dada classe estão centradas no ponto médio daquela classe. Conseqüentemente, o valor total da freqüência da classe da *i*-ésima classe é simplesmente o produto  $f_i m_i$ , onde  $f_i$  é a freqüência (absoluta simples) da classe *i* e  $m_i$  é ponto médio da classe *i*. Sob essa hipótese, a média aproximada para uma distribuição de uma amostra com *k* classes vem a ser:

$$\begin{aligned}\bar{X} &\cong \frac{f_1 m_1 + f_2 m_2 + \dots + f_k m_k}{f_1 + f_2 + \dots + f_k} \cong \frac{\sum fm}{\sum f} \\ &= \frac{\sum fm}{n}\end{aligned}$$

É importante notar que todos os somatórios na equação acima referem-se às classes e não às observações individuais. Consideremos a seguinte tabela de distribuição de freqüência para dados de gasto com alimentação extraídos de uma pesquisa de orçamentos familiares.

Classe	f	m	fm
R\$ 120,00 - R\$139,99	5	130,0	650,0
140,00 - 159,99	26	150,0	3900,0
160,00 - 179,99	24	170,0	4080,0
180,00 - 199,99	15	190,0	2850,0
200,00 - 219,99	8	210,0	1680,0

<sup>10</sup> Isto é, principalmente se no agrupamento dos dados originais em uma tabela de distribuição de freqüência, empregou-se um número adequado de classes de freqüência.

220,00 - 239,99	2	230,0	460,0
Total	80		13620,0

$$\bar{x} = \frac{13620,00}{80} = R\$170,25$$

Ao utilizar essa aproximação estamos considerando a hipótese de que todas as observações em cada classe estão uniformemente distribuídas nessa classe. Por exemplo, se tivermos um intervalo de tamanho 100 e com frequência igual a 6 observações, a localização dessas observações seria 0,20,40,60,80 e 100, com distância constante entre cada par de observações, de forma que:

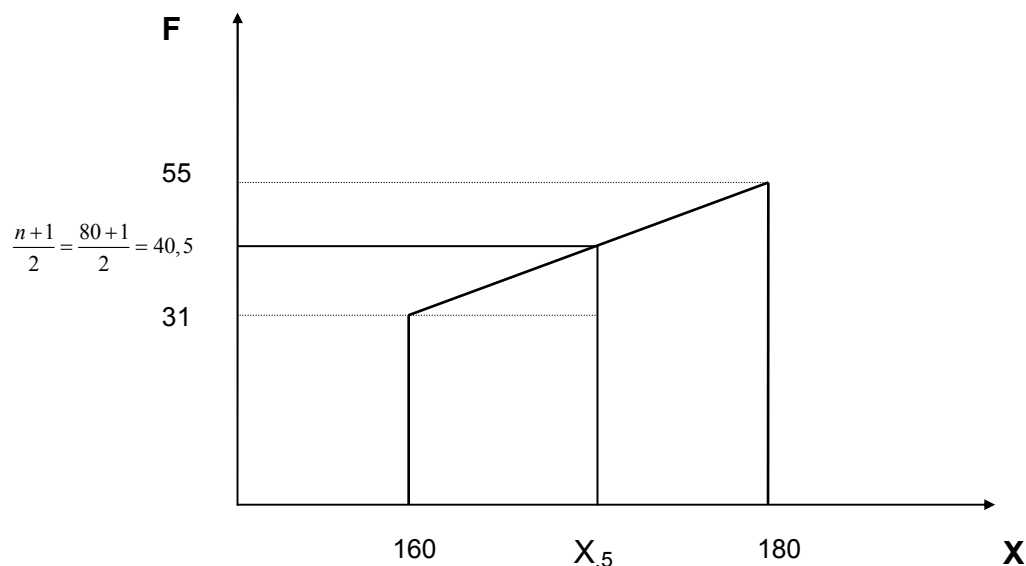
$0+20+40+60+80+100 = 300 = m \times 6$  e  $m = 50$ , ou seja, o ponto médio do intervalo de 0 a 100. Conclui-se que se a distribuição das observações for uniforme em cada intervalo, o somatório dos valores das observações de cada intervalo é igual ao produto da frequência no intervalo pelo valor do ponto médio desse intervalo. Supõe-se que com uma conveniente construção de intervalos de classe os eventuais erros nos intervalos compensam-se mutuamente.

### 2.5.9 A Mediana para dados Agrupados

Assim como é possível estabelecer uma aproximação da média aritmética para dados agrupados, o mesmo pode ser feito para a mediana. O método usado é o da interpolação utilizando-se a distribuição de frequência acumulada ou ogiva. Inicialmente determina-se a classe que contém a mediana. Essa será a classe cuja frequência acumulada relativa correspondente a seu limite inferior é menor que 0,50 (ou 50 %) e a frequência acumulada relativa correspondente a seu limite superior é maior que 0,50 (ou 50 %). O próximo passo é a determinação do ponto exato onde se localiza a mediana naquela classe. Para o exemplo anterior de gastos com alimentação de famílias, temos:

Classe	freq. absoluta	freq.acumulada	freq. relativa acumulada
R\$ 120,00 - R\$139,99	5	5	0,0625
140,00 - 159,99	26	31	0,3875
160,00 - 179,99	24	55	0,6875
180,00 - 199,99	15	70	0,8750
200,00 - 219,99	8	78	0,9750
220,00 - 239,99	2	80	1,0000
Total	80		

A classe que contém a mediana é a terceira classe, pois a frequência relativa acumulada da classe anterior (segunda classe) é menor que 0,5 e a frequência relativa acumulada da terceira classe é maior do que 0,5.<sup>11</sup> Na figura a seguir, F é a frequência acumulada (representada no eixo vertical) e X é o valor da variável (representada no eixo horizontal).



<sup>11</sup> - A frequência relativa acumulada da classe anterior à classe corrente é a frequência relativa acumulada do limite inferior da classe corrente. A frequência relativa acumulada da classe corrente é a frequência relativa acumulada do limite superior dessa mesma classe.

Por semelhança de triângulos, verifica-se que:

$$\frac{X_{.5} - 160}{180 - 160} = \frac{40,5 - 31}{55 - 31}$$

$$\therefore X_{.5} = 167,92$$

Este procedimento é o mesmo que a seguinte fórmula de interpolação:

$$X_{.5} = LI_{.5} + \left[ \frac{(n+1)/2 - F_a}{f_{.5}} \right] c$$

onde:

$LI_{.5}$  = limite de classe inferior da classe da mediana,

$F_a$  = frequência acumulada da classe imediatamente anterior à classe da mediana,

$f_{.5}$  = frequência absoluta simples da classe da mediana,

$c$  = amplitude (tamanho) da classe da mediana.

### 2.5.10 A Moda para dados Agrupados

A moda de uma distribuição de frequência pode muitas vezes ser aproximada pelo ponto médio da classe modal - a classe com maior densidade de frequência.<sup>12</sup> Então, para os dados de gastos com alimentação do exemplo anterior,  $x_m = \text{R\$ } 150$ , o ponto médio da segunda classe, que possui a maior frequência. Esse método de localizar a moda é totalmente satisfatório quando as densidades de frequência da classe imediatamente

---

<sup>12</sup> Definimos densidade de frequência de um intervalo de classe como sendo o quociente entre a frequência absoluta simples desse intervalo e o seu tamanho (amplitude). Quando os intervalos de classe possuem amplitudes desiguais, existe uma tendência de os intervalos maiores apresentarem maiores frequências. Dessa forma a classe modal não é a classe de maior frequência mas a classe de maior densidade de frequência. Naturalmente, quando todos os intervalos têm a mesma amplitude, como no caso do exemplo anterior e como geralmente são construídos para não distorcer a distribuição, a classe modal é a classe de maior densidade de frequência assim como também a classe de maior frequência. Esse conceito de densidade de frequência será muito útil, quando definirmos, mais adiante, a função densidade de probabilidade e para a sua compreensão intuitiva.

anterior à classe modal (a classe premodal) e da classe imediatamente posterior à classe modal (classe posmodal) são aproximadamente iguais. Quando isso não ocorre, como sugerido pela figura a seguir, resultados mais precisos podem ser obtidos com a seguinte fórmula, para uma amostra:

$$X_m \cong L_m + \left( \frac{\Delta_1}{\Delta_1 + \Delta_2} \right) c$$

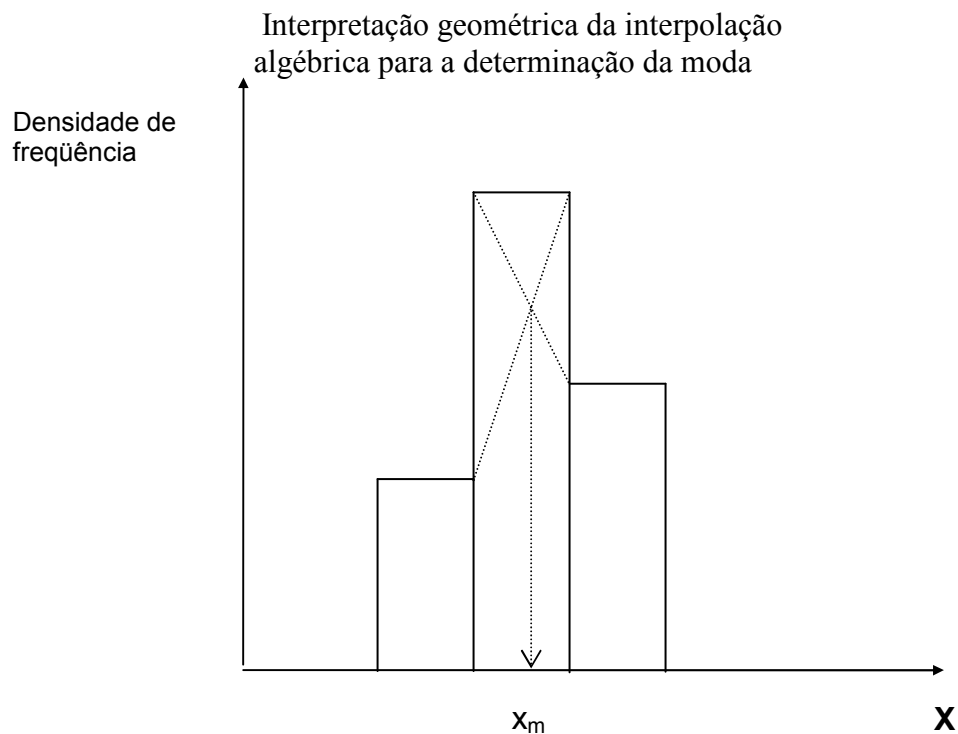
onde:

$L_m$  = o verdadeiro<sup>13</sup> limite inferior de classe da classe modal

$\Delta_1$  = da diferença entre das densidades de frequência da classe modal e classe premodal.

$\Delta_2$  = da diferença entre das densidades de frequência da classe modal e classe posmodal.

$C$  = a verdadeira amplitude de classe da classe modal.



<sup>13</sup> Para determinar os limites de classe verdadeiros para uma variável contínua, temos que escrever os limites de classe com uma casa decimal a mais do que os dados originais. Por exemplo, se o conjunto de dados consiste de medidas de peso arredondadas para um décimo de grama, os limites nominais de classe (também chamados de limites aparentes podem ser 11,0 - 11,2; 11,3 - 11,5; 11,6 - 11,8; ... Os limites

No exemplo anterior de gastos com alimentos de 80 famílias, como a amplitude de todos os intervalos são iguais, podemos utilizar as frequências absolutas de classe no lugar das densidades de frequência, para o cálculo do valor aproximado da mediana.

$$L_m = 140,00 \quad \Delta_1 = 26 - 15 = 11$$

$$c = 20 \quad \Delta_2 = 26 - 24 = 2$$

$$x_m \cong 140,00 + \left(\frac{11}{11+2}\right)20 = 156,92$$

Uma observação é aqui necessária. É possível calcular os valores aproximados da mediana e da moda para dados agrupados quando o último intervalo de classe tem limite superior indeterminado. No caso da mediana isso é imediato e no caso da moda, o seu cálculo somente pode ser feito se a última classe não for a classe modal e é preciso primeiramente calcular as densidades de frequência. Como exemplo, suponhamos que a distribuição de renda de uma certa região é dada pela seguinte distribuição de frequência:

renda (R\$) limites nominais	limites reais	frequência absoluta	densidade de frequência
0 - 120	0 - 120,50	40	$40/120,50 = 0,332$
121 - 605	120,50 - 605,50	170	$170/485 = 0,350$
606 - 1200	605,50 - 1200,50	220	$220/595 = 0,370$
1201 - 2400	1250,50 - 2400,50	15	$15/1150 = 0,013$
mais de 2400	mais de 2450,50	97	indeterminado
Total		542	

---

verdadeiros de classe (também conhecidos como limites reais ou efetivos) seriam 10,95 - 11,25; 11,25 - 11,55; 11,55 - 11,85;...



A mediana está localizada na terceira classe:<sup>14</sup>

$$x_s \cong 605,50 + \left[ \frac{(542 + 1)/2 - 210}{220} \right] (1200,50 - 605,50) = 772$$

A classe modal também é a terceira classe:<sup>15</sup>

$$x_m = 605,50 + \frac{(0,370 - 0,350)}{(0,370 - 0,350) + (0,370 - 0,013)} (1200,50 - 605,50) = 637$$

Infelizmente, para esse exemplo não é possível o cálculo da média, o que demonstra que para algumas situações temos que contar com a mediana como medida de posição (ou de tendência central) de uma distribuição estatística.

Discutiremos agora comparativamente algumas das características das três principais medidas de posição:

### **A Média Aritmética**

- 1) Ela é afetada por todas as observações e é influenciada pelas magnitudes absolutas dos valores extremos na série de dados.
- 2) Ela é das três medidas de posição a que possibilita maiores manipulações algébricas, dadas as características de sua fórmula.
- 3) Em amostragem, a média é uma estatística estável. Isso será aprofundado posteriormente.

---

<sup>14</sup> Observe-se que os dados originais estão, de acordo com o sugerido pela tabela acima, com aproximação igual a unidades de gramas. Os limites verdadeiros (ou reais) de classe) passam, portanto, a ter aproximação de uma casa decimal de grama. O valor final dos cálculos da mediana e da moda são aproximados para unidades de grama, já que essa é a aproximação dos dados originais (que se refere ao instrumento de medida).

<sup>15</sup> Já que esta classe é a que apresenta maior densidade de freqüência. Como a última classe não tem limite superior definido não foi possível calcular sua densidade de freqüência, já que não podemos determinar sua amplitude. Dependendo dessa amplitude ela poderia ter uma densidade de freqüência maior que a da

### **A Mediana**

- 1) Seu valor é afetado pelo número de observações e como elas estão distribuídas mas ela não é afetada pelos valores das observações extremas.
- 2) Sua fórmula não é passível de manipulação algébrica.
- 3) Seu valor pode ser obtido, como vimos, em distribuições, com limites superiores indeterminados para a sua última classe.
- 4) A mediana é a estatística mais adequada para descrever observações que são ordenadas ao invés de medidas.

### **A Moda**

- 1) A moda é o valor mais típico e representativo de uma distribuição. Ela representa o seu valor mais provável.
- 2) Como a mediana, a moda também não é influenciada pelos valores extremos da distribuição e não permite manipulações algébricas como a fórmula da média.

Existem algumas relações entre as diversas medidas de posição:

- 1) Para qualquer série, exceto quando no caso de todas as observações coincidirem em um único valor, a média aritmética é sempre maior que a média geométrica, a qual, por sua vez, é maior que a média harmônica.
- 2) Para uma distribuição simétrica e unimodal,  $\text{média} = \text{mediana} = \text{moda}$ .
- 3) Para uma distribuição positivamente assimétrica,  $\text{média} > \text{mediana} > \text{moda}$ .  
A distância entre a mediana e a média é cerca de um terço da distância entre a moda e a média.

---

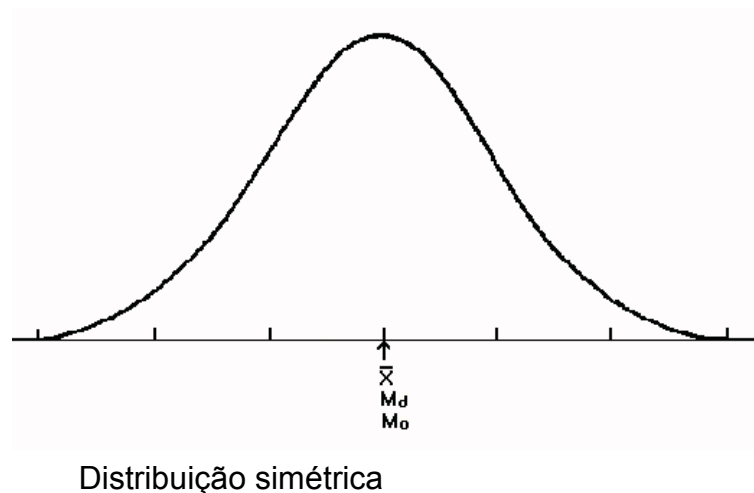
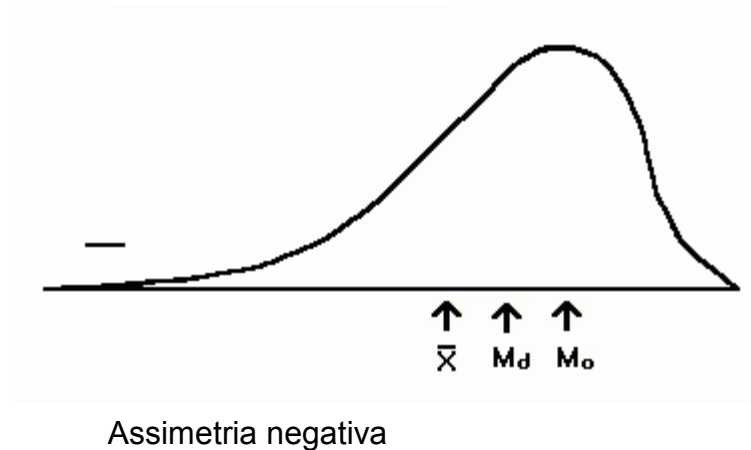
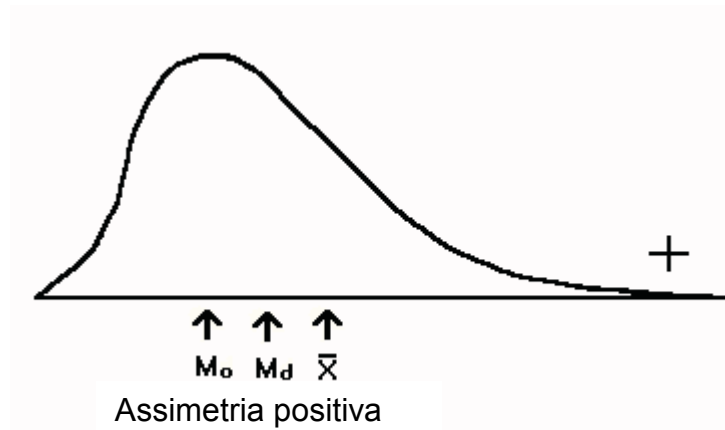
terceira classe. Mas mesmo nesse caso, a terceira classe ainda seria modal, já que sua densidade de frequência é maior que a das suas classes vizinhas, e a distribuição passaria a ser bimodal.

4) Para uma distribuição negativamente assimétrica, média < mediana < moda.

A distância entre a mediana e a média é cerca de um terço da distância entre a moda e a média.

Essas últimas características são apresentadas graficamente, a seguir

### POSIÇÕES RELATIVAS DA MÉDIA, MEDIANA E MODA EM FUNÇÃO DA ASSIMETRIA DAS DISTRIBUIÇÕES



## **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 homogêneos (menos dispersos em relação a média) do que a outra. Um país, por exemplo, com uma distribuição de renda mais equânime, terá uma dispersão de suas rendas menor do que um país com estrutura de renda 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 inadequação das médias**

A importância das médias é com frequência exagerada. Se dizemos que a renda familiar média de um determinado país é de US\$ 5.000 por ano não sabemos muita coisa sobre a distribuição de renda desse país. Uma média, como um simples valor adotado 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 tem o mesmo valor da média. Quase sem exceção, as observações incluídas em uma 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 qual distribuição tem maior ou menor grau de dispersão da informação dada pela tabela abaixo.

	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; outras não. Assim, para descrever uma distribuição precisamos também de uma medida do grau de simetria ou assimetria. A estatística descritiva 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.

### 2.5.11 O Intervalo (ou amplitude)

A medida de dispersão mais simples é a *amplitude*, a diferença entre o maior e o menor valor nos dados. Para uma distribuição de frequência 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 ações e de outros ativos financeiros são freqüentemente descritos em termos de sua amplitude, com a apresentação pelas Bolsas de Valores do maior valor e do menor valor da ação em um determinado período de tempo.

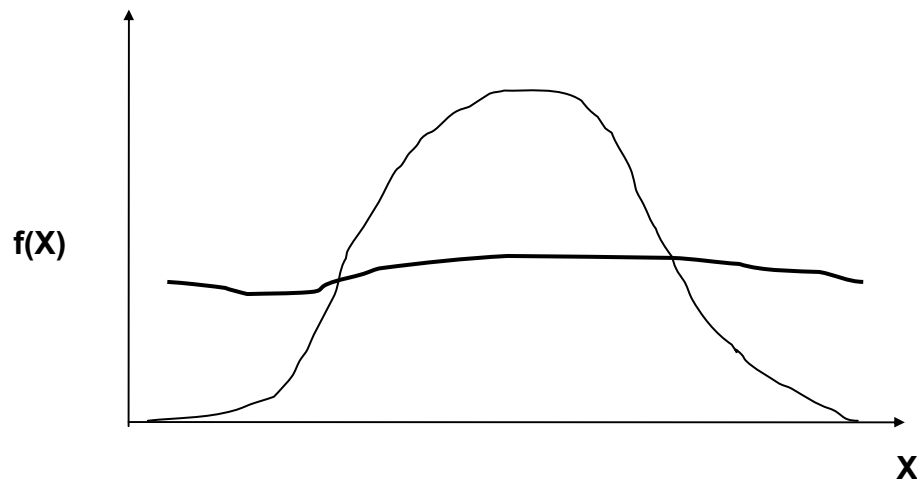
Para algumas distribuições simétricas a média pode ser aproximada tomando-se a semi-soma dos dois valores extremos,<sup>16</sup> que é freqüentemente chamada de semi-amplitude. Por exemplo, é prática entre os meteorologistas derivar a média diária de temperatura

---

<sup>16</sup> Foi o que fizemos ao calcular a média para valores agrupados em classes de frequência. Nesse caso utilizamos 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.

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 independe do que ocorre no interior da distribuição, já que somente depende dos valores extremos. Este defeito é ilustrado na figura a seguir:



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

#### **2.5.12 Percentis, Decis e Quartis**

Podemos tentar responder a seguinte pergunta: “que proporção dos valores de uma variável é menor ou igual a um dado valor? Ou maior ou igual a um dado valor? Ou entre dois valores?” Quando construímos uma distribuição de frequência acumulada, tais questões somente podem ser respondidas com relação aos limites de classe exatos. Por exemplo, a partir da distribuição de frequência relativa acumulada da página 28, podemos dizer que 38,75 % das observações são menores do que 159,99. Mas não podemos responder a pergunta: “qual é o gasto familiar tal que a proporção da amostra tendo este valor ou menos é 35 %?”. Mas é visível da tabela que 6,25 % das famílias gastam com alimentação até R\$ 139,99 e 38,75 % das famílias gastam até R\$ 159,99. Portanto, como 35 % está entre estes dois valores, o gasto familiar tal que a proporção da amostra tendo

este valor ou menos é 35 % está situado entre R\$ 139,99 e R\$ 159,99. Este valor é chamado de percentil 35.

O percentil 40 é o valor da variável que é maior do que 40 % das observações. Generalizando, o percentil  $x$ , é o valor da variável que é maior do que  $x$  % das observações. Em outras palavras, o percentil  $x$  é o valor da variável correspondente ao valor de frequência relativa acumulada de  $x$  %.<sup>17</sup> O primeiro decil é o valor da variável que supera um décimo (ou 10 %) do total de observações. Se tivermos 200 observações, o segundo decil será aproximadamente a observação de posto 40.

O primeiro quartil é o valor da variável cuja frequência relativa acumulada é 0,25 (ou 25 %). O terceiro quartil é o valor da variável cuja frequência relativa acumulada é 0,75 (ou 75 %). O primeiro quartil é maior do que um quarto dos valores observados e menor do que três quartos destes valores. O terceiro quartil é maior do que três quartos dos valores observados e menor do que um quarto destes valores. O segundo quartil confunde-se com a mediana.

Uma medida de dispersão é o chamado desvio interquartilico que é a diferença entre o terceiro e o primeiro quartis.

### 2.5.13 Variância e Desvio Padrão

A variância é definida como a média dos desvios ao quadrado em relação à média da distribuição. Para uma amostra,

$$S^2 = \frac{\sum (x - \bar{X})^2}{n - 1}$$

---

<sup>17</sup> Para o cálculo do valor exato do percentil  $x$  para dados agrupados utiliza-se o mesmo método para a determinação da mediana, ou seja, a interpolação linear. Como no caso da mediana, podemos empregar uma

fórmula de interpolação 
$$X_p = LI_p + \left[ \frac{p \times (n + 1) / 100 - F_a}{f_p} \right] c$$

onde  $X_p$  é o percentil  $p$ ,  $LI_p$  é o limite inferior real da classe que contem o percentil,  $F_a$  é a frequência relativa acumulada da classe anterior à classe que contem o percentil,  $f_p$  é a frequência relativa (simples) da classe que contem o percentil,  $c$  é a amplitude do intervalo de classe que contem o percentil e  $n$  é o número de observações. O mesmo método pode ser empregado também para os decis e quartis.

Para uma população finita,

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

Na penúltima equação, **n-1** é chamado de número de “graus de liberdade” de **S<sup>2</sup>**, um conceito a ser definido mais tarde. Existe uma restrição para esta equação: **n > 1** (não se pode calcular a variância para uma amostra de uma observação apenas). O desvio padrão é a raiz quadrada da variância, e é denotado **S** (para amostra) e **σ** (para população). Existem fórmulas que facilitam os cálculos para **S<sup>2</sup>** e **σ<sup>2</sup>**:

$$S^2 = \frac{n \sum x^2 - (\sum x)^2}{n(n-1)}$$
$$\sigma^2 = \frac{\sum x^2}{N} - \left( \frac{\sum x}{N} \right)^2$$

Com estas duas últimas fórmulas, podemos calcular a variância somente com a soma dos valores ( $\sum x$ ) e a soma dos quadrados dos valores ( $\sum x^2$ ); não é mais necessário calcular a média, em seguida os desvios em relação às médias e finalmente os quadrados destes desvios.

Para ilustrar o processo de cálculo da variância e desvio padrão e para mostrar o uso destas medidas, considere o seguinte exemplo. Dois tipos diferentes de máquina, **X** e **Y** são projetadas para produzir o mesmo produto. Elas têm o mesmo preço de venda. Um fabricante está tentando decidir qual delas comprar e observou 10 máquinas distintas de cada tipo em operação por uma hora. A tabela seguinte mostra as produções horárias nas primeiras duas colunas. As médias são  $\bar{x} = 403/10 = 40,3$  unidades por hora e  $\bar{y} = 408/10 = 40,8$  unidades por hora. Portanto, com base nestes dados, o tipo **Y** é um pouco mais rápida. Podemos retirar mais alguma informação a partir destes dados?



Podemos medir e comparar as dispersões das produções horárias dos dois tipos de máquina. Usando a penúltima fórmula para os dados da tabela, obtemos:

$$S_X^2 = \frac{10(16.405) - (403)^2}{10(10-1)} = 18,23$$

$$S_X = \sqrt{18,23} = 4,27 \text{ unidades por hora}$$

$$S_Y^2 = \frac{10(17.984) - (408)^2}{10(10-1)} = 135,11$$

$$S_Y = \sqrt{135,12} = 11,62 \text{ unidades por hora}$$

x	y	x <sup>2</sup>	y <sup>2</sup>
35	25	1.225	625
36	26	1.296	676
49	55	2.401	3.025
44	52	1.936	2.704
43	48	1.849	2.304
37	24	1.369	576
38	34	1.444	1.156
42	47	1.764	2.209
39	50	1.521	2.500
40	47	1.600	2.209
Soma 403	408	16.405	17.984

O tipo X tem menor dispersão que o tipo Y. Apesar de ter maior preço que o tipo Y, a máquina X é mais precisa.

### 2.5.14 Variância e Desvio Padrão para Dados Agrupados

A *variância* e o desvio padrão (como a média, mediana, moda, quartis, percentis, decis) podem ser calculados para dados agrupados, ou seja, distribuições de frequência com intervalos de classe. Entretanto, os resultados podem ser apenas aproximadamente precisos. Utiliza-se, como no caso da média, a hipótese do ponto médio: a de que toda observação está localizada no ponto médio de sua classe. Cada ponto médio entra nos cálculos quantas vezes são as observações naquele intervalo de classe. As equações para as variâncias são:

$$S^2 = \frac{\sum f(m - \bar{X})^2}{n - 1}, \text{ para a amostra;}$$
$$\sigma^2 = \frac{\sum f(m - \mu)^2}{N}, \text{ para a população.}$$

Os símbolos utilizados nestas equações já foram definidos anteriormente. Para facilitar os cálculos podemos utilizar as seguintes fórmulas mais convenientes para as variâncias:

$$S^2 = \frac{\sum fm^2 - (\sum fm)^2 / n}{n - 1}$$

e

$$\sigma^2 = \frac{\sum fm^2 - (\sum fm)^2 / N}{N}$$

para a amostra e população, respectivamente. Aqui, como antes, assumimos que a população é finita.

Os somatórios em todas estas equações são para todas as  $k$  classes, não para as observações individuais. Estas equações podem ser aplicadas tanto para intervalos de classe iguais como para intervalos de classe desiguais. Entretanto, elas não podem ser empregadas quando existem um ou mais intervalos sem limites. Como para os dados não

agrupados, a raiz quadrada destas equações são os desvios padrões para a amostra e para a população, respectivamente.

Aplicando as últimas equações para o exemplo de consumo de alimentos, temos:

Classe	(1) m	(2) f	(3) fm (2)(1)	(4) fm <sup>2</sup> (3)(1)
R\$ 120,00 - R\$139,99	130	5	650	84.500
140,00 - 159,99	150	26	3.900	585.000
160,00 - 179,99	170	24	4.080	693.000
180,00 - 199,99	190	15	2.850	541.500
200,00 - 219,99	210	8	1.680	352.800
220,00 - 239,99	230	2	460	105.800
Total		80	13.620	2.363.200

$$S^2 = \frac{\sum fm^2 - (\sum fm)^2 / n}{n - 1} = \frac{2.363.200 - (13.620)^2 / 80}{80 - 1} = 561,96$$

$$S = \sqrt{561,96} = 23,71$$

### 2.5.15 Interpretando e Aplicando o Desvio Padrão

O desvio padrão é mais a mais usada das medidas de variabilidade. Infelizmente, o desvio padrão não tem uma interpretação intuitivamente óbvia. Por exemplo, no exemplo anterior das máquinas,  $S_X = 4,27$  unidades por hora, mas não é óbvio o que isto quer dizer para a máquina X. Para muitas séries de dados há dois teoremas para a interpretação do desvio padrão que são muito úteis. Eles são chamados de Desigualdade de Chebyshev e a Regra de Gauss, as quais introduzimos a seguir.

**Teorema: Desigualdade de Chebyshev.** Para qualquer conjunto de dados e qualquer constante  $h > 1$ , no mínimo  $1 - 1/h^2$  dos dados estarão situados dentro de um intervalo formado por  $h$  desvios padrões abaixo e acima da média.

Por este teorema temos certeza de que no mínimo  $\frac{3}{4}$ , ou 75 % dos dados irão situar-se dentro do intervalo  $\bar{X} \pm 2S$ . Neste caso  $h = 2$  e  $1 - 1/h^2 = 1 - 1/2^2 = 3/4$ . No mínimo  $8/9$ , ou 88,9 % dos dados estarão no intervalo  $\bar{X} \pm 3S$ ; e no mínimo  $15/16$ , ou cerca de 94 % dos valores de qualquer variável estarão incluídos dentro do intervalo  $\bar{X} \pm 4S$ .

Considere o exemplo anterior das máquinas. Temos  $\bar{X} = 40,3$  e  $S_X = 4,27$ . Que percentagem das máquinas terá produção entre  $\bar{X} \pm 1,5S_X = 40,3 \pm 1,5 \times 4,27$ , ou seja, entre 33,9 e 46,7? Resposta: no mínimo  $1 - \frac{1}{1,5^2} = 0,56$ , ou aproximadamente 56 %. Da tabela anterior encontramos 9 das 10 máquinas tipo X que estão dentro deste intervalo e claramente  $9/10$  é maior do que 56 %.

A vantagem da Desigualdade de Chebyshev é que ela pode ser aplicada à variáveis com qualquer padrão de distribuição (não importa que sejam simétricas, assimétricas, mesocúrticas, platicúrticas, leptocúrticas, etc.). Entretanto, ela tem a desvantagem de não ser muito precisa, já que a porcentagem efetiva que caem dentro do intervalo em torno da média é quase sempre muito maior do que o mínimo dado por  $1 - 1/h^2$ , especialmente quando as amostras são pequenas, como no nosso exemplo anterior.

**Teorema: A Regra de Gauss.** Se os dados são amostrais e se são, de forma aproximada, distribuídos normalmente, ou seja, o histograma dos dados é aproximadamente simétrico e tem a forma de um sino, então:

1.  $\bar{X} \pm 1S$  incluirá aproximadamente 68 % dos dados
2.  $\bar{X} \pm 2S$  incluirá aproximadamente 95 % dos dados
3.  $\bar{X} \pm 3S$  incluirá aproximadamente 100 % dos dados

Chamamos isto de Regra de Gauss, porque é baseada na distribuição de probabilidade gaussiana (ou distribuição de probabilidade normal). Esta distribuição será discutida em detalhe em um capítulo posterior.

### 2.5.16 Coeficiente de Variação

Com freqüência, como no caso do exemplo das duas máquinas, queremos comparar a variabilidade de dois ou mais conjuntos de dados. Podemos fazer isto facilmente usando as variâncias ou os desvios padrões quando, primeiro, todas as observações individuais têm a mesma unidade de medida e, segundo, as médias dos conjuntos de dados são aproximadamente iguais. Quando qualquer uma destas condições não é satisfeita, uma medida relativa de dispersão deve ser usada. Uma medida relativa de variabilidade freqüentemente usada é chamada de *coeficiente de variação*, denotada por CV para uma amostra. Esta medida é o valor do desvio padrão em relação à média:

$$CV = \frac{S}{\bar{X}}$$

Suponha que um cientista na Índia obteve os seguintes dados referentes aos pesos de elefantes e ratos.

Elefantes	Ratos
$\bar{x}_E = 6.000 \text{ kg}$	$\bar{x}_R = 0,150 \text{ kg}$
$s_E = 300 \text{ kg}$	$s_R = 0,04 \text{ kg}$

Se calcularmos os respectivos coeficientes de variação, teremos:

$$cv(X_E) = \frac{s_E}{\bar{x}_E} = \frac{300}{6000} = 0,050 \text{ ou } 5,0 \%$$
$$cv(X_R) = \frac{s_R}{\bar{x}_R} = \frac{0,04}{0,150} = 0,266 \text{ ou } 26,7 \%$$

Portanto, a variabilidade relativa dos pesos dos ratos é mais do que 5 vezes maior do que a variabilidade dos pesos dos elefantes. Para o exemplo anterior das máquinas, teremos:

$$cv(X) = \frac{4,27}{40,30} = 0,1060 \quad \text{ou } 10,60 \%$$

$$cv(Y) = \frac{11,62}{40,80} = 0,2848 \quad \text{ou } 28,48 \%$$

Assim, a dispersão relativa da produção da máquina Y é quase três vezes maior do que a dispersão relativa da máquina X.

## 2.6 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 freqüentemente 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 em uma posição intermediária, a medida que aumenta a assimetria da distribuição. Conseqüentemente, a distância entre a média e a moda poderia ser usada para medir a assimetria. Precisamente,

$$\text{Assimetria} = \text{média} - \text{moda}$$

Quanto maior é a 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 **SK<sub>P</sub>** e dado por:

$$SK_P = \frac{\bar{X} - X_m}{S}$$

A aplicação desta expressão envolve outra dificuldade, que surge devido ao fato de que o 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

$$X_m = \bar{X} - 3(\bar{X} - X_{.5})$$

é adequada (não envolve imprecisão muito grande). A partir disto, vemos que:

$$\bar{X} - X_m = \bar{X} - [\bar{X} - 3(\bar{X} - X_{.5})] = 3(\bar{X} - X_{.5})$$

Com este resultado, podemos rescrever o coeficiente de assimetria de Pearson como:

$$SK_P = \frac{3(\bar{X} - X_{.5})}{S}$$

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  $\pm 3$ . Aplicando **SK<sub>P</sub>** 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 econômicos, 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 ciências sociais.

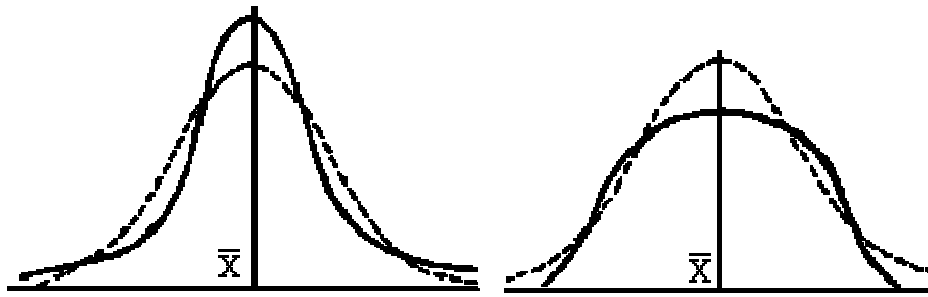
## 2.7 Curtose: uma medida de achatamento

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-interquartilico, ou seja, a metade do valor do desvio interquartilico, 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 concentrada no centro. Conseqüentemente, 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  $\frac{1}{2}(Q_3 - Q_1) < (D_9 - D_1)$  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 interquartilico. 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 fato de que para a variável normal padronizada, **k** = 0,2630 (veremos este ponto em capítulo posterior).





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

### 3. Probabilidade

Objetivos do capítulo:

- Definir o termo probabilidade.
  - Descrever as abordagens clássica, da frequência relativa e subjetiva da probabilidade.
  - Entender os termos experimento, espaços 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
- 
- Probabilidade: é uma medida de possibilidade de ocorrência de um determinado evento; ela pode assumir um valor entre **0** e **1**
  - Evento: Uma coleção de um ou mais resultados de um experimento
  - Exemplo: Experimento → jogar uma moeda duas vezes

Possíveis resultados (espaço amostral) → { KK, KC, CK, CC }

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

Como 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 números reais que vai de **0** a **1**, inclusive as extremidades deste intervalo.

- Quanto mais uma probabilidade é próxima de **0**, o evento a ela associado é mais improvável de ocorrer.
- Quanto mais uma probabilidade é próxima de **1**, o evento a ela associado é mais provável de ocorrer.

### 3.1 Definição Clássica de Probabilidade

- Probabilidade Clássica: é baseada na hipótese de que os resultados de um experimento são igualmente prováveis.

Usando o ponto de vista clássico:

$$\text{Probabilidade de um evento} = \frac{\text{número de resultados favoráveis}}{\text{número total de possíveis resultados}}$$

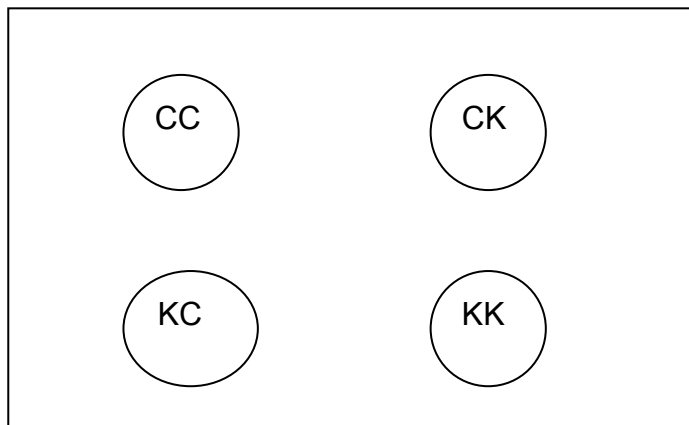
Considere o experimento de jogar duas moedas.

- O espaço amostral deste experimento é  $S = \{ CC, CK, KC, KK \}$
- Considere o evento: uma cara

$$\text{Probabilidade de um evento} = \frac{\text{número de resultados favoráveis}}{\text{número total de possíveis resultados}} = \frac{2}{4} = \frac{1}{2}$$

### Definições

- Eventos mutuamente exclusivos: a ocorrência de qualquer um evento significa que **nenhum** dos outros pode ocorrer ao mesmo tempo.
- No caso do experimento de jogar duas moedas, os quatro possíveis resultados são mutuamente exclusivos.



- Eventos Coletivamente Exaustivos: no mínimo um dos eventos deve ocorrer quando o experimento é conduzido.

No experimento de jogar 2 moedas, os quatro possíveis resultados são coletivamente exaustivos.

Soma das probabilidades = 1

- Desde que cada resultado no experimento de jogar 2 moedas tem probabilidade igual a  $\frac{1}{4}$ , então a soma das probabilidades dos resultados possíveis é  $\frac{1}{4} + \frac{1}{4} + \frac{1}{4} + \frac{1}{4} = 1$

### 3.2 Conceito da Frequência Relativa

- A probabilidade de um evento ocorrer “no longo prazo” é determinada pela observação de que fração de vezes o evento ocorreu no passado.
- A probabilidade pode ser calculada pela fórmula:

$$\text{Probabilidade do evento} = \frac{\text{número de vezes em que o evento ocorreu no passado}}{\text{número total de observações}}$$

## Exemplo 2

- A questão de ser ou não um réu culpado: em uma amostra de 500 estudantes em um determinado campus, 275 indicaram que o réu era culpado. Qual é a probabilidade de que um estudante neste campus indicará que o réu neste caso era culpado?
- Nota: Neste problema podemos aplicar a fórmula para a probabilidade baseada na frequência relativa.

Assim,  $P(\text{culpado}) = 275/500 = 0,55$

## 3.3 Probabilidade Subjetiva

- Probabilidade Subjetiva : é a probabilidade de que um particular evento ocorra atribuída por um indivíduo e baseada em um conjunto de informação disponível.

Exemplos de probabilidade subjetiva são:

- Estimar a probabilidade de que o time de futebol da Ponte Preta disputará a final do campeonato nacional.
- Estimar a probabilidade de que você obtenha conceito A neste curso.

## 3.4 Algumas Regras Básicas de Probabilidade

- Regra da Adição: Se dois eventos **A** e **B** são mutuamente exclusivos, a regra especial da adição estabelece que a probabilidade de que A ou B ocorram é igual a soma de suas respectivas probabilidades. A regra é dada pela seguinte fórmula:

$$P(A \text{ ou } B) = P(A) + P(B)$$

### Exemplo 3

A companhia de aviação X recentemente forneceu a seguinte informação para o Departamento de Aviação Civil (DAC) sobre os vôos da origem A ao destino B

Chegada	Frequência
Adiantada	100
No horário	800
Atrasada	75
Cancelado	25
Total	1000

- Seja **A** o evento: o vôo chega adiantado

$$\text{Então } P(A) = 100 / 1000 = 0,1$$

- Seja **B** o evento: o vôo chega atrasado

$$\text{Então } P(B) = 75 / 1000 = 0,075$$

- Nota: os eventos A e B são mutuamente exclusivos. Por quê?
- Qual é a probabilidade de que um vôo chegue adiantado ou atrasado?

$$P(A \text{ ou } B) = P(A) + P(B) = 0,1 + 0,075 = 0,175$$

### 3.5 A Regra do Complemento

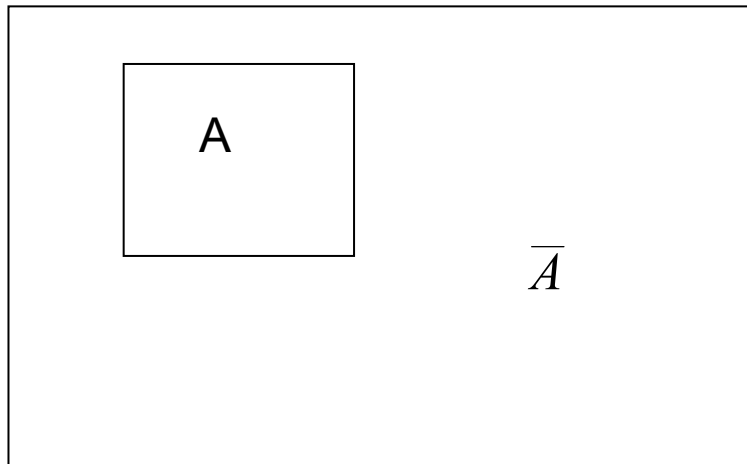
A regra do complemento é usada para determinar a probabilidade de um evento ocorrer subtraindo-se a probabilidade do evento não ocorrer de 1.

Seja  $P(A)$  a probabilidade do evento  $A$  e  $P(\bar{A})$  a probabilidade do evento não  $A$  (complemento de  $A$ ).

$$P(A) + P(\bar{A}) = 1$$

$$P(A) = 1 - P(\bar{A})$$

Um diagrama de Venn pode ilustrar a Regra do Complemento:



#### Exemplo 3

- Reconsidere os dados do exemplo 2. Seja **C** o evento: o voo chega no horário. Então  $P(C) = 800 / 1000 = 0,8$

- Seja **D** o evento: o voo é cancelado.

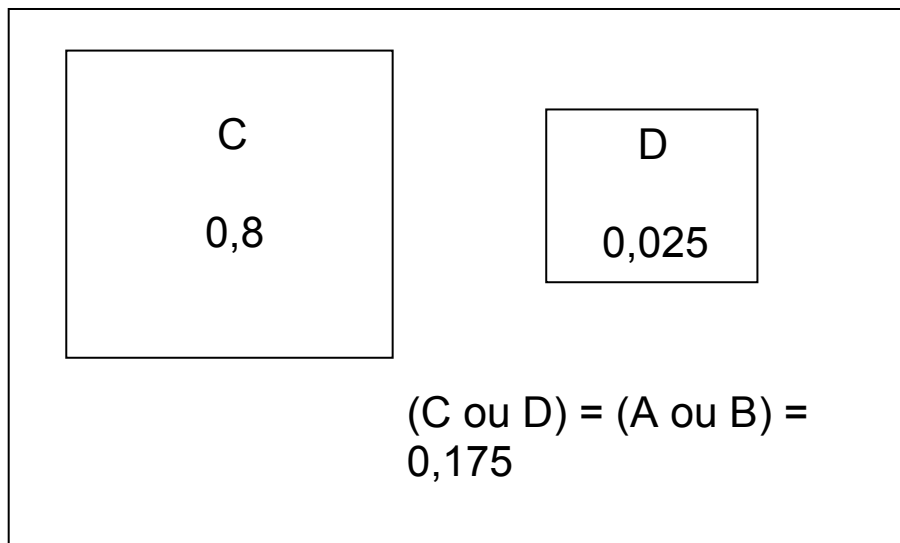
Então  $P(D) = 25 / 1000 = 0,025$

- Nota: os eventos  $C$  e  $D$  são mutuamente exclusivos. Por quê?

Use a regra do complemento para mostrar que a probabilidade do voo chegar adiantado (A) ou atrasado (B) é 0,175

- $P(A \text{ ou } B) = 1 - P(C \text{ ou } D) = 1 - [0,8 + 0,025] = 0,175$

O diagrama de Venn abaixo ilustra esta situação:



- A regra do complemento é muito importante no estudo de probabilidade. Com frequência, é mais eficiente calcular a probabilidade de um evento ocorrer determinando-se a probabilidade do evento não ocorrer e subtraindo o resultado de 1.

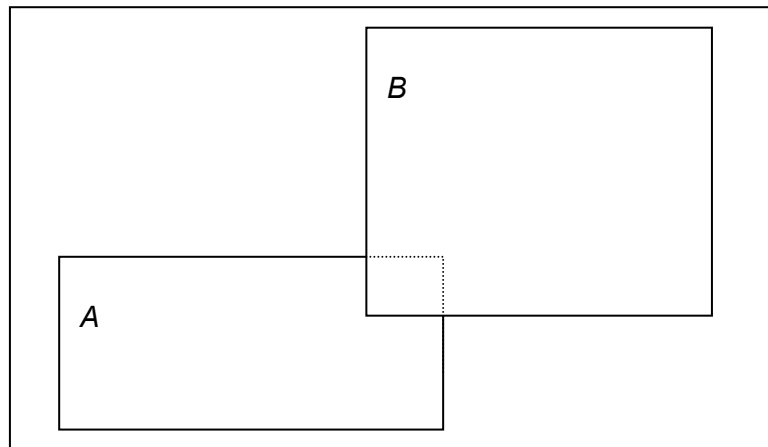
### 3.6 A Regra Geral da Adição

- Sejam **A** e **B** dois eventos que não são mutuamente exclusivos. Então  $P(A \text{ ou } B)$  é dado pela seguinte fórmula:

$$P(A \text{ ou } B) = P(A) + P(B) - P(A \text{ e } B)$$

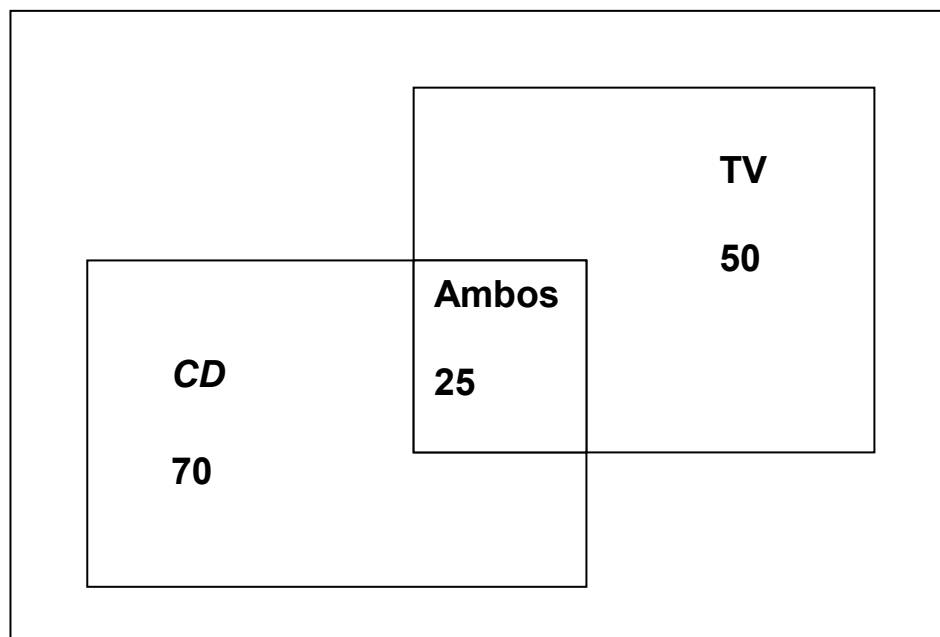


O Diagrama de Venn abaixo ilustra esta regra:



### Exemplo 5

- Em uma amostra de 150 estudantes, 70 disseram que somente têm um aparelho de CD, 50 disseram que somente têm uma TV e 25 disseram que têm ambos. O Diagrama de Venn a seguir descreve esta situação.



Se um estudante é selecionado ao acaso, qual é a probabilidade de que ele tenha somente um aparelho de CD ? De somente uma TV ? De tanto uma TV como um aparelho de CD?

- Seja **C** o evento “o estudante tem um aparelho de CD” e **T** o evento “o estudante tem uma TV”

$$P(C) = 70 / 150 = 0,4667$$

$$P(T) = 50 / 150 = 0,3333$$

$$P(C \text{ e } T) = 25 / 150 = 0,1667$$

- Se um estudante é selecionado ao acaso, qual é a probabilidade de que ele tenha tanto um aparelho de CD ou uma TV? (Nota: isto inclui a probabilidade de ter ambos os aparelhos).

Desde que:

$$P(C \text{ ou } T) = P(C) + P(T) - P(C \text{ e } T)$$

$$\text{Então, } P(C \text{ ou } T) = 0,4667 + 0,3333 - 0,1667 = 0,6333$$

### **3.7 Regras de Multiplicação**

#### **Regra Especial de Multiplicação**

- A regra especial de multiplicação requer que dois eventos **A** e **B** sejam independentes.

- Definição: Dois eventos **A** e **B** são independentes se a ocorrência de um não tem efeito sobre a probabilidade de ocorrência do outro.
- A regra especial é escrita simbolicamente como:

$$P(A \text{ e } B) = P(A) \cdot P(B)$$

- 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:

$$P(A \text{ e } B \text{ e } C) = P(A) \cdot P(B) \cdot P(C)$$

### Exemplo 6

Um investidor possui duas ações. Uma é de uma companhia de produção de petróleo e a outra é de uma cadeia de supermercados, de forma que podemos assumir que suas cotações são independentes. A probabilidade de que a ação da companhia de petróleo 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.

- Qual é a probabilidade de que ambas as ações cresçam em valor no próximo ano?
- Seja **A** o evento: a cotação da companhia de petróleo cresce no próximo ano e seja **B** o evento: a cotação da cadeia de supermercados cresce no próximo ano.

$$P(A \text{ e } B) = (0,50) \cdot (0,70) = 0,35$$

- Qual é a probabilidade de que ao menos uma destas ações aumentem em valor no próximo ano?

Isto implica que tanto uma pode aumentar (sem que a outra aumente) assim como ambas. Portanto,

$$P(\text{no mínimo uma}) = (0,50) \cdot (0,30) + (0,50) \cdot (0,70) + (0,70) \cdot (0,50) = 0,85$$

### Exemplo 7

Um estudo recente constatou que 60 % das mães com crianças de idade de até 10 anos empregam-se em tempo integral. Três mães são selecionadas aos acaso. Assumiremos que as mães são empregadas de forma independente umas das outras.

- Qual é a probabilidade de que todas sejam empregadas em período integral?

$$P(\text{todas as três empregadas em período integral}) = (0,60).(0,60).(0,60) = 0,216$$

- Qual é a probabilidade de que no mínimo umas das mães sejam empregadas em período integral?

$$P(\text{no mínimo uma}) = 1 - P(\text{nenhuma empregada em período integral}) =$$

$$1 - [(0,40).(0,40).(0,40)] = 0,936$$

### 3.8 Probabilidade Condicional

É a probabilidade de que um evento particular ocorra, dado que outro evento tenha ocorrido.

- Notação: A probabilidade do evento A dado que o evento B ocorreu é denotada por  $P(A/B)$

### Regra Geral da Multiplicação

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

- A regra 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.

A probabilidade conjunta,  $P(A \text{ e } B)$  é dada pela seguinte fórmula:

$$P(A \text{ e } B) = P(A) \cdot P(B/A)$$

Alternativamente, podemos também escrever:

$$P(A \text{ e } B) = P(B) \cdot P(A/B)$$

### Exemplo 8

Uma faculdade coletou a seguinte informação sobre seus estudantes de graduação:

Curso	Homens	Mulheres	Total
Contabilidade	120	80	200
Finanças	110	70	180
Marketing	70	50	120
Administração	110	100	210
Estatística	50	10	60
Computação	140	90	230
Total	600	400	1000

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

- Seja **A** o evento: o(a) estudante está cursando Contabilidade e **F** o evento: o(a) estudante é mulher.

$$P(A \text{ e } F) = 80 / 1000$$

- Qual é a probabilidade de selecionar uma mulher ?

$$P(F) = 400 / 1000$$

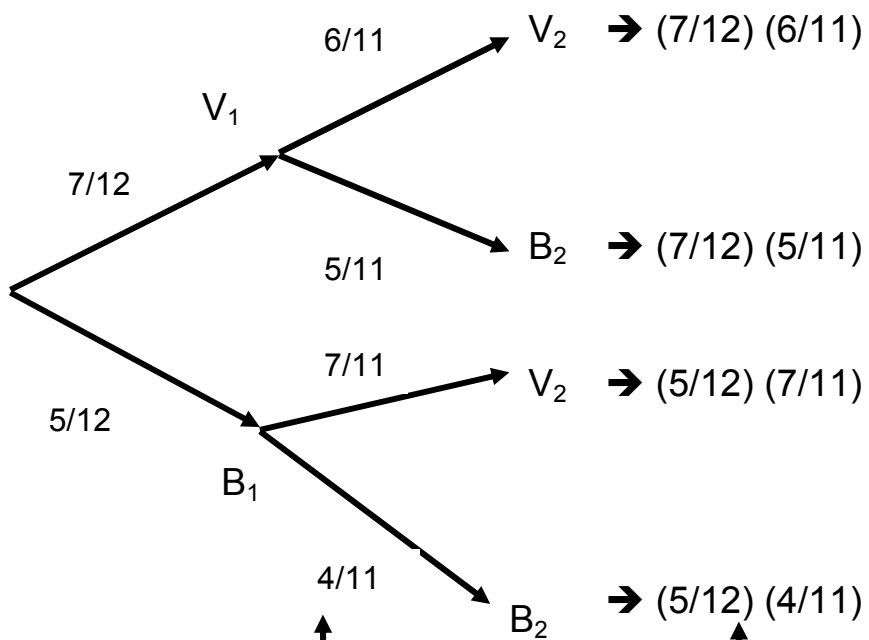
- Dado que o(a) estudante é mulher, qual é a probabilidade de que esteja cursando Contabilidade ?

Precisamos calcular  $P(A / F)$ .

$$P(A / F) = P(A \text{ e } F) / P(F) = [80 / 1000] / [400 / 1000] = 0,20$$

### 3.9 Diagramas em Árvore

- Um diagrama em árvore é muito útil para representar probabilidades conjuntas e probabilidades condicionais. Ele é particularmente útil para analisar decisões quando há diversos estágios no problema.
- Exemplo: Suponha que há 7 peças vermelhas e 5 peças azuis em uma sacola. Suponha que você selecione duas peças, uma após a outra e sem reposição. Construa um diagrama em árvore para esta informação.

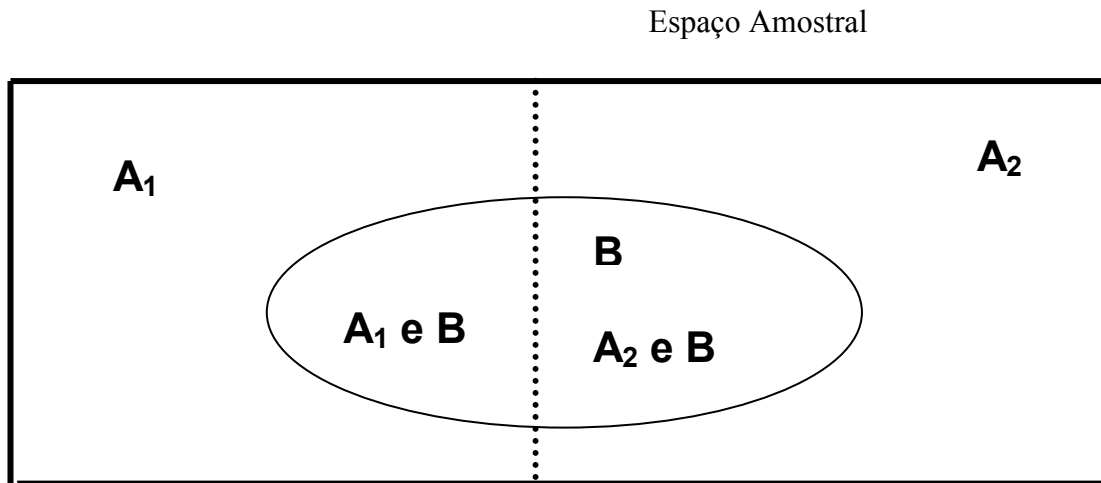


Probabilidades  
Condicionais

Probabilidades  
Conjuntas

### 3.10 Teorema de Bayes

- Considere o seguinte diagrama com as partições  $A_1$  e  $A_2$  :



$$P(A_1 / B) = P(A_1 \text{ e } B) / P(B); \quad P(A_1 \text{ e } B) = P(A_1) \cdot P(B / A_1)$$

$$P(B) = P(A_1 \text{ e } B) + P(A_2 \text{ e } B);$$

$$P(A_2 \text{ e } B) = P(A_2) P(B / A_2)$$

A partir disto, temos a fórmula seguinte (Teorema de Bayes):

$$P(A_1 / B) = \frac{P(A_1) \times P(B / A_1)}{P(A_1) \times P(B / A_1) + P(A_2) \times P(B / A_2)}$$

Nota: Este teorema pode ser estendido para diversas partições do espaço amostral (  $A_1$ ,  $A_2$ ,  $A_3$ , etc.)



### Exemplo 9:

A Companhia C & W tem recebido recentemente 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? Seja S o evento: a garrafa foi preenchida com conteúdo abaixo do especificado.

	% da Produção Total	% de garrafas com pouco preenchimento
A	55	3
B	45	4

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

### *Anexo 1 – Recordando Definições e Conceitos*

Uma moeda mostra cara 50% do tempo, em média. Depois de muitos lances, o número de caras é aproximadamente igual ao número de coroas.

#### **Um conceito de Probabilidade**

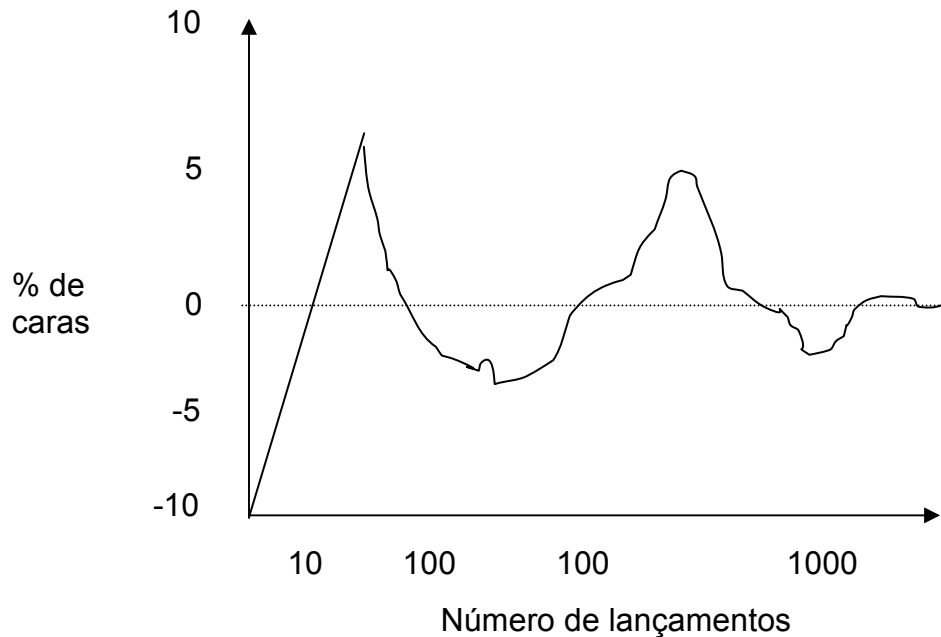
No limite quando o número de lances  $\rightarrow$  infinito  $\frac{\text{número de caras}}{\text{número de lançamentos}} \rightarrow 0,5$

Dizemos que a probabilidade de aparecer uma cara em qualquer lance é 1/2. Isto ilustra o conceito de probabilidade que será usada neste curso.

#### **Exemplo - 10 000 lances de moeda**

John Kerrich, um matemático sul africano estava visitando Copenhague quando a Segunda Guerra Mundial começou. Dois dias antes de seu vôo marcado para a Inglaterra,

os alemães invadiram a Dinamarca. Kerrich passou o resto da guerra internado em um acampamento em Jutland e para passar o tempo ele levou a cabo uma série de experimentos em teoria da probabilidade. Em um destes experimentos, lançou uma moeda 10.000 vezes. Seus resultados são mostrados no gráfico seguinte.



(O eixo horizontal está em uma escala logarítmica)

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

### Exemplo - 2 lançamentos de uma moeda

Lance uma moeda duas vezes e registre para cada lance se o resultado era uma cara (C) ou uma coroa (K). Exercício: Liste os **possíveis** resultados.

Seja A o evento deu uma ou mais caras. Quais resultados pertencem ao evento A? (CK, KC, CC).

Seja B o evento não aparece nenhuma cara. (KK)

Neste exemplo, os eventos A e B são ditos disjuntos ou mutuamente exclusivos, pois eles não têm nenhum resultado em comum. Eles também são exaustivos, já que eles cobrem todos os possíveis resultados do experimento.

Exercício: Defina um evento C que não é disjunto em relação a A.

## DEFINIÇÕES

Um **espaço amostral** é o conjunto de **todos os possíveis resultados** de um experimento.

Um **evento** é um conjunto de um ou mais **resultados** no espaço amostral.

Dois eventos são **disjuntos** ou **mutuamente exclusivos** se eles não têm nenhum resultado em comum.

A **variação aleatória** ocorre quando é impossível prever com certeza o resultado exato de um experimento individual, mas como o experimento é repetido um número grande de vezes uma distribuição regular de frequências relativas surge.

A **probabilidade** de um resultado ou evento **pode ser determinada tanto empiricamente** (baseado em dados) ou **teoricamente (baseado em um modelo matemático do processo)**. A definição empírica é a seguinte: Suponha que um resultado (ou evento) A ocorre f vezes em n observações. Então

$$\text{frequência relativa de A} = \frac{\text{número de vezes em que A ocorre}}{\text{número de observações}} = \frac{f}{n}$$

O conceito da probabilidade de um evento A é uma idealização da frequência relativa. É o valor limite da frequência relativa quando n fica muito grande, i.e. quando  $n \Rightarrow \infty$

$$\frac{f}{n} \rightarrow P(A) \text{ quando } n \rightarrow \infty$$

(P(A) denota a probabilidade de A ocorrer).

Estimativas teóricas de probabilidade estão baseadas em suposições plausíveis. A suposição mais comum é a de que todos os possíveis resultados são igualmente prováveis. Então

$$P(A) = \frac{\text{número de resultados correspondendo a A}}{\text{número total de resultados no espaço amostral}}$$

Por analogia com frequências relativas, as probabilidades têm as seguintes propriedades:

1.  $P(A)$  é um valor entre 0 e 1.
2.  $P(A) = 0$  significa A nunca acontece (correspondendo a  $f = 0$ )
3.  $P(A) = 1$  significa A sempre acontece (correspondendo a  $f = n$ )
4. O conjunto S de todos os possíveis resultados tem probabilidade 1.  $P(S) = 1$ , os quais se agrupam em 5 eventos.

## ***Anexo 2 - Independência e Modelos de Árvore para Calcular Probabilidades***

Se eventos X e Y são mutuamente exclusivos, então,

$$P(X \text{ ou } Y) = P(X) + P(Y)$$

**Em geral, se eventos X e Y não são mutuamente exclusivos então**

$$P(X \text{ ou } Y) = P(X) + P(Y) - P(X \text{ e } Y).$$

### **Exemplo - Fruta em 2 distritos**

Um certo tipo de fruta é produzido em 2 distritos, A e B. Ambas as áreas às vezes são atacadas por uma praga (mariposa que ataca as frutas).

Suponha que as probabilidades são

$$P(A) = 1/10, P(B) = 1/20, P(A \text{ e } B) = 1/50$$

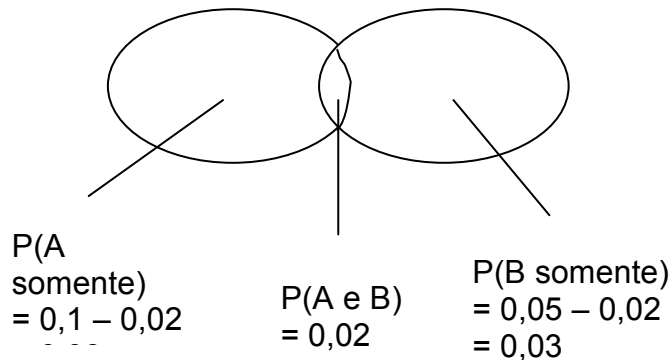
Qual é a probabilidade de que um ou outro (ou ambos) distrito estão infetados em um determinado momento?

$$P(A \text{ ou } B) = P(A) + P(B) - P(A \text{ e } B)$$

$$= 0.1 + 0.05 - 0.02$$

$$= 0.13$$

Alternativamente, considere partes mutuamente exclusivas



A ou B consiste em 3 partes mutuamente exclusivas: A somente, B somente, A e B.

$$P(A \text{ ou } B) = P(A \text{ somente}) + P(B \text{ somente}) + P(A \text{ e } B)$$

$$= 0.08 + 0.02 + 0.03 = 0.13 .$$

Dois eventos X e Y são ditos **independentes** se a probabilidade de que X aconteça não é afetada pelo fato de Y acontecer ou não. Pode ser mostrado que isto implica:

$$P(X \text{ e } Y) = P(X)P(Y)$$

Esta é chamado a **regra de multiplicação** para eventos independentes.

### Exemplo - 2 guardas de segurança e o seus aparelhos de controle

Há dois guardas de segurança para um grande estabelecimento. Cada um carrega um aparelho de controle ativado por detectores nos edifícios. O Guarda 1 é consciencioso e está atento ao aparelho 80% do tempo. O Guarda 2 não é tão confiável e só responde ao aparelho 50% do tempo.

Se os guardas relatam independentemente qualquer alerta para a polícia ou o corpo de bombeiros, qual é a probabilidade de que pelo menos um informará um alerta?

Seja X o evento o Guarda 1 relata o alerta.  $P(X) = 0.8$

Seja Y o evento o Guarda 2 relata o alerta.  $P(Y) = 0.5$

São os eventos X e Y mutuamente exclusivos? - Não, ambos podem informar.

X e Y são independentes? - Considere por hipótese que Sim.

$P(\text{no mínimo um Guarda informa})$

$$= P(X \text{ ou } Y)$$

$$= P(X) + P(Y) - P(X \text{ e } Y)$$

Mas  $P(X \text{ e } Y) = P(X) P(Y)$  (independentes)

$$= 0.8 \times 0.5 = 0.4$$

$$\text{assim } P(X \text{ ou } Y) = 0.8 + 0.5 - 0.4 = 0.9$$

Assim embora Y é só fidedigno 50% do tempo, empregá-lo aumenta a probabilidade de informar um alerta.

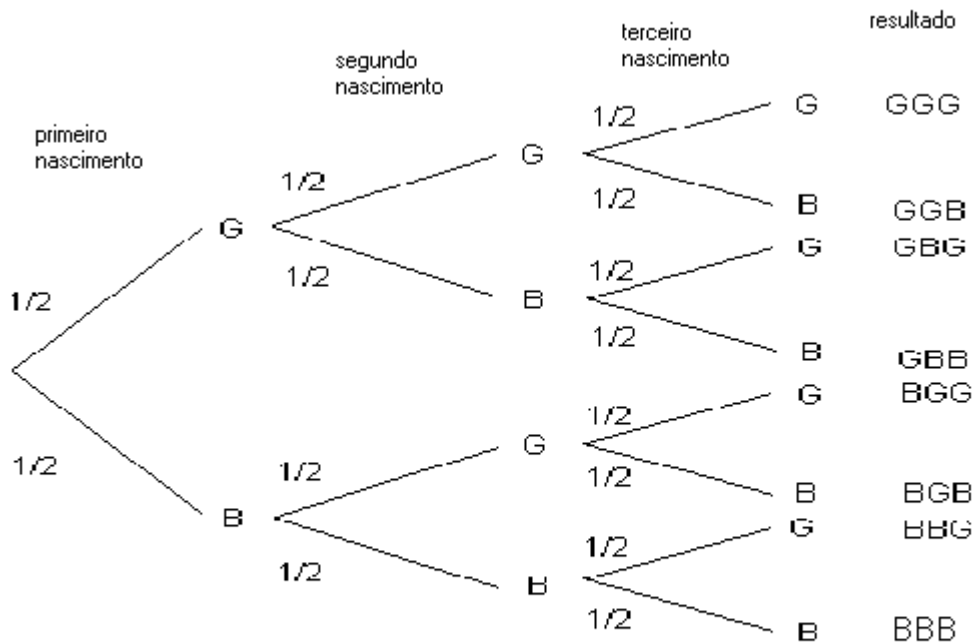
Diagramas de árvore são úteis em cálculos que envolvem várias fases. Cada segmento na árvore é uma fase do problema e as probabilidades nos ramos a partir de cada ponto tem que somar 1. A probabilidade de alcançar o fim de qualquer caminho completo é o produto das probabilidades escritas em seus ramos.

### **Exemplo - Meninos e meninas em uma família de 3 filhos**

Modelo de árvore para meninos (B) e meninas (G) em uma família de tamanho 3.

(ver figura a seguir)

**Figura 1**



Cada caminho representa um resultado ( família de 3 filhos). Há 8 resultados. Se você assume que estes são igualmente prováveis então a probabilidade de cada é  $1/8$ .

por exemplo  $P(BGB) = 1/8$ .

Outro modo de calcular isto é assumir que para cada nascimento

$$P(B) = P(G) = 1/2.$$

Então por exemplo  $P(BGB) = 1/2 \times 1/2 \times 1/2 = 1/8 = 0.125$

i.e. assumindo que sexo é independente dos nascimentos prévios e multiplicando probabilidades ao longo dos ramos da árvore.

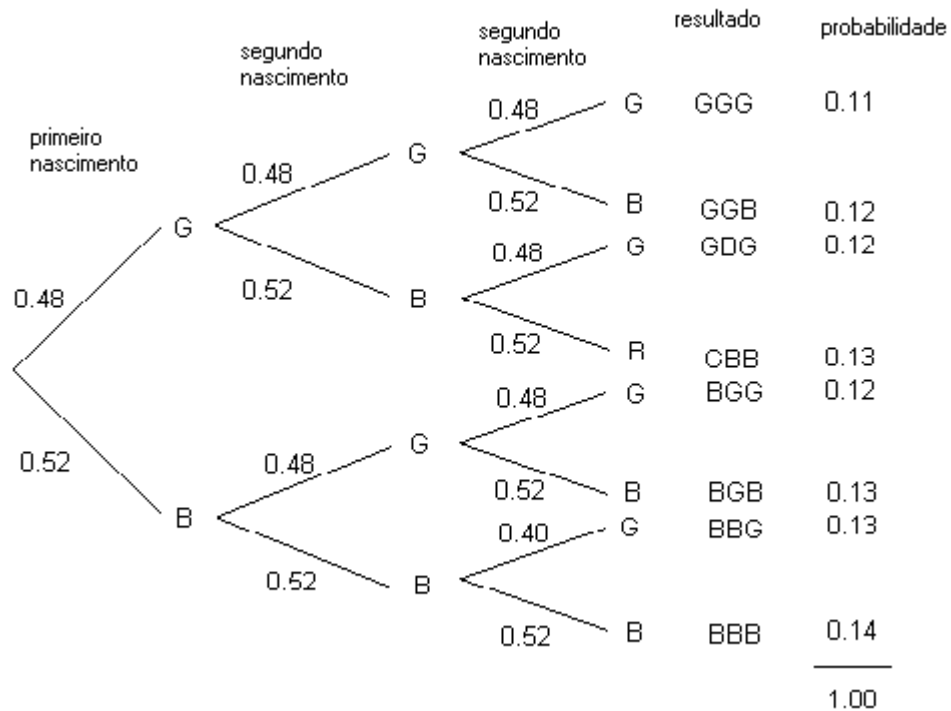
Modelos de árvore são úteis para analisar qualquer processo "passo por passo".

### **Exemplo - Gênero em populações humanas**

Em populações humanas aproximadamente 52% de nascimentos são meninos e 48% são meninas. Assim um modelo mais realista é usar

$$P(B) = 0.52 \quad P(G) = 0.48$$

**Figura 2 - modelo mais realista**



por exemplo  $P(BGB) = 0.52 \times 0.48 \times 0.52 = 0.13$

Um *evento* é qualquer subconjunto de resultados.

Calcule probabilidades para os eventos seguintes que usam o "modelo realista".

C: todas as crianças têm o mesmo sexo

D: menos de 2 meninos

E: C e D  $\implies$  todas meninas

F: C ou D  $\implies$  não 2 meninos



<b>GGG</b>	<b>GGG</b>	<b>GGG</b>	<b>GGG</b>
GGB	<b>GGB</b>	GGB	<b>GGB</b>
GBG	<b>GGB</b>	GGB	<b>GGB</b>
GBB	GBB	GBB	GBB
BGG	<b>BGG</b>	BGG	<b>BGG</b>
BGB	BGB	BGB	BGB
BBG	BBG	BBG	BBG
BBG	BBG	BBG	BBG
<b>BBB</b>	BBB	BBB	<b>BBB</b>
C	D	E	F

$$P(C) = P(GGG) + P(BBB) = 0.11 + 0.14 = 0.25$$

$$P(D) = 0.11 + 3 \times 0.12 = 0.47$$

$$P(E) = P(C \text{ e } D) = P(GGG) = 0.11$$

$$P(F) = P(C \text{ ou } D) = 0.11 + 3 \times 0.12 + 0.14 = 0.61$$

Os eventos C e D não são mutuamente exclusivos (disjuntos) porque o resultado GGG está em ambos. C e D podem acontecer simultaneamente.

Então  $P(C \text{ ou } D) = P(F)$  não é igual a  $P(C) + P(D)$ , porque isto contaria o resultado comum (GGG) duas vezes.

[compare isto com a regra de adição para probabilidades de eventos mutuamente exclusivos].

Ao invés, use a regra mais geral para  $P(C \text{ ou } D)$

$$P(C \text{ or } D) = P(C) + P(D) - P(C \text{ and } D)$$

$\uparrow$   
GGG,  
BBB

$\uparrow$   
GGG, GGB,  
GBG, GBB

$\uparrow$   
GGG

$$= 0.25 + 0.47 - 0.11$$

$$= 0.61 \text{ como requerido}$$

### **Anexo 3 - Probabilidade Condicional**

A probabilidade de um evento A pode ter que ser recalculada se nós sabemos com certeza que outro evento B já aconteceu e A e B não são independentes.

#### **Exemplo - Uma família de 3 crianças**

Em uma família de 3 crianças suponha se sabe que há menos que 2 meninos. Qual é a probabilidade que todas as 3 crianças são do mesmo sexo?

Usando a anotação prévia

C: todas as crianças do mesmo sexo

D: menos que 2 meninos.

Nós queremos a probabilidade de C dado que D aconteceu. Usaremos notação  $P(C|D)$  descrever isto.

	'C'	'D'
	<b>GGG</b>	<b>GGG</b>
	GGB	<b>GGB</b>
Cada coluna lista todo os resultados.	GBG	<b>GBG</b>
Aqueles que incluem o eventos	GBB	GBB
C e D estão em negrito.	BGG	<b>BGG</b>
	BGB	BGB
	BBG	BBG
	<b>BBB</b>	BBB

Como D aconteceu, só 4 resultados são agora possíveis: GGG, GGB, GBG e BGG. As suas probabilidades devem somar 1. Para obter estas probabilidades calculadas previamente elas precisam ser "recalculadas" dividindo pelo seu total que era  $P(D) = 0.47$ . A probabilidade de C, dado que D aconteceu, é chamada de probabilidade condicional e é escrita como  $P(C|D)$ . Lembre-se que a probabilidade de GGG era 0.11:

$$P(C / D) = \frac{P(C \text{ ou } D)}{P(D)} = \frac{0,11}{0,47} = 0,23$$

Em geral para eventos X e Y a probabilidade condicional de X dado que Y aconteceu é

$$P(X / Y) = \frac{P(X \text{ e } Y)}{P(Y)}$$

$P(X|Y) = P(X \text{ e } Y)/P(Y)$  Isto pode ser rearranjado como:

$$P(X \text{ e } Y) = P(X|Y)P(Y)$$

$$P(X \text{ e } Y) = P(Y|X) P(X)$$

### Exemplo - Gênero de empregados

A tabela abaixo mostra as probabilidades de homens (M) e mulheres (F) sendo empregados (E) ou desempregados (U) em alguma população (exclui aqueles que não desejam ser empregado).

	M	F	
E	0.52	0.41	0.93
U	0.05	0.02	0.07
	0.57	0.43	1.00

Ache

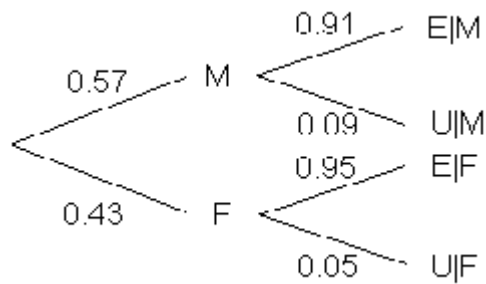
(a)  $P(E|M)$ , a probabilidade condicional de emprego dado que a pessoa é masculina

(b)  $P(M|E)$ , a probabilidade condicional de ser masculino dado que a pessoa é empregada.

Respostas:

$$P(E / M) = \frac{P(E \text{ e } M)}{P(M)} = \frac{0,52}{0,57} = 0,91$$

Figure 3: Modelo de Árvore que mostra probabilidades condicionais



por exemplo  $P(E) = P(E \text{ e } M) + P(E \text{ e } F)$

$$= P(E|M)P(M) + P(E|F)P(F)$$

$$= 0.91 \times 0.57 + 0.95 \times 0.43 = 0.93$$

$$\begin{aligned}
 P(M / E) &= \frac{P(M \text{ e } E)}{P(E)} \\
 &= \frac{P(E / M) \times P(M)}{P(E / M) \times P(M) + P(E / F) \times P(F)} \\
 &= \frac{0,52}{0,93} = 0,56
 \end{aligned}$$

### Independência Revisitada

A regra para a interseção de dois eventos é

$$P(X \text{ e } Y) = P(X)P(Y|X) = P(Y)P(X|Y)$$

Se  $P(X|Y) = P(X)$ , então diríamos que X é independente de Y que a probabilidade de X ocorrer não é afetada se Y acontece ou não. Substituindo isto na equação acima dá  $P(X \text{ e } Y) = P(X) \cdot P(Y)$ , a regra para eventos independentes.

#### ***Anexo 4 – Revisando os conceitos***

Vimos até o momento alguns conceitos referentes à Teoria das Probabilidades. Vamos recordar alguns deles.

A definição clássica de probabilidade diz que a probabilidade de um evento é calculada como a razão existente entre o número de eventos favoráveis a este particular evento e o número de eventos possíveis e equiprováveis. Nesta definição temos que contar com uma situação em que podemos desdobrar o evento (aquele para o qual queremos calcular a sua probabilidade) em diversos sub-eventos. Por exemplo, queremos calcular a probabilidade de que saia um número par em uma jogada de um dado. Temos 3 sub-eventos favoráveis ao resultado (evento) par: 2, 4 e 6 e temos 6 eventos possíveis e equiprováveis: 1,2,3,4,5, e 6. Neste caso a probabilidade do evento par é:

$$Prob(E) = \frac{\text{numero de eventos favoraveis a E}}{\text{numero de eventos possiveis e equiprovaveis}} = \frac{3}{6} = 0,5$$

Se quiséssemos calcular a probabilidade de ocorrência de 2 números pares na jogada (simultânea ou não) de dois dados, o número de eventos favoráveis ao evento  $E = \{\text{saem dois pares}\}$  é 3: os eventos (2,2), (4,4) e (6,6). E temos neste caso 36 eventos possíveis e equiprováveis, pois na jogada de 2 dados podemos ter 36 “combinações” possíveis dentro de dois conjuntos de 1 a 6, ou seja 6 vezes 6 = 36 “combinações”. O espaço amostral de um experimento aleatório é definido como o conjunto de todos os possíveis resultados deste experimento. No experimento de jogar apenas um dado o espaço amostral é  $S = \{1,2,3,4,5,6\}$  e no experimento de jogar dois dados o espaço amostral é  $S = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(2,2),(2,3),\dots,(6,6)\}$ .

A probabilidade de sair dois números pares no lançamento de dois dados é, portanto:

$$\text{Prob}(E) = \frac{\text{numero de eventos favoráveis a } E}{\text{numero de eventos possíveis e equiprováveis}} = \frac{9}{36} = \frac{1}{4} = 0,25 = 25\%$$

Nos dois exemplos anteriores de lançamentos de dados não precisamos realizar os experimentos para calcular as probabilidades. Estas foram calculadas de uma forma “*a priori*” e para isto só precisamos definir os eventos e fazer algumas contagens, tomando-se o cuidado de enumerar completamente o espaço amostral, sem esquecer nenhum evento. Outro cuidado que devemos tomar é que o espaço amostral seja definido em termos de eventos possíveis e equiprováveis, ou seja, todos os eventos que estão no espaço amostral devem ser equiprováveis. Muitas vezes um espaço amostral pode ser definido em termos de uma relação de eventos não equiprováveis. Por exemplo, realizamos o experimento de uma família com duas crianças e contamos o numero de meninas. O espaço amostral pode ser definido neste caso de duas formas. A primeira é  $S_1 = \{0 \text{ meninas}, 1 \text{ menina}, 2 \text{ meninas}\}$  e a segunda é  $S_2 = \{(\text{menino}, \text{menino}), (\text{menino}, \text{menina}), (\text{menina}, \text{menina}), (\text{menina}, \text{menino})\}$ . Observe que no segundo espaço amostral todos os eventos tem a mesma probabilidade que é igual a  $\frac{1}{4}$  (supondo-se que a probabilidade de nascer um menino é igual a probabilidade de nascer uma menina).

Se a probabilidade de nascer um menino é igual a probabilidade de nascer uma menina são iguais, ambas são iguais a  $\frac{1}{2}$ . Agora, para nascer duas meninas em dois nascimentos seguidos (e supondo-se que os resultados destes nascimentos são independentes, um conceito que iremos definir mais precisamente mais adiante) a probabilidade deve ser  $\frac{1}{2}$  vezes  $\frac{1}{2} = \frac{1}{4}$ . O mesmo podemos dizer para qualquer um dos eventos do espaço amostral  $S_2$ , todos tendo probabilidade igual a  $\frac{1}{4}$ . Desta forma podemos dizer que os eventos do espaço amostral  $S_2$  são equiprováveis. Não podemos dizer o mesmo a respeito do espaço amostral  $S_1$ , pois, por exemplo, a probabilidade do evento  $\{0 \text{ meninas}\}$  é igual a  $\frac{1}{4}$ , mas a probabilidade do evento  $\{1 \text{ menina}\}$  é igual a  $\frac{1}{2}$ . O espaço amostral  $S_1$  não é, portanto, definido em termos de eventos equiprováveis.

Na segunda definição de probabilidade consideramos que a mesma é também uma fração, mas neste caso realizamos o experimento para calcular esta probabilidade. Por exemplo,

se desejarmos calcular a probabilidade de sair um número par em um lançamento de um dado, lançamos este dado um numero bastante grande de vezes e contamos quantas vezes saiu resultado par. À medida que o numero de repetições do experimento vai se tornando maior o resultado da divisão entre o numero de resultados favoráveis ao evento (o dado saiu par) e o numero de lançamentos (numero repetições do experimento) vai se aproximando da probabilidade “teórica”  $\frac{1}{2}$ . Podemos enunciar esta definição de forma rigorosa através da seguinte relação:

$$Prob(E) = \lim_{n \rightarrow \infty} \frac{f(E)}{n} = \lim_{n \rightarrow \infty} f_r(E)$$

onde  $f(E)$  é a frequência de ocorrências do evento  $E$  (o numero de vezes em que o evento  $E$  ocorre em  $n$  repetições do experimento) e  $n$  é o numero de repetições do experimento.  $f_r(E)$  é a frequência relativa do evento  $E$ , ou seja, a proporção de vezes em que o evento  $E$  ocorre (em relação ao numero de repetições do experimento). A definição diz a probabilidade de um evento  $E$  é o limite, quanto  $n$  tende ao infinito, da razão entre a frequência de ocorrência do evento  $E$  e o numero de realizações do experimento. A medida que o numero de repetições do experimento vai se tornando maior a razão  $f(E)/n$  vai se aproximando tendencialmente do verdadeiro valor (obtido através da primeira definição) da probabilidade.

Podemos ver, portanto que probabilidade pode ser calculada de duas formas, uma forma “teórica” e uma forma “empírica”. Na primeira não precisamos realizar o experimento para calculá-la. Na segunda, aguardamos o resultado da realização do experimento com uma repetição considerável de vezes para calcular a probabilidade. Qual deve ser o numero de repetições necessário para chegarmos a um valor bem aproximado da probabilidade? Esta é uma questão que não tem uma resposta teórica. Apenas podemos dizer que quanto maior o numero de repetições do experimento mais tendemos a nos aproximar do valor teórico obtido pela primeira definição. Pode ser que a verdadeira probabilidade de um evento seja, digamos, 0,2 e ao repetirmos o experimento 1000 vezes tenhamos como resultado 0,21 e ao repetirmos o experimento 10000 vezes tenhamos

0,18. Isto não quer dizer que na tendência não estejamos nos aproximando do verdadeiro valor. Tal situação é muito difícil de ocorrer.

### **Axiomas da Teoria das Probabilidades**

1)  $0 \leq P(E) \leq 1$

2)  $P(S)=1$

3)  $P(\emptyset) = 0$

4)  $P(A) = 1 - P(\bar{A})$

### **Regras da Teoria das Probabilidades**

1)  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Uma ampliação desta regra para 3 eventos é:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

2)  $P(A \cap B) = P(A) \times P(B|A) = P(B) \times P(A|B)$

Onde  $P(B|A)$  é a probabilidade do evento B condicionada a ocorrência do evento A. Por exemplo, queremos calcular a probabilidade de que um aluno tire nota maior do que 7 se ele estudar. O evento A neste caso é  $A = \{\text{o aluno tira nota maior do que 7}\}$  e o evento B é  $B = \{\text{o aluno estuda}\}$ . Temos neste exemplo um típico caso de eventos dependentes, pois certamente a ocorrência de A irá depender da ocorrência de B. A probabilidade do evento A depende da ocorrência (ou não ocorrência) do evento B. De forma corriqueira, a probabilidade do evento A (o aluno tirar nota maior do que 7) terá um valor caso ocorra o evento B (o aluno estuda) e terá outro valor caso não ocorra B (o aluno não estuda). Se os dois eventos A e B são independentes a probabilidade do evento A não depende da ocorrência do evento B, e vice versa, a probabilidade do evento B não depende da



ocorrência do evento A. No conhecido exemplo do “efeito borboleta” suponhamos que uma borboleta bata asas no coração do Estado de Minas Gerais e ocorra um terremoto em Tóquio. Certamente estes dois eventos são independentes e sua ocorrência depende unicamente da vontade divina.

Se quisermos calcular a probabilidade da ocorrência simultânea de dois eventos independentes basta multiplicar as probabilidades dos dois eventos individuais. Por exemplo, suponhamos que um avião tem 4 motores e desejamos calcular a probabilidade de que ocorra uma falha simultânea em todos os motores. Se supusermos que a ocorrência de falhas nos quatro motores são eventos independentes, podemos dizer que a ocorrência de falha simultânea nos 4 será igual ao produto de suas probabilidades individuais (que neste caso iremos supor que são todas as mesmas). Se a probabilidade de qualquer um dos 4 motores falhar é igual a 0,001, a probabilidade dos quatro motores falharem simultaneamente será  $0,001^4 = (10^{-3})^4 = 10^{-12}$ .

A suposição de independência entre dois eventos é uma hipótese bastante forte. Muitas vezes podemos estar lidando com uma quase independência, mas que não é na prática independência. No exemplo anterior, bastaria que a mesma equipe de mecânicos realizasse a manutenção dos 4 motores para que não houvesse uma independência perfeita entre os 4 eventos.

Vimos anteriormente que pode ocorrer que dois eventos A e B não tenham intersecção, ou seja,  $A \cap B = \emptyset$ . Neste caso, dizemos que os dois eventos A e B são eventos mutuamente exclusivos ou eventos disjuntos. Uma confusão freqüente em teoria das probabilidades é assumir que dois eventos mutuamente exclusivos são necessariamente independentes. Independência de dois eventos não implica necessariamente a condição de serem mutuamente exclusivos. E vice-versa, a condição de dois eventos serem mutuamente exclusivos não implica necessariamente os mesmos eventos serem independentes. Se dois eventos A e B são mutuamente exclusivos então podemos dizer que  $P(A \cap B) = P(\emptyset) = 0$ . Mas não podemos dizer que  $P(A \cap B) = P(A) \times P(B)$ . Seria assim se pelo menos uma das probabilidades dos dois eventos fossem nulas, ou seja,

$P(A)=P(B)=0$ . Certamente existe um único caso em que os dois eventos A e B são iguais ao conjunto vazio e assim  $P(A \cap B) = P(A) \times P(B)$ .

Podemos dizer também que existem muitos casos de eventos independentes para os quais  $A \cap B \neq \emptyset$ , ou seja, estes eventos têm uma intersecção não vazia. O exemplo da borboleta e do terremoto é um caso típico. Os dois eventos podem ocorrer simultaneamente e também são independentes. Mas também podemos dizer que existem eventos mutuamente exclusivos e que são também independentes. Neste caso a probabilidade da ocorrência simultânea dos dois eventos será nula, pois sua intersecção é o conjunto vazio. E se são também independentes necessariamente  $P(A \cap B) = P(A) \times P(B)$  e para que isto ocorra pelo menos um dos dois eventos (ou ambos) tem probabilidade nula.

De fato, a única maneira que dois eventos podem ser ambos mutuamente exclusivos e independentes é se no mínimo um deles tem probabilidade igual à zero. Se A e B são mutuamente exclusivos, sabendo-se que B ocorreu podemos dizer que A não ocorreu. Isto é bastante claro: a probabilidade condicional de A dado B é zero! Isto muda a probabilidade (condicional) de A a não ser que sua probabilidade (não condicional) seja zero. Podemos concluir esta distinção entre eventos mutuamente exclusivos e eventos independentes com duas afirmativas:

- 1) Se dois eventos são mutuamente exclusivos, eles não podem ocorrer na mesma tentativa: a probabilidade de sua intersecção é zero. A probabilidade de sua união é a soma de suas probabilidades.
- 2) Se dois eventos são independentes, ambos podem ocorrer na mesma tentativa (exceto possivelmente, se no mínimo um deles tem probabilidade zero). A probabilidade de sua intersecção é o produto de suas probabilidades. A probabilidade de sua união é menos que a soma de suas probabilidades, a menos que no mínimo um dos eventos tenha probabilidade igual à zero.

Um ultimo ponto sobre a interpretação de uma probabilidade condicional. Quando dizemos  $P(A|B)$  estamos nos referindo a probabilidade de ocorrer o evento A sabendo-se que ocorreu o evento B. Se A e B são eventos mutuamente exclusivos é evidente que  $P(A|B) = 0$ , pois A e B não podem ocorrer simultaneamente. A probabilidade condicional pode ser interpretada como uma probabilidade calculada em um espaço amostral restrito. Quando dizemos  $P(A|B)$  o espaço amostral passa a ser B (anteriormente quando dizíamos simplesmente  $P(A)$ , ou seja, uma probabilidade não condicional, o espaço amostral para o calculo desta probabilidade era S, o espaço amostral original do experimento que estamos considerando). Vamos supor um exemplo numérico em que  $S = \{1,2,3,4,5,6,7,8,9,10\}$ ,  $A = \{3,4,5,6\}$   $B = \{5,6,7,8\}$ . Suponhamos que todos os algarismos de S possam ser selecionados com a mesma probabilidade de 1/10. Portanto neste caso  $P(A) = 4/10$ ,  $P(B) = 4/10$ ,  $A \cap B = \{5,6\}$ ,  $P(A \cap B) = 2/10$ .

$$\text{Pela regra do produto } P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{2/10}{4/10} = 1/2$$

Mas podemos interpretar a coisa da seguinte forma: quando calculamos  $P(A|B)$  o novo espaço amostral passa a ser B. Se aplicarmos a definição clássica de probabilidade temos que:

$$P(A|B) = \frac{\text{numero de eventos favoraveis a A e que pertecem tambem a B}}{\text{numero de eventos possiveis e equiprovaveis de B}} = \frac{2}{4} = 1/2$$

Já  $P(A)$ , a probabilidade de A não condicional é calculada como:

$$P(A) = \frac{\text{numero de eventos favoraveis a A}}{\text{numero de eventos possiveis e equiprovaveis de S}} = \frac{4}{10}$$

Se dois eventos A e B são independentes então  $P(A|B) = P(A)$ ,  $P(B|A) = P(B)$  e  $P(A \cap B) = P(A).P(B)$

Podemos generalizar isto dizendo que se  $k$  eventos  $A_i$ ,  $i=1,2,...,k$  são todos independentes entre si, então:

$$P(A_1 \cap A_2 \cap \dots \cap A_k) = P(A_1).P(A_2)....P(A_k)$$

Se os eventos não fossem independentes a probabilidade simultânea de todos os  $k$  eventos seria:

$$P(A_1 \cap A_2 \cap \dots \cap A_k) = P(A_1).P(A_2|A_1).P(A_3|A_1 \cap A_2)...P(A_k|A_1 \cap A_2 \cap \dots \cap A_{k-1})$$

### Solução de Exercícios de Probabilidade

- 1) Durante o mês de agosto a probabilidade de chuva em um dia determinado é de  $4/10$ . O Fluminense ganha um jogo em um dia com chuva com probabilidade de  $6/10$  e em um dia sem chuva com probabilidade de  $4/10$ . Sabendo-se que o Fluminense ganhou um jogo naquele dia de agosto, qual a probabilidade de que choveu neste dia?

Enumeramos os seguintes passos metodológicos para a solução de um problema de probabilidades:

- Enunciar o experimento aleatório que está sendo tratado.
- Definir o espaço amostral referente ao experimento aleatório.
- Definir os eventos relevantes do problema e apresentar os dados do problema e as perguntas em termos de expressões de probabilidades
- Aplicar as regras e princípios da Teoria das Probabilidades

No caso deste exercício, o experimento aleatório refere-se a jogos de futebol com um determinado time observando-se as condições de tempo (se chove ou não) e qual é o resultado observado do jogo. O espaço amostral deste experimento pode ser definido como segue:

$$S = \{F \cap C, \bar{F} \cap C, F \cap \bar{C}, \bar{F} \cap \bar{C}\}$$

onde os eventos relevantes do problema são:

$F = \{\text{o Fluminense ganha o jogo}\}$

$C = \{\text{chove no dia do jogo}\}$

Os dados deste problema (apresentados no enunciado) são:

$$P(C) = 0,4$$

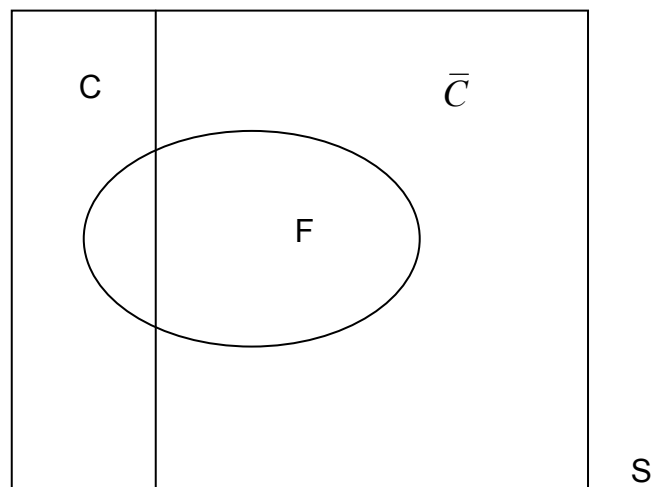
$$P(F/C) = 0,6$$

$$P(F/\bar{C}) = 0,4$$

E a pergunta feita no enunciado do problema é:

$$P(C/F) = ?$$

Façamos um Diagrama de Venn para auxiliar na compreensão do problema:



Podemos ver que a figura oval corresponde ao evento  $F$  e a parte desta figura que está à esquerda do traço vertical (que divide o evento  $C$  e  $\bar{C}$ ) representa o evento  $C \cap F$ , ou seja o conjunto de resultados do espaço amostral em que o Fluminense ganha o jogo em um dia de chuva. À direita do traço vertical, temos a parcela da superfície oval que corresponde ao evento  $\bar{C} \cap F$ , ou seja, o conjunto de resultados em que o Fluminense ganha o jogo em dias que não chove. Observe que devemos fazer uma importante

distinção entre dois eventos diferentes:  $F \cap C$  e  $F/C$ . No primeiro caso temos o evento que corresponde a ocorrência simultânea dos eventos F e C. No segundo caso temos um evento a ocorrência do evento F condicionado a ocorrência do evento C. Podemos perguntar qual é a probabilidade de que o Fluminense ganhe em um dia de chuva. Neste caso queremos saber  $P(F \cap C)$ . Outra coisa é perguntarmos qual é a probabilidade de que o Fluminense ganhe caso ocorra um dia de chuva. Neste caso, estamos perguntado o valor de  $P(F/C)$ . A diferença é bastante sutil, mas ela mostra a distinção entre uma probabilidade de um evento simultâneo e a probabilidade de um evento condicional.

Podemos perguntar, por exemplo, qual é a probabilidade de que o pneu do meu carro fure se ele é novo (neste caso é a probabilidade de um evento conjunto ou simultâneo) e qual é a probabilidade de que ele fure sabendo-se que ele é novo (neste caso temos a probabilidade de um evento condicional).

Continuemos o exemplo do jogo de futebol. O que significa a expressão  $P(C) = 0,4$ ?

Significa que de cada cem partidas que o Fluminense joga aproximadamente em 40 chovem. O que significa a expressão  $P(F/C)=0,6$ ? Significa que de cada 100 partidas em que o Fluminense joga com chuva, em 60 destas partidas chuvosas ele ganha. O que significa a expressão  $P(F/\bar{C})=0,4$ . Significa que em 100 partidas que o Fluminense joga em dias não chuvosos, em aproximadamente 40 ele ganha. Agora, se quisermos calcular  $P(F \cap C)$ , procedemos da seguinte forma. Aplicando a regra do produto temos que:

$$P(F \cap C) = P(C).P(F/C) = 0,4 \times 0,6 = 0,24$$

O que significa esta ultima expressão? Significa que em cada cem partidas que o Fluminense joga (incluindo-se nestas 100 partidas dias chuvosos e não chuvosos) em aproximadamente 24 partidas o Fluminense ganha em dias chuvosos. Repare que esta probabilidade é bem distinta de  $P(F/C)=0,6$ , pois aqui estamos afirmando que em cada 100 partidas que o Flu joga em dias chuvosos (estamos condicionando a este tipo de dia) em aproximadamente 60 destas partidas o time ganha.

Continuemos a solução do problema. Reparem que a metade esquerda do ovo mais a metade direita formam o evento F (o Flu ganha o jogo). Isto pode ser dito assim:

$$F = (F \cap C) \cup (F \cap \bar{C})$$

Esta expressão diz que o evento {o Flu ganha} é formado pela união de dois eventos: o Flu ganha em dia com chuva e o Flu ganha em dia sem chuva. Observem pelo Diagrama de Venn que estes dois eventos são disjuntos, ou seja, eles tem intersecção igual ao conjunto vazio. Isto ocorre porque ou o Flu ganha em um dia chuvoso ou ele ganha em um dia não chuvoso e não existe meio-termo. Se dois eventos são iguais as suas probabilidades também serão iguais. Então podemos dizer que:

$$P(F) = P((F \cap C) \cup (F \cap \bar{C}))$$

Sabemos que a probabilidade do conjunto união de dois eventos disjuntos é a soma das probabilidades de cada um dos eventos unidos. Então:

$$P(F) = P(F \cap C) + P(F \cap \bar{C})$$

Vamos aplicar o teorema do produto aos dois termos do lado direito da expressão acima.

$$P(F) = P(C).P(F / C) + P(\bar{C}).P(F / \bar{C})$$

Todas as probabilidades do lado direito da expressão acima são dados do problema.

$$P(F) = 0,4 \times 0,6 + 0,6 \times 0,4 = 0,48$$

Então temos aqui que de cada 100 partidas (não importando se chove ou não, pois neste caso estamos lidando com uma probabilidade não condicional) em aproximadamente 48 o Flu ganha.

Agora, estamos em condição de dar uma resposta a pergunta do problema. Pelo teorema do produto, sabemos que:

$$P(F \cap C) = P(F) \times P(C/F)$$

Portanto, manipulando os termos desta equação, temos que:

$$P(C/F) = \frac{P(F \cap C)}{P(F)} = \frac{0,24}{0,48} = 0,50$$

que é a resposta do problema. Um dos segredos para um bom entendimento da Teoria das Probabilidades é não nos contentarmos simplesmente com a solução de um problema, mas continuarmos a esmiuçar o seu resultado tentando melhor compreendê-lo e interpretar o que conseguimos alcançar. Descobrimos que  $P(C/F) = 0,50$  o que significa que de cada 100 partidas e que o Flu ganha o jogo, em aproximadamente 50 destas partidas chove. Isto é bem diferente de dizer que  $P(F/C) = 0,6$ , pois aqui estamos afirmando que em 100 partidas em dia chuvosos o Flu ganha em aproximadamente 60 partidas. Poderíamos também não satisfeitos com a simples solução do problema, perguntarmos qual é o valor de  $P(\bar{C}/F)$ ,  $P(\bar{C}/\bar{F})$  e  $P(\bar{C} \cap \bar{F})$ .

Observe que  $C/F$  e  $\bar{C}/F$  são dois eventos complementares. Quando consideramos um evento C condicionado a um evento F estamos considerando que C ocorreu se F ocorreu. O espaço amostral aqui deixa de ser S (o espaço amostral original do problema) e passa a ser F. Agora estamos apenas considerando os jogos em que o Fluminense ganha (e nos abstraindo – retirando fora – dos jogos em que ele perde). Então, quando estamos dizendo  $C/F$  e  $\bar{C}/F$  estamos nos referindo a duas possibilidades que perfazem todos os resultados em que o Flu ganha. Em uma parte destas possibilidades (eventos) o jogo é em dia de chuva e em outra parte é em dia sem chuva. Vimos pela solução do problema que o Flu ganha em 48 % dos jogos. Destes 48 jogos em 100, uma parte deles ocorre em dias de chuva e outra parte ocorre em dias sem chuva. Então quando dizemos  $C/F$  estamos nos referindo aos eventos que são parte das 48 partidas que o Flu ganha e nas quais ocorre



chuva. A outra parcela corresponde aos dias não chuvosos em que o Flu ganha o jogo. Desta forma podemos compreender que estes dois eventos são complementares, ou seja,  $(C/F) \cap (\bar{C}/F) = F$ . Quando dois eventos unidos perfazem a totalidade do espaço amostral dizemos que eles são complementares em relação a este espaço amostral. Da mesma forma podemos dizer que  $F$  e  $\bar{F}$  são complementares em relação ao espaço amostral  $S$  como também o são  $C$  e  $\bar{C}$ . Já que  $C/F$  e  $\bar{C}/F$  são complementares, temos que:

$$P(\bar{C}/F) = 1 - P(C/F) = 1 - 0,50 = 0,50$$

Observando o Diagrama de Venn podemos dizer que:

$$\bar{C} = (F \cap \bar{C}) \cup (\bar{F} \cap \bar{C})$$

Em linguagem corriqueira, a expressão acima diz que as partidas que o Flu joga em dias não chuvosos são constituídas por dois grupos: o grupo das partidas em dias não chuvosos em que ele ganha e o grupo das partidas em dias não chuvosos em que ele perde. Fazendo o mesmo procedimento já utilizado de tomar probabilidades de ambos os termos da equação, temos:

$$P(\bar{C}) = P((F \cap \bar{C}) \cup (\bar{F} \cap \bar{C}))$$

Aplicando a regra da soma das probabilidades, temos:

$$P(\bar{C}) = P(F \cap \bar{C}) + P(\bar{F} \cap \bar{C})$$

Aplicando agora a regra do produto aos dois termos do membro direito da equação acima, temos:

$$P(\bar{C}) = P(F).P(\bar{C}/F) + P(\bar{F} \cap \bar{C})$$

$$\text{ou } P(\bar{F} \cap \bar{C}) = P(\bar{C}) - P(F).P(\bar{C}/F) = 0,60 - 0,48 \times 0,50 = 0,36$$

Novamente, pelo teorema do produto temos que:

$$P(\bar{C}/\bar{F}) = \frac{P(\bar{C} \cap \bar{F})}{P(\bar{F})} = \frac{0,36}{0,52} = 0,6923$$

Ou seja, em aproximadamente 70 % dos jogos em que o Flu perde ocorre de não chover.

Então temos a seguinte “contabilidade final” para o problema. De cada 1000 partidas em que o Flu joga, em 400 chove e em 600 não chove, pois  $P(C) = 0,4$ . Em 480 o Flu ganha e em 520 ele perde, pois  $P(F) = 0,48$ . Das 400 partidas chuvosas, em  $400 \times 0,6 = 240$  partidas o Flu ganha, pois  $P(F/C) = 0,6$  e das 600 partidas não chuvosas, em  $600 \times 0,4 = 240$  partidas ele ganha pois  $P(F/\bar{C}) = 0,6$ . Das 480 partidas em que o Flu ganha em 240 chove, pois  $P(C/F) = 0,5$ . Então temos o seguinte esquema:

	$F$	$\bar{F}$	
$C$	240	160	400
$\bar{C}$	240	360	600
	480	520	1000

Veja no quadro acima que todas as probabilidades discutidas e resolvidas no problema estão representadas. Por exemplo, a probabilidade do Flu ganhar está representada pelos 480 em 1000, ou seja, 0,48. A probabilidade condicional do Flu ganhar em um dia chuvoso está representada pelas 240 partidas que o Flu ganha em dias chuvosos em um total de 400 dias chuvosos (primeira célula a esquerda da tabela) o que dá  $240 / 400 = 0,6$

que é  $P(F/C)$ , um dado do problema. Podemos ver também que  $P(\bar{C} \cap \bar{F}) = 0,36$  pois o cruzamento da linha  $\bar{C}$  e  $\bar{F}$  representam 360 partidas em que ocorre simultaneamente do Flu ganhar e chover ao mesmo tempo em um total de 1000 partidas jogadas pelo Flu. Em resumo, na tabela acima as probabilidades dos eventos conjuntos (intersecção de dois eventos) podem ser calculadas como a divisão das células correspondentes e o total geral de partidas jogadas pelo Flu (1000) e as probabilidades condicionais podem ser vistas pela divisão de cada célula e o total das linhas ou colunas. Por exemplo, se quisermos ver quanto é  $P(\bar{C}/\bar{F})$  basta vermos a divisão do elemento linha  $\bar{C}$  com a coluna  $\bar{F}$  que é 360 com o total da coluna  $\bar{F}$  que é 520 o que dá 0,6923. Finalmente as probabilidades não condicionais são obtidas através da divisão entre os totais das linhas ou colunas pelo total geral (1000). Por exemplo, se quisermos calcular a probabilidade de não chover basta dividirmos o total da linha  $\bar{C}$  por 1000 o que dá 0,6.

2) Num exame há 3 respostas para cada pergunta e apenas uma delas é certa. Portanto, para cada pergunta, um aluno tem probabilidade de  $1/3$  de escolher a resposta certa se ele está adivinhando e 1 se sabe a resposta. Um estudante sabe 30 % das respostas do exame. Se ele deu a resposta correta para uma das perguntas, qual é a probabilidade de que a adivinhou?

No caso deste problema qual é o experimento? O experimento pode ser enunciado da seguinte forma: “Um aluno vai responder as perguntas de um exame de três respostas para cada pergunta – sendo apenas uma certa – e ele pode tentar adivinhar ou ele sabe com inteira confiança a resposta certa para cada pergunta”. Este experimento tem muitos resultados para cada uma de suas realizações. Vamos enumerar estes resultados:

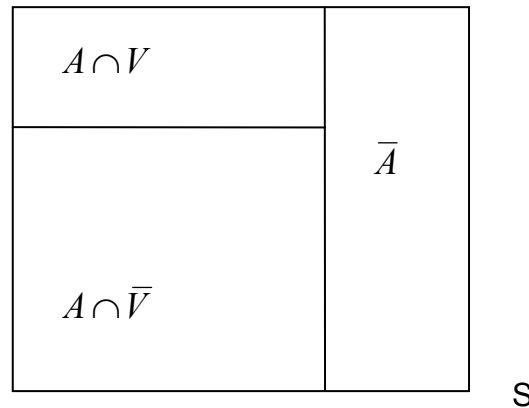
$$S = \{V \cap A, \bar{V} \cap A, \bar{A}\}$$

Onde:

$V = \{\text{o estudante acerta a questão}\}$

$A = \{\text{o estudante tenta adivinhar a questão}\}$

Observem que quando ocorre o evento  $\bar{A}$ , ou seja, o estudante sabe com segurança a resposta certa da pergunta não existe nenhuma possibilidade dele errar. Isto quer dizer que  $\bar{A} \cap \bar{V} = \emptyset$ . Por isso, no espaço amostral acima somente colocamos  $\bar{A}$  e não desmembramos este evento como fizemos com o evento A. Vejamos agora a coisa em termos de uma representação espacial através do Diagrama de Venn:



Quais são os dados do problema? Se em cada questão do exame há apenas uma resposta certa em cada três afirmativas, então  $P(V / A) = 1/3$ . Se o estudante sabe 30 % das respostas do exame, então  $P(\bar{A}) = 0,30$ . E se existe certeza absoluta quando ele sabe a resposta podemos dizer que  $P(V / \bar{A}) = 1$ . A pergunta do problema é:  $P(A / V) = ?$ .

Tentando resolver este problema de forma “intuitiva”, isto é, sem apelar para nossos conhecimentos de Teoria das Probabilidades (sem aplicar as regras, teoremas e princípios discutidos em sala de aula) podemos perceber que de 100 questões do exame, supondo-se que seja um super-exame, 70 questões o aluno tenta adivinhar e 30 questões ele sabe. Das 70 questões que ele tenta adivinhar  $1/3$  que é 23,33 ele acerta e  $70 - 23.33 = 46.67$  ele

erra. Portanto, no total ele acerta  $30 + 23.33 = 53.33$  questões e erra 46.67 questões. Se eu quero saber quanto é  $P(A/V)$  basta saber quantas respostas adivinhadas corresponde em percentagem do total que ele acerta. Do total que ele acerta (53.33) em 23.33 ele tentou (com final feliz) adivinhar. Portanto,  $P(A/V) = \frac{23.33}{53.33} = 0,4375 = 43,75\%$ .

Mas o professor de estatística é um “espírito de porco engarrafado” e não dará o gostinho ao aluno para que ele exercite a sua poderosa intuição e vai exigir que ele resolva “formalmente” o problema.

Sabemos pelo teorema do produto que  $P(A/V) = \frac{P(A \cap V)}{P(V)}$ . Sabemos também, ao observar o diagrama de Venn acima que  $V = (A \cap V) \cup \bar{A}$ . E que portanto,

$$P(V) = P(A \cap V) + P(\bar{A})$$

Aplicando novamente o Teorema do Produto para a expressão acima, temos que:

$$P(V) = P(A) \times P(V/A) + P(\bar{A}) = 0,70 \times 1/3 + 0,30 = 0,5333$$

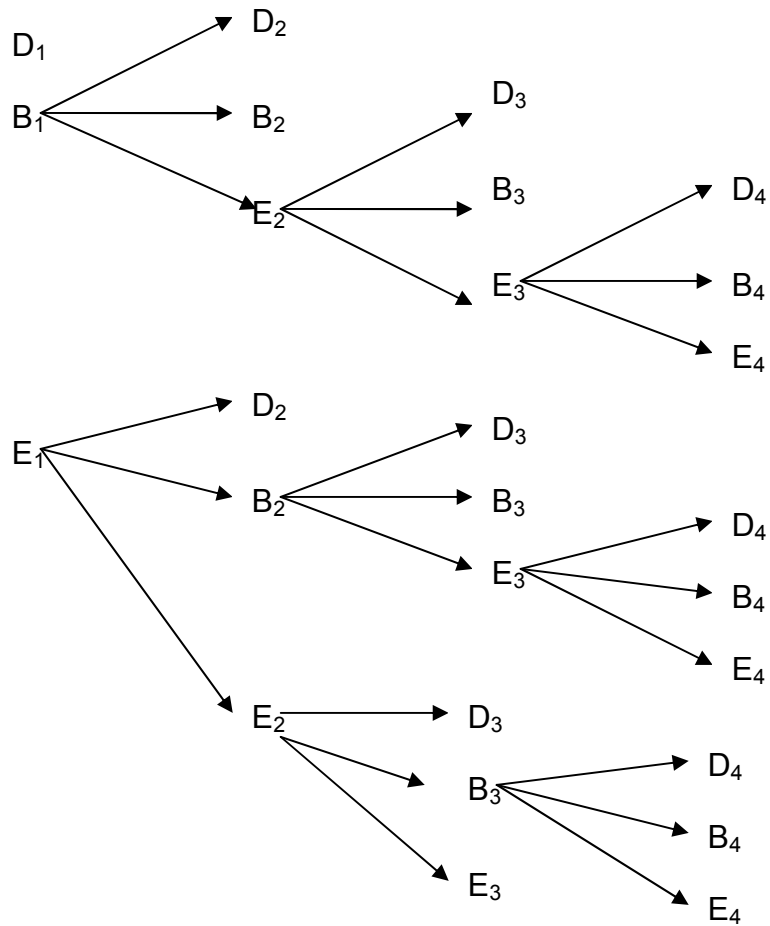
$$\text{Portanto } P(A/V) = \frac{P(A \cap V)}{P(V)} = \frac{0,70 \times 1/3}{0,5333} = 0,4375$$

Não satisfeitos novamente com a simples solução do problema (porque somos insaciáveis estudantes de Estatística) vamos tentar responder a outras probabilidades de eventos e interpretar mais profundamente tudo que fizemos. Qual é a probabilidade do aluno saber a questão dado que ele acertou, ou seja,  $P(\bar{A}/V)$ ? E qual é a probabilidade de ocorrerem simultaneamente os eventos A e V? E os eventos A e  $\bar{V}$ ? Como poderíamos interpretar estas duas probabilidades e como poderíamos distingui-las de probabilidades condicionais “semelhantes” a elas? Como poderíamos construir uma tabela semelhante ao

exercício anterior em que fosse mais fácil, visualizar as diversas probabilidades de eventos conjuntos, de eventos condicionais e eventos não condicionais? Deixo ao aluno criativo e amante da Estatística a resposta a estas questões que discutiremos em sala de aula. Talvez esta ciência não tenha a forma concreta de um objeto comum de amor, mas dada a sua coerência e lógica ela atrai muito aqueles que sempre procuram à verdade. Sócrates, Platão e muitos outros pensadores que não tiveram a oportunidade de virem tudo isto, talvez hoje estejam no seu Hades observando com imenso prazer as nossas belas tentativas.

- 2) Um simples míssil de certa variedade tem uma probabilidade de  $\frac{1}{4}$  de derrubar um bombardeiro, uma probabilidade de  $\frac{1}{4}$  de danificá-lo e uma probabilidade de  $\frac{1}{2}$  de errá-lo. Além disso, dois tiros danificadores derrubarão o avião. Se quatro destes mísseis são lançados, qual é a probabilidade de derrubar um avião?

Podemos definir diversas seqüências de quatro lançamentos. Se chamarmos o evento  $D = \{\text{o míssil derruba o bombardeiro}\}$ , o evento  $B = \{\text{o míssil danifica o bombardeiro}\}$  e o evento  $E = \{\text{o míssil erra o bombardeiro}\}$  teoricamente temos  $3^4$  seqüências de 4 tiros, ou, seja 81 seqüências. As seqüências que tiverem dois ou mais tiros danificadores ou apenas um tiro que derruba o avião são consideradas os eventos favoráveis a derrubar o avião.



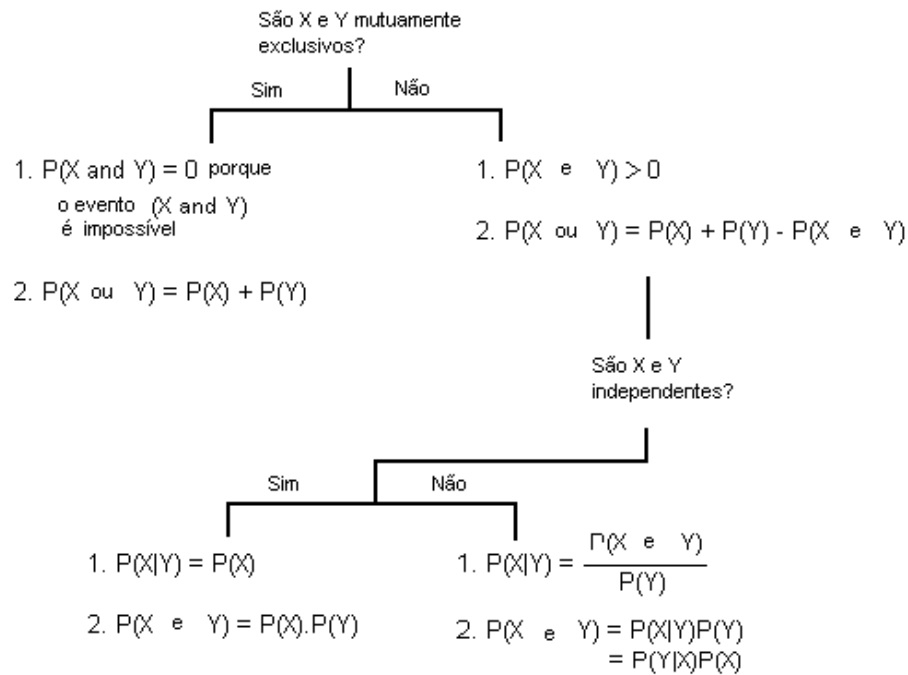
Então, para este experimento temos o seguinte espaço amostral:

$S = \{D_1, B_1D_2, B_1B_2, B_1E_2D_3, B_1E_2B_3, B_1E_2E_3D_4, B_1E_2E_3B_4, B_1E_2E_3E_4, E_1D_2, E_1B_2D_3, E_1B_2B_3, E_1B_2E_3D_4, E_1B_2E_3B_4, E_1B_2E_3E_4, E_1E_2D_3, E_1E_2B_3D_4, E_1E_2B_3B_4, E_1E_2B_3E_4, E_1E_2E_3D_4, E_1E_2E_3B_4, E_1E_2E_3E_4\}$

Com exceção das seqüências  $B_1E_2E_3E_4$ ,  $E_1B_2E_3E_4$ ,  $E_1E_2B_3E_4$ ,  $E_1E_2E_3B_4$  e  $E_1E_2E_3E_4$ , todas as demais derrubam o avião. Portanto, utilizando o evento complementar e considerando os resultados de cada tiro independentes, temos:

$$P(\text{derrubar}) = 1 - (1/4 \times 1/2 \times 1/2 \times 1/2 + 1/2 \times 1/4 \times 1/2 \times 1/2 + 1/2 \times 1/2 \times 1/4 \times 1/2 + 1/2 \times 1/2 \times 1/2 \times 1/4 + 1/2 \times 1/2 \times 1/2 \times 1/2) = \dots$$

## Resumo do Cálculo de Probabilidades





## Exercícios de Probabilidade

- 3) Três moedas são jogadas simultaneamente. Qual é a probabilidade de obter 2 caras? Qual é a probabilidade de obter pelo menos 2 caras?
- 4) Dois dados são jogados simultaneamente. Calcular a probabilidade de que a soma dos números mostrados nas faces de cima seja 7.
- 5) Dois dados são jogados simultaneamente. Calcular a probabilidade de que o máximo seja maior ou igual a 3.
- 6) Para a Copa do Mundo 24 países são divididos em seis grupos, com 4 países cada um. Supondo que a escolha do grupo de cada país é feita ao acaso, calcular a probabilidade de que dois países determinados A e B se encontrem no mesmo grupo. ( Na realidade a escolha não é feita de forma completamente aleatória).
- 7) Uma loteria tem N números e só um prêmio. Um jogador compra n bilhetes em uma extração. Outro compra só um bilhete em n extrações diferentes. ( Ambos os jogadores apostam portanto a mesma importância). Qual deles tem maior probabilidade de ganhar o prêmio?
- 8) Seis bolas são colocadas em três urnas diferentes. Qual é a probabilidade de que todas as urnas estejam ocupadas?
- 9) Um número entre 1 e 300 é escolhido aleatoriamente. Calcular a probabilidade de que ele seja divisível por 3 ou por 5.
- 10) Um torneio é disputado por 4 times A,B, C e D. É 3 vezes mais provável que A vença do que B, duas vezes mais provável que B vença do que C e é 3 vezes mais provável que C vença do que D. Quais as probabilidades de ganhar para cada um dos times?
- 11) Uma caixa contém 20 peças em boas condições e 15 em más condições. Uma amostra de 10 peças é extraída. Calcular a probabilidade de que ao menos uma peça na amostra seja defeituosa.
- 12) Uma cidade tem 30 000 habitantes e três jornais A, B e C. Uma pesquisa de opinião revela que:  
  
12 000 lêem A;  
8 000 lêem B;  
7 000 lêem A e B;  
6 000 lêem C;  
4 500 lêem A e C;

1 000 lêem B e C;  
500 lêem A,B e C.

Qual é a probabilidade de que um habitante leia:

- a) Pelo menos um jornal;
- b) Só um jornal.

13) Os algarismos 1,2,3,4,5 são escritos em 5 cartões diferentes. Estes cartões são escolhidos (sem reposição) aleatoriamente e os algarismos que vão aparecendo são escritos da esquerda para a direita, formando um número de 5 algarismos.

- a) calcular a probabilidade de que o número escrito seja par
- b) Se a escolha fosse com reposição qual seria a probabilidade?

14) Colocam-se aleatoriamente  $b$  bolas em  $b$  urnas. Calcular a probabilidade de que exatamente uma urna seja deixada desocupada.

15) Dez pessoas são separadas em dois grupos de 5 pessoas cada um. Qual é a probabilidade de que duas pessoas determinadas A e B façam parte do mesmo grupo?

16) 5 homens e 5 mulheres compram 10 cadeiras consecutivas na mesma fila de um teatro. Supondo que se sentaram aleatoriamente nas 10 cadeiras, calcular:

- a) a probabilidade de que homens e mulheres se sentem em cadeiras alternadas;
- b) A probabilidade de que as mulheres se sentem juntas.

17) Um número entre 1 e 200 é escolhido aleatoriamente. Calcular a probabilidade de que seja divisível por 5 ou por 7.

18) Uma moeda foi cunhada de tal forma que é 4 vezes mais provável de dar cara do que coroa. Calcular as probabilidades de cara e coroa.

19) Aos números inteiros entre 1 e  $n$  são designadas probabilidades proporcionais aos seus valores. Calcular  $P(i)$  para  $1 \leq i \leq n$

20) Três dados são jogados simultaneamente. Calcular a probabilidade de obter 12 como a soma dos resultados.

21) Sejam A e B eventos tais que

$$P(A) = \frac{1}{2}, P(B) = \frac{1}{4} \text{ e } P(A \cap B) = \frac{1}{5}$$

*Calcular :*

a)  $P(A \cup B)$

b)  $P(\bar{A})$

c)  $P(\bar{B})$

d)  $P(A \cap \bar{B})$

e)  $P(\bar{A} \cap B)$

f)  $P(\bar{A} \cap \bar{B})$

g)  $P(\bar{A} \cup \bar{B})$

22) No jogo da Sena são sorteadas 6 dezenas distintas entre as dezenas 01 – 02 - ...- 50. O apostador escolhe 6 dessas 50 dezenas e é premiado se são sorteadas 4 (quadra), 5 (quina), 6 (Sena Principal) das dezenas por ele escolhidas ou se as dezenas sorteadas são escolhidas aumentadas (Sena Anterior) ou diminuídas (Sena Posterior) de uma unidade ( $50 + 1 = 01$ ,  $01 - 1 = 50$ ). Determine a probabilidade de uma apostador fazer:

- a) uma quadra
- b) uma quina
- c) a Sena Principal
- d) A Sena Anterior ou a Posterior.

23) No jogo da Loto são sorteadas 5 dezenas distintas entre as dezenas 01 – 02 - ...- 99 - 00. O apostador escolhe 6,7,8,9 ou 10 dezenas e é premiado se são sorteadas 3 (terno), 4 (quadra) ou 5 (quina) das dezenas escolhidas. Determine a probabilidade de uma apostador que escolheu 10 dezenas fazer:

- a) um terno
- b) uma quadra
- c) a quina

24) Na Loteria Esportiva há 13 jogos e o apostador deve indicar em cada um deles a vitória do time 1, a vitória do time 2 ou o empate. Um jogador é premiado:

- a) com 10 pontos, se acerta os resultados dos 10 primeiros jogos e erra os dos 3 últimos;
- b) com 11 pontos, se acerta os resultados dos 10 primeiros jogos e acerta apenas um dos resultados dos 3 últimos;
- c) com 12 pontos, se acerta os resultados dos 10 primeiros jogos e acerta apenas 2 dos resultados dos 3 últimos;

d) com 13 pontos, se acerta os resultados dos 13 jogos.

Supondo que em cada jogo os resultados possíveis tenham probabilidades iguais, determine a probabilidade de um apostador ser premiado:

- a) com 10 pontos;
- b) com 11 pontos;
- c) com 12 pontos;
- d) com 13 pontos.

25) Escolhem-se ao acaso duas peças de um dominó. Qual é a probabilidade delas possuírem um número comum?

26) Em um armário há  $n$  pares de sapatos. Retiram-se ao acaso  $p$  pares de sapatos desse armário. Qual a probabilidade de haver entre esses pés exatamente  $k$  pares de sapatos?

27) Colocam-se ao acaso  $n$  botões em um tabuleiro  $n \times n$ , não sendo permitido haver dois botões em uma mesma casa. Qual é a probabilidade de não haver dois botões nem na mesma linha nem na mesma coluna?

28) Um polígono regular de  $2n + 1$  lados está inscrito em um círculo. Escolhem-se 3 dos seus vértices, formando-se um triângulo. Qual é a probabilidade do centro do círculo ser interior ao triângulo?

29) Tem-se  $n$  urnas. Bolas são colocadas ao acaso nas urnas, uma de cada vez, até que alguma urna receba duas bolas. Qual é a probabilidade de colocarmos exatamente  $p$  bolas nas urnas?

30) João e Pedro lançam, cada um, um dado não-tendencioso. Qual é a probabilidade do resultado de João ser maior ou igual ao resultado de Pedro?

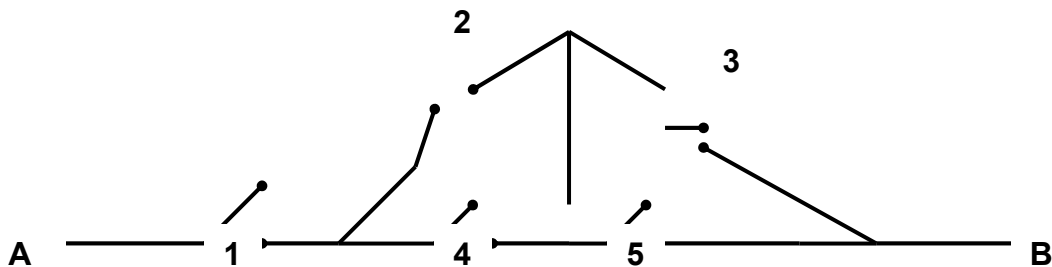
31) Numa prova há 7 perguntas do tipo verdadeiro-falso. Calcular a probabilidade de acertarmos todas as 7 se:

- a) escolhermos aleatoriamente as 7 respostas,
- b) escolhermos aleatoriamente as respostas mas sabendo que há mais respostas “verdadeiro” do que “falso”.

32) Sabe-se que 80 % dos pênaltis marcados a favor do Brasil são cobrados por jogadores do Flamengo. A probabilidade de um pênalti ser convertido é 40 % se o cobrador for do Flamengo e de 70 % em caso contrário. Um pênalti a favor do Brasil acabou de ser marcado:

- a) Qual a probabilidade do pênalti ser cobrado por um jogador do Flamengo e ser convertido?

- b) Qual a probabilidade do pênalti ser convertido?
- c) Um pênalti foi marcado a favor do Brasil e acabou de ser desperdiçado. Qual é a probabilidade de que o cobrador tenha sido um jogador do Flamengo?
- 33) Marina quer enviar uma carta a Verônica. A probabilidade de que Marina escreva a carta é de  $\frac{8}{10}$ . A probabilidade de que o correio não perca é de  $\frac{9}{10}$ . A probabilidade de que o carteiro entregue é de  $\frac{9}{10}$ . Dado que Verônica não recebeu a carta, qual é a probabilidade condicional de que Marina não a tenha escrito?
- 34) Durante o mês de agosto a probabilidade de chuva em um dia determinado é de  $\frac{4}{10}$ . O Fluminense ganha um jogo em um dia com chuva com probabilidade de  $\frac{6}{10}$  e em um dia sem chuva com probabilidade de  $\frac{4}{10}$ . Sabendo-se que o Fluminense ganhou um jogo naquele dia de agosto, qual a probabilidade de que choveu neste dia?
- 35) Num exame há 3 respostas para cada pergunta e apenas uma delas é certa. Portanto, para cada pergunta, um aluno tem probabilidade de  $\frac{1}{3}$  de escolher a resposta certa se ele está adivinhando e 1 se sabe a resposta. Um estudante sabe 30 % das respostas do exame. Se ele deu a resposta correta para uma das perguntas, qual é a probabilidade de que a adivinhou?
- 36) Um jogador deve enfrentar, em um torneio, dois outros A e B. Os resultados dos jogos são independentes e as probabilidades dele ganhar de A e de B são  $\frac{1}{3}$  e  $\frac{2}{3}$  respectivamente. O jogador vencerá o torneio se ganhar dois jogos consecutivos, de uma série de 3. Que série de jogos é mais favorável ao jogador: ABA ou BAB?
- 37) A probabilidade de fechamento de cada relé do circuito apresentado na figura abaixo é igual a  $p$ ,  $0 < p < 1$ .



Se todos os relés funcionam independentemente, qual é a probabilidade de que haja corrente circulando entre os terminais A e B?

- 38) Escolhe-se ao acaso um número entre 1 e 50. Se o número é primo qual é a probabilidade de que seja ímpar?

- 39) Uma moeda é jogada 6 vezes. Sabendo-se que no primeiro lançamento deu coroa, calcular a probabilidade condicional de que o número de caras nos 6 lançamentos supere o número de coroas.
- 40) Uma moeda é jogada 4 vezes. Sabendo que o primeiro resultado foi cara, calcular a probabilidade condicional de obter pelo menos 2 caras.
- 41) Joga-se um dado duas vezes. Calcule a probabilidade condicional de obter 3 na primeira jogada, sabendo que a soma dos resultados foi 7.
- 42) Duas máquinas A e B produzem 3000 peças em um dia. A máquina A produz 1000 peças, das quais 3 % são defeituosas. A máquina B produz as restantes 2000, das quais 1 % são defeituosas. Da produção total em um dia uma peça é escolhida ao acaso e, examinando-a, constata-se que é defeituosa. Qual é a probabilidade de que a peça tenha sido produzida pela máquina A?
- 43) Um estudante resolve um teste do tipo verdadeiro-falso. Ele sabe dar a solução correta para 40 % das questões. Quando ele responde uma questão cuja solução conhece, dá a resposta correta, e nos outros casos decide na cara ou coroa. Se uma questão foi respondida corretamente, qual é a probabilidade que ele sabia a resposta?

44) Sejam A e B dois eventos independentes tais que

$$P(A) = 1/3 \text{ e } P(B) = 1/2$$

Calcule  $P(A \cup B)$ ,  $P(\bar{A} \cup B)$  e  $P(\bar{A} \cap B)$

45) Sejam A e B dois eventos independentes tais que

$$P(A) = 1/4 \text{ e } P(A \cup B) = 1/3$$

Calcule  $P(B)$

46) Uma moeda equilibrada é jogada duas vezes. Sejam A e B os eventos:

A: cara na primeira jogada;

B: cara na segunda jogada

Verifique que A e B são independentes

47) Jogue um dado duas vezes. Considere os eventos:

A = o resultado do 1º lançamento é par;

B = o resultado do 2º lançamento é par;

$C$  = a soma dos resultados é par.

A e B são independentes? e A e C? e B e C? e A, B e C?

48) Uma pessoa com um molho de  $n$  chaves tenta abrir uma porta. Apenas uma das chaves consegue abrir a porta. Qual é a probabilidade dela só conseguir abrir a porta na  $k$ -ésima tentativa:

- a) supondo que após cada tentativa mal sucedida ela descarta a chave usada;
- b) supondo que ela não faz isso.

49) (Problema de Chevalier de Méré) Determine a probabilidade de obter:

- a) ao menos um 6 em 4 lançamentos de um dado;
- b) ao menos um duplo 6 em 24 lançamentos de um par de dados.

50) A probabilidade de um homem ser canhoto é  $1/10$ . Qual é a probabilidade de, em um grupo de 10 homens, haver pelo menos um canhoto?

51) Sacam-se, sucessivamente e sem reposição, duas cartas de um baralho comum (52 cartas). Calcule a probabilidade de a 1ª carta ser uma dama e a 2ª ser de copas.

52) Um exame de laboratório têm eficiência de 95 % para detectar uma doença quando essa doença existe de fato. Entretanto o teste aponta um resultado “falso positivo” para 1 % das pessoas sadias testadas. Se 0,5 % da população tem a doença, qual é a probabilidade de uma pessoa ter a doença dado que seu exame foi positivo?

53) A lança uma moeda  $n+1$  vezes e B lança a mesma moeda  $n$  vezes. Qual é a probabilidade de A obter mais caras que B?

54) Quantas pessoas você deve entrevistar para ter probabilidade igual ou superior a 0,5 de encontrar pelo menos uma que aniversarie hoje?

55) Uma urna contém 3 bolas vermelhas e 7 bolas brancas. A e B sacam alternadamente, sem reposição, bolas dessa urna até que uma bola vermelha seja retirada. A saca a primeira bola. Qual é a probabilidade de A sacar a bola vermelha?

56) Em uma cidade com  $n+1$  habitantes, uma pessoa conta um boato para outra pessoa, a qual por sua vez conta para uma terceira pessoa, etc. Calcule a probabilidade do boato ser contado  $m$  vezes:

- a) sem retornar à primeira pessoa;
- b) sem repetir nenhuma pessoa.

- 57) Sacam-se, com reposição,  $n$  ( $n > 1$ ) bolas de uma urna que contem 9 bolas numeradas de 1 a 9. Qual é a probabilidade do produto dos números das  $n$  bolas extraídas ser divisível por 10?
- 58) Quantas vezes, no mínimo, se deve lançar um dado não tendencioso para que a probabilidade de obter algum 6 seja superior a 0,9?
- 59) Um júri de 3 pessoas tem dois jurados que decidem corretamente (cada um) com probabilidade  $p$  e um terceiro jurado que decide por cara ou coroa. As decisões são tomadas por maioria. Outro júri tem probabilidade  $p$  de tomar uma decisão correta. Qual dos júris tem maior probabilidade de acerto?
- 60) Um dia você captura 10 peixes em um lago, marca-os e coloca-os no lago novamente. Dois dias após, você captura 20 peixes no mesmo lago e constata que 2 desses peixes haviam sido marcados por você.
- se o lago possui  $k$  peixes, qual era a probabilidade de, capturando 20 peixes, encontrar dois peixes marcados?
  - para que valor de  $k$  essa probabilidade é máxima?
- 61) Qual é a probabilidade de, em um grupo de 4 pessoas:
- haver alguma coincidência de signos zodiacais?
  - as quatro terem o mesmo signo?
  - duas terem o mesmo signo, e as outras duas, outro signo?
  - três terem o mesmo signo e, a outra, outro signo?
  - todas terem signos diferentes?
- 62) Deseja-se estimar a probabilidade  $p$  de um habitante de determinada cidade ser um consumidor de drogas. Para isso realizam-se entrevistas com alguns habitantes da cidade. Não se deseja perguntar diretamente ao entrevistado se ele usa drogas, pois ele poderia se recusar a responder ou, o que seria pior, mentir. Adota-se então o seguinte procedimento: propõe-se ao entrevistado duas perguntas do tipo SIM-NÃO:
- Você usa drogas?
  - Seu aniversário é anterior ao dia 2 de julho?

Pede-se ao entrevistado que jogue uma moeda, longe das vistas do entrevistador, e que se o resultado for cara, responda à primeira pergunta e, se for coroa, responda à segunda pergunta.

- sendo  $p_1$  a probabilidade de um habitante da cidade responder sim, qual é a relação entre  $p$  e  $p_1$ ?
- se forem realizadas 1000 entrevistas e obtidos 600 sim é razoável imaginar que  $p_1 \approx 0,6$ . Qual seria, então, sua estimativa de  $p$ ?



- 63) Uma firma fabrica “chips” de computador. Em um lote de 1000 “chips”, uma amostra de 10 “chips” revelou 1 “chip” defeituoso. Supondo que no lote houvesse  $k$  “chips” defeituosos:
- Calcule a probabilidade de em uma amostra de 20 “chips” haver exatamente 1 “chip” defeituoso.
  - Determine o valor de  $k$  que maximiza a probabilidade calculada no item a).
- 64) Jogamos uma moeda não viciada 10 vezes. Qual é a probabilidade de obtermos exatamente 5 caras?
- 65) Um aluno marca ao acaso as respostas em um teste múltipla-escolha com 10 questões e 5 alternativas por questão. Qual é a probabilidade dele acertar exatamente 4 questões?
- 66) Joga-se uma moeda não viciada. Qual é a probabilidade de serem obtidas 5 caras antes de 3 coroas?
- 67) Lança-se um dado não viciado até a obtenção do terceiro 6. Seja  $X$  o número do lançamento em que isto ocorre. Calcule:
- $P(X = 10)$ ; b)  $P(X > 10)$ ; c)  $P(X = 10)$ .
- 68) Dois adversários A e B disputam uma série de partidas. A probabilidade de A ganhar uma partida é 0,6 e não há empates. Qual é a probabilidade de A ganhar a série?
- 69) Dois adversários A e B disputam uma série de partidas. O primeiro que obtiver 12 vitórias ganha a série. No momento o resultado é 6 x 4 a favor de A. Qual é a probabilidade de A ganhar a série sabendo que em cada partida as probabilidades de A e B vencerem são respectivamente 0,4 e 0,6?
- 70) Motores de avião funcionam independentemente e cada motor tem uma probabilidade  $p$  de falhar durante o voo. Um avião voa com segurança se a maioria de seus motores funciona. Para que valores de  $p$  um avião com 3 motores é preferível a um avião com 5 motores?
- 69) Suponha que uma característica (como a cor dos olhos, por exemplo) dependa de um par de genes. Representemos por A um gen dominante e por a um gen recessivo. Assim um indivíduo com genes AA é dominante puro, um com genes aa é um recessivo puro e um com genes Aa é um híbrido. Dominantes puros e híbridos são semelhantes em relação à característica. Filhos recebem um gen do pai e um da mãe. Suponha que pai e mãe sejam híbridos e tenham 4 filhos.
- Qual é a probabilidade do primeiro filho ser um recessivo puro?
  - Qual é a probabilidade de exatamente um dos 4 filhos ser um recessivo puro?

- 70) (O problema das caixas de fósforos de Banach<sup>18</sup>) Um matemático sai de casa todos os dias com duas caixas de fósforos, cada uma com  $n$  palitos. Toda vez que ele que acender um cigarro, ele pega (ao acaso) uma das caixas e retira daí um palito. O matemático é meio distraído, de modo que quando ele retira o último palito de uma caixa, ele não percebe que a caixa está vazia. Como ele fuma muito, em certa hora ele pega uma caixa e constata que ela está vazia. Qual é a probabilidade de nesse momento a outra caixa conter exatamente  $k$  ( $0 \leq k \leq n$ ) palitos?
- 71) Lança-se repetidamente um par de dados não tendenciosos. Qual é a probabilidade de obtermos duas somas iguais a 7 antes de obtermos três somas iguais a 3?
- 72) Uma moeda tem probabilidade 0,4 de dar cara. Lançando-a 12 vezes qual o mais provável valor do número de caras obtidas?
- 73) Suponha que uma variável aleatória  $T$  tem a seguinte distribuição de probabilidade

$T$	0	1	2
$P(T=t)$	0,5	0,3	0,2

- Ache  $P(T \leq 0)$
  - Ache  $P(T \geq 0 \text{ and } T < 2)$
  - Calcule  $E(T)$ , a média da variável aleatória  $T$ .
- 74) Suponha que você escolha uma bola de uma urna contendo 7 bolas vermelhas, 6 bolas brancas, 5 bolas azuis e 4 bolas brancas. Qual é a probabilidade de que você escolha uma bola vermelha?
- 75) Suponha que você escolha uma bola aleatoriamente de uma urna 7 bolas vermelhas, 6 bolas brancas, 5 bolas azuis e 4 bolas amarelas. Qual é a probabilidade de que você escolha uma bola branca?
- 76) Um dado não viciado é jogado duas vezes. Ache a probabilidade de sair um 5 ou 6 no primeiro lance e um 1, 2 ou 3 no segundo lance.
- 77) Ache a probabilidade de não sair um 5 ou 6 em qualquer uma de duas jogadas de um dado não viciado.
- 78) Você tem um baralho de 52 cartas bem embaralhadas. Qual é a probabilidade de escolher dois valetes consecutivos se a primeira carta não é recolocada no baralho?

---

<sup>18</sup> Stefan Banach (1892-1945), matemático polonês

- 79) Uma urna contém 5 bolas vermelhas, 3 bolas brancas e 6 bolas azuis. Determine a probabilidade de que elas sejam escolhidas na ordem azul, branca e vermelha dado que cada bola é recolocada na urna depois de escolhida.
- 80) Uma urna contém 5 bolas vermelhas, 3 bolas brancas e 6 bolas azuis. Determine a probabilidade de que elas sejam escolhidas na ordem azul, branca e vermelha dado que cada bola não é recolocada na urna depois que ela é escolhida.
- 81) A urna A contém 2 bolas vermelhas e 3 azuis. A urna B contém 8 bolas vermelhas e 2 azuis. Você joga uma moeda honesta. Se a moeda mostra cara você escolhe uma bola da urna A. Se a moeda mostra coroa você escolhe uma bola da urna B. Determine a probabilidade de que você escolha uma bola vermelha.
- 82) Você tem 6 bolas, cada uma de cor diferente. De quantas maneiras distintas você pode dispo-las em uma fila?
- 83) De quantas maneiras possíveis 8 pessoas podem sentar-se em um banco se apenas estão disponíveis 3 assentos?
- 84) De quantas maneiras números de 3 algarismos podem ser formados com os dígitos 0,1,2,...,9 se repetições são permitidas?
- 85) De quantas maneiras números de 3 algarismos podem ser formados com os dígitos 0,1,2,...,9 se repetições não são permitidas?
- 86) Três diferentes livros de Ciências, 5 diferentes livros de Inglês e 4 diferentes livros de Economia são arranjados em uma estante. De quantas maneiras é possível dispo-los se todos os livros de cada assunto precisam ficar juntos?
- 87) Três diferentes livros de Ciências, 5 diferentes livros de Inglês e 4 diferentes livros de Economia são arranjados em uma estante. De quantas maneiras é possível dispo-los se somente os livros de Ciências precisam ficar juntos?
- 88) Calcule  $C(8,3)$
- 89) De quantas maneiras pode um comitê de 6 ser escolhido de 10 pessoas?
- 90) A partir de 4 médicos e de 6 enfermeiras, um comitê consistindo de 3 médicos e 4 enfermeiras precisa ser formado. De quantas maneiras isto pode ser feito se um particular médico deve ser incluído e se qualquer enfermeira pode ser incluída?
- 91) A partir de 4 médicos e de 6 enfermeiras, um comitê consistindo de 3 médicos e 4 enfermeiras precisa ser formado. De quantas maneiras isto pode ser feito se uma particular enfermeira não pode ser incluída no comitê?
- 92) De quantas maneiras diferentes saladas de frutas podem ser feitas de maçã, laranja, tangerina e banana?

93) A partir de 6 consoantes e 4 vogais, quantas combinações distintas de letras podem ser feitas?

94) Quais dos seguintes pares de eventos são mutuamente exclusivos?

- a. A: os números pares ;                      B: o número 5;
- b. A: os números ímpares;                      B: os números maiores do que 10;
- c. A: os números menores que 5;              B: todos os números negativos
- d. A: os números maiores do que 100;      B: os números menores do que 200;
- e. A: os números negativos;                      B: os números pares

95) Uma carta é escolhida de um baralho padrão de 52 cartas. Ao descrever a ocorrência de dois possíveis eventos, um Ás e um Rei, estes dois eventos são:

- (a) independentes
- (b) mutuamente exclusivos
- (c) variáveis aleatórias
- (d) aleatoriamente independentes.

96) Suponha que certa característica oftalmológica é associada com a cor dos olhos. 300 indivíduos selecionados aleatoriamente são estudados e apresentam os seguintes resultados:

Característica	Cor dos olhos			
	Azuis	Castanhos	Outra	Total
Sim	70	30	20	120
Não	20	110	50	180
Total	90	140	70	300

A. Qual é a probabilidade de que uma pessoa tenha olhos azuis ?

B. O que você espera que seja o valor de  $P(\text{Ter a característica e olhos azuis})$  se a cor dos olhos e a existência da característica são independentes ?

C. Quais das seguintes expressões descrevem a relação entre os eventos  $A$  = a pessoa tem olhos castanhos e  $B$  = a pessoa tem olhos azuis ? (marque a resposta correta).

i. independente    ii. exaustivo

iii. simples    iv. mutuamente exclusivos

97) Uma amostra de 1000 pessoas diagnosticada com certa doença é distribuída de acordo com a altura e o status (evolução) da doença a partir de um exame clínico de acordo com a seguinte tabela:

	Sem doença	Fraca	Moderada	Severa	Totais
Alta	122	78	139	61	400
Média	74	51	90	35	250
Baixa	104	71	121	54	350
Totais	300	200	350	150	1000

Como você estimaria, a partir dessa tabela, a probabilidade de ser média ou baixa em altura e ter moderado ou severo grau de evolução da doença ?

a.  $600/1000 * 500/1000$     d.  $300/600$

b.  $300/500$     e.  $800/1000$

c.  $300/1000$

98) De cerca de 25 artigos, nove são defeituosos, seis tem defeitos superficiais e três tem defeitos importantes. Determine a probabilidade de que um artigo selecionado aleatoriamente tenha defeitos importantes dado que ele tem defeito.

a.  $1/3$

b. 0,25

c. 0,24

d. 0,08

99) A seguinte tabela de duas entradas mostra as frequências de ocorrência de uma exposição hipotética e a doença em um grupo de 1000 pessoas.

<b>Doença Exposição</b>	Presente	Ausente	Totais
Presente	75	325	400
Ausente	25	575	600
Totais	100	900	1000

- Qual é a probabilidade de exposição no grupo ?
- Qual é a probabilidade conjunta de tanto exposição como de doença estar presente no grupo ?
- Calcule a probabilidade de doença estar presente condicionada a presença de exposição e condicionada a ausência de exposição.

100) Um epidemiologista acredita que as rodovias têm alguma relação com o desenvolvimento de uma nova doença porque a probabilidade de uma pessoa estar morando a menos de uma milha das rodovias, dado que ela tem a doença, é 0,80. Você concorda com ele ? Porque ou porque não ?

101) Um dormitório de um campus universitário abriga 200 estudantes. 120 são homens, 50 são dos graus mais avançados e 40 são homens dos graus mais avançados. Um estudante é selecionado ao acaso. A probabilidade de selecionar um estudante de grau menos elevado, dado que o estudante é mulher, é:

- (a)  $7/8$               (d)  $7/20$   
(b)  $7/15$             (e)  $1/4$   
(c)  $2/5$

102) Uma amostra de 2000 indivíduos é distribuída de acordo com a cor de olho e a presença ou ausência de uma certa característica oftalmológica como segue:

Característica	Cor dos olhos			
	Castanho	Azul	Outro	
Sim	400	270	130	800
Não	200	650	350	1200
Total	600	920	480	2000

Em uma seleção aleatória de um indivíduo da população em estudo, Qual é sua estimativa da probabilidade de:

- a pessoa tem olhos azuis? \_\_\_\_\_
- a característica está presente e a pessoa tem castanhos? \_\_\_\_\_
- a pessoa nem não tem olhos castanhos nem olhos azuis dados que a característica está ausente? \_\_\_\_\_
- a pessoa nem não tem olhos de outra cor nem olhos azuis e a característica está presente \_\_\_\_\_
- a pessoa não tem olhos castanhos? \_\_\_\_\_
- a pessoa tem olhos azuis ou nem não tem olhos azuis nem olhos castanhos? \_\_\_\_\_
- a pessoa não tem a característica ou não tem olhos castanhos? \_\_\_\_\_

103) Um sindicato de trabalhadores local consiste de associados encanadores e eletricitas, classificado de acordo com grau:

	Aprendiz	Jornaleiro	Oficial	
Encanadores	25	20	30	75
Eletricistas	15	40	20	75
	40	60	50	

Um associado do sindicato é selecionado ao acaso. Dado que o pessoa selecionada é um encanador, a probabilidade de que ele é um jornaleiro é:

- $1/2$
- $1/3$
- $4/15$
- $2/15$
- nenhuma das anteriores.

104) Entre vinte e cinco artigos, nove são defeituosos, seis tem somente um defeito não importante e três têm um defeito importante. Determine a probabilidade de que

um artigo selecionado ao acaso tenha defeitos importantes dado que ele tenha defeitos.

- a.  $1/3$
- b. 0,25
- c. 0,24
- d. 0,08

105) Os depositantes do Banco X são categorizados por idade. Selecionaremos aleatoriamente um indivíduo desse grupo de 2.000 depositantes

Idade	Sexo	
	Homem	Mulher
30 ou menos	800	600
31 ou mais	400	200

- i) Então  $P(\text{mulher de 30 ou menos}) =$   
 a)  $2/5$    b)  $3/4$    c)  $3/7$    d)  $3/10$    e) nenhuma das anteriores
- ii) Então  $P[\text{homem ou (31 ou mais)}] =$   
 a)  $1/5$    b)  $3/10$    c)  $1/2$    d)  $7/10$    e) nenhuma das anteriores
- iii) Então  $P(\text{mulher}) =$   
 a)  $3/10$    b)  $2/5$    c)  $3/5$    d)  $2/3$    e) nenhuma das anteriores

iv) Qual é a probabilidade condicional de que um depositante escolhido tenha idade de 30 anos ou menos, dado que ele é homem?

- a)  $2/3$    b)  $7/10$    c)  $4/7$    d)  $2/5$    e) nenhuma das anteriores
- v) São as idades e sexos dos depositantes independentes para o Banco X? Porque?

105) Um epidemiologista sente que as rodovias tem alguma relação com o desenvolvimento de uma nova doença porque a probabilidade de que uma pessoa esteja morando a uma milha ou menos da rodovia, dado que ela tem a doença é 0,80. Você concorda com ele? Explique porque.

106) Existem duas urnas marcadas com H e T. A urna H contem 2 bolas vermelhas e 1 bola azul. A urna T contem 1 bola vermelha e 2 azuis. Uma moeda é jogada ao acaso. Se sai cara é escolhida uma bola da urna H. Se sai coroa, uma bola é escolhida da urna T. Ache as seguintes probabilidades.



- a.  $P(\text{cara e vermelha})$       b.  $P(\text{coroa})$       c.  $P(\text{vermelha})$   
 d.  $P(\text{azul})$       e.  $P(\text{cara}|\text{vermelha})$

107) O número de paradas de máquinas em uma grande fábrica durante uma semana tem a seguinte distribuição de probabilidade:

B	5	10	15	20	25
$P(B = b)$	0,25	0,30	0,25	0,15	0,05

Usando essa distribuição, Calcule  $E[B]$  e  $V[B]$

- 108) A Companhia Beta comprou 80 componentes eletrônicos de um fornecedor que declara que somente 2 % dos componentes que ele vende são defeituosos e que os componentes defeituosos são misturados aleatoriamente com os componentes bons. Cada componente defeituoso custará a Beta US\$ 250 em custos de reparo. Se o fornecedor está certo, qual será o número esperado de componentes defeituosos ? E qual é o custo esperado de reparo?
- 109) Um vendedor de carros oferece a todos os seus clientes potenciais uma corrida de 30 milhas no tipo de carro que o cliente está interessado em comprar, mais um almoço ou jantar gratuitos. Todos estes custos são cerca de US\$ 50. Se o cliente não compra o carro, o vendedor perde US\$ 50, mas se o cliente comprar o carro, o lucro médio do vendedor é de cerca de US\$ 500 (dos quais os custos da corrida e da refeição devem ser deduzidos). No passado, 20 % dos clientes compraram o carro depois da corrida e da refeição gratuita. Qual é o lucro esperado para o vendedor nessa situação?
- 110) Um processo de produção é paralisado para ajuste toda vez que uma amostra aleatória de cinco itens, selecionada com reposição, apresenta dois ou mais defeituosos. Ache a probabilidade de que o processo será paralisado após uma inspeção se ele está produzindo:
- a) 20 % de defeituosos  
 b) 10 % de defeituosos  
 c) 5 % de defeituosos
- 111) Um simples míssil de certa variedade tem uma probabilidade de  $\frac{1}{4}$  de derrubar um bombardeiro, uma probabilidade de  $\frac{1}{4}$  de danificá-lo e uma probabilidade de  $\frac{1}{2}$  de errá-lo. Além disso, dois tiros danificadores derrubarão o avião. Se quatro destes mísseis são lançados, qual é a probabilidade de derrubar um avião?

112) De acordo com um cientista político, a população votante de certa cidade consiste de 46 % do candidato A, 40 % do candidato B, 11 % do candidato C e 3 % do candidato D. Em uma amostra aleatória de 5 votantes, qual é a probabilidade de que a amostra contenha:

- a) Dois votantes para o candidato A e um de cada das outras categorias?
- b) Três votantes para o candidato A e dois para o candidato B?
- c) Nenhum votante para o candidato D?

## 4. Variáveis Aleatórias Discretas

### Objetivos do Capítulo:

- Distinguir entre uma distribuição de probabilidade discreta e 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
- Descrever as características das distribuições Binomial, Hipergeométrica e de Poisson.
- Definição: Uma variável aleatória é um valor numérico determinado pelo resultado de um experimento (é uma quantidade resultante de um experimento aleatório que, por acaso, pode assumir diversos valores).

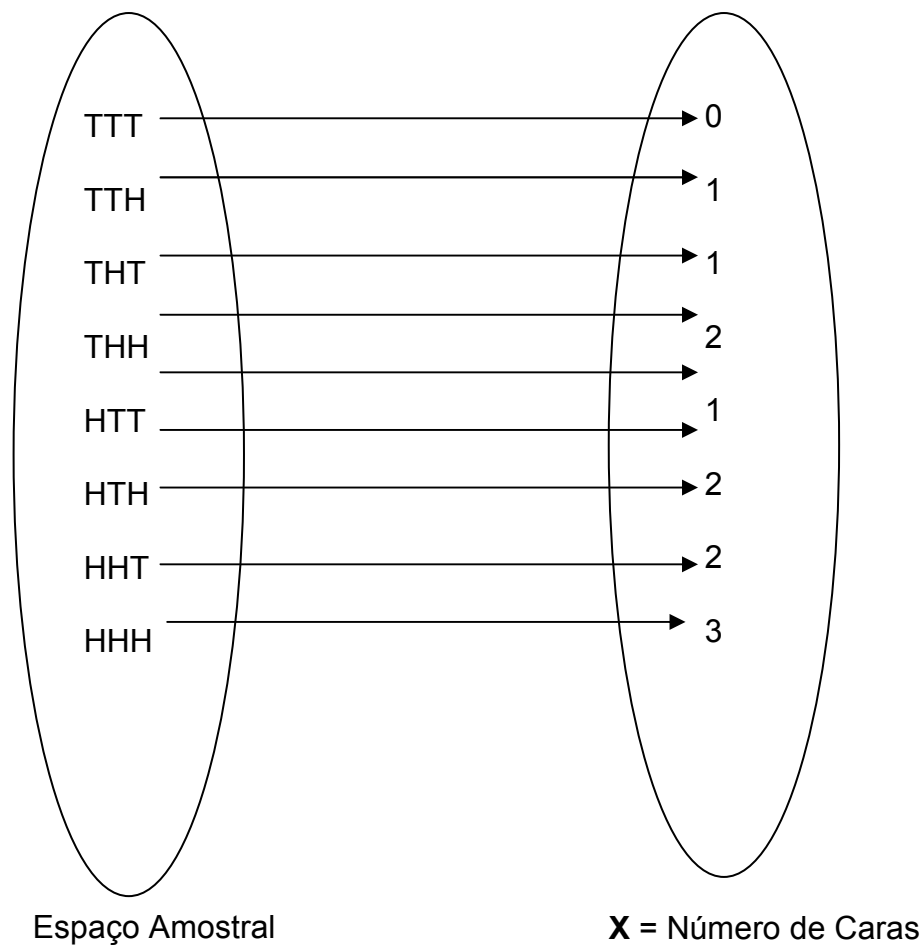
Exemplo 1: Considere um experimento aleatório no qual uma moeda é jogada 3 vezes. Seja X o número de caras. Seja H o resultado cara e T o resultado coroa.

- O espaço amostral para este experimento será:

TTT, TTH, THT, THH, HTT, HTH, HHT, HHH

- Assim, os possíveis valores de  $X$  (número de caras) serão:  
 $X = 0, 1, 2, 3$ .
- Nota: Neste experimento, há 8 possíveis resultados no espaço amostral. Desde que eles são todos igualmente prováveis de ocorrer, cada resultado tem uma probabilidade de  $1/8$  de ocorrer.

A figura a seguir ilustra a associação existente entre resultados do experimento (no espaço amostral) e os valores assumidos pela variável  $X$ .



- O resultado zero caras ocorre somente uma vez
- O resultado 1 cara ocorre três vezes
- O resultado 2 caras ocorre três vezes

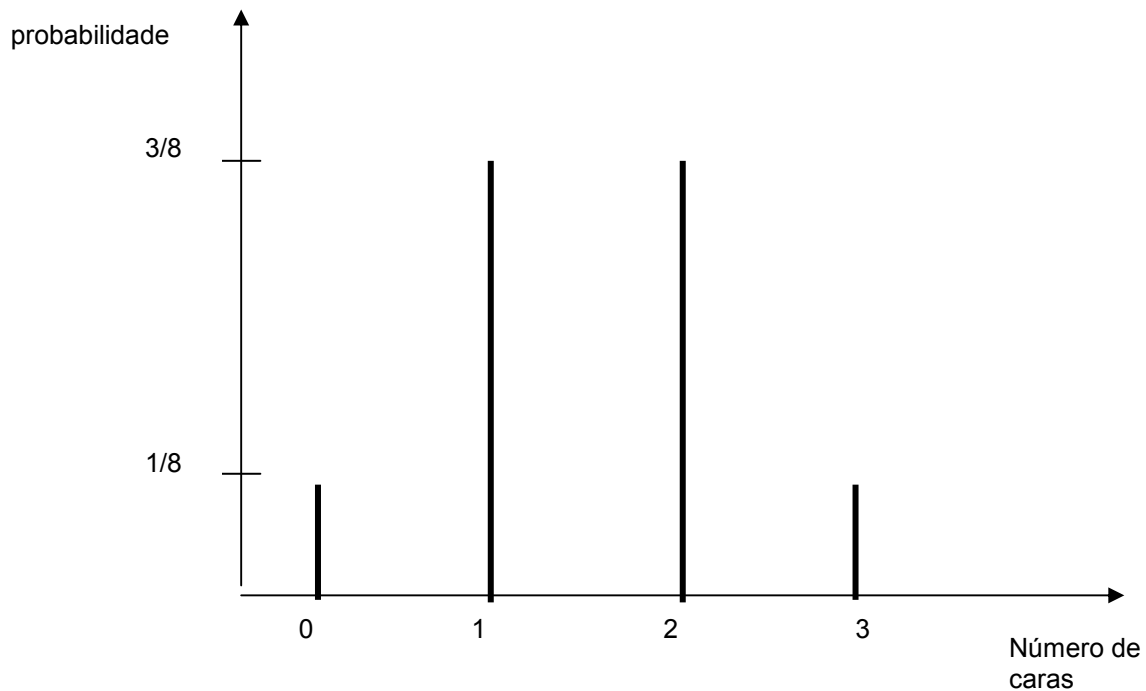
- O resultado 3 caras ocorre somente uma vez
- Da definição de uma variável aleatória,  $X$ , tal como é definida neste experimento, é uma variável aleatória. Seus valores são determinados pelos resultados do experimento.
- Nota: A variável aleatória  $X$  é uma associação de pontos no espaço amostral com pontos na reta 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.
- **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.

### **Distribuição de Probabilidade para Três Jogadas de uma 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	$8/8 = 1$
-------	-----------



### **CARACTERÍSTICAS DE UMA DISTRIBUIÇÃO DE PROBABILIDADE**

- A probabilidade de um resultado deve estar sempre situada entre 0 e 1.  
Exemplo:  $P(0 \text{ caras}) = 0,125$ ,  $P(1 \text{ cara}) = 0,375$ , etc. no experimento de jogar 3 moedas.
- A soma das probabilidades de todos os resultados mutuamente exclusivos é sempre 1 (veja a tabela de distribuição de probabilidade no texto).

### **VARIÁVEL ALEATÓRIA DISCRETA**

- **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 de interesse.

- 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 (são claramente separados, em descontinuidade).

**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.

## **VARIÁVEL ALEATÓRIA CONTÍNUA**

- **Definição:** Uma variável aleatória contínua é uma variável que pode assumir um número infinitamente grande de valores (com certas limitações práticas).

Exemplo: (a) Peso de um estudante

(b) comprimento de um carro

### **4.1 O 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  $\mu$  (letra grega, mi) representa a média (ou valor esperado) e  $P(X)$  é a probabilidade dos vários resultados de X.

#### 4.2 A 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  $\sigma^2$  (sigma ao quadrado).
- O desvio padrão é obtido através da raiz quadrada de  $\sigma^2$ .
- 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 2**

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

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

- Os dados acima podem ser considerados como uma distribuição de probabilidade? Porque ou porque não?

- Converta o número de carros alugados por semana em uma distribuição de probabilidade.

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

- Calcule o número médio de carros alugados por semana.

A média

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

- Calcule a variância do número de carros alugados por semana.

A variância

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

Cálculo de E(X):



Número de Carros alugados	Probabilidade, P(X)	XP(X)
10	0,25	2,5
11	0,30	3,3
12	0,35	4,2
13	0,10	1,3
Total	1,00	E(X) = 11,3

Cálculo de  $\sigma^2$

Número de Carros Alugados	Prob. P(X)	$(X - \mu)$	$(X - \mu)^2$	$(X - \mu)^2 P(X)$
10	0,25	10-11,3	1,69	0,4225
11	0,30	11-11,3	0,09	0,0270
12	0,35	12-11,3	0,49	0,1715
13	0,10	13-11,3	2,89	0,2890
Total				$\sigma^2 = 0,9135$

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

### 4.3 A Distribuição de Probabilidade Binomial

A Distribuição Binomial tem as seguintes características:

- Considere um experimento que apresenta apenas dois resultados possíveis que são categorias mutuamente exclusivas: sucesso e falha.
- São repetidos diversas vezes este mesmo experimento.

- A probabilidade de sucesso permanece constante para cada tentativa (consequentemente, a probabilidade de falha também permanece constante).
- As tentativas são independentes, significando que o resultado de uma tentativa não afeta o resultado de qualquer outra tentativa.

Para construir uma distribuição binomial, consideremos:

- $n$  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 falha em cada tentativa, que é igual a  $1-p$

### **FÓRMULA PARA A DISTRIBUIÇÃO DE PROBABILIDADE BINOMIAL**

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

onde  $n!$  é lido como  $n$  fatorial. Por exemplo,  $4! = (4).(3).(2).(1)=24$ .

$0!$  é igual a 1, por definição e  $1! = 1$ .

#### **Exemplo 3**

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

- Três estão desempregados na amostra. (Nota:  $n = 14$  e  $p = 0,2$ )

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

- No mínimo um dos trabalhadores da amostra estão 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 estão desempregados.

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

#### Exemplo 4

Uma companhia fabrica rolamentos para serem usados em bicicletas. Sabe-se que 5 % dos diâmetros dos rolamentos estarão fora dos limites de aceitação (defeituosos). Se 6 rolamentos são selecionados ao acaso, qual é a probabilidade de que:

Exatamente zero sejam defeituosos? Exatamente um seja defeituoso? Exatamente dois sejam defeituosos? Exatamente três sejam defeituosos? Exatamente quatro sejam defeituosos? Exatamente cinco sejam defeituosos? Exatamente seis sejam defeituosos?

- Note que as condições de uma distribuição binomial estão satisfeitas neste exemplo:
  - Há uma probabilidade constante de sucesso (0,05)
  - Há um número fixo de tentativas (6)
  - As tentativas são independentes (Por quê?)
  - Há somente dois possíveis resultados (um rolamento é defeituoso ou não defeituoso).

### **DISTRIBUIÇÃO DE PROBABILIDADE BINOMIAL para $n = 6$ e $p = 0,05$**

<b>Número de rolamentos defeituosos, X</b>	<b>Probabilidade de ocorrência, P(X)</b>
0	0,735
1	0,232
2	0,031
3	0,002
4	0,000
5	0,000
6	0,000

- Verifique os cálculos para os valores da tabela acima
- Mostre a representação gráfica para a Distribuição de Probabilidade Binomial com  $n = 6$  e  $p = 0,05$
- Para um  $n$  fixo (10) e  $p$  pequeno (0,05), a distribuição é positivamente assimétrica
- Para um  $n$  fixo (10) e  $p$  aproximando-se de 0,5, a distribuição torna-se simétrica.
- Para um  $n$  fixo (10) e  $p$  grande (0,95), a distribuição torna-se negativamente assimétrica.
- Para um  $p$  fixo e para valores cada vez maiores de  $n$ , a distribuição torna-se cada vez mais simétrica

Nota: Como os procedimentos de cálculo tornam-se repetitivos (e monótonos) faremos a seguir uma simulação no computador para diversos valores dos parâmetros  $n$  e  $p$  de uma distribuição binomial.

#### 4.4 A Média e Variância De Uma Distribuição Binomial

- A média é dada por:

$$\mu = np$$

- A variância é dada por:

$$\sigma^2 = np(1 - p)$$

Nota: A demonstração teórica para estes valores será desenvolvida em sala de aula e encontra-se na maioria dos livros de Introdução a Estatística.

- Para o exemplo anterior:

$$p = 0,05 \text{ e } n = 6$$

$$\mu = np = 6 \times 0,05 = 0,3$$

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

#### Distribuição Cumulativa de Probabilidade

Um engenheiro estimou que 60 % das pontes de um Estado necessitam de reparos. Uma amostra de 10 pontes no Estado foi aleatoriamente selecionada.

- Qual é a probabilidade de que exatamente 6 destas pontes necessitem de reparos? Esta situação (deste exemplo) satisfaz as condições para uma distribuição binomial ? Porque?
- Verificar:

$$n = 10, p = 0,6 \quad P(X = 6) = 0,251$$

- Qual é a probabilidade de que 7 ou menos destas pontes necessitem de reparos ?

$$P(X \leq 7) = P(X = 0) + P(X = 1) + \dots + P(X = 7) = 0,833 \text{ (verificar)}$$

Este é um exemplo de probabilidade cumulativa.

### Apêndice 1 (Recordação)

Uma variável aleatória (v.a.) é um valor numérico que é definido em ou é determinado pelos resultados ou eventos de um experimento. Variáveis aleatórias normalmente são denotadas por letras maiúsculas, X, Y etc e podem ser **discretas** ou **contínuas**.

Seja a v.a. X o número de sementes que germinam em 100 plantadas. Possíveis valores para X são 0,1,2,100, (discreta)

Seja a v.a. X a temperatura máxima diária em Uberlândia. Possíveis valores são 0 - 50 C por exemplo 26.1276(contínua).

Seja X a resposta a uma questão com ' Sim', ' Não', 'Não Sei'. X não é uma v.a (não numérica).

Seja Y o número de 'Sim's. Y é uma v.a. discreta.

### Distribuição de probabilidade de um v.a. Discreta.

Esta é uma lista dos possíveis valores da v.a. e as probabilidades correspondentes (que tem que somar 1). As probabilidades podem ser escritas:

$$P(X = x_i) = p_i \text{ para } i = 1, 2, \dots, k \text{ e } 0 \leq p_i \leq 1$$

$$\sum_{i=1}^k p_i = 1$$

## Apendice 2 (Recordação)

### Variável Aleatória discreta

Uma variável aleatória discreta é uma variável aleatória que toma valores discretos com probabilidades especificadas.

#### Exemplo - uma Família de 3 crianças.

Seja  $X$  uma Variável Aleatória (VA) = número de meninas

Possíveis valores:

$X = 3$       GGG  
 $X = 2$       GGB    GBG    BGG  
 $X = 1$       BBG    BGB    GBB  
 $X = 0$       BBB

Considere que os 8 resultados são igualmente prováveis de forma que

x	0	1	2	3
Probabilidade $P(X = x)$	1/8	3/8	3/8	1/8

A lista de valores que  $X$  pode assumir e as suas probabilidades é chamada de **distribuição de probabilidade discreta** para  $X$ .

Convenção de notação - use letras maiúsculas para variáveis aleatórias e letras minúsculas para valores específicos

#### Exemplo - tentativas de Bernoulli

Cada tentativa é um 'experimento' com exatamente 2 possíveis resultados, "sucesso " e " fracasso " com probabilidades  $p$  e  $1-p$ .

Seja  $X = 1$  se sucesso, 0 se fracasso

A Distribuição de probabilidade é

<b>x</b>	<b>0</b>	<b>1</b>
$P(X = x)$	$p$	$1-p$

### Exemplo - são lançados 2 dados

Seja  $X$  a soma dos resultados.

Resultados:

11	21	31	41	51	61
12	22	32	42	52	62
13	23	33	43	53	63
14	24	34	44	54	64
15	25	35	45	55	65
16	26	36	46	56	66

Considere que os 36 resultados são igualmente prováveis. Portanto cada um tem probabilidade =  $1/36$ .

Possíveis valores de  $X$  são 2, 3, ..., 12

por exemplo  $P(X = 4) = P(1,3 \text{ ou } 2,2 \text{ ou } 3,1) = 3/36$ .

A distribuição de probabilidade é

<b>x</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>...</b>	<b>10</b>	<b>11</b>	<b>12</b>
$P(X=x)$	$1/36$	$2/36$	$3/36$	$...$	$3/36$	$2/36$	$1/36$

### Apêndice 3 (Recordação)

#### A distribuição Binomial

Considere  $n$  tentativas Bernoulli.

Assuma que a probabilidade de sucesso ( $S$ ) é a **mesma para todas as tentativas**,  $P(S) = p$ .

Assuma que as tentativas são **independentes** e portanto a probabilidade para qualquer determinada combinação de sucessos e fracassos, por exemplo para 5 tentativas, a



probabilidade do resultado SSFSF, pode ser obtida **multiplicando as probabilidades** para cada resultado de tentativa.

por exemplo  $P(SSFSF) = p.p. (1-p) .p.(1-p) = p^3(1-p)^2$

De fato, a probabilidade de obter três sucessos e dois fracassos em cinco tentativas é  $p^3(1-p)^2$  para cada um dos modos diferentes que isto poderia acontecer, isto é, SSSFF, SSFSF,... etc.

O número de arranjos "distintos" de 3 sucessos e 2 fracassos pode ser facilmente calculado usando o coeficiente binomial  $\binom{n}{x}$  onde  $n$  é o número de tentativas e  $x$  é o número de sucessos requerido.

O coeficiente binomial (leia-se como "binomial de  $x$  em  $n$ ") é definido como  $\binom{n}{x}$

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Neste exemplo,  $\binom{5}{3} = \frac{5!}{3!2!} = \frac{5.4.3.2.1}{(3.2.1)(2.1)} = 10$ , portanto

há 10 maneiras distintas de se obter 3 sucessos em 5 tentativas, com cada arranjo tendo uma probabilidade  $p^3(1-p)^2$

Seja  $X$  a V.A. igual ao **número total de sucessos** em  $n$  tentativas .

Para calcular a probabilidade de obter  $x$  sucessos, pode ser mostrado que

$$P(X = x) = \binom{n}{x} \times p^x \times (1-p)^{n-x}$$

número de arranjos de x S's e (n-x) F's      prob. de x S's      prob. de (n-x) F's

onde o número mínimo de sucessos é 0 e o máximo é  $n$ .

A distribuição do número de sucessos é chamada distribuição binomial com dois parâmetros,  $n$  e  $p$ , necessários para determinar  $P(X=x)$ . Dizemos, de forma abreviada,  $X \sim B(n,p)$

### Exemplo - Um time de futebol joga 3 jogos

Assuma que cada jogo é uma tentativa Bernoulli com  $\text{prob}(\text{ganhar}) = 0,5$

Seja  $X$  a V.A. = número de vitórias

Então  $X$  tem distribuição binomial com  $n=3$  e  $p=0,5$ , com resultado vitória (W) ou derrota (L) em cada tentativa.

(Isto é abreviado como  $X \sim B(3;0,5)$ )

Qual é a probabilidade de que o time ganhe 2 jogos exatamente?

$$P(X=2) = P(WWL) + P(WLW) + P(LWW)$$

$$= 3/8 \text{ (desenhe um diagrama de árvore)}$$

ou usando a fórmula para probabilidades binomiais, a  $\text{prob}(WWL) = p^2(1-p)$  e o número

de distintos arranjos de 2 vitórias em três jogos é  $\binom{3}{2} = \frac{3!}{2!!} = 3$ . Portanto a resposta é

$$P(X = 2) = \binom{3}{2} (0,5)^2 (1 - 0,5)^1 \text{ usando } n = 3, x = 2, p = 0,5$$

$3p^2(1-p)$ . Assim

$$= \frac{3}{8}$$

As distribuições binomiais são usadas para modelar situações que podem ser pensadas como tentativas repetidas e " independentes " cada uma com somente 2 possíveis resultados. Nós os usaremos posteriormente para fazer inferências estatísticas sobre proporções.

### Exemplo - Um Sistema de Controle de Qualidade

Um Sistema de Controle de Qualidade requer que de cada lote de itens uma amostra de 10 é selecionada e é testada. Se 2 ou mais itens da amostra são defeituosos o lote inteiro é rejeitado.

Se a probabilidade de um item ser defeituoso é 0,05

(i) qual é a probabilidade de 2 defeituosos na amostra?

(ii) Qual é a probabilidade do lote ser rejeitado?

Seja X a V.A. = número de defeituosos na amostra de  $n = 10$  itens.

Portanto,  $X \sim \text{Binomial}(10; 0,05)$

$$(i) \quad P(X = 2) = \binom{10}{2} (0,05)^2 (0,95)^8 = 0,0746$$

$$(ii) \quad P(\text{rejeitar o lote}) = P(X \geq 2) = \sum_{x=2}^{10} \binom{10}{x} (0,05)^x (0,95)^{10-x} \text{ o que é muito}$$

trabalhoso de calcular. Mas:

$$\begin{aligned} P(X \geq 2) &= 1 - P(X < 2) = 1 - P(X = 0 \text{ ou } X = 1) \\ &= 1 - [P(X = 0) + P(X = 1)] \text{ mutuamente exclusivos} \\ &= 1 - \left[ \binom{10}{0} (0,05)^0 (0,95)^{10} + \binom{10}{1} (0,05)^1 (0,95)^9 \right] \\ &= 0,0862 \end{aligned}$$

#### **Apêndice 4 (Recordação) *Valor Esperado e Variância de uma Variável Aleatória***

##### **Análise de decisão**

##### **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.

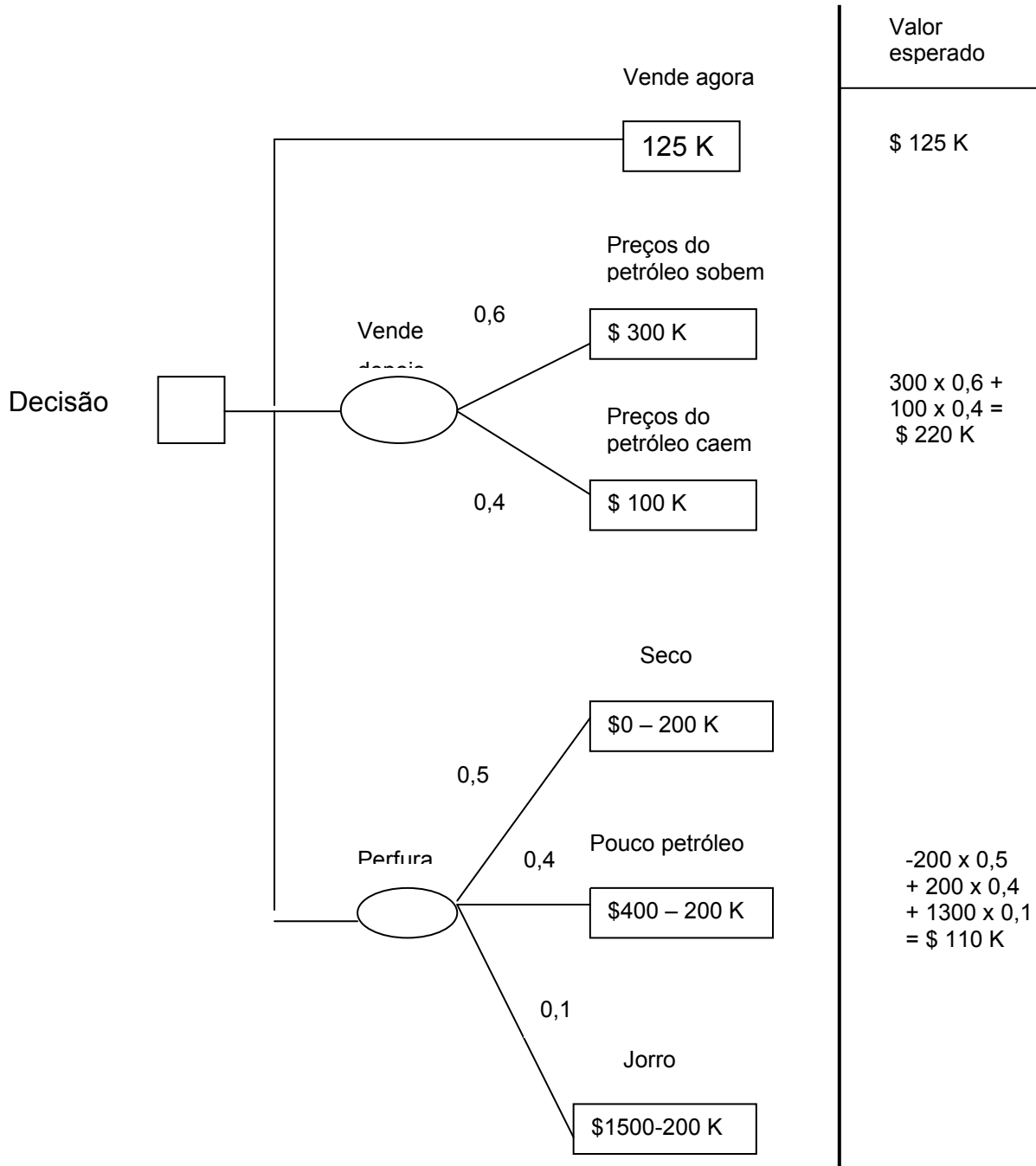
O custo de perfurar é \$200,000 (\$200K).

Perfurando conduzirá a um dos resultados seguintes

Resultado	Probabilidade	Receita
Poço Seco	0.5	\$0
Poço com pouco petróleo	0.4	\$400K
Poço com jorro	0.1	\$1500K

Se vende agora, a companhia pode adquirir \$125K.

Se segura durante um ano e os preços do petróleo sobem (probabilidade = 0.6) pode vender por \$300K ou se os preços do petróleo caem (probabilidade = 0.4) pode adquirir \$100K. O que deveria fazer?



A melhor decisão é segurar durante um ano e então vender. Este é um exemplo de usar um diagrama de árvore para Análise de Decisão. Também ilustra o conceito do valor esperado de uma variável aleatória .

Se a distribuição de probabilidade de uma variável aleatória X é

Valores de X	x <sub>1</sub>	x <sub>2</sub>	...	x <sub>k</sub>
Probabilidades	p <sub>1</sub>	p <sub>2</sub>	...	p <sub>k</sub>

seu valor esperado é

$$E(X) = x_1 p_1 + x_2 p_2 + \dots + x_k p_k = \sum_{i=1}^K x_i p_i$$

exemplo Perfuração de Petróleo

Resultado	Probabilidade	Receita
Poço seco	0.5	0
Poço com pouco petróleo	0.4	\$400K
Poço com jorro	0.1	\$1500K

Seja X a variável aleatória lucro financeiro

= Receita – custo de perfuração

= Receita - \$200K

A distribuição de probabilidade para X é

x	-200	200	1300
P(X=x)	0.5	0.4	0.1

Portanto, o valor esperado (média) de X é

$$E(X) = -200 \times 0.5 + 200 \times 0.4 + 1300 \times 0.1 = \$110K$$

Isto é diretamente análogo à média amostral.

E(X) pode ser considerada como uma idealização de, ou um valor teórico para, a média da amostra.

E(X) é denotado freqüentemente pela letra grega  $\mu$  (pronuncia-se mu).

### Variância de uma Variável Aleatória

Recorde que a variância é uma medida de dispersão. Para uma amostra de observações de uma população a variância ao redor da média é definida como

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

A variância de uma Variável Aleatória X é definida como

$$\text{var}(x) = p_1(x_1 - \mu)^2 + p_2(x_2 - \mu)^2 + \dots + p_K(x_K - \mu)^2$$

ou

$$\text{var}(x) = \sum_{i=1}^K p_i(x_i - \mu)^2 = E(x - \mu)^2$$

Ela representa o limite teórico da variância amostral  $s^2$  quando o tamanho da amostra (n) fica muito grande.

$\text{var}(X)$  é denotada freqüentemente por  $\sigma^2$  (sigma quadrado).

Uma fórmula mais simples para  $\text{var}(X)$  é

$$\begin{aligned}\text{var}(X) &= (p_1 x_1^2 + \dots + p_K x_K^2) - \mu^2 \\ &= E(X^2) - [E(X)]^2\end{aligned}$$

#### **Exemplo - Gênero em uma classe de 5**

Assuma que a probabilidade de um estudante em uma classe ser masculino é um meio. Seja a variável aleatória X o número de estudantes masculinos em um grupo da classe de tamanho 5.

Qual é o valor de  $E(X)$ , o número esperado de homens no grupo, e qual é a variância de X?

Considere  $X \sim \text{binomial}(5; 0,5)$ .

Então a distribuição de probabilidade de X é

x	0	1	2	3	4	5
P(X=x)	1/32	5/32	10/32	10/32	5/32	1/32

(Confira isto usando a fórmula para probabilidades binomiais e desenhe um diagrama de árvore para analisar a estrutura dos resultados.)

$$\begin{aligned}
 E(X) = \sum xp(x) &= 0 \times \frac{1}{32} + 1 \times \frac{5}{32} + 2 \times \frac{10}{32} + 3 \times \frac{10}{32} \\
 &\quad + 4 \times \frac{5}{32} + 5 \times \frac{1}{32} = \frac{80}{32} = 2,5 = \mu
 \end{aligned}$$

isto é, em média tais grupos têm 2,5 homens.

$$\begin{aligned}
 \text{var}(X) &= \sum x^2 p(x) - \mu^2 = 0^2 \times \frac{1}{32} + \dots + 5^2 \times \frac{1}{32} - (2,5)^2 \\
 &= 7,5 - (2,5)^2 \\
 &= 1,25
 \end{aligned}$$

$$\text{Portanto, } \sigma = \sqrt{\text{var}(X)} = \sqrt{1,25} = 1,12$$

Esta é uma medida da variabilidade de X.

Em geral se  $X \sim \text{binomial}(n,p)$  pode ser mostrado que

$$E(X) = np \text{ e } \text{var}(X) = npq$$

$$\text{onde } q = 1 - p$$

[Confira os valores de  $E(X)$  e  $\text{var}(X)$  calculados acima para  $X \sim \text{binomial}(5;0,5)$  usando estas fórmulas.]



EMPÍRICO (baseado nos dados) QUANTIDADE	TEÓRICO (MATEMÁTICO) QUANTIDADE	
(a) Frequência relativa $x_i = \frac{f_i}{n}$	PROB[X = x <sub>i</sub> ] = p <sub>i</sub>	$\frac{f_i}{n} \rightarrow \infty$ <i>quando</i> $n \rightarrow 0$
(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, $\mu =$ $E(X) = \sum_i p_i x_i$	$\bar{x} \rightarrow E(X)$ <i>quando</i> $n \rightarrow \infty$
(d) VARIÂNCIA $S^2 =$ $\sum_{i=1}^n \frac{(x_i - \bar{x})^2 f_i}{n - 1}$	$VAR(X) =$ $\sum_{ii=1}^n (x_i - x)^2 p_i$	$S^2 \rightarrow VAR(X)$ <i>quando</i> $n \rightarrow \infty$

### 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,

$$E(Y) = a E(X) + b$$

$$\text{var}(Y) = a^2 \text{var}(X)$$

$$\text{Portanto, } \sigma_Y = \sqrt{a^2 \text{var}(X)} = \sqrt{a^2 \sigma_x^2} = a \sigma_x$$

e

Semelhantemente se  $T = aX + bY + c$  onde  $X$  e  $Y$  são variáveis aleatórias e  $a$ ,  $b$  e  $c$  são constantes conhecidas, então,

$$E(T) = a E(X) + b E(Y) + c.$$

$$e \quad Var(T) = a^2 var(X) + b^2 var(Y) + 2ab cov(X, Y)$$

Em particular, se  $X$  e  $Y$  são independentes então a covariância  $cov(X, Y)$  é zero. Portanto

$$Var(T) = a^2 var(X) + b^2 var(Y)$$

Prova: Segue das definições de  $E(X)$  e  $var(X)$ .

### **Exemplo - Lucro previsto estimado**

Uma companhia faz produtos para mercados locais e de exportação.

O número de vendas do próximo ano não pode ser predito exatamente mas estimativas podem ser feitas como a seguir

unidades de X, local	1,000	3,000	5,000	10,000
Probabilidade	0.1	0.3	0.4	0.2

unidades Y, export.	300	500	700
Probabilidade	0.4	0.5	0.1

Conseqüentemente  $E(X) = 1000 \times 0.1 + 3000 \times 0.3 + 5000 \times 0.4 + 10000 \times 0.2$

$= 5000$  (= esperou vendas locais)

$E(Y) = 300 \times 0.4 + 500 \times 0.5 + 700 \times 0.1$

$= 440$  (= 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

$$E(T) = 2000 E(X) + 3500 E(Y)$$

$$= 2000 \times 5000 + 3500 \times 440$$

$$= \$11,540,000$$

- este é o lucro estimado (previsto) durante o próximo ano.

Exemplo - Fabricação de um componente de metal

Um componente é feito cortando um pedaço de metal de comprimento  $X$  e reduzindo este valor da quantidade  $Y$ . Ambos estes 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

$$E(T) = a E(X) + b E(Y) = 1 E(X) + (-1)E(Y)$$

$$= E(X) - E(Y)$$

$$Var(T) = a^2 var(X) + b^2 var(Y)$$

$$\text{Por tanto } var(T) = 1^2 var(X) + (-1)^2 var(Y) \\ = var(X) + var(Y)$$

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  $T$ .

### ***Variáveis Aleatórias Independentes***

Lembremos que dois eventos  $A$  e  $B$  são independentes se e somente se  $P(A \text{ e } B) = P(A)P(B)$  – 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 seqüências que não se sobrepõem seja de jogadas de moedas, de jogadas de

dados são alguns exemplos. O segundo e terceiro exemplo acima mostram porque existe a necessidade das seqüências serem não sobrepostas (ou seja, não tenham intersecção).

### **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.

### **Apêndice 4 (recordação)**

#### **A DISTRIBUIÇÃO BINOMIAL DE PROBABILIDADE**

Suponhamos que um experimento consista de tentativas repetidas, cada uma com dois possíveis resultados que podem ser vistos como **sucesso** ou **fracasso**. Uma aplicação óbvia na área de ciências sociais aplicadas refere-se a um experimento que se refere a selecionar repetidas vezes um elemento de amostra de uma população que contenha apenas duas categorias, por exemplo, pessoas que votarão em um determinado candidato ou não. Consideremos que se a pessoa for votar no candidato teremos um resultado de sucesso e se não for votar teremos um resultado de fracasso. Outro exemplo seria um jogo de baralho em que extraímos repetidas vezes uma carta do conjunto de 52 cartas. Neste caso poderemos considerar como sucesso o resultado ser uma carta numérica e fracasso o resultado ser uma carta de figura. A um experimento definido desta forma damos o nome de **processo Bernoulli**. Podemos também definir uma variável aleatória que terá valor  $X = 1$  se ocorrer sucesso e valor  $X = 0$  se ocorrer fracasso. Desta forma também podemos chamar tal variável de variável aleatória Bernoulli. Cada tentativa do experimento é denominada tentativa Bernoulli.

Podemos observar que tanto no exemplo do candidato como no exemplo das cartas se a pessoa selecionada da população de eleitores ou se a carta selecionada do baralho não for repostada a probabilidade de um sucesso para as repetidas tentativas muda. Suponhamos que a nossa população de eleitores tenha 1000 pessoas e dentro desta

população 300 votarão no candidato e 700 não votarão. Na primeira tentativa do experimento (seleção da primeira pessoa) temos uma probabilidade de sucesso igual a  $300 / 1000 = 0,3$ . Na segunda tentativa, se não for feita a reposição da primeira pessoa na população de origem, teremos uma probabilidade de sucesso igual a  $299/999$  caso tenha ocorrido sucesso na primeira tentativa e igual a  $300/999$  caso tenha ocorrido fracasso na primeira tentativa. Fica mais complicado de ver o que ocorrerá na terceira tentativa, pois o resultado irá depender do que ocorreu na primeira e na segunda tentativas. Neste caso não teremos tentativas Bernoulli porque a probabilidade de sucesso não se mantém constante no decorrer das tentativas sequenciais.

Iremos definir um **processo Bernoulli** da forma como segue. Estritamente falando, um processo Bernoulli deve ter as seguintes propriedades:

- 1) O experimento consiste de  $n$  tentativas repetidas.
- 2) Cada tentativa tem um resultado que pode ser classificado como um sucesso ou um fracasso.
- 3) A probabilidade de sucesso, denotada por **p**, permanece constante de tentativa para tentativa.
- 4) As tentativas repetidas são independentes.

Considere um conjunto de tentativas Bernoulli onde três itens são selecionados aleatoriamente de um processo de fabricação. A seguir eles são inspecionados e classificados como defeituosos ou não defeituosos. Um item defeituoso é designado como um sucesso. O numero de sucesso é uma variável aleatória  $X$  com valores de 0 a 3. O espaço amostral destes experimento é definido por oito eventos:

$$S = \{NNN, NDN, NND, DNN, NDD, DND, DDN, DDD\}$$

Temos então a seguinte tabela de resultados para esta variável aleatória:

Resultado	X
NNN	0
NDN	1

NND	1
DNN	1
NDD	2
DND	2
DDN	2
DDD	3

Como os itens são selecionados independentemente de um processo que digamos, produz 25 % de defeituosos, teremos, por exemplo:

$$P(\text{NDN}) = P(N).P(D).P(N) = \left(\frac{3}{4}\right) \cdot \left(\frac{1}{4}\right) \cdot \left(\frac{3}{4}\right) = \frac{9}{64} = 0,14$$

Se não houvesse independência estatística entre as tentativas Bernoulli e se tivéssemos um lote de produção de 1000 peças, teríamos o seguinte resultado:

$$P(\text{NDN}) = P(N_1).P(D_2 / N_1).P(N_3 / N_1 \cap D_2) =$$

$$\frac{750}{1000} \times \frac{250}{999} \times \frac{749}{998} = 0,14086$$

Como pode ser visto, existe uma pequena diferença no valor da probabilidade quando calculamos considerando como tentativas Bernoulli (tentativas independentes e com probabilidade de sucesso constante) e quando consideramos que as tentativas não são independentes (neste ultimo caso não são tentativas Bernoulli).

Vamos agora calcular o valor das probabilidades para cada valor da variável aleatória X (numero de sucessos). Para isto construímos a seguinte tabela de distribuição de probabilidades:

<b>x</b>	<b>f(x) = P(X=x)</b>
0	27/64
1	27/64
2	9/64

3	1/64
---	------

A variável aleatória  $X$  que é definida como o numero de sucessos é chamada de variável aleatória Binomial. A distribuição de probabilidade é chamada de distribuição binomial.

Podemos generalizar este resultado com a seguinte definição:

---

---

### **Distribuição binomial**

Uma tentativa Bernoulli pode resultar em um sucesso com probabilidade  $p$  e com fracasso com probabilidade  $q = 1 - p$ . Então a distribuição de probabilidade de uma variável aleatória binomial  $X$ , o numero de sucessos em  $n$  tentativas independentes, é

$$b(x; n, p) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, 2, \dots, n.$$


---

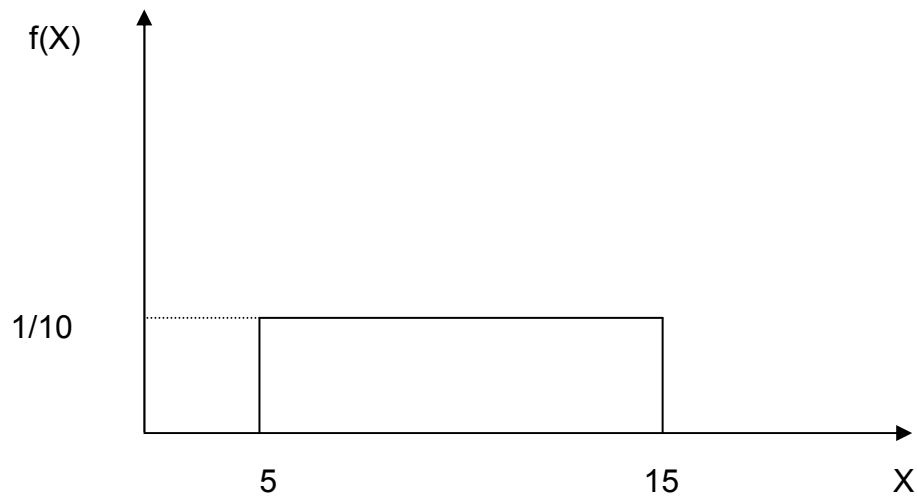
---

## **5. Variáveis Aleatórias Contínuas e Distribuição Normal**

### **5.1 Variáveis Aleatórias Contínuas**

Diferentemente de uma variável aleatória discreta, para uma variável aleatória contínua não podemos definir uma função de distribuição 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:





Observe que  $f(X)$  é uma função constante assumindo sempre o valor  $1/10$  no intervalo fechado  $5 \leq X \leq 10$ . Essa função goza das seguintes propriedades:

- 1) ela é sempre positiva ou nula. Ou seja,  $f(X) \geq 0$  para qualquer valor de  $X$ .
- 2) se integrarmos esta função no intervalo  $5 \leq X \leq 10$  o valor desta integral definida será igual a 1. Ou seja,

$$\int_5^{15} f(X) dx = \int_5^{15} 1/10 dx = [x/10]_5^{15} = \frac{15}{10} - \frac{5}{10} = 1$$

Toda função que satisfizer essas duas propriedades chamaremos de função densidade de probabilidade. Essa função é apenas um instrumento matemático para que possamos calcular probabilidades para variáveis aleatórias contínuas (assim como utilizamos a função distribuição de probabilidade para as variáveis aleatórias discretas). Por exemplo, para o exemplo acima, se quisermos calcular a probabilidade da variável aleatória contínua  $X$  estar contida no intervalo  $10 \leq X \leq 12$  será:

$$P(10 \leq X \leq 12) = \int_{10}^{12} f(X) dx = \int_{10}^{12} (1/10) dx = [x/10]_{10}^{12} = \frac{12}{10} - \frac{10}{10} = 2/10$$

Dessa forma, podemos calcular a probabilidade para qualquer intervalo sendo esta probabilidade o valor da integral definida da função densidade de probabilidade sendo

que os limites de integração são as extremidades do intervalo. De uma forma geral, podemos dizer que se  $f(X)$  é a função densidade de probabilidade de uma variável aleatória contínua, então:

$$P(a \leq X \leq b) = \int_a^b f(X)dx$$

## 5.2 Média e Variância de uma Variável Aleatória Contínua

A média (ou valor esperado) de uma variável aleatória contínua é dada pela expressão:

$$E[X] = \int_{-\infty}^{+\infty} Xf(X)dx$$

No exemplo anterior, o valor esperado da variável aleatória  $X$  será:

$$E[X] = \int_{-\infty}^{+\infty} Xf(X)dx = \int_5^{15} X(1/10)dx = \left[ \frac{x^2}{20} \right]_5^{15} = \frac{225}{20} - \frac{25}{20} = 10$$

A variância de uma variável aleatória contínua é dada pela expressão:

$$V[X] = \int_{-\infty}^{+\infty} (X - E[X])^2 f(X)dx$$

No exemplo anterior, a variância da variável aleatória  $X$  será:

$$V[X] = \int_{-\infty}^{+\infty} (X - E[X])^2 f(X)dx = \int_5^{15} (X - 10)^2 (1/10)dx = \int_5^{15} \left( \frac{X^2}{10} - 2X + 10 \right) dx = \left[ \frac{X^3}{30} - X^2 + 10X \right]_5^{15} = 8,333$$

## COVARIÂNCIA ENTRE DUAS VARIÁVEIS ALEATORIAS

Covariância é uma medida de associação (relação) linear entre duas variáveis aleatórias. Se  $X$  e  $Y$  são duas v.a., a covariância entre elas é definida por:

$$\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))]$$

Desta forma a covariância entre duas variáveis  $X$  e  $Y$  é igual a média de uma variável aleatória  $Z$  que por sua vez é o produto dos desvios de cada uma das duas variáveis  $X$  e  $Y$  em relação às suas respectivas médias.

Exemplifiquemos com o seguinte quadro de distribuição conjunta de duas variáveis aleatórias discretas  $X$  e  $Y$ :

X \ Y	0	1	2	P(y)
1	3/20	3/20	2/20	8/20
2	1/20	1/20	2/20	4/20
3	4/20	1/20	3/20	8/20
P(x)	8/20	5/20	7/20	1

Para interpretar este quadro, podemos dizer que a probabilidade conjunta de  $X = 1$  e  $Y = 2$  é  $P(X=1, Y=2) = 1/20$ . A probabilidade marginal de  $X = 1$  é  $P(X=1) = 5/20$ . A probabilidade condicional de  $X = 2$  dado que  $Y = 1$  é

$$P(Y = 2 / X = 1) = \frac{P(X = 1, Y = 2)}{P(X = 1)} = \frac{1/20}{5/20} = \frac{1}{5}$$

A distribuição de probabilidade da variável aleatória  $Z = (X - E(X))(Y - E(Y))$  é a própria distribuição de probabilidade conjunta dada no quadro acima para as variáveis  $X$  e  $Y$ . Como a covariância é uma esperança temos que:

$$\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))] = \sum (X - E(X))(Y - E(Y))p(X, Y)$$

Ou seja, a covariância é o somatório do produto da variável  $Z = (X - E(X))(Y - E(Y))$  pelas probabilidades conjuntas. Para calcular a covariância devemos calcular as esperanças (médias) de  $X$  e  $Y$ . Estas são:

$$E(X) = 0 \times \frac{8}{20} + 1 \times \frac{5}{20} + 2 \times \frac{7}{20} = \frac{19}{20}$$

$$E(Y) = 1 \times \frac{8}{20} + 2 \times \frac{4}{20} + 3 \times \frac{8}{20} = 2$$

No exemplo do quadro acima a covariância é igual a:

$$\begin{aligned} Cov(X, Y) &= (0 - \frac{19}{20}) \times (1 - 2) \times \frac{3}{20} + (1 - \frac{19}{20}) \times (1 - 2) \times \frac{3}{20} + (2 - \frac{19}{20}) \times (1 - 2) \times \frac{2}{20} \\ &+ (0 - \frac{19}{20}) \times (2 - 2) \times \frac{1}{20} + (1 - \frac{19}{20}) \times (2 - 2) \times \frac{1}{20} + (2 - \frac{19}{20}) \times (2 - 2) \times \frac{2}{20} \\ &+ (0 - \frac{19}{20}) \times (3 - 2) \times \frac{4}{20} + (1 - \frac{19}{20}) \times (3 - 2) \times \frac{1}{20} + (2 - \frac{19}{20}) \times (3 - 2) \times \frac{3}{20} = 0 \end{aligned}$$

Um outro método (mais fácil) de se calcular a covariância é dado pela expressão:

$$Cov(X, Y) = E(XY) - E(X).E(Y)$$

Exercício: Demonstre a validade da expressão acima

Sabemos que a definição de covariância é:

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))]$$

Podemos desenvolver o segundo termo desta expressão da seguinte forma:

$$\begin{aligned} Cov(X, Y) &= E(XY - E(X)Y - E(Y)X + E(X)E(Y)) = \\ &E(XY) - E(X).E(Y) - E(Y).E(X) + E(X).E(Y) = E(XY) - E(X)E(Y) \end{aligned}$$

Aplicamos esta expressão aos dados do quadro acima para calcular a covariância:

Para isto precisamos calcular  $E(XY)$ . Para fazer isto devemos para cada valor do quadro (para cada dupla de valores de X e Y) calculamos o valor do produto XY e multiplicamos pela probabilidade conjunta.

$$\begin{aligned} E(XY) &= 0.1.3/20 + 1.1.3/20 + 2.1.2/20 + 0.2.1/20 + 1.2.1/20 + 2.2.2/20 \\ &+ 0.3.4/20 + 1.3.1/20 + 2.3.3/20 = 1,9 \end{aligned}$$

Portanto a covariância será:

$$Cov(X, Y) = 1,9 - (0,95).(2) = 0$$

Concluimos que as duas variáveis aleatórias X e Y são não correlacionadas.

**Se X e Y são duas variáveis aleatórias independentes, então  $Cov(X, Y) = 0$**

**Mas a recíproca não é verdadeira. O fato de  $\text{Cov}(X,Y) = 0$  não implica necessariamente que X e Y sejam independentes.**

Para o ultimo exemplo, verificamos que  $\text{Cov}(X,Y) = 0$ . No entanto vamos verificar que estas duas variáveis não são independentes. Para que X e Y sejam independentes é estritamente necessário que  $P(X,Y) = P(X).P(Y)$  para todos os valores de X e Y. Ou seja, para todas as células da distribuição de probabilidade conjunta, o valor da probabilidade conjunta deve ser igual ao produto das probabilidades marginais respectivas. Verifiquemos esta propriedade para o quadro de distribuição de probabilidade conjunta anterior.

X	0	1	2	P(y)
Y				
1	<b>3/20</b> $8/20 \cdot 8/20 = 16/400$	<b>3/20</b> $8/20 \cdot 5/20 = 40/400$	<b>2/20</b> $8/20 \cdot 7/20 = 56/400$	8/20
2	<b>1/20</b> $4/20 \cdot 8/20 = 32/400$	<b>1/20</b> $4/20 \cdot 5/20 = 20/400$	<b>2/20</b> $4/20 \cdot 7/20 = 28/400$	4/20
3	<b>4/20</b> $8/20 \cdot 8/20 = 56/400$	<b>1/20</b> $8/20 \cdot 5/20 = 40/400$	<b>3/20</b> $8/20 \cdot 7/20 = 56/400$	8/20
P(x)	8/20	5/20	7/20	1

No quadro acima os valores em negrito são as probabilidades conjuntas e logo em seguida vem o calculo do produto das probabilidades marginais respectivas. Observe-se que para a primeira célula temos  $P(X=0,Y=1) = 3/20 = 0,15$  e  $P(X=0).P(Y=1) = 16/400 = 0,04$ . Na segunda célula da primeira linha temos  $P(X=1,Y=1) = 3/20 = 0,15$  e  $P(X=1).P(Y=1) = 40/400 = 0,1$ . Portanto em nenhuma destas duas células a probabilidade conjunta coincide com o produto das probabilidades marginais respectivas. Bastava que para apenas uma das células não ocorresse a igualdade de probabilidades e as variáveis aleatórias já seriam dependentes. Para que ocorra independência perfeita entre as variáveis aleatórias é necessário que para todas as células da distribuição de probabilidade conjunta ocorra a igualdade entre a probabilidade conjunta e o produto das probabilidades marginais respectivas.

**Sejam X e Y duas variáveis quaisquer. Então**

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y)$$

**No caso de X e Y serem independentes temos o caso particular de**

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) \quad \text{já que } \text{Cov}(X, Y) = 0$$

Exercício: Demonstre teoricamente a expressão acima.  **$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y)$**

Para mais de duas variáveis independentes:

$$\text{Var}(X_1 + X_2 + \dots + X_n) = \text{Var}(X_1) + \text{Var}(X_2) + \dots + \text{Var}(X_n)$$

## **COVARIÂNCIA PARA VARIÁVEIS ALEATORIAS CONTÍNUAS**

Se X e Y são duas variáveis aleatórias contínuas a covariância é dada pela seguinte expressão:

$$\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))] = \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) dx dy$$

Exemplo: suponhamos que duas variáveis aleatórias contínuas X e Y tenham a seguinte função de densidade conjunta.

$$f(x, y) = \begin{cases} 8xy, & 0 \leq x \leq 1, 0 \leq y \leq x \\ 0, & \text{para outros valores} \end{cases}$$

Em primeiro lugar calculamos as funções de densidade marginais.

$$g(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_{-\infty}^{\infty} 8xy dy = 8x \int_0^x y dy = 8x \cdot \frac{x^2}{2} = 4x^3$$

Portanto a função de densidade de probabilidade marginal para a variável aleatória X é:

$$g(x) = \begin{cases} 4x^3 & 0 \leq x \leq 1 \\ 0 & \text{para outros valores} \end{cases}$$

Da mesma forma podemos determinar a função de densidade de probabilidade marginal de Y da seguinte forma:

$$h(y) = \int_{-\infty}^{\infty} f(x, y) dx = \int_{-\infty}^{\infty} 8xy dx = 8y \int_y^1 x dx = 8y \cdot \frac{1^2 - y^2}{2} = 4y(1 - y^2)$$

Portanto a função de densidade de probabilidade marginal para a variável aleatória Y é:

$$g(y) = \begin{cases} 4y(1 - y^2) & 0 \leq y \leq 1 \\ 0 & \text{para outros valores} \end{cases}$$

As esperanças matemáticas (medias) de X e Y são calculadas como:

$$\mu_X = E(X) = \int_0^1 x \cdot 4x^3 dx = 4/5$$

$$\mu_Y = E(Y) = \int_0^1 y \cdot 4y(1 - y^2) dy = 8/15$$

$$E(XY) = \int_0^1 \int_y^1 8x^3 y^2 dx dy = \frac{4}{9}$$

$$Cov(X, Y) = \sigma_{XY} = E(XY) - \mu_X \mu_Y = \frac{4}{9} - \left(\frac{4}{5}\right)\left(\frac{8}{15}\right) = \frac{4}{225}$$

É importante destacar que a variância de uma variável aleatória pode ser interpretada como a covariância desta variável com relação a ela mesma. Ou seja,

$$Cov(X, X) = E[(X - \mu_X)(X - \mu_X)] = E[(X - \mu_X)^2] = Var(X)$$

Qual é a interpretação prática da covariância?

A covariância serve para verificar se duas variáveis aleatórias movimentam-se ou não no mesmo sentido. Por exemplo, se quando uma variável X aumenta a variável Y também aumenta e se quando X diminui, Y também diminui (as variáveis) movimentam-se, covariam no mesmo sentido, a covariância é positiva. Ao contrario, quando X aumenta, Y diminui ou quando X diminui, Y aumenta, ou seja, as variáveis covariam em sentidos opostos, a covariância é negativa.

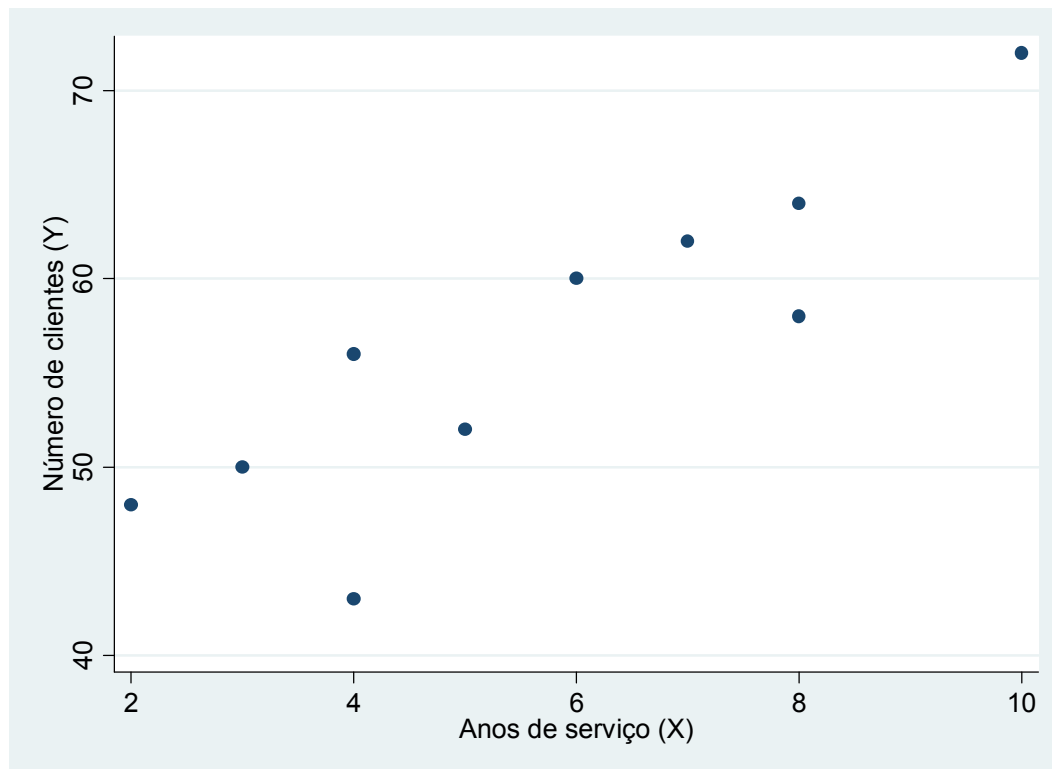
Podemos calcular a covariância, quando temos a distribuição de probabilidade conjunta como foi mostrado anteriormente. Mas podemos também calcular a covariância se

tivermos um conjunto de dados observados para as variáveis X e Y. Por exemplo, utilizemos o exemplo 4.4 da pagina 81 do Morettin.

Na figura a seguir temos o gráfico de dispersão das variáveis X e Y referente a tabela a seguir. Neste gráfico temos os pares de valores (x,y).

Agente	Anos de serviço (X)	Número de clientes (Y)
A	2	48
B	3	50
C	4	56
D	5	52
E	4	43
F	6	60
G	7	62
H	8	58
I	8	64
J	10	72

Diagrama de dispersão para as variáveis X (anos de serviço) e Y (numero de clientes)





Na figura a seguir apresentamos a planilha Excel para o cálculo da covariância e do coeficiente de correlação. Na quarta e quinta colunas da planilha temos os valores das variáveis subtraídos das suas médias. Na sexta coluna temos o produto destas duas últimas colunas. A média desta última coluna é o valor da covariância. Finalmente nas duas últimas colunas calculamos os desvios ao quadrado das variáveis em relação às suas médias (para podermos calcular a variância e o desvio padrão de cada uma delas). Finalmente, calculamos o coeficiente de correlação como sendo a divisão entre a covariância e o produto dos desvios padrões de X e de Y.



## Problemas Resolvidos

### Problema 31, pg 228 (Morettin)

Casal	Rendimento do Homem (X)	Rendimento da Mulher (Y)
1	10	5
2	10	10
3	5	5
4	10	5
5	15	5
6	10	10
7	5	10
8	15	10
9	10	10
10	5	10

- (a) Construa a distribuição de probabilidade conjunta de X e Y
- (b) Determine as distribuições marginais de X e Y
- (c) X e Y são v.a. Independentes? Justifique.
- (d) Calcule as medias e variâncias de X e Y e a covariância entre elas.
- (e) Considere a variável aleatória Z igual a soma dos rendimentos de cada homem e cada mulher. Calcule a media e variância de Z.
- (f) Supondo que todos os casais tenham a renda de um ano disponível, e que se oferecera ao casal escolhido a possibilidade de comprar uma casa pelo preço de 20, qual a probabilidade de que o casal escolhido possa efetuar a compra?

### Solução

- (a) Para a variável X temos a ocorrência de 3 valores (5, 10 e 15) e para a variável Y temos a ocorrência de 2 valores (5 e 10). Iremos construir a distribuição de probabilidade conjunta de acordo com a frequência relativa de ocorrência destes valores conjuntos. Por exemplo, em 10 observações (casais) vemos que o par  $(X=10, Y=5)$  ocorre 2 vezes. Portanto a frequência relativa é  $2/10 = 0,2$  e este é o valor da probabilidade conjunta de  $X = 10$  e  $Y = 5$ . Desta forma a nossa distribuição de probabilidade conjunta é:\
- (b)

Y \ X	5	10	15	p(Y)
5	1/10	2/10	1/10	4/10
10	2/10	3/10	1/10	6/10
p(X)	3/10	5/10	2/10	1

(b) As distribuições marginais já foram calculadas na Tabela acima

(c) Para verificar se X e Y são independentes, verifiquemos se o produto das probabilidades marginais é igual a probabilidade conjunta para todas as células. Para a primeira célula no canto superior esquerdo, temos  $P(X=5).P(Y=5) = 3/10 \cdot 4/10 = 12/100$  que é diferente de  $1/10$  e desta forma já constatamos a não independência entre X e Y.

(d) A média de X é  $E[X] = 5 \cdot 3/10 + 10 \cdot 5/10 + 15 \cdot 2/10 = 15/10 + 50/10 + 30/10 = 95/10 = 9,5$ . A media de Y é  $E[Y] = 5 \cdot 4/10 + 10 \cdot 6/10 = 20/10 + 60/10 = 8$ .

A variância de X pode ser calculada como  $E[X^2] - (E[X])^2$   
 $= 5^2 \cdot \frac{3}{10} + 10^2 \cdot \frac{5}{10} + 15^2 \cdot \frac{2}{10} - 9,5^2 = 12,25$

A Variância de Y é  $E[Y^2] - (E[Y])^2 = 5^2 \cdot 4/10 + 10^2 \cdot 6/10 - 8^2 = 10 + 60 - 64 = 6$

A covariância pode ser calculada de três formas. Na primeira forma, consideramos a expressão da própria definição de covariância.

$$\begin{aligned} Cov(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] = \sum (X - \mu_X)(Y - \mu_Y)p_{XY} \\ &= (5 - 9,5)(5 - 8)\frac{1}{10} + (10 - 9,5)(5 - 8)\frac{2}{10} + \dots + (15 - 9,5)(10 - 8)\frac{1}{10} = -1 \end{aligned}$$

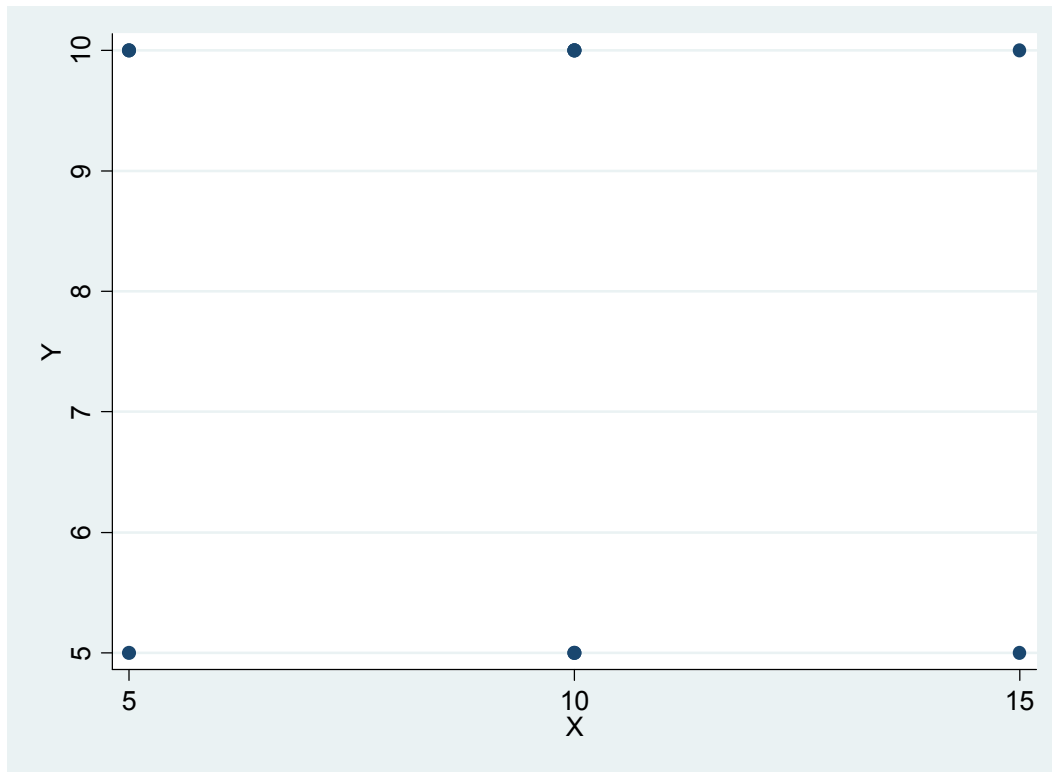
Na segunda forma, utilizamos a expressão

$$\begin{aligned} Cov(X, Y) &= E(XY) - E(X)E(Y) = \sum XYp_{XY} - E(X)E(Y) = \\ &= 5 \cdot 5 \cdot \frac{1}{10} + 5 \cdot 10 \cdot \frac{2}{10} + 5 \cdot 15 \cdot \frac{1}{10} + 10 \cdot 5 \cdot \frac{2}{10} + 10 \cdot 10 \cdot \frac{3}{10} + 10 \cdot 15 \cdot \frac{1}{10} - 9,5 \times 8 = -1 \end{aligned}$$

O coeficiente de correlação é calculado como:

$$\rho = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{-1}{\sqrt{12,25 \times \sqrt{6}}} = -0,1166$$

**Diagrama de dispersão para as variáveis X e Y**



## Planilha Excel para o calculo da covariância pelos 3 metodos:

Microsoft Excel - CALCULO DA COVARIANCIA E DO COEFICIENTE DE CORRELACAO												
Arquivo Editar Exibir Inserir Formatar Ferramentas Dados Janela Ajuda												
M75												
A	B	C	D	E	F	G	H	I	J	K	L	M
PROBLEMA 31 PAGINA 228 MORETTIN												
METODO COM OS VALORES DA TABELA ORIGINAL (ESTATISTICA DESCRITIVA)												
Casal	Rendimento do Homem (X)	Rendimento da Mulher (Y)	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$	$(y - \bar{y})^2$					
1	10	5	0.5	-3	-1.5	0.25	9					
2	10	10	0.5	2	1	0.25	4					
3	5	5	-4.5	-3	13.5	20.25	9					
4	10	5	0.5	-3	-1.5	0.25	9					
5	15	5	5.5	-3	-16.5	30.25	9					
6	10	10	0.5	2	1	0.25	4					
7	5	10	-4.5	2	-9	20.25	4					
8	15	10	5.5	2	11	30.25	4					
9	10	10	0.5	2	1	0.25	4					
10	5	10	-4.5	2	-9	20.25	4					
medias		9.5	8	cov =		-1	12.25	6				
X	5	10	15	p(Y)								
Y												
5	0.10	0.20	0.10	0.40								
10	0.20	0.30	0.10	0.60								
p(X)	0.30	0.50	0.20	1.00								
Plan1 / Plan2 / Plan3 /												
Pronto												
MAIU												

### Problema 35 pagina 229 do MORETTIN

Se  $E[X] = \mu$  e  $\text{Var}(X) = \sigma^2$ , escreva em função de  $\mu$  e  $\sigma^2$  as seguintes expressões:

(a)  $E(X^2)$                       (b)  $E[(X(X-1))]$

**Solução:**

(a)  $E(X^2) = \text{Var}(X) + (E(X))^2 = \sigma^2 + \mu^2$

(b)  $E[X(X-1)] = E[X^2 - X] = E[X^2] - E[X] = \text{Var}(X) + (E[X])^2 - E[X] = \sigma^2 + \mu^2 - \mu = \sigma^2 + \mu(\mu - 1)$

### Problema 39 pagina 229 do MORETTIN

Se  $\rho(X, Y)$  for o coeficiente de correlação entre X e Y, e se tivermos  $Z = AX + B$ ,  $W = CY + D$ , com  $A > 0$ ,  $C > 0$ , prove que  $\rho(X, Y) = \rho(Z, W)$

**Solução:**

$$\begin{aligned} \text{Cov}(Z, W) &= \text{Cov}(AX + B, CY + D) = E[(AX + B)(CY + D)] - E[AX + B].E[CY + D] \\ &= E[ACXY + ADX + BCY + BD] - (E[AX] + E[B]).(E[CY] + E[D]) = \\ \rho(Z, W) &= \frac{\text{cov}(Z, W)}{\sigma_Z \sigma_W} \quad \text{Mas } AC.E[XY] + AD.E[X] + BC.E[Y] + E[BD] - (A.E[X] + B).(C.E[Y] + D) = \\ &AC.E[XY] + AD.E[X] + BC.E[Y] + BD - (AC.E[X].E[Y] + AD.E[X] + BC.E[Y] + BD) = \\ &AC.E[XY] - ACE[X].E[Y] = AC.\text{Cov}(X, Y) \end{aligned}$$

$$\sigma_Z = \sigma_{AX+B} = A\sigma_X \text{ pois } \text{Var}(AX+B) = A^2 \text{Var}(X) \text{ e da mesma forma } \sigma_W = C\sigma_Y$$

$$\text{Portanto } \rho(Z, W) = \frac{\text{cov}(Z, W)}{\sigma_Z \sigma_W} = \frac{AC \text{cov}(X, Y)}{A\sigma_X C\sigma_Y} = \rho(X, Y)$$

É interessante notar que a condição imposta pelo MORETTIN de que  $A > 0$  e  $C > 0$  não é necessária.

### Problema 41 pagina 230 do MORETTIN

Suponha que X e Y sejam v.a. com  $\text{Var}(X) = 1$ ,  $\text{Var}(Y) = 2$  e  $\rho(X, Y) = 1/2$ . Determine  $\text{Var}(X - 2Y)$

**Solução:**

$$\begin{aligned}
Var(X - 2Y) &= Var(X) + Var(2Y) - 2Cov(X, 2Y) = \\
&= Var(X) + 4Var(Y) - 2(E[2XY] - E[X].E[2Y]) = \\
&= Var(X) + 4Var(Y) - 2(2E[XY] - 2E[X].E[Y]) = \\
&= Var(X) + 4Var(Y) - 4Cov(X, Y) = \\
&= Var(X) + 4Var(Y) - 4\rho(X, Y).\sigma_X\sigma_Y = \\
1 + 4.2 - 4.\frac{1}{2}.1.\sqrt{2} &= 11,83
\end{aligned}$$

## Recordação: VARIANCIA E COVARIANCIA

A media ou valor esperado de uma variável aleatória  $X$  é de especial importância na estatística porque ela descreve onde a distribuição de probabilidade está centrada. Por ela mesma, entretanto, a media não dá nenhuma noção adequada da distribuição ou da forma da distribuição da variável aleatória. Precisamos caracterizar a variabilidade da distribuição.

### Definição

Seja  $X$  uma variável aleatória com distribuição de probabilidade  $f(x)$  e media  $\mu$ . A **variância** de  $X$  é

$$\sigma^2 = E[(X - \mu)^2] = \sum_x (x - \mu)^2 f(x) \quad \text{se } X \text{ é uma variavel aleatoria discreta}$$

$$\sigma^2 = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx \quad \text{se } X \text{ é uma variavel aleatoria continua}$$

A raiz quadrada da variância,  $\sigma$ , é chamada de **desvio padrão** de  $X$ .

Estas são as definições de variância quando estamos considerando a estatística antes da realização do experimento aleatório, ou seja, trata-se de um calculo baseado em um modelo probabilístico (a distribuição de probabilidade no caso de variável aleatória discreta ou a função de densidade de probabilidade no caso de variável aleatória continua). Quando estamos enfocando a estatística após a realização do experimento aleatório temos uma formula distinta para a variância que é:

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N} \quad \text{para o caso de uma população}$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n - 1} \quad \text{para o caso de uma amostra}$$



A quantidade  $x - \mu$  é chamada desvio de uma observação em relação a media da distribuição. Como estes desvios são elevados ao quadrado e depois é calculada a media destes valores ao quadrado.  $\sigma^2$  será menor se os valores de  $x$  estiverem mais próximos da media  $\mu$  e será maior se os valores de  $x$  estiverem mais afastados desta media  $\mu$ .

Exemplo:

Seja  $X$  uma variável aleatória representada pelo numero de automóveis que são usados para propósitos de negócios oficiais em qualquer dia da semana. A distribuição de probabilidade para a companhia A é dada por

x	1	2	3
f(x)	0,3	0,4	0,3

e para a companhia B é dada por

x	0	1	2	3	4
f(x)	0,2	0,1	0,3	0,3	0,1

Mostre a companhia A, a média é dada por

$$\mu = E(X) = (1)(0,3) + (2)(0,4) + (3)(0,3) = 2,0$$

E a variância é dada por

$$\sigma^2 = \sum_{i=1}^3 (x - 2)^2 f(x) = (1 - 2)^2 (0,3) + (2 - 2)^2 (0,4) + (3 - 2)^2 (0,3) = 0,6$$

Para a companhia A, temos

$$\mu = E(X) = (0)(0,2) + (1)(0,3) + (2)(0,3) + (3)(0,3) + (4)(0,1) = 2,0$$

E a variância é dada por

$$\sigma^2 = \sum_{i=1}^4 (x - 2)^2 f(x) = (0 - 2)^2 (0,2) + (1 - 2)^2 (0,1) + (2 - 2)^2 (0,3) + (3 - 2)^2 (0,3) + (4 - 2)^2 (0,1) = 1,6$$

A variância para o numero de automóveis é maior para a companhia B do que para a companhia A.

A variância pode ser obtida também através de uma formula mais simples.

$$\sigma^2 = \sum_x (x - \mu)^2 f(x) = \sum_x (x^2 - 2\mu x + \mu^2) f(x) = \sum_x x^2 f(x) - 2\mu \sum_x x f(x) + \mu^2 \sum_x f(x)$$

Como  $\mu = \sum_x x f(x)$  por definição e  $\sum_x f(x) = 1$  para qualquer distribuição de probabilidade discreta, segue-se que

$$\sigma^2 = \sum_x x^2 f(x) - \mu^2 = E(X^2) - \mu^2$$

Exemplo: Seja a variável aleatória X que representa o numero de peças defeituosas para uma maquina quando 3 peças são selecionadas da linha de produção e testadas. Temos então a seguinte distribuição de probabilidade para X:

x	0	1	2	3
f(x)	0,51	0,38	0,10	0,01

Utilizando a ultima expressão, calcule  $\sigma^2$ .

Solução:

$$\mu = (0)(0,51) + (1)(0,38) + (2)(0,10) + (3)(0,01) = 0,61$$

$$E(X^2) = (0)(0,51) + (1)(0,38) + (4)(0,10) + (9)(0,01) = 0,87$$

$$\text{Portanto: } \sigma^2 = 0,87 - (0,61)^2 = 0,4979$$

Exemplo: A demanda semanal por uma marca de refrigerante, em milhares de litros é uma variável aleatória continua X com densidade de probabilidade

$$f(x) = \begin{cases} 2(x-1), & 1 < x < 2 \\ 0, & \text{para quaisquer outros valores} \end{cases}$$

Ache a media e a variância de X.

Solução:

$$\mu = E(X) = 2 \int_1^2 x(x-1) dx = \frac{5}{3}$$

e

$$E(X^2) = 2 \int_1^2 x^2(x-1) dx = \frac{17}{6}$$

Portanto:

$$\sigma^2 = \frac{17}{6} - \left(\frac{5}{3}\right)^2 = \frac{7}{18}$$

Neste ponto a variância ou desvio padrão somente tem significado quando comparamos duas ou mais distribuições que tem as mesmas unidades de medida. Portanto, podemos comparar as variâncias das distribuições dos conteúdos, medidos em litros, para duas companhias engarrafadoras de suco de laranja e o maior valor indicará a companhia cujo produto é mais variável ou menos uniforme. Não tem significado comparar a variância da distribuição de alturas de pessoas com a variância da distribuição dos coeficientes de inteligência (Q.I.).

Agora vamos estender nosso conceito de variância de uma variável aleatória  $X$  para também incluir variáveis aleatórias relacionadas a  $X$ . Para a variável aleatória  $g(X)$ , a variância será denotada por  $\sigma_{g(X)}^2$  e é calculada por meio dos seguintes teoremas.

---

Teorema: Seja  $X$  uma variável aleatória com distribuição de probabilidade  $f(x)$ . A média ou valor esperado da variável aleatória  $g(X)$  é

$$\mu_{g(X)} = E[g(X)] = \sum_x g(x)f(x) \text{ se } X \text{ é discreta, e}$$

$$\mu_{g(X)} = E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx \text{ se } X \text{ é contínua}$$

Exemplo: Suponha que o numero de carros,  $X$ , que passam através de uma lavadora de carros entre 4 hs e 5 hs da tarde em qualquer sexta-feira ensolarada tem a seguinte distribuição de probabilidade:

x	4	5	6	7	8	9
P(X=x)	1/12	1/12	1/4	1/4	1/6	1/6

Seja  $g(X) = 2X - 1$  que representa a quantidade de dinheiro em reais, paga ao lavador pelo proprietário. Encontre o ganho esperado do lavador para este período de tempo particular.

Solução:

$$\mu_{g(X)} = E[g(X)] = E[2X - 1] = \sum_4^9 (2x - 1)f(x) = (7)\left(\frac{1}{12}\right) + (9)\left(\frac{1}{12}\right) + (11)\left(\frac{1}{4}\right) \\ (13)\left(\frac{1}{4}\right) + (15)\left(\frac{1}{6}\right) + (17)\left(\frac{1}{6}\right) = R\$12,67$$

---

Teorema.

Seja X uma variável aleatória com distribuição de probabilidade f(x). A variância da variável aleatória g(X) é

$$\sigma_{g(X)}^2 = E\left\{ [g(X) - \mu_{g(X)}]^2 \right\} = \sum_x [g(X) - \mu_{g(X)}]^2 f(x)$$

Se X é discreta, e

$$\sigma_{g(X)}^2 = E\left\{ [g(X) - \mu_{g(X)}]^2 \right\} = \int_{-\infty}^{\infty} \sum_x [g(X) - \mu_{g(X)}]^2 f(x) dx$$

Se X é contínua.

Não vamos nos preocupar com a demonstração deste teorema.

---

Exemplo: Calcule a variância de  $g(X) = 2X + 3$ , onde X é uma variável aleatória com distribuição de probabilidade

x	0	1	2	3
f(x)	1/4	1/8	1/2	1/8

Solução:

$$\mu_{2X+3} = E[2X + 3] = \sum_{x=0}^3 (2x + 3)f(x) = 6 \\ \sigma_{2X+3}^2 = E\left\{ [(2X + 3) - \mu_{2X+3}]^2 \right\} = E\left\{ [2X + 3 - 6]^2 \right\} = E[4X^2 - 12X + 9] \\ = \sum_{x=0}^3 (4x^2 - 12x + 9)f(x) = 4$$

Exemplo: Seja X uma variável aleatória que tem função densidade

$$f(x) = \begin{cases} \frac{x^3}{3}, & -1 < x < 2 \\ 0, & \text{para quaisquer outros valores} \end{cases}$$

Ache a variância da variável aleatória  $g(X) = 4X+3$

$$E(4X + 3) = \int_{-1}^2 \frac{(4x + 3)x^2}{3} dx = \frac{1}{3} \int_{-1}^2 (4x^3 + 3x^2) dx = 8$$

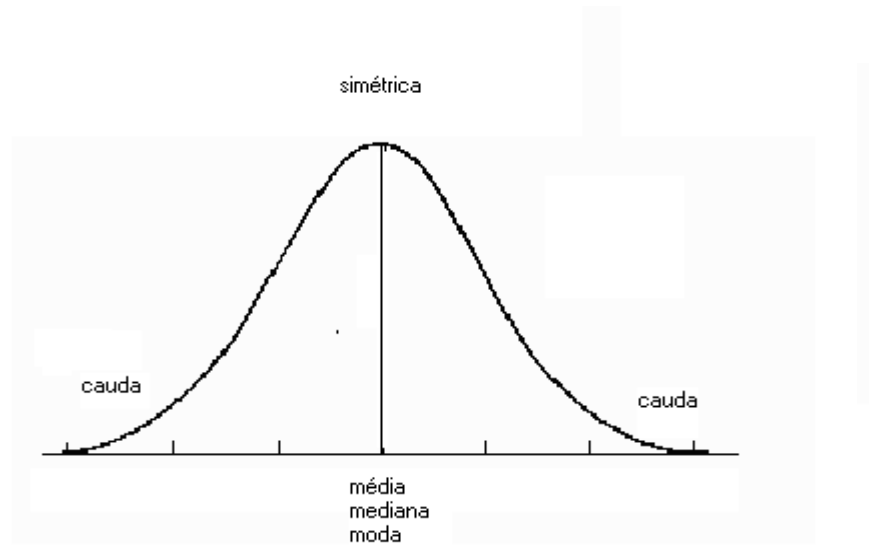
$$\begin{aligned}\sigma_{4X+3}^2 &= E\left\{ \left[ (4X + 3) - 8 \right]^2 \right\} = E\left[ (4X - 5)^2 \right] = \int_{-1}^2 (4x - 5)^2 \frac{x^2}{2} dx \\ &= \frac{1}{3} \int_{-1}^2 (16x^4 - 40x^3 + 25x^2) dx = \frac{51}{5}\end{aligned}$$

### 5.3 Variável Aleatória Normal

- A mais importante (e mais utilizada na prática) variável aleatória contínua é a variável aleatória normal.
- A variável aleatória normal tem uma função densidade de probabilidade (chamada de curva normal) que apresenta a forma de um sino e é unimodal no centro exato da distribuição.
- A média, mediana e a moda da distribuição normal são iguais e localizadas no pico da distribuição.
- Metade da área sob a curva está acima do ponto central (pico) e a outra metade está abaixo dele.
- A distribuição de probabilidade normal é simétrica em relação a sua média.
- Ela é assintótica → a curva aproxima-se cada vez mais do eixo X mas nunca toca efetivamente ele.

Figura 1 – Características de uma Função Densidade de Probabilidade Normal (Distribuição Normal)

Figura 2 – Duas Distribuições Normais com mesma média mas distintos desvios padrões



Podemos também ter distribuições normais com o mesmo desvio padrão, mas com distintas médias ou com médias e desvios padrões distintos. Na realidade a distribuição normal é um nome genérico para definir 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 o seu desvio padrão. A função densidade de probabilidade de uma variável aleatória normal é dada por:

$$f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(X-\mu)^2}{2\sigma^2}}$$

É possível demonstrar matematicamente que a média (ou valor esperado) dessa variável aleatória é igual ao seu parâmetro  $\mu$  e o seu desvio padrão é igual ao seu segundo parâmetro (da equação acima)  $\sigma$ . O que quer dizer que se aplicarmos as definições de valor esperado e de variância de uma variável aleatória contínua a expressão acima

chegaremos aos resultados  $\mu$  e  $\sigma^2$ . O problema é recaímos em integrais mais difíceis de serem resolvidas:

$$E[X] = \int_{-\infty}^{+\infty} X f(X) dx = \int_{-\infty}^{+\infty} X \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}} dx = \mu$$

e

$$V[X] = \int_{-\infty}^{+\infty} (X - E[X])^2 f(X) dx = \int_{-\infty}^{+\infty} (X - \mu)^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}} dx = \sigma^2$$

(talvez um bom matemático possa fazer essa demonstração, mas não é o nosso caso pois pretendemos ser bons em estatística aplicada tão somente).

É possível também demonstrar matematicamente que as duas abscissas no eixo X de valor  $+\sigma$  e  $-\sigma$  correspondem a pontos de inflexão da curva normal. Para isto basta obter a segunda derivada da função densidade e provar que o seu valor muda de sinal no ponto de inflexão mostrando que aí a curvatura muda de sentido de côncava para convexa ou vice-versa.

## 5.4 Distribuição Normal Padrão

É muito difícil ficarmos calculando probabilidades para distribuições normais através de cálculos de integração. Para evitar este trabalho foi definida uma distribuição normal particular chamada de distribuição normal padrão. Esta distribuição tem as características de ser uma distribuição normal com média (valor esperado) igual a zero e desvio padrão igual a 1. Em notação matemática dizemos que:

$$Z \sim N(0,1)$$

Se X é uma variável aleatória normal com média  $\mu$  diferente de zero e desvio padrão  $\sigma$  diferente de 1 podemos “converter” essa distribuição em uma distribuição normal padrão através da transformação linear:

$$Z = \frac{X - \mu}{\sigma}$$

Para que serve essa distribuição Z? Nada melhor que um exemplo para explicar isso.

Exemplo: As rendas mensais dos graduados em um curso de especialização em uma grande empresa são normalmente distribuídas com uma média de R\$ 2000 e um desvio padrão de R\$ 200. Qual é o valor de Z para uma renda X de R\$ 2200? R\$ 1700?

- Para  $X = 2200 \Rightarrow Z = \frac{X - \mu}{\sigma} = \frac{2200 - 2000}{200} = 1$
- Para  $X = 1700 \Rightarrow Z = \frac{X - \mu}{\sigma} = \frac{1700 - 2000}{200} = -1,5$
- Um valor de  $Z = 1$  indica que o valor de R\$ 2200 está localizado 1 desvio padrão acima da média de R\$ 2000.
- Um valor de  $Z = -1,5$  indica que o valor de R\$ 1700 está localizado 1,5 desvio padrão abaixo da média de R\$ 2000.

### 5.5 Áreas Abaixo da Curva Normal

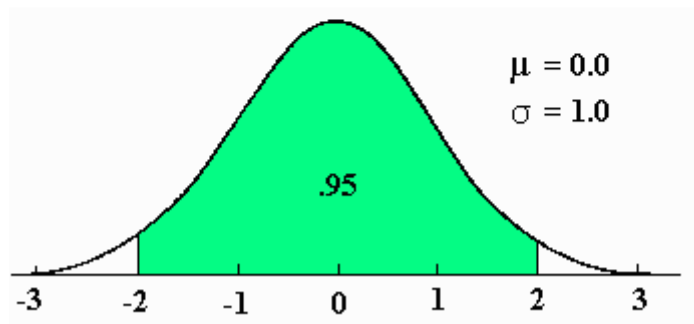
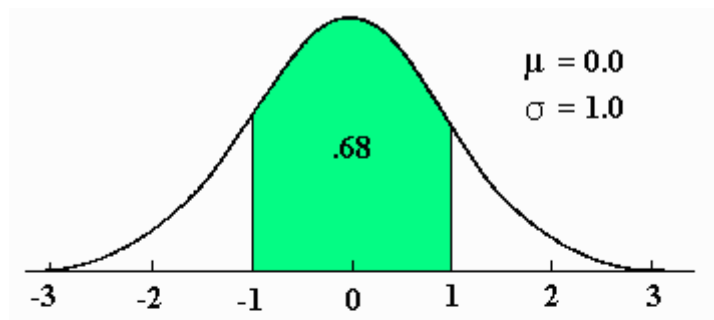
- Cerca de 68 % da área sob a curva normal está entre menos um e mais um desvio padrão da média. Isto pode ser escrito como  $\mu \pm 1\sigma$ .
- Cerca de 95 % da área sob a curva normal está entre menos dois e mais dois desvios padrões da média, escrito como  $\mu \pm 2\sigma$ .
- Praticamente toda (99,74 %) a área sob a curva normal está entre menos três e mais três desvios padrões da média, escrito como  $\mu \pm 3\sigma$ .



Exemplo 2:

O uso diário de água por pessoa em uma determinada cidade é normalmente distribuído com média  $\mu$  igual a 20 litros e desvio padrão  $\sigma$  igual a 5 litros. O uso diário de cerca de 68 % das pessoas nesta cidade caem entre que valores?

- $\mu \pm 1\sigma = 20 \pm 1(5)$  . Ou seja, cerca de 68 % das pessoas usam de 15 a 25 litros de água por dia.



- 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 selecionada ao acaso usará menos do que 20 litros por dia ?

- O valor de Z é  $Z = (20 - 20) / 5 = 0$ . Portanto  $P(X < 20) = P(Z < 0) = 0,5$ .

Qual é a probabilidade de que uma pessoa selecionada ao acaso use mais do que 20 litros por dia ?

- O valor de  $Z$  é  $Z = (20 - 20) / 5 = 0$ . Portanto  $P(X > 20) = P(Z > 0) = 0,5$ .

Que percentagem da população usa entre 20 e 24 litros por dia ?

$$X = 20 \rightarrow Z = 0$$

$$X = 24 \rightarrow Z = \frac{24 - 20}{5} = 0,8$$

$$P(20 < X < 24) = P(0 < Z < 0,8) = 0,2881 \text{ (28,81 \%)}.$$

Que percentagem usa entre 16 e 20 litros ?

$$X = 16 \rightarrow Z = \frac{16 - 20}{5} = -0,8$$

$$X = 20 \rightarrow Z = 0$$

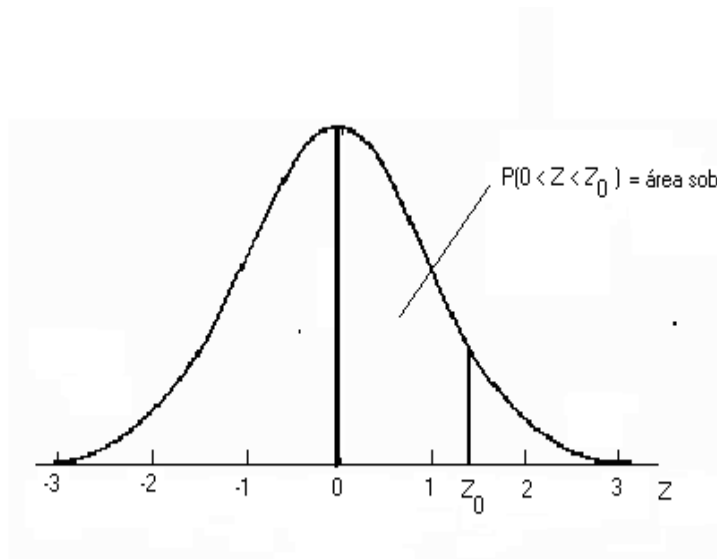
$$P(16 < X < 20) = P(-0,8 < Z < 0) = (\text{porque ?}) P(0 < Z < 0,8) = 0,2881 = 28,81$$

Para a obtenção das probabilidades para a curva normal padrão  $Z$  consulta-se uma tabela que pode ser encontrada em anexo em praticamente todos os livros de estatística. Reproduziremos a seguir integralmente essa tabela (para que possa ser mostrado para os exemplos anteriores como foram obtidas as áreas (que são probabilidades) abaixo da curva normal  $Z$ . Resolvemos colocar a tabela no corpo do texto devido a sua grande importância em estatística aplicada (e achamos que ela não deve ser relegada a um anexo que poucos alunos tem a curiosidade de consultar).

**Tabela 1** – Valor de  $P(0 < Z < Z_0)$  onde  $Z$  é variável normal padrão

$Z_0$	Segunda decimal de $Z_0$									
	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						0,1103	0,1141
0,3	0,1179	0,1217	0,1255						0,1480	0,1517
0,4	0,1554	0,1591	0,1628						0,1844	0,1879
0,5	0,1915	0,1950	0,1985						0,2190	0,2224
0,6	0,2257	0,2291	0,2324						0,2517	0,2549
0,7	0,2580	0,2611	0,2642						0,2823	0,2852
0,8	0,2881	0,2910	0,2939						0,3106	0,3133
0,9	0,3159	0,3186	0,3212						0,3365	0,3389
1,0	0,3413	0,3438	0,3461						0,3599	0,3621
1,1	0,3643	0,3665	0,3686						0,3810	0,3830
1,2	0,3849	0,3869	0,3888						0,3997	0,4015
1,3	0,4032	0,4049	0,4066						0,4162	0,4177
1,4	0,4192	0,4207	0,4222						0,4306	0,4319
1,5	0,4332	0,4345	0,4357						0,4429	0,4441
1,6	0,4452	0,4463	0,4474						0,4535	0,4545
1,7	0,4554	0,4564	0,4573						0,4625	0,4633
1,8	0,4641	0,4649	0,4658						0,4699	0,4706
1,9	0,4713	0,4719	0,4726						0,4761	0,4767
2,0	0,4772	0,4778	0,4783						0,4812	0,4817
2,1	0,4821	0,4826	0,4830						0,4854	0,4857
2,2	0,4861	0,4864	0,4868						0,4887	0,4890
2,3	0,4893	0,4896	0,4898						0,4913	0,4916
2,4	0,4918	0,4920	0,4922						0,4934	0,4936
2,5	0,4938	0,4940	0,4941						0,4951	0,4952
2,6	0,4953	0,4955	0,4956						0,4963	0,4964
2,7	0,4965	0,4966	0,4967						0,4973	0,4974
2,8	0,4974	0,4975	0,4976						0,4980	0,4981
2,9	0,4981	0,4982	0,4982						0,4986	0,4986

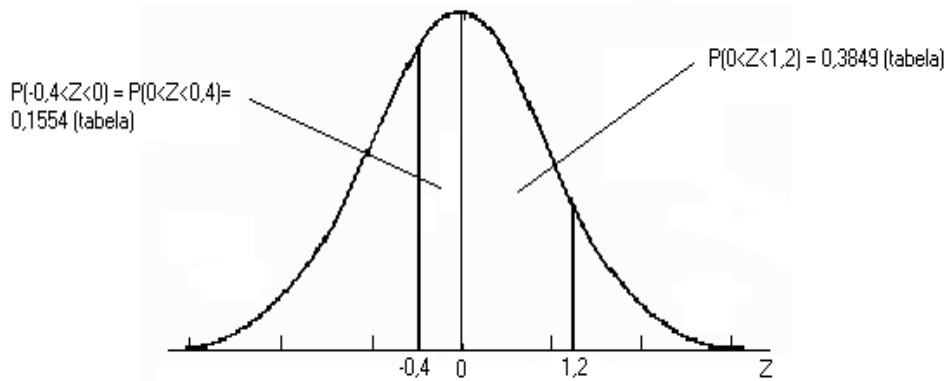
3,0	0,4987	0,4987	0,4987						0,4990	0,4990
-----	--------	--------	--------	--	--	--	--	--	--------	--------



Qual é a probabilidade de que uma pessoa selecionada ao acaso use mais do que 28 litros ?

$$X = 28 \rightarrow Z = (28 - 20) / 5 = 1,6$$

$$P(X > 28) = P(Z > 1,6) = 0,5 - 0,4452 = 0,0548$$



Qual é a porcentagem entre 18 e 26 litros ?

$$X = 18 \rightarrow Z = \frac{18 - 20}{5} = -0,4$$

$$X = 26 \rightarrow Z = \frac{26 - 20}{5} = 1,2$$

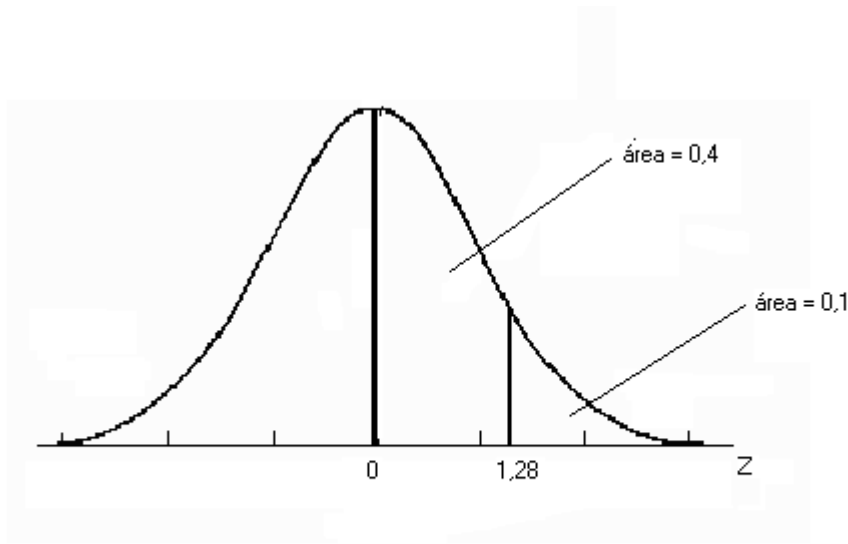
$$P(18 < X < 26) = P(-0,4 < Z < 1,2) = 0,1554 + 0,3849 = 0,5403$$

- Quantos litros ou mais 10 % da população usam ? Em outras palavras, para os 10 % da população que mais consomem água qual é o valor mínimo desse consumo ?

Seja  $X'$  a quantidade mínima. Portanto, precisamos encontrar  $X'$  tal que  $P(X \geq X') = 0,1$ . Para achar o valor de  $Z$  correspondente veja no corpo (miolo) da tabela o valor de  $Z_0$  que deixa uma área entre 0 e  $Z_0$  igual a  $(0,5 - 0,1) = 0,4$ . O valor correspondente de  $Z_0$  é 1,28 (aproximadamente). Portanto, temos:

$$\frac{X' - 20}{5} = 1,28 \rightarrow X' = 26,4. \text{ Ou seja, 10 \% da população usa no mínimo 26,4 litros por dia (ver figura).}$$

#### Exemplo 4



Um professor verificou que as médias finais em seu curso de Estatística tem distribuição normal com uma média igual a 72 e desvio padrão 5. Ele decide atribuir conceitos para o seu curso tal que os melhores 15 % recebem grau A . Qual é a mínima média que o estudante precisa receber para obter um A ?

Seja  $X'$  a mínima média.  $P(X \geq X') = 0,15$

O Z correspondente é 1,04 (aproximadamente)

$$\frac{X' - 72}{5} = 1,04 \quad X' = 77,2$$

## 6. Métodos de Amostragem e Distribuições Amostrais

### OBJETIVOS DO CAPÍTULO:

- Explicar porque em muitas situações uma amostra é a única forma plausível de aprender alguma coisa sobre uma população.
- Explicar os métodos de selecionar uma amostra
- Distinguir entre amostragem probabilística e amostragem não probabilística
- Definir e construir uma distribuição amostral de médias amostrais
- Explicar o Teorema do Limite Central e sua importância para a Inferência Estatística
- Calcular Intervalos de Confiança para Médias e Proporções
- Determinar que tamanho uma amostra deve ter para estimar médias e proporções

### Porque amostrar uma população

- Natureza destrutiva de certos testes
- A impossibilidade física de checar todos os itens na população
- O custo de estudar todos os itens em uma população é frequentemente proibitivo
- Muitas vezes as estimativas baseadas em uma amostra são mais precisas do que os resultados obtidos através de um levantamento censitário
- Tempo muito elevado para a apuração de resultados em censos

## INFERENCIA ESTATISTICA

O nome inferência refere-se a operações estatísticas em que com base em uma amostra estabelecem-se afirmações sobre uma população. Uma população é qualquer conjunto de objetos (pessoas, animais, coisas – não querendo naturalmente “coisificar” os seres humanos). Geralmente é um conjunto bastante amplo de objetos, por exemplo as pessoas de um determinado país ou região, os representantes de uma espécie animal no planeta e assim por diante. Os elementos de uma população podem ser enumeráveis ou não

enumeráveis. São enumeráveis quando podemos identificar cada um dos elementos e relacionar, por exemplo, um numero inteiro a cada um deles. Assim se por exemplo estivermos estudando uma variável  $X$ , a renda pessoal de cada uma das pessoas desta população, a renda da primeira pessoa será  $X_1$ , a renda da pessoa será  $X_2$  e a renda da ultima pessoa da população será  $X_N$  se a população tiver  $N$  elementos.

Muitas vezes, na maior parte dos casos não se examinam todos os elementos da população, por razoes de custo da pesquisa e de tempo. Então seleciona-se uma amostra de elementos da população. No caso em que investigamos todos os elementos da população a pesquisa é comumente chamada de censo, como é o caso do Censo Demográfico do IBGE que é realizado de 10 em 10 anos.

Os levantamentos por amostragem estão sujeitos a um erro denominado erro de amostragem ou erro amostral. Estes erros tem natureza probabilística, pois não podemos prever ou antecipar qual será o erro exato da amostra mas podemos calcular probabilidades de que o erro seja de um determinado valor. E mesmo isto só é possível quando nossa amostra é probabilística. Uma amostra probabilística é uma amostra cuja seleção é definida em termos de probabilidades de seleção dos elementos da população. Se definimos um regra de seleção da amostra a partir dos elementos da população atribuindo valores bem precisos de probabilidades de seleção para todos os elementos da população, dizemos que a amostra é probabilística. Por exemplo, podemos dizer que a probabilidade de seleção de cada elemento da população na amostra é um numero que deve ser diretamente proporcional a idade dos elementos. Então se um individuo tiver 30 anos ele terá o dobro da chance de ser selecionado que um individuo de 15 anos. Alem disto as probabilidades somadas de todos os indivíduos da população deve perfazer um total exatamente igual a 1. Dizemos então que  $P(X_i) = k \cdot \text{Idade}_i$  e

$\sum_{i=1}^N P(X_i) = 1$ . Se tivermos uma população com 10 indivíduos com as seguintes idades (10,20,30,40,50,60,70,80,90,100) a probabilidade de seleção do primeiro individuo será igual a  $k \cdot 10$ , a probabilidade de seleção do segundo individuo será  $k \cdot 20$  e então  $k \cdot 10 + k \cdot 20 + \dots + k \cdot 100 = 1$  e portanto  $k = 1/550$  e a probabilidade de seleção do primeiro individuo será igual a  $10/550 = 0,01818$  , a probabilidade de seleção do segundo



indivíduo será  $20/550 = 0,03636$  e a probabilidade de seleção do último indivíduo cuja idade é de 100 anos será de  $100/550 = 0,1818$ .

A amostra probabilística mais comumente utilizada aquela chamada de amostra aleatória simples que doravante chamaremos de AAS. Nesta todos os elementos da população tem a mesma probabilidade de serem selecionados. Por exemplo, suponhamos que temos uma população com  $N = 100$  e desejamos selecionar uma amostra de tamanho  $n = 30$ . Se cada elemento da população tem a mesma chance de entrar na amostra então a probabilidade de seleção de cada um deles será um número constante e igual a  $1/100$ . Esta será a probabilidade do elemento  $X_i$  ser selecionado na primeira extração da amostra. Como a seleção da amostra envolve 30 extrações aleatórias e independentes temos aqui um problema mais complexo. Dizemos com maior rigor que uma amostra é AAS se para cada uma das extrações todos os elementos da população tem idêntica probabilidade.

Aprofundemos esta questão para o caso mais simples, uma AAS selecionada com reposição. Neste exemplo de  $N = 100$  e  $n = 30$  a probabilidade do número de vezes em que cada um dos elementos da população está contido na amostra é uma variável aleatória binomial com parâmetros  $p = 100$  e  $n = 30$ . Por exemplo, qual é a probabilidade de que o décimo elemento da população (poderia ser qualquer um) ser selecionado 3 vezes na amostra de 30 elementos? Esta probabilidade é igual a

$$P(X = 3) = \binom{10}{3} \left(\frac{1}{100}\right)^3 \left(1 - \frac{1}{100}\right)^{30-3} = 0,00009148$$

Não vamos complicar para o caso de uma AAS selecionada sem reposição. Fica a cargo do aluno interessado e curioso tentar calcular qual seria a probabilidade de selecionar ao menos uma vez um dos 100 elementos da população em uma amostra de 30 elementos. Fica apenas a sugestão de utilizar a distribuição hipergeométrica (deve ser justificada).

Qual é a vantagem de uma amostra ser aleatória, frente a amostras não aleatórias, escolhidas por critérios subjetivos e ao gosto do pesquisador. Em primeiro lugar, porque como veremos adiante, uma amostra escolhida por critérios rigorosamente objetivos tende a não introduzir vieses que ocorrem quando a seleção é obtida por critérios discutíveis.

Um pesquisador de campo pode selecionar apenas domicílios que tem menos quantidade de pessoas para facilitar o seu trabalho e pode também evitar os domicílios de difícil acesso. Mas a principal vantagem de uma amostra probabilística (frente a uma amostra “subjetivista”) é que através de seus resultados é possível realizar cálculos probabilísticos.

Uma boa amostra deve apresentar 3 virtudes: 1) deve ser aleatória (pelos motivos que já introduzimos e que iremos aprofundar mais adiante). 2) deve ser precisa e 3) deve ser representativa e não geradora de vieses para as suas estimativas de parâmetros da população. A precisão de uma amostra irá depender de seu tamanho. Geralmente quanto maior o tamanho de uma amostra maior será a sua precisão. A representatividade da amostra dependerá não de seu tamanho, mas da maneira como os elementos são selecionados da população. Por exemplo, se tivermos uma população constituída de 30 % de mulheres e 70 % de homens e estivermos interessados em estimar a quantidade media de horas de estudo. Se selecionarmos uma amostra com 50 % de homens e 50 % de mulheres a quantidade media de horas de estudo desta amostra não será um bom estimador da quantidade media de horas de estudo da população (parâmetro), se a quantidade de horas de estudo for uma variável que depende do sexo da pessoa. Se as mulheres tenderem a estudar mais do que os homens esta amostra não representativa irá conduzir a um valor superestimado da quantidade de horas de estudo media. Já uma amostra que tenha o mesmo percentual de homens e de mulheres que existe na população será considerada representativa desta população e a media de horas estudo obtida desta amostra será uma estimativa não viesada da media de horas de estudo da população (parâmetro).

Podemos neste momento fazer uma importante distinção entre: 1) parâmetro; 2) estimador e 3) estimativa. Parâmetro é uma grandeza fixa e que se refere a uma população. No nosso exemplo anterior, o numero de horas de estudo médio na população é nosso parâmetro. O valor deste parâmetro é na maior parte dos casos um valor desconhecido, pois raramente podemos medir toda uma população. Geralmente desconhecemos o valor deste parâmetro e tentamos estimá-lo. Para estimá-lo utilizamos um estimador que se refere a uma formula matemática que será aplicada aos dados de

uma amostra. A media de uma amostra é um estimador do parâmetro – media da população. A estimativa é o valor que o estimador assume para uma dada amostra particular e somente para aquela amostra. As estimativas variam de amostra para amostra. Para cada amostra diferente temos diferentes estimativas e estas são calculadas de acordo com uma mesma fórmula, que é o estimador. No caso do numero de horas de estudo médio, podemos selecionar de uma população de 1000 estudantes diversas amostras de tamanho  $n = 50$ . Em uma primeira amostra, temos uma estimativa de media da amostra, em uma segunda amostra temos outro valor para a estimativa. Este valor da estimativa que oscilará de amostra para amostra poderá ser considerado resultado (valores observados) de uma mesma variável aleatória que é justamente o estimador. Então de forma geral podemos dizer que um estimador da media populacional ( $\mu_X$ ) é a media da amostra ( $\bar{X}$ ) e este estimador se comporta como uma variável aleatória sendo que cada um de seus valores de amostra para amostra é uma estimativa.

A seguir mostramos uma tabela com uma listagem de diversos estimadores e parâmetros.

Nome do estimador	Estimador	Nome do parâmetro	Parâmetro
Media amostral	$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$	Media populacional	$\mu = \frac{\sum_{i=1}^N X_i}{N}$
Variância amostral	$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$	Variância populacional	$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N}$
Proporção amostral	$\hat{p} = \frac{\sum_{i=1}^n X_i}{n}$ onde $X_i = 0$ ou $1$	Proporção populacional	$p = \frac{\sum_{i=1}^N X_i}{N}$ onde $X_i = 0$ ou $1$
Total amostral expandido	$\hat{T} = \frac{N}{n} \sum_{i=1}^n X_i$	Total populacional	$T = \sum_{i=1}^N X_i$

Na primeira linha desta tabela temos a media amostral que é um estimador da media populacional. Na segunda linha temos a variância amostral que é um estimador da

variância populacional. Observe o detalhe que o denominador da formula da variância amostral é  $n-1$  e não  $n$  (ao contrario o denominador da formula da variância populacional é  $N$  e não  $N-1$ ). Isto porque é necessário que o estimador seja não viesado o que quer dizer que  $E[s^2] = \sigma^2$ . Para entender melhor este conceito suponhamos que de uma população de tamanho  $N = 1000$  selecionemos todas as amostras possíveis de tamanho  $n = 50$ . Para cada uma destas amostras calculamos o valor de  $s^2$  (utilizando a formula da segunda linha, segunda coluna da tabela acima). Os valores de  $s^2$  irão variar de amostra para amostra e podemos assim considerar que  $s^2$  é uma variável aleatória pois depende de cada amostra selecionada (sendo que todas estas amostras tem o mesmo tamanho  $n = 50$ ). A media de todos estes valores de  $s^2$  deverá ser igual ao valor de  $\sigma^2$  (calculado utilizando-se a formula da segunda linha, quarta coluna). Caso não ocorresse isto o estimador seria viesado.

Vamos supor uma população de 4 elementos  $\{2,3,4,5\}$  tendo media  $\mu = 3,5$  e variância  $\sigma^2 = 1,25$

## 6.1 Amostragem Probabilística

- O que é uma amostragem probabilística ?
- É uma amostra selecionada de tal forma que cada item ou pessoa na população estudada têm uma probabilidade (não nula) conhecida de ser incluída na amostra.

Métodos de Amostragem Probabilística:

- **Amostragem Aleatória Simples (AAS)**

Uma amostra escolhida de tal forma que cada item ou pessoa na população tem a mesma probabilidade de ser incluída.

Se a população tem um tamanho  $N$ , cada pessoa desta população tem a mesma probabilidade igual a  $1/N$  de entrar na amostra. Utilizamos uma tabela de números aleatórios para sortear (com mesma probabilidade) os elementos da amostra. Também

pode ser utilizada uma função randômica: No Excel, por exemplo, temos a função ALEATÓRIO ENTRE.

- **Amostragem Aleatória Sistemática**

Os itens ou indivíduos da população são ordenados de alguma forma – alfabeticamente ou através de algum outro método. Um ponto de partida aleatório é sorteado, e então cada k-ésimo membro da população é selecionado para a amostra.

- **Amostragem Aleatória Estratificada**

A população é inicialmente dividida em subgrupos (estratos) e uma subamostra é selecionada a partir de cada estrato da população

- **Amostragem aleatória Estratificada com Repartição Proporcional**

Suponhamos que a população é subdividida em **k** estratos. Sejam:

$N$  = o número de indivíduos na população

$n$  = o número de indivíduos na amostra

$N_i$  = o número de indivíduos contidos no  $i$ -ésimo estrato da população

$n_i$  = o número de indivíduos contidos no  $i$ -ésimo estrato na amostra

$$n_i = n \times \frac{N_i}{N} \quad i = 1, 2, \dots, k$$

os estratos devem ser o mais homogêneos possíveis com relação às características relevantes da pesquisa (variáveis que se correlacionam fortemente com a variável estudada) para um mesmo tamanho amostral, a amostragem aleatória estratificada com

repartição proporcional é mais precisa (menor variância do estimador) do que a amostragem aleatória simples (AAS).

- **Amostragem Aleatória Estratificada com Repartição de Neyman (ou repartição ótima)**

Se conhecermos a variância de cada estrato populacional referente a variável que estamos desejando estimar o seu parâmetro, um método mais adequado é o da repartição de Neyman.

$$n_i = n \times \frac{w_i \sigma_i}{\sum_{i=1}^k W_i \sigma_i} = n \times \frac{N_i \sigma_i}{\sum_{i=1}^k N_i \sigma_i}$$

para um mesmo tamanho amostral a precisão é maior para amostra aleatória estratificada com repartição de Neyman (repartição ótima) do que para a amostra aleatória estratificada com repartição proporcional que por sua vez é maior do que a amostra aleatória simples

- **Amostragem por Conglomerados**

A população é inicialmente subdividida inicialmente em subgrupos (estratos) e uma amostra de estratos é selecionada (por exemplo, com probabilidade proporcional ao tamanho de cada estrato). A seguir, amostras são selecionadas dos estratos selecionados previamente.

A principal vantagem da amostra por conglomerados é a de possibilitar considerável redução de custos (em relação, por exemplo, a uma amostragem aleatória estratificada) para um mesmo tamanho amostral.

O método costuma ser empregado quando não dispomos de um cadastro da população (como no caso da amostragem sistemática) e os custos de ser elaborado um cadastro para toda a população é muito elevado.

- Erro amostral: A diferença entre a estatística amostral e seu correspondente parâmetro.
- Uma distribuição de probabilidade consiste de uma lista de todos os possíveis valores das médias amostrais de um dado tamanho amostral constante selecionado da população e a probabilidade de ocorrência associada a cada média amostral.
- **Exemplo 1** – Uma empresa tem 5 sócios. Semanalmente, os sócios relatam o número de horas de atendimento a clientes

Sócio	Horas
1	22
2	26
3	30
4	26
5	22

- Dois sócios são selecionados aleatoriamente. Quantas amostras ‘distintas são possíveis?
- O número de amostras distintas de dois elementos tomados em 5 objetos corresponde a:

$${}_5C_2 = \frac{5!}{(2!)(3!)} = 10$$

Sócios	Total	Média
1,2	48	24
1,3	52	26
1,4	48	24
1,5	44	22
2,3	56	28
2,4	52	26
2,5	48	24
3,4	56	28
3,5	52	26
4,5	48	24

- Organize as médias amostrais em uma distribuição de frequências.

Média Amostrai	frequência	Frequência Relativa (Probabilidade)
22	1	1/10
24	4	4/10
26	3	3/10
28	2	2/10

- Calcule a média das médias amostrais e compare-a com a média da população.

- A média da população é:

$$\mu = \frac{22 + 26 + 30 + 26 + 22}{5} = 25,2$$

- A média das médias amostrais é:

$$\frac{(22)(1) + (24)(4) + (26)(3) + (28)(2)}{10} = 25,2$$

- Observe que a média das médias amostrais é igual a média populacional

## 6.2 Teorema do Limite Central

- Para uma população com média  $\mu$  e uma variância  $\sigma^2$ , a distribuição amostral das médias de todas as possíveis amostras de tamanho  $n$ , geradas a partir da população,



será aproximadamente normalmente distribuída – com a média da distribuição amostral igual  $\mu$  e variância igual  $\sigma^2/n$  - **assumindo que o tamanho amostral é suficientemente grande**, ou seja,  $n \geq 30$ .

- Em outras palavras, se a população tem qualquer distribuição (**não precisa ser necessariamente normal**) com média igual a  $\mu$  e variância igual a  $\sigma^2$ , então a distribuição amostral dos valores médios amostrais é **normalmente distribuída** com

a **média das médias** ( $\mu_{\bar{X}}$ ) igual a **média da população** ( $\mu_X$ ) e o **erro padrão das médias amostrais** igual a  $\frac{\sigma}{\sqrt{n}}$ , desde que  $n \geq 30$ .

- Note que o erro padrão das médias amostrais mostra quão próximo da média da população a média amostral tende a ser.
- O erro padrão das médias amostrais é calculado por:

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$$

$\sigma_{\bar{X}}$  é o símbolo para o erro padrão das médias amostrais

$\sigma_X$  é o desvio padrão da população

$n$  é o tamanho da amostra

Se  $\sigma$  não é conhecido e  $n \geq 30$  (considerada uma amostra grande), o desvio padrão da amostra, designado por  $s$ , é usado para aproximar o desvio padrão da população,  $\sigma$ . A fórmula para o erro padrão torna-se:

$$S_{\bar{X}} = \frac{S}{\sqrt{n}}$$

onde  $s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$

### 6.3 Estimativa de Ponto

- Estimativa de ponto é um valor (chamado um ponto) que é usado para estimar um parâmetro populacional
- Exemplos de estimativas de ponto são a média amostral, o desvio padrão amostral, a variância amostral, a proporção populacional, etc.

Exemplo: O número de itens defeituosos produzidos por uma máquina foi registrado em cinco horas selecionadas aleatoriamente durante uma semana de trabalho de 40 horas. O número observado de defeituosos foi 12,4,7,14 e 10. Portanto, a média amostral é 9,4. Assim a estimativa de ponto para a média semanal do número de defeituosos é 9,4.

### 6.4 Estimativa de Intervalo

- Uma Estimativa de Intervalo estabelece uma faixa de valores dentro da qual um parâmetro populacional provavelmente cai.
- O intervalo dentro do qual um parâmetro populacional é esperado ocorrer é chamado de intervalo de confiança.
- Os intervalos de confiança que são extensivamente usados são os de 95 % e 99 %.
- Um intervalo de confiança de 95 % significa que cerca de 95 % dos intervalos construídos similarmente conterão o parâmetro que está sendo estimado.

- Outra interpretação do intervalo de confiança de 95 % é que 95 % das médias amostrais para um tamanho de amostra especificado cairão a uma distância máxima de 1,96 desvios padrões da média populacional.
- Para o intervalo de confiança de 99 %, 99 % das médias amostrais para um tamanho amostral especificado cairão a uma distância máxima de 2,58 desvios padrões da média populacional.

Os intervalos de confiança para 95 % e 99 % são construídos como segue, para  $n \geq 30$ :

- O IC de 95 % para a média populacional  $\mu$  é dado por:

$$\bar{X} \pm 1,96 \frac{s}{\sqrt{n}}$$

- O IC de 99 % para a média populacional  $\mu$  é dado por:

$$\bar{X} \pm 2,58 \frac{s}{\sqrt{n}}$$

- Em geral, um intervalo de confiança para a média, é calculado por:

$$\bar{X} \pm Z \frac{s}{\sqrt{n}}$$

onde  $Z$  é obtido da tabela de distribuição normal padrão.

### Exemplo 2

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 de 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 ?

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.

### **Interprete os resultados**

- Se nós tivéssemos tempo para selecionar aleatoriamente 100 amostras de tamanho 49 da população de alunos do campus e calcular as médias amostrais e os intervalos de confiança para cada uma destas 100 amostras, a média populacional (parâmetro) do número de horas trabalhadas estaria contida em cerca de 95 dos 100 intervalos de confiança. Cerca de 5 dos 100 intervalos de confiança não conteriam a média populacional.

## **6.5 Intervalo de Confiança para Uma Proporção Populacional**

Um intervalo de confiança para uma proporção populacional é dado por:

$$\bar{p} \pm Z\sigma_{\bar{p}}$$

onde:

$\bar{p}$  é a proporção amostral

$\sigma_{\bar{p}}$  é o erro padrão da proporção amostral e é dado por:

$$\sigma_{\bar{p}} = \sqrt{\frac{\bar{p}(1 - \bar{p})}{n}}$$

O intervalo de confiança é construído por:

$$\bar{p} \pm Z\sqrt{\frac{\bar{p}(1 - \bar{p})}{n}}$$

onde:

$\bar{p}$  é a proporção amostral

Z é o valor da variável normal padrão para o grau de confiança adotado.

n é o tamanho amostral

### Exemplo 3

Um planejador financeiro está estudando os planos de mudança de jovens executivos. Uma amostra de 500 jovens executivos que possuem suas próprias casas revelou que 175 planejam vendê-las e retirarem-se para o interior do País. Construa um intervalo de confiança de 98 % para o parâmetro proporção populacional de executivos que planejam mudar para o interior.

- Aqui  $n = 500$ ,  $\bar{p} = 175/500 = 0,35$   
e  $Z = 2,33$  (para  $\alpha = 0,98$  – nível de confiança adotado )
- O CI de 98 % é  $0,35 \pm 2,33 \sqrt{\frac{(0,35) \times (0,65)}{500}}$  ou  $0,35 \pm 0,0497$

### Interprete a resposta

## 6.6 Fator de Correção de População Finita

- Uma população que tem um limite superior definido é chamada de finita. Em estatística, considera-se como população finita quando  $n/N > 0,05$  (ou seja, quando a fração amostral é maior do que 5 %).
- Para uma população finita, onde o número total de objetos é N e o tamanho da amostra é n, o seguinte ajuste é feito para os erros padrões da média amostral e da proporção amostral.
- Erro padrão da média amostral:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

- Erro padrão da proporção amostral:

$$\sigma_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \sqrt{\frac{N-n}{N-1}}$$

- Este ajuste é chamado de Fator de Correção de População Finita (FCPF)

Nota: se  $\frac{n}{N} \leq 0,05$ , o fator de correção de população finita é ignorado.

#### Exemplo 4

A universidade do exemplo 2 quer estimar o número médio de horas trabalhadas por semana pelos estudantes. Uma amostra de 49 estudantes mostrou uma média de 24 horas e um desvio padrão de 4 horas. Construa um intervalo de confiança para o número médio de horas trabalhadas se há somente 500 estudantes no campus.

- Agora  $\frac{n}{N} = \frac{49}{500} = 0,098 > 0,05$ . Portanto, temos que usar o FCPF
- $24 \pm 1,96 \times \frac{4}{\sqrt{49}} \times \sqrt{\frac{500-49}{500-1}} = [22,93 ; 25,11]$

#### 6.7 Selecionando uma Amostra

- Há 3 fatores que determinam o tamanho de uma amostra, nenhum dos quais tendo uma relação direta com o tamanho da população. Eles são:
  1. O grau de confiança adotado
  2. O máximo erro permissível
  3. A variabilidade da população

Uma fórmula de cálculo conveniente para determinar o tamanho amostral **n** é:

$$n = \left( \frac{Zs}{E} \right)^2$$

onde:

E é o erro permissível

Z é o valor da variável normal padrão associado ao grau de confiança adotado

s é o desvio padrão da amostra piloto

### **Exemplo 5**

Um grupo de consumidores deseja estimar a média de gasto mensal em eletricidade para um domicílio familiar simples em Julho. Baseado em estudos similares o desvio padrão é estimado como sendo R\$ 20,00. Deseja-se construir um intervalo de confiança de 99 % com um erro máximo admissível de  $\pm R\$5,00$ . Qual deve ser o tamanho da amostra?

$$n = \left( \frac{(2,58) \times (20)}{5} \right)^2 = 106,50 \cong 107$$

### **6.8 Tamanho Amostral para Estimativa de Proporções**

A fórmula para determinar o tamanho amostral no caso de estimativa de proporções é:

$$n = \bar{p}(1 - \bar{p}) \left( \frac{Z}{E} \right)^2 \quad \text{onde}$$

$\bar{p}$  é a proporção estimada, baseada na experiência passada ou em uma amostra piloto

Z é o valor da variável normal padrão associado ao grau de confiança adotado.

E é o máximo erro permissível que o pesquisador tolera.

### Exemplo 6

- Um clube deseja estimar a proporção de crianças que tem um cachorro. Se o clube deseja que a estimativa esteja no máximo afastada 3 % da proporção populacional, quantas crianças devem conter a amostra? Assuma um intervalo de confiança de 95 % e que o clube estimou, com base em experiência anterior, que aproximadamente 30 % das crianças têm um cachorro.

$$n = (0,30)(0,70) \left( \frac{1,96}{0,03} \right)^2 = 893,4 \cong 893$$

## 7. Teste de Hipóteses – Amostras Grandes

### OBJETIVOS:

- 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
- Realizar um teste para a diferença entre duas médias ou proporções populacionais
- Descrever os erros estatísticos associados aos testes de hipóteses

### Nota:

- Se nada é conhecido acerca da população, a estimação 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 Teste de Hipóteses é usado para determinar a plausibilidade desta informação.

### O que é uma hipótese?

- Hipótese: uma sentença sobre o valor de um parâmetro populacional desenvolvida para o propósito de teste.
- Exemplos de hipóteses, ou sentenças, feitas acerca de um parâmetro populacional são:



- A renda média mensal proveniente de todas as fontes para os analistas de sistemas é de US 3625
- Vinte por cento de todos os transgressores juvenis são presos e sentenciados a prisão.

### O que é um Teste de Hipóteses?

- Teste de Hipóteses: 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.
- A seguir são propostos 5 passos para um teste de hipóteses:

Passo 1: Estabeleça a Hipótese Nula e a Hipótese Alternativa

Passo 2: Selecione um nível de significância

Passo 3: Identifique a Estatística de teste

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 aceitar  $H_1$

- Hipótese Nula  $H_0$ : Uma afirmação (sentença) sobre o valor de um parâmetro populacional
- Hipótese Alternativa  $H_1$ : Uma afirmação (sentença) que é aceita se os dados amostrais fornecem evidência de que a hipótese nula é falsa.
- Nível de Significância: A probabilidade de rejeitar a hipótese nula quando ela é efetivamente verdadeira, ou seja, valor de  $\alpha$  (alfa)
- **Erro Tipo I:** Rejeitar a Hipótese Nula,  $H_0$ , quando ela é efetivamente verdadeira. A probabilidade do erro tipo I é igual ao nível de significância,  $\alpha$  (alfa).
- **Erro Tipo II:** Aceitar a Hipótese Nula,  $H_0$ , quando é efetivamente falsa. A probabilidade do erro tipo II é igual a  $\beta$  (beta)

## Tipos de Erros

	Aceita $H_0$	Rejeita $H_0$
$H_0$ é verdadeira	Decisão Correta	Erro Tipo I
$H_0$ é falsa	Erro Tipo II	Decisão Correta

$\alpha$  = erro tipo I  $\beta$  = erro tipo II

Estatística de Teste (ou z efetivo ou valor de t): Um valor, determinado a partir da informação amostral, usado para determinar se devemos ou não rejeitar a hipótese nula.

- Valor Crítico (ou z crítico ou valor de t): O ponto divisor entre a região onde a hipótese nula é rejeitada e a região onde ela não é rejeitada. Este valor é obtido a partir da tabela de z (normal padrão) ou da tabela de t (t de Student).

### 7.1 Testes de Significância Unicaudais

- Um teste é unicaudal quando a hipótese alternativa,  $H_1$ , estabelece uma direção tal como:
  - $H_0$ : A renda média das mulheres é menor que ou igual a renda média dos homens.
  - $H_1$ : A renda média das mulheres é maior que a renda média dos homens.
  - A região de rejeição neste caso é a cauda direita (superior) da curva.

Figura com distribuição normal mostrando a região de rejeição para um teste unicaudal

### 7.2 Testes de Significância Bicaudais

- Um teste é bicaudal quando não existe uma direção especificada para a hipótese alternativa  $H_1$ , tal com:

- $H_0$ : A renda média das mulheres é igual a renda média dos homens.
- $H_1$ : A renda média das mulheres não é igual a renda média dos homens.
- A região de rejeição neste caso é dividida igualmente em duas caudas da curva.

Figura com distribuição normal mostrando a região de rejeição para um teste bicaudal

(distribuição amostral para a estatística  $z$  para um teste bicaudal, 0.05 de nível de significância)

Testando a Média Populacional: Amostra Grande, Desvio Padrão da População é conhecido.

- Neste caso a estatística de teste ( $z$  efetivo) é dado por:

$$z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

### Exemplo 1

- Os processadores de uma indústria indicam o ponto (marca) que a garrafa contem 16 onças (medida inglesa de peso) do produto. O Departamento de Controle de Qualidade é responsável pelo controle da quantidade incluída na garrafa. Uma amostra de 36 garrafas é selecionada por hora e o seu conteúdo pesado. Na última hora uma amostra de 36 garrafas apresentou um peso médio de 16,12 onças com um desvio padrão de 0,5 onças.
- Ao nível de significância de 0,05 podemos concluir que o processo está fora de controle?

**Passo 1:** Estabelecer a Hipótese Nula e a Hipótese Alternativa:

$$H_0: \mu=16 \quad H_1: \mu \neq 16$$

**Passo 2:** Estabelecer a regra de decisão:

$H_0$  é rejeitado se o  $z$  (efetivo – calculado com base nos valores amostrais)  $< -1,96$  ou  $z > 1,96$ .

**Passo 3:** calcule o valor da estatística de teste (  $z$  efetivo)

$$z = \frac{[16,12 - 16]}{[0,5 / \sqrt{36}]} = 1,44$$

**Passo 4:** Qual é a decisão sobre  $H_0$ ?

$H_0$  não é rejeitada, porque 1,44 é menor que o valor crítico de 1,96.

### 7.3 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 (efetivo) para o teste.
- Se o p-value é menor que o nível de significância (alfa),  $H_0$  é rejeitada.
- Se o p-value é maior que o nível de significância (alfa),  $H_0$  não é rejeitada.

### 7.4 Cálculo do p-value

- Teste Unicaudal (para a direita ou cauda superior):  
 $p\text{-value} = P\{z \geq \text{valor da estatística de teste calculada}\}$
- Teste Unicaudal (para a esquerda ou cauda inferior):

p-value =  $P\{z \leq \text{valor da estatística de teste calculada}\}$

- Teste Estatístico Bicaudal

p-value =  $2P\{z \geq \text{valor absoluto do valor da estatística de teste calculado}\}$

Para o exemplo anterior,  $z = 1,44$ , e desde que era um teste bicaudal, então o

p-value =  $2P\{z \geq 1,44\} = 2(0,5 - 0,4251) = 0,1498$ . Desde que  $0,1498 > 0,05$ , não é rejeitada  $H_0$ .

Testando para a Média Populacional: Grandes Amostras, Desvio Padrão Populacional desconhecido

- Aqui  $\sigma$  é desconhecido, portanto o estimamos com o desvio padrão amostral  $\underline{s}$ .
- Quanto maior for o tamanho amostral for  $n \geq 30$ , o  $z$  efetivo pode ser aproximado com

$$z = \frac{\bar{X} - \mu}{s / \sqrt{n}}$$

## Exemplo 2

- A cadeia de Lojas Arjo emite o seu próprio cartão de crédito. O administrador de crédito quer verificar se o saldo não pago mensal é maior do que US\$ 400. O nível de significância é fixado em 0,05. Uma amostra aleatória de 172 saldos não pagos revelou uma média amostral de US\$ 407 e o desvio padrão amostral de US\$ 38. O administrador de crédito pode concluir que a média populacional é maior que US\$ 400, ou é razoável assumir que a diferença de US\$ 7 (US\$ 407 – US\$ 400 é devido a chance (variação aleatória))?
- Etapa 1: Estabeleça a Hipótese Nula e a Hipótese Alternativa.

$$H_0: \mu \leq 400$$

*contra*

$$H_1: \mu > 400$$

- Etapa 2: Estabeleça a regra de decisão.

$H_0$  é rejeitada se o  $z$  (efetivo)  $> 1,645$ .

- Etapa 3: Calcule o valor da estatística de teste.

$$z = \frac{407 - 400}{38 / \sqrt{172}} = 2,42$$

- Etapa 4: Qual é a decisão sobre  $H_0$ ?

$H_0$  é rejeitada. O administrador conclui que a média dos saldos não pagos é maior do que US\$ 400.

Figura ilustrando a região de rejeição do exemplo

### 7.5 Teste de Hipóteses: Duas Médias Populacionais

- Assuma que os parâmetros para duas populações são:  $\mu_1, \mu_2, \sigma_1$  e  $\sigma_2$ .
- Caso I: Quando  $\sigma_1, \sigma_2$  são conhecidos, a estatística de teste ( $Z$  efetivo) é:

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

- Caso II: Quando  $\sigma_1, \sigma_2$  não são conhecidos mas os tamanhos amostrais  $n_1$  e  $n_2$  são maiores ou iguais a 30, a estatística de teste (Z efetivo) é:

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

### Exemplo 3

- Na indústria X foi realizado um estudo para comparar o número médio de anos de serviço para aqueles que se aposentaram em 1975 com aqueles que se aposentaram no último ano. Os seguintes dados amostrais foram obtidos. A um nível de significância de 0,01 podemos concluir que os trabalhadores que se aposentaram no último ano tiveram mais anos de serviço?

Característica	1975	Último ano
Média Amostral	25,6	30,4
Desvio Padrão Amostral	2,9	3,6
Tamanho amostral	40	4,5

- Estabeleça a Hipótese Nula e a Hipótese Alternativa

Considere que a população 2 é aquela dos que se aposentaram no último ano.

$$H_0 : \mu_2 \leq \mu_1 \quad H_1 : \mu_2 > \mu_1$$

- Estabeleça a regra de decisão

Rejeitar  $H_0$  se o  $z$  (efetivo)  $> 2,33$ .

- Calcule o valor da estatística de teste (valor de  $z$  efetivo):

$$z = \frac{30,4 - 25,6}{\sqrt{\frac{3,6^2}{45} + \frac{2,9^2}{40}}} = 6,80$$

- Nota: Desde que neste problema estamos testando para:

- $H_0: \mu_2 \leq \mu_1$

Precisamos trocar as posições das variáveis na equação do  $z$  efetivo (a seguinte equação).

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$z$  efetivo

- Qual é a decisão sobre a hipótese nula? Interprete os resultados?

Desde que o  $Z$  efetivo  $= 6,80 > Z$  crítico  $= 2,33$ ,  $H_0$  é rejeitada. Aqueles que se aposentaram no último ano tiveram mais anos de serviço.



## 7.6 Testes Referentes à Proporção

- **Proporção:** Uma fração ou porcentagem que indica uma parte da população ou amostra que tem um particular traço de interesse.

A proporção amostral é denotada por  $\bar{p}$  onde:

$$\bar{p} = \frac{\text{número de sucessos na amostra}}{\text{tamanho da amostra}}$$

### Estatística de teste para testar uma Proporção Simples de uma População

$$z = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

$p \equiv$  proporção populacional

$\bar{p} \equiv$  proporção amostral

#### Exemplo 4

- No passado, 15 % das solicitações postais feitas por uma instituição de caridade resultaram em contribuição financeira. Uma nova carta de solicitação foi redigida. Esta nova carta elevou a taxa de contribuição? A nova carta é enviada a uma amostra de 200 pessoas e 45 responderam com uma contribuição.
- Ao nível de significância de 0,05 pode-se concluir que a nova carta é mais efetiva?
- Estabeleça a Hipótese Nula e a Hipótese Alternativa:

$$H_0 : p \leq 0,15$$

$$H_1 : p > 0,15$$

- Estabeleça a regra de decisão

$H_0$  é rejeitada se o Z (efetivo)  $> 1,645$ .

- Calcule o valor da estatística de teste ( valor do Z efetivo):

$$z = \frac{\frac{45}{200} - 0,15}{\sqrt{\frac{(0,15)(0,85)}{200}}} = 2,97$$

- Qual é a decisão sobre a hipótese nula? Interprete os resultados.

Desde que o z efetivo = 2,97  $>$  z crítico (1,645),  $H_0$  é rejeitada. A nova carta é mais efetiva.

### Um Teste envolvendo a Diferença entre duas Proporções Populacionais

- A Estatística de teste (Z efetivo) neste caso é :

$$z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\frac{\bar{p}_c(1 - \bar{p}_c)}{n_1} + \frac{\bar{p}_c(1 - \bar{p}_c)}{n_2}}}$$

$n_1$  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 5

- Os trabalhadores solteiros são mais prováveis de faltar ao trabalho do que os trabalhadores casados?

Uma amostra de 250 trabalhadores casados mostrou que 22 faltaram mais do que 5 dias no último ano por alguma razão. Uma amostra de 300 trabalhadores solteiros mostrou que 35 faltaram mais do que 5 dias. Use o nível de significância de 0,05.

- Estabeleça a hipótese nula.

$$H_0 : p_2 \leq p_1 \quad H_1 : p_2 > p_1$$

onde o subscrito 2 refere-se a população dos trabalhadores solteiros.

- Estabeleça a regra de decisão.

Rejeitar  $H_0$  se  $z > 1,645$ .

- Calcular o valor da estatística de teste,  $Z$  efetivo:

$$\bar{p}_c = \frac{22+35}{250+300} = 0,1036$$

$$Z = \frac{\frac{22}{250} - \frac{35}{300}}{\sqrt{\frac{0,1036(1-0,1036)}{300} + \frac{0,1036(1-0,1036)}{250}}} = 1,10$$

Nota: Novamente, trocamos a posição das duas variáveis

- Qual é a decisão referente a hipótese nula?

$H_0$  é rejeitada. Não há diferença na proporção de ausências para trabalhadores casados e solteiros.

- Qual é o p-value?

p-value =  $P\{z > 1,1\} = 0,1357$ , ( a hipótese nula não é rejeitada).

## EXERCÍCIOS :

(incluem recordação de tópicos anteriores)

1. A Associação Nacional de Educação coleta e publica dados sobre o número de anos de experiência em sala de aula dos professores do curso secundário. Uma amostra é obtida neste ano de 10 professores de curso secundário e foram publicados os seguintes dados sobre o número de anos de experiência.

33	18	21	12	2
18	9	16	15	17

- a. Calcule a média amostral,  $\bar{X}$ , dos dados.

- b. Calcule a amplitude dos dados.
- c. Calcule o desvio padrão amostral,  $s$ , dos dados.
- d. Pelo Teorema de Chebychev, no mínimo \_\_\_\_\_ % dos dados caem dentro de dois desvios padrões de cada lado da média.

2. A seguinte tabela de contingência fornece uma distribuição de freqüências conjunta para os votos populares apurados na eleição presidencial de 1984 por região e por partido político. Os dados estão em milhares, arredondados para o mais próximo milhar.

		Democrata	Republicano	Outros	
		P1	P2	P3	Total
Nordeste	R1	9,056	11,336	101	20,493
Meio Oeste	R2	10,511	14,761	169	25,441
Sul	R3	10,998	17,699	136	28,833
Oeste	R4	7,022	10,659	214	17,895
Total		37,587	54,455	620	92,662

- a. Quantas pessoas votaram no partido Republicano?
- b. Quantas pessoas no Meio Oeste votaram?
- c. Quantas pessoas no Sul votaram no partido Democrata?
- d. Determine a probabilidade dos eventos R3 e P2 (simultâneos).
- e. Calcule  $\Pr(R3 \text{ ou } P2)$ , usando a tabela de contingência diretamente
- f. Calcule  $\Pr(R3 \text{ ou } P2)$ , usando a regra geral da adição de probabilidade, isto é,  $\Pr(A \text{ ou } B) = \Pr(A) + \Pr(B) - \Pr(A \text{ e } B)$ .
- g. Ache  $\Pr(R3 | P2)$ .
- h. Calcule  $\Pr(P1)$  e  $\Pr(P1 | R4)$ .
- i. São os eventos P1 e R4 independentes? Explique sua resposta.  
São os eventos P1 e R4 mutuamente exclusivos? Explique sua resposta.

3. Em um levantamento recente, a probabilidade de que um acidente de carro é causado por um motorista embriagado é cerca de 0,229. Nos próximos três acidentes, qual é a probabilidade de que:
- exatamente um acidente seja causado por um motorista embriagado?
  - No mínimo um acidente seja causado por um motorista embriagado?
  - Se você tem os seguintes resultados de probabilidade de acidentes causados por motoristas embriagados nos 10 próximos acidentes

	pdf (*)	Cdf (**)
0	0,0742	0,0742
1	0,2205	0,2947
2	0,2947	0,5893
3	0,2334	0,8227
4	0,1213	0,9440
5	0,0432	0,9873
6	0,0107	0,9980
7	0,0018	0,9998
8	0,0002	1,0000
9	0,0000	1,0000
10	0,0000	1,0000

(\*) Pdf = Probability Distribution Function (Função de Distribuição de Probabilidade)

(\*\*) Cdf = Cumulative Distribution Function (Função de Distribuição Cumulativa)

- Ache  $\Pr(x = 3)$ .
- Ache  $\Pr(5 < x \leq 9)$ .
- Qual é a média e a variância da distribuição tabulada acima?

4. Um dentista tem 5 cadeiras disponíveis para pacientes em sua sala de espera. A distribuição de probabilidade do número de cadeiras ocupadas,  $x$ , é dada por

$x$	$p(x)$
0	0,304
1	0,228
2	0,171
3	0,128
4	0,096
5	0,073

- Ache a média  $\mu$  da variável aleatória  $x$ .
- Calcule o desvio padrão,  $\sigma$ , da variável aleatória  $x$ .
- Calcule  $\Pr(2 \leq x \leq 5)$ .
- Desenvolva (no formato tabular a cdf (Cumulative Distribution Function - Função de Distribuição Acumulada) dessa distribuição.

5. Seja  $X$  normalmente distribuída com média  $\mu = 100$  e desvio padrão  $\sigma = 7$  (daqui em diante indicaremos tal distribuição como  $X \sim N(100;7)$ ). Determinar:

- $P(X = 80)$
- $P(X > 100)$
- $P(|X - 95| < 5)$
- $P(|X - 100| < 10)$

6. Dado que  $X$  é uma variável aleatória normal com média  $\mu = 10$  e  $P(X > 12) = 0,1587$ , qual é a probabilidade de que  $X$  esteja incluído no intervalo  $(9,11)$  ?

7. Os pesos de certos produtos em quilogramas são normalmente distribuídos com média  $\mu = 180$  e desvio padrão  $\sigma^2 = 4$ . Se uma unidade deste produto é escolhida aleatoriamente, qual é o peso desta unidade se a probabilidade de ocorrência :
- De um peso maior é igual a 0,10 ?
  - De um peso menor é igual a 0,05 ?
8. Se  $W$  é uma variável aleatória normal e se  $P(W < 10) = 0,8413$  e  $P(W < -10) = 0,0668$ , qual é  $E(W)$  e  $V(W)$  respectivamente ?
9. Há dois procedimentos para possibilitar que um determinado tipo de avião esteja pronto para a decolagem. O procedimento A requer um tempo médio de 27 minutos com desvio padrão de 5 minutos. Para o procedimento B,  $\mu = 30$  e  $\sigma = 2$  minutos, respectivamente. Qual procedimento deve ser utilizado se o tempo disponível é de 30 minutos? 34 minutos?
10. Suponha que os dividendos anuais de quatro ações sejam respectivamente \$ 2,00, \$ 4,00, \$ 6,00 e \$ 8,00. Deduza a distribuição amostral de  $\bar{X}$  considerando as seguintes hipóteses :

1. tamanho amostral  $n = 2$ .

2. método de amostragem: amostragem aleatória simples com reposição

Para a distribuição amostral deduzida de  $\bar{X}$ , verifique por demonstração que

a.  $E(\bar{X}) = \mu$

b.  $V(\bar{X}) = \sigma^2 / n$

c. Se a amostragem for sem reposição deduza a distribuição de  $\bar{X}$  e demonstre que

$$E(\bar{X}) = \mu \text{ e } V(\bar{X}) = \left( \frac{\sigma}{\sqrt{n}} \right) \left[ \sqrt{(N-n) / (N-1)} \right]$$

d. Se a amostragem fosse realizada com reposição, qual é o valor de  $V(\bar{X})$ ?



11. Uma população consta de 4 números: 3, 7, 11 e 15. Considerar todas as amostras possíveis que podem ser retiradas com reposição. Determinar: a) a média populacional; b) o desvio padrão da população; c) a média da distribuição amostral das médias; d) o desvio padrão da distribuição amostral das médias. Verificar (c) e (d) diretamente e por meio de (a) e (b) através das fórmulas apropriadas.
12. Certas válvulas fabricadas por uma companhia têm uma vida média de 800 horas e desvio padrão de 60 horas. Determinar a probabilidade de uma amostra aleatória de 16 válvulas, retiradas do grupo, ter a vida média: (a) entre 790 e 810 horas; (b) inferior a 785 horas. Para realizar esses cálculos, o que é necessário supor? Explique a razão de sua afirmativa.
13. De acordo com o exercício 8. Se for tomada uma amostra de 64 válvulas, como será resolvido? Explicar a diferença.
14. Os pesos de fardos recebidos por um depósito têm média de 150 kg e um desvio padrão de 25 kg. Qual é a probabilidade de 25 fardos, recebidos ao acaso e carregados em um elevador, não exceder o limite específico desse último, que é de 4100 kg? Neste caso, para a solução do problema, é necessário especificar a forma da distribuição estatística (função densidade de probabilidade) dos pesos dos fardos na população?

15. Questão teórica. Demonstre que  $s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$  é um estimador viesado para a

variância populacional  $\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N}$ , onde  $n$  é o tamanho da amostra e  $N$  é o tamanho da população. Calcule o valor do viés. O que ocorre com esse valor quando  $n$  tende ao infinito. (Lembrar que um estimador  $\hat{\Theta}$  de um parâmetro  $\Theta$  é dito não viesado se  $E[\hat{\Theta}] = \Theta$ )

16. Questão teórica

- a. Enuncie o Teorema do Limite Central e o interprete da melhor forma possível
- b. O que é considerado população finita (e infinita) para fins estatísticos ?
- c. Assinale as condições em que é necessário realizar a correção de população finita, justificando a resposta:

- ☐ quando a população é infinita, não importando se a amostragem é feita com ou sem reposição
- ☐ quando a população é finita, não importando se a amostragem é feita com ou sem reposição
- ☐ quando a população é finita e a amostragem é feita com reposição
- ☐ quando a população é finita e a amostragem é feita sem reposição
- ☐ quando a população é infinita e a amostragem é feita com reposição
- ☐ quando a população é infinita e a amostragem é feita sem reposição
- ☐ quando a população é finita ou a amostragem é feita com reposição
- ☐ existem outras alternativas não enumeradas acima

17. Uma função de probabilidade é uma regra de correspondência ou uma equação que:

- a) Acha o valor médio da variável aleatória
- b) Atribui valores de  $x$  a eventos de um experimento probabilístico
- c) Atribui probabilidades para valores de  $x$
- d) Define a variabilidade no experimento
- e) Nenhuma das anteriores é correta

18. Suponha que a variável aleatória  $T$  tenha a seguinte distribuição de probabilidade:

$t$		0	1	2
-----	--	---	---	---

-----  
 $P(T = t) \mid \begin{matrix} .5 \\ .3 \\ .2 \end{matrix}$

- a. Ache  $P(T \leq 0)$
- b. Ache  $P(T \geq 0 \text{ e } T < 2)$

Calcule  $E(T)$ , a média da variável aleatória  $T$ .

19. Uma centena de estudantes realizou um teste no qual o escore médio foi de 73 com uma variância de 64. Um grau A foi dado para quem obteve um escore de 85 ou mais. Quantos As foram obtidos aproximadamente, assumindo que os escores São normalmente distribuídos? (escolha o mais próximo)

- 1. 42
- 2. 7
- 3. 58
- 4. 5
- 5. 22

20. Se uma distribuição normal tem média 200 e desvio padrão 20, ache  $K$  tal que a probabilidade de que um valor amostral seja menor do que  $K$  é 0,975.

- a. 239    b. 204    c. 210    d. 215    e. 220
- f. 230    g. 239    h. 250

21. Se  $\bar{X}$  é a média de uma amostra extraída de uma distribuição normal com  $\mu = 10$ ,  $\sigma^2_X = 25$  e  $n = 9$ , então  $P(\bar{X} > 15)$  é:

- (a) 0,001350                      (c) 0,98778
- (b) 0,998650                      (d) 0,15866

22. A distribuição do tempo de vida de certo tipo de lâmpada elétrica é normalmente distribuída com média de 1000 horas e um desvio padrão de 100 horas. Ache o 33º Percentil da distribuição de tempo de vida.

- a. 560
- b. 330
- c. 1044
- d. 1440
- e. nenhuma das anteriores

23. O valor de Z correspondente ao 52º percentil é:

- a. 2,06
- b. 2,05
- c. 1,99
- d. 0,48
- e. 0,05

24.  $\Pr(Z > +1.96 \text{ ou } Z < -1.65)$  é

- 1) 0,025
- 2) 0,05
- 3) 0,0745
- 4) 0,0495
- 5) Nenhuma das anteriores

25. Em uma distribuição normal com média 3 e variância 49, quais são o limite superior e inferior para os 50 % dos dados centrais?

- a. -29,83 e 35,83

- b. -1,31 e 7,69
- c. -1,69 e 7,69
- d. 3,00 e 24,00
- e. nenhuma das anteriores

26. Uma amostra aleatória de tamanho 25 é escolhida de uma população com média 7 e variância 4. A média amostral é calculada como 8. Qual é o valor da variável normal padrão (z) correspondente a média amostral?

- a. 25
- b. 1,25
- c. -1,25
- d. +2,5
- e. nenhuma das anteriores

27. Suponha que para uma amostra de 36 Auxiliares de Enfermagem de diversos hospitais similares, uma avaliação de competência com intervalo entre 0 e 100 foi obtida a partir de um teste clínico. Suponha que a média populacional da avaliação para todas as Auxiliares de Enfermagem destes hospitais foi de 80 e a variância populacional foi de 100. Para uma amostra de 36 Auxiliares de Enfermagem, qual é a probabilidade de que a nota média esteja entre 75 e 80?

- a. 0,4987   b. 0,1915   c. 0,5013   d. 0,2287   e. 0,5115

28. Uma companhia fabrica cilindros que tem uma média de 2 polegadas de diâmetro. O desvio padrão dos diâmetros dos cilindros é de 10 polegadas. Os diâmetros de uma amostra de 4 cilindros são medidos todas as horas. A média amostral é usada para decidir se o processo de fabricação está operando satisfatoriamente ou não. A seguinte

regra de decisão é aplicada: se diâmetro médio da amostra de 4 cilindros é maior ou igual a 2,15 polegadas, ou menor ou igual a 1,85 polegadas, interrompe-se o processo.

- a. Qual é a probabilidade de parar o processo se a média do processo  $\mu$  permanece constante no valor de 2,00 polegadas ?
- b. Qual é a probabilidade de parar o processo se a média do processo muda para  $\mu = 2,10$  polegadas ?
- c. Qual é a probabilidade do processo continuar operando se a média do processo mudar para  $\mu = 2,15$  polegadas ?

29. Qual (ou quais) das seguintes sentenças descreve “inferência estatística” ?

- a. uma sentença verdadeira sobre uma população feita através de uma informação amostral de uma população
- b. uma conjectura acerca de uma população feita a partir da informação contida em uma amostra daquela população
- c. uma sentença verdadeira acerca de uma amostra feita a partir da informação contida em uma população.

30. Para uma certa população normalmente distribuída, o valor do desvio padrão é conhecido, mas o valor da média é desconhecido. Qual será o efeito de mudanças no tamanho amostral e do grau de confiança no comprimento do intervalo de confiança da estimativa da média populacional?

- a. Aumentando o tamanho amostral aumenta o comprimento dado um grau de confiança fixo.
- b. Aumentando o grau de confiança reduz o comprimento, dado um tamanho amostral fixo.
- c. Aumentando o tamanho amostral reduz o comprimento, dado um grau de confiança fixo.

d. Nenhuma das anteriores.

31. A distribuição das médias de todas as possíveis amostras de tamanho (n) escolhidas de uma população se aproximará de uma curva normal se

- a. n é grande o bastante
- b. a população é grande
- c. a população é simétrica
- d. a média de cada amostra é igual a média da população
- e. nenhuma das anteriores é correta

32. A distribuição amostral das médias de amostras aleatórias de tamanho n extraídas de uma população se aproximará de uma distribuição normal se

- a. somente se a população é normalmente distribuída e se n é grande
- b. somente se a população é normalmente distribuída não importando o valor de n
- c. se n é grande não importando a forma da distribuição da população
- d. não importa o valor de n e não importa a forma da distribuição da população original

33. Em um estudo sobre que relação existente entre uma atitude de criança e a idade na qual ela fala primeiro, os pesquisadores registraram a idade (em meses) da primeira fala da criança e o número de pontos (“escore”) obtido pela criança em um teste sobre a atitude. Seguem-se os dados para 21 crianças:

criança	1	2	3	4	5	6	7	8	9	10	11
Idade	15	2	10	9	15	20	18	11	8	20	7
Escore	95	71	83	91	102	87	93	100	104	94	113

Criança	12	13	14	15	16	17	18	19	20	21	
Idade	9	10	11	11	10	12	42	17	11	10	
Escore	96	83	84	102	100	105	57	121	86	100	

A linha de mínimo quadrado para a predição do “score” a partir da idade da primeira fala é:

$\text{escore} = 110 - 1,13 * \text{idade}$  ; o valor do coeficiente de correlação é  $-0,640$ .

- Que proporção da variabilidade nos escores da atitude é explicada pela reta de mínimos quadrados ?
- Qual seria a predição de mínimos quadrados para os escore de uma criança que fala primeiro aos 20 meses ?
- Calcule o resíduo para a criança 6.
- A partir do diagrama de dispersão, qual criança tem o maior (em valor absoluto) resíduo? O que é incomum para esta criança?
- Qual criança tem o menor valor ajustado?

34. Uma amostra no ano de 1989 de 130 mulheres que visitaram um ginecologista em uma determinada universidade do Noroeste dos EUA indicou que 113 tiveram experiência sexual.

- Assumindo que essas mulheres são uma amostra aleatória simples da população de todas as mulheres daquela universidade, calcule um intervalo de confiança para a proporção da população que é sexualmente ativa.
- O intervalo seria mais largo, mais estreito ou da mesma largura se 520 mulheres fossem amostradas? (Você não precisa fazer nenhum cálculo) Explique.



- c. O intervalo seria mais largo, mais estreito ou da mesma largura se resultassem 73 mulheres com experiência sexual 130 mulheres amostradas? (Você não precisa fazer nenhum cálculo) Explique.
  - d. Você acha que é razoável assumir que essas mulheres formam uma amostra aleatória? Explique.
35. Não execute nenhum cálculo para responder o seguinte. Explique seu raciocínio em cada caso.
- a. Tres pesquisadores Alex, Bob e Chuck selecionam de maneira independente amostras aleatórias da mesma população. Os tamanhos amostrais são 1000 para Alex, 4000 para Bob e 250 para Chuck. Cada pesquisador constrói um intervalo de confiança de 95 % para  $\mu$  a partir de seus dados. A semi-amplitude dos três intervalos são 0,015; 0,031 e 0,062. Relacione cada semi-amplitude com o pesquisador.
  - b. Cada um dos dois pesquisadores Donna e Eileen selecionam amostras aleatórias de tamanho 1000 de populações diferentes e constróem intervalos de confiança de 95 % para  $p$  (a proporção populacional). A semi-amplitude do intervalo de Donna é 0,030 e a de Eileen é 0,025. Dado que as proporções amostrais foram  $\bar{p}_1 = .20$  e  $\bar{p}_2 = .40$ , relacione cada pesquisadora com a sua proporção amostral.
  - c. Um pesquisador de nome Fran seleciona 100 indivíduos aleatoriamente de uma população, observa 50 sucessos e calcula 5 intervalos de confiança. Os níveis de confiança são 80 %, 90 %, 95 %, 98 % e 99 % e os cinco intervalos são (0,402 ; 0,598), (0,371 ; 0,629), (0,418 ; 0,582), (0,436 ; 0,564) e (0,384 ; 0,616). Relacione cada intervalo com o seu nível de confiança.
36. Suponha que 80 % de todos os habitantes da Pensilvânia comam Peru no Dia de Ação de Graças. Suponha além disso que você planeja selecionar uma amostra aleatória simples (AAS) de 300 habitantes da Pensilvânia visando determinar a sua proporção que come peru no Dia de Ação de Graças.

- a. 80 % é uma parâmetro ou uma estatística? Que símbolo você deve usar para representá-lo?
- b. De acordo com o Teorema do Limite Central, como a proporção amostral de quem come peru no Dia de Ação de Graças varia de amostra para amostra ?
- c. Determine a probabilidade de que menos do que 3 quartos da amostra comam peru no Dia de Ação de Graças.
- d. Seria a resposta a (c) menor, maior ou a mesma se o tamanho amostral de 800 fosse usado? (você não precisa executar o cálculo). Explique.
- d. Podemos mostrar que nesse contexto  $P(\bar{p} \leq 0,80) = 0.15$ . Se essa afirmativa não estiver correta escreva uma verdadeira que a substitua. Escreva uma ou duas sentenças explicando para um leigo o que essa afirmativa significa.

37. A seguinte tabela lista a temperatura média mensal e minha conta de eletricidade para aquele mês.

mês	temp	conta	mês	temp	Conta
Abr-91	51	\$41.69	Jun-92	66	\$40.89
Mai-91	61	\$42.64	Jul-92	72	\$40.89
Jun-91	74	\$36.62	Ago-92	72	\$41.39
Jul-91	77	\$40.70	Set-92	70	\$38.31
Ago-91	78	\$38.49	Out-92	*	*
Set-91	74	\$37.88	Nov-92	45	\$43.82
Out-91	59	\$35.94	Dez-92	39	\$44.41
Nov-91	48	\$39.34	Jan-93	35	\$46.24
Dez-91	44	\$49.66	Fev-93	*	*
Jan-92	34	\$55.49	Mar-93	30	\$50.80
Fev-92	32	\$47.81	Abr-93	49	\$47.64
Mar-92	41	\$44.43	Mai-93	*	*
Abr-92	43	\$48.87	Jun-93	68	\$38.70

Mai-92	57	\$39.48	Jul-93	78	\$47.47
--------	----	---------	--------	----	---------

A linha de mínimos quadrados é desenhada no diagrama de dispersão; a equação dessa reta é : conta = 55,1 – 0,214 temp. média

- Estime o valor do coeficiente de correlação entre a conta de eletricidade e a temperatura média.
- Qual é a predição de mínimos quadrados para a conta de energia elétrica em uma temperatura média de 60 graus F?
- Sem fazer cálculos, identifique que mês tem o maior (em valor absoluto) resíduo.
- Que mês tem o menor valor ajustado?

### Exercícios Resolvidos

### EXERCICIOS

1) Em quatro leituras experimentais de um “comercial” de 30 segundos, um locutor levou em media 29,2 segundos com desvio padrão de 5,76 segundos. Construir os limites de confiança para a media, dado  $\alpha = 10 \%$ , supondo que a população tem distribuição normal. Resp. (22,42; 35,98)

Solução:

Os limites de confiança para a estimativa por intervalo do parâmetro media populacional  $\mu$  é dado pela seguinte expressão:

$$\bar{X} - Z_{\alpha} \cdot \frac{\sigma_x}{\sqrt{n}} < \mu < \bar{X} + Z_{\alpha} \cdot \frac{\sigma_x}{\sqrt{n}}$$

No caso  $n = 4$  (o tamanho da amostra), a media amostral é  $\bar{X} = 29,2$  e o desvio-padrão amostral é  $s = 5,76$  (que utilizamos na expressão acima no lugar do  $\sigma_x$ , o desvio padrão da população, já que  $s$  é um estimador não viesado para este parâmetro da população). Para  $\alpha = 10\%$  vemos que o valor de  $z_\alpha$  correspondente é 1,6448. Assim o intervalo de confiança de 90 % de probabilidade ( $1 - \alpha = 0,90$ ) será:

$$29,2 - 1,6448 \cdot \frac{5,76}{\sqrt{4}} < \mu < 29,2 + 1,6448 \cdot \frac{5,76}{\sqrt{4}}$$

$$24,46 < \mu < 33,94$$

2) De 50.000 válvulas fabricadas por uma companhia, retira-se uma amostra aleatória de 400 válvulas e obtém-se a vida média de 800 horas, sendo o desvio padrão populacional de 100 horas.

- a) Qual é o intervalo de confiança de 99 % para a estimativa da media populacional?
- b) Com que confiança dir-se-ia que a vida média é de  $800 \pm 9,8$ ?
- c) Que tamanho deve ter a amostra para que seja de 95 % a confiança na estimativa do intervalo de  $800 \pm 7,84$ ?

Resp. a) (787,1; 812,9) b) 95 % c) 625

Solução:

$$\bar{X} = 800 \quad \sigma^2 = 100 \quad n = 400 \quad N = 50.000$$

- a)  $1 - \alpha = 0,99$

$$\bar{X} - Z_{\alpha} \cdot \frac{\sigma_X}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} < \mu < \bar{X} + Z_{\alpha} \cdot \frac{\sigma_X}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

$$800 - 2,5758 \times \frac{100}{\sqrt{400}} \times \sqrt{\frac{5000-400}{5000-1}} < \mu < 800 + 2,5758 \times \frac{100}{\sqrt{400}} \times \sqrt{\frac{5000-400}{5000-1}}$$

$$787,6 < \mu < 812,3$$

b)  $800 \pm 9,8$  Portanto:

$$z_{\alpha} \times \frac{100}{\sqrt{400}} \times \sqrt{\frac{5000-400}{5000-1}} = 9,8$$

$$z_{\alpha} = 2,0432$$

$$1 - \alpha = 0,9590$$

3) Uma amostra aleatória de 625 donas de casa revelou que 70 % delas preferem a marca X de detergente. Construir um intervalo de confiança para p = proporção das donas de casa que preferem X, com uma confiança de 90 %.

Resp. (67 %, 73 %)

4) Antes de uma eleição, um determinado partido político está interessado em estimar a proporção p de eleitores favoráveis a seu candidato. Uma amostra piloto de tamanho 100 revelou que 60 % dos eleitores eram favoráveis ao candidato em questão.

a) Determine o intervalo de confiança para a proporção de votos favoráveis para o conjunto dos eleitores, com uma confiança de 95 %.

b) Determine o tamanho da amostra necessário para que o erro cometido na estimação seja de, no máximo, 0,01 com probabilidade de 80 %.

Resp. a) (0,504; 0,696) b) n = 3933

5) Suponha que estamos interessados em estimar a percentagem de consumidores de certo produto. Se uma amostra de tamanho 300 forneceu 100 indivíduos que consomem o dado produto, determine:

- a) O intervalo de confiança de  $p$ , com coeficiente de 95 %.
- b) O tamanho da amostra para que o erro da estimativa não exceda a 0,02 unidade com probabilidade de 95 %.

Resp. a) 0,2800; 0,3866) b)  $n = 2134$

## EXERCICIOS RESOLVIDOS

1) Considere a função dada por:

$$f(x) = \begin{cases} 0 & \text{se } x < 1,5 \\ x & \text{se } 1,5 \leq x \leq 2 \\ 0,25 & \text{se } 2 \leq x \leq 2,5 \\ 0 & \text{se } x > 2,5 \end{cases}$$

- a. Mostre que  $f(x)$  é uma função densidade de probabilidade.
- b. Escolhido um valor ao acaso para  $x$ , qual é a probabilidade de  $x$  pertencer ao intervalo  $[1,5;2]$ ?

Solução:

a)  $f(x)$  é uma função densidade de probabilidade se

$$f(x) \geq 0 \text{ para qualquer } x \text{ e } \int_{-\infty}^{+\infty} f(x)dx = 1$$

dessa forma:

$$\int_{1,5}^2 x dx + \int_2^{2,5} 0,25 dx = x^2/2 \Big|_{1,5}^2 + 0,25x \Big|_2^{2,5} = 4/2 - 2,25/2 + 0,25 \times 2,5 - 0,25 \times 2 = 2 - 1,125 + 0,625 - 0,5 = 1$$

Portanto  $f(x)$  é uma função densidade de probabilidade

$$b) P(1,5 \leq x \leq 2) = \int_{1,5}^2 f(x) dx = \int_{1,5}^2 x dx = x^2/2 \Big|_{1,5}^2 = 4/2 - 2,25/2 = 0,875$$

2) Uma pessoa dispõe de R\$ 100.000 e pode comprar terrenos ou investir no mercado financeiro. Uma avaliação preliminar mostrou que:

- a. O cenário futuro para os terrenos indica que eles deverão valorizar em média 25 % em dois anos, mas fatores não controláveis transferem grande variabilidade para esta previsão. Acredita-se que a valorização tenha uma variância de 12 %.
- b. O mercado financeiro é mais estável e acena com uma taxa de ganho de 20 %, com variância de 4 % em dois anos.

O investidor se satisfaz com um ganho de 16 % nessa operação e pretende decidir pelo investimento mais confiável neste sentido. Qual deve ser sua decisão, se supõe as distribuições normais?

Solução:

Seja  $X_1$  a variável aleatória que representa a taxa de valorização dos terrenos e  $X_2$  a taxa de valorização do mercado financeiro.

$$X_1 \sim N(25;12) \text{ e } X_2 \sim N(20;4)$$

$$P(X_1 \geq 16) = P\left(z \geq \frac{16-25}{\sqrt{12}}\right) = P(z \geq -2,598) = 1 - 0,0047 = 0,9953$$

$$P(X_2 \geq 16) = P\left(z \geq \frac{16-20}{\sqrt{4}}\right) = P(z \geq -2) = 1 - 0,0227 = 0,9773$$

O investidor deve escolher pela compra dos terrenos pois  $P(X_1 \geq 16) > P(X_2 \geq 16)$

3) Com a finalidade de estabelecer o custo de um novo produto, o encarregado de custos levantou os possíveis fornecedores de um componente desse produto. Dos 60 fornecedores cadastrados foram sorteados e consultados 6 deles. Os preços fornecidos apresentam uma média de 4,83 u.m. A experiência do encarregado indica que o desvio padrão para o preço médio é de 10 % deste preço. Qual deve ser o intervalo de confiança de 93 % para o preço médio desse componente?

Solução:

$$\bar{X} = 4,83 \quad \sigma_{\bar{X}} = 0,10\bar{X} = 0,10 \times 4,83 = 0,483 \quad n = 6 \quad N = 60$$

Observe que a amostragem é realizada considerando-se população finita já que  $n/N > 0,05$ . Além disso como  $n$  é menor que 30 e como no enunciado nada é dito sobre a distribuição de  $X$  não podemos aplicar o Teorema do Limite Central. Dessa forma, para solucionarmos o problema, temos que supor que o custo tem distribuição normal (caso contrário teríamos que trabalhar com a teoria das amostras pequenas e aplicar uma outra distribuição chamada "t" de Student). No cálculo do intervalo de confiança temos que fazer a correção de população finita.

O intervalo de confiança para um nível de confiança de  $1 - \alpha$  % é dado por:

$$P\left(\bar{X} - z_{\alpha} \times \sigma_{\bar{X}} \times \sqrt{\frac{N-n}{N-1}} \leq \mu \leq \bar{X} + z_{\alpha} \times \sigma_{\bar{X}} \times \sqrt{\frac{N-n}{N-1}}\right) = 1 - \alpha$$

Para  $1 - \alpha = 93$  % temos:

$$P\left(4,83 - 1,48 \times 0,483 \times \sqrt{\frac{60-6}{60-1}} \leq \mu \leq 4,83 + 1,48 \times 0,483 \times \sqrt{\frac{60-6}{60-1}}\right) = 0,93$$

$$P(4,146 \leq \mu \leq 5,514) = 0,93$$



4) Uma amostra de 50 elementos foi retirada de uma população de 500 elementos, para a avaliação da média populacional, fornecendo  $s(x) = 4$ . Qual deve ser o tamanho de uma amostra que avalie a média com erro máximo de 2 unidades, ao nível de confiança de 90 %?

**Solução:**

$$N = 500 \quad n = 50 \quad s_x = 4$$

**O tamanho da amostra (sem considerar correção de população finita) para a estimativa da média populacional é dado pela expressão:**

$$Z_\alpha \times \sigma_{\bar{x}} = erro$$

**Como**

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} \cong \frac{s_x}{\sqrt{n}}$$

**Temos:**

$$Z_\alpha \times \frac{s_x}{\sqrt{n}} = erro \quad \text{ou} \quad n = \left( \frac{Z_\alpha \times s_x}{erro} \right)^2 = \left( \frac{1,64 \times 4}{2} \right)^2 = 10,758 \cong 11$$

**Como  $n/N > 0,05$  temos que considerar a correção de população finita, ou seja, temos que empregar a seguinte relação:**

$$Z_\alpha \times \sigma_{\bar{x}} \times \sqrt{\frac{N-n}{N-1}} = erro$$

**elevando ao quadrado ambos os termos da equação acima temos:**

$$Z_{\alpha}^2 \times \sigma_{\bar{X}}^2 \times \frac{N-n}{N-1} = erro^2$$

$$Z_{\alpha}^2 \times \frac{s_x^2}{n} \times (N-n) = erro^2 \times (N-1)$$

**isolando n no primeiro membro temos:**

$$n = \frac{z_{\alpha}^2 \times s_x^2 \times N}{e^2 \times N - e^2 + z_{\alpha}^2 \times s_x^2} = \frac{1,64^2 \times 16 \times 500}{4 \times 500 - 4 + 1,64^2 \times 16} =$$

$$10,55 \cong 11$$

**Portanto observa-se que mesmo com a correção de população finita o tamanho amostral necessário não se altera.**