UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Instituto de Matemática
Cadernos de Matemática e Estatística
Série B: Trabalho de Apoio Didático

INTRODUÇÃO À ESTATÍSTICA

NOTAS DE AULA

Liane Werner
Márcia Echeveste

Série B, Número 37
Porto Alegre, setembro de 1997.

SUMÁRIO

Quando se ouve a palavra "estatística", logo se imagina: taxa de acidentes, índices de mortalidade, quilometragem por litro. Esse pensamento popular, relaciona a estatística com a descrição de fatos. A noção usual da estatística prende-se apenas à parte de organização e representação do dados, através de gráficos e tabelas.

Evidentemente que a parte de organização e descrição são importantes, mas a estatística vai além, sendo necessário também analisar e interpretar os dados.

A estatística é a ciência que se ocupa com a organização, descrição, análise e interpretação de dados. É uma ciência rica em ferramentas para auxiliar na tomada de decisão. O seu uso é de grande importância e muito difundido nos últimos tempos, uma vez que é aplicável em qualquer ramo do conhecimento que trabalhe com dados experimentais, tais como: economia, engenharia, medicina, química, biologia, ciências sociais, entre outros.

1. ANÁLISE ESTATÍSTICA PRELIMINAR:

1.1. Conceitos Básicos de Estatística:

**ESTATÍSTICA:* É a ciência que compreende a coleta, a organização, análise e interpretação de dados. Pode ser dividida em duas grandes áreas:

* *Estatística Descritiva*: Esta área se interessa em descrever dados geralmente associados a contagens e gráficos, a informação contém os dados. A idéia é remover os detalhes estranhos e focar a características de interesse. Onde estão os valores centrais? Como os valores se estendem? Que forma tem a distribuição dos valores? Existe alguma mudança nos valores com o passar do tempo? O objetivo da estatística descritiva é providenciar respostas para este tipo de perguntas.

* *Estatística Inferencial*: É o ramo da Estatística que se preocupa em obter conclusões sobre o todo a partir de parte deste todo, isto é, tomar decisões com base nos dados colhidos de uma amostra. Como o processo de indução não é exato, estamos sujeitos a um certo grau de incerteza. A Estatística Inferencial irá dizer até que ponto podemos estar errando em nossas induções, e com que *probabilidade*.

**POPULAÇÃO*: Conjunto de elementos que possui alguma característica em comum. Pode ser finito (quando se conhece o número total de elementos) ou infinito.

**AMOSTRA*: É um subconjunto da população, isto é, uma parte da população retirada segundo alguns critérios estatísticos.

* *RECENSEAMENTO*: É o estudo estatístico realizado em toda a população.

* *CENSO*: É o resultado do recenseamento.

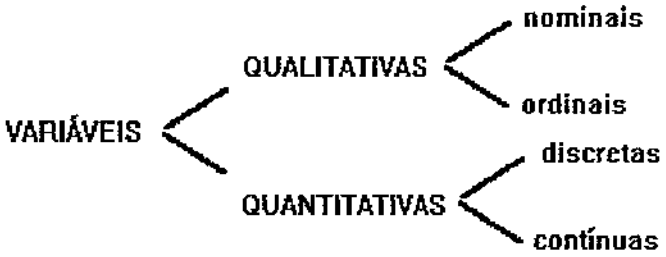
* *AMOSTRAGEM*: É o processo de obtenção de uma amostra, são técnicas, planos a fim de tornar representativa a amostra extraída da população.

**PARÂMETRO*: É uma medida característica da população em estudo.

Exemplo: Temos por população todos os veículos da marca W em Porto Alegre, sendo que uma podemos compor amostra dos veículos que são táxis dessa marca e podemos estar interessados em estudar a idade *média* (parâmetro) dos veículos dessa marca.

**VARIÁVEIS*: É a característica de interesse de uma população escolhida de acordo com o estudo.

Exemplo: Idade dos veículos da frota de Porto Alegre.



**Variáveis Qualitativas*: Expressam uma qualidade, podem ser chamada de ATRIBUTO, subdividem-se em:

NOMINAIS: Fornecem categorias ou nomes a alguma variável.
Exemplo: Sexo, estado civil, profissão.

ORDINAIS: As categorias de uma variável são ordenadas de acordo com a intensidade do fenômeno. Exemplo: classe social, grau de instrução.

**Variáveis Quantitativas*: Expressam uma quantidade, subdividem-se em:

DISCRETAS: Podem ter valores observados somente em pontos isolados ao longo de uma escala. Exemplo: nº de pessoas, nº carros fabricados por dia.

CONTÍNUAS: Podem assumir qualquer valor ao longo de uma escala.
Exemplo: Altura, idade, velocidade.

1.2. Descrição dos Dados:

Um conjunto de números pode reduzir-se a algumas medidas numéricas que resumem os dados. Quando analisamos um conjunto de dados é necessário encontrar um ponto que represente a localização dos dados (medidas de tendência central) e estudar a dispersão deste grupo (medidas de variabilidade).

**MEDIDAS DE TENDÊNCIA CENTRAL*: São valores que geralmente se localizam em torno do meio ou do centro de uma distribuição, onde a maior parte dos dados está concentrada.

**Média Aritmética*: É o ponto de equilíbrio dos dados, é dada pela soma de todos os elementos divido pelo número de parcelas.

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} \text{ (média da população)}$$

Exemplo: Suponha que ao passar pelo de acabamento de certo processo de manufatura, observe-se o tempo que um operário leva para examinar sete embalagens do mesmo produto. Considere o tempo em segundos:

50 s 51 s 49 s 52 s 51 s 49 s 50 s 51 s 49 s 48 s

Então: $T = \sum_{i=1}^{10} X_i = 500s \quad \Rightarrow \quad \mu = 50 \text{ s}$

**Mediana*: É a medida estatística de tendência Central que divide a distribuição dos dados ordenados em duas partes de igual freqüência, de forma que 50% das observações a antecedem.

No exemplo: Ordenamos os dados: 48 49 49 49 50 50 51 51 51 52

Calculamos a Posição da mediana dada por: P = N+1 / 2

$$P = \frac{10+1}{2} = 5{,}5$$

A mediana se encontra entre o 5° e o 6°### elemento

$$Md = \frac{X_5 + X_6}{2} = \frac{50+50}{2} = 50 \text{ s}$$

Caso N seja ímpar a mediana será o elemento posicionado em P = N+1 / 2.

**Moda:* É o valor que ocorre com maior freqüência. Podemos classificar as distribuições de acordo com o número de modas, conforme segue:

uma moda = unimodal;
duas modas = bimodal;
várias modas = multimodal.
sem moda = amodal

No exemplo: O conjunto é bimodal: 49 s e 51 s.

**MEDIDAS DE VARIABILIDADE*: Um aspecto fundamental da natureza é o fato que os objetos fisicos não se repetem com precisão, pelo contrário são caracterizados por uma certa diferença entre os elementos, a variabilidade.

Exemplo: Suponhamos que se deseja comparar o desempenho de dois funcionários, com base no número de formulários preenchidos corretamente durante uma semana:

Empregado A: 800, 810, 790, 800, 800 => $\mu\#\#\#_A = 800$ formulários

Empregado B: 700, 900, 800, 720, 930 => $\mu\#\#\#_B = 810$ formulários

Baseados nestes únicos resultados obtidos, diriamos que o desempenho de B é melhor do que de A, já que B produz, em média, um maior número de formulários diariamente. No entanto, se formos um pouco cuidadosos, percebemos que a produção de A varia de 790 a 810 formulários, ao passo que a de B varia de 700 a 930 formulários, o que indica que o desempenho de A é bem mais uniforme do que de B. É evidente que um alto grau de uniformidade costuma ser considerado como uma qualidade desejável nessa situação.

**Amplitude*: É a medida estatística de variabilidade ou dispersão mais simples, definida pela diferença entre o maior e o menor valor.

$$H = X_{máx} - X_{mín}$$

No exemplo: Para o empregado A temos: H = 810 - 790 = 20 formulários

**Variância*: É uma medida estatística que leva em consideração todas as informações do conjunto em análise, fazendo uso da soma de quadrados dos desvios em torno de μ. Denotada pelo símbolo σ^2 (na população).

$$\boxed{\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N} \text{ (fórmula conceitual)}}$$

$$\sigma^2 = \frac{\sum_{i=1}^{N}\left(X_i^2 - 2X_i\mu + \mu^2\right)}{N}$$

$$\sigma^2 = \frac{\sum_{i=1}^{N} X_i^2 - 2\mu\sum_{i=1}^{N} X_i + \sum_{i=1}^{N}\mu^2}{N}$$

$$\sigma^2 = \frac{\sum_{i=1}^{N} X_i^2 - 2\mu\sum_{i=1}^{N} X_i + N\mu^2}{N}$$

$$\sigma^2 = \frac{\sum_{i=1}^{N} X_i^2}{N} - 2\mu\frac{\sum_{i=1}^{N} X_i}{N} + \frac{N\mu^2}{N}$$

$$\sigma^2 = \frac{\sum_{i=1}^{N} X_i^2}{N} - 2\mu\mu + \mu^2$$

$$\boxed{\sigma^2 = \frac{\sum_{i=1}^{N} X_i^2}{N} - \mu^2 \text{ (fórmula operacional)}}$$

No exemplo: O empregado B tem variância

$$\sigma^2 = \frac{(700-810)^2 + (900-810)^2 + (800-810)^2 + (720-810)^2 + (930-810)^2}{5} = 95{,}52 \text{ form.}^2$$

OBS.: Aqui a unidade de medida é ao quadrado.

**Desvio Padrão*: Para resolver o problema da unidade de medida utilizamos outra medida estatística que consiste em extrair a raiz quadrada da variância. Denotado pelo símbolo σ (na população).

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}} \text{ (fórmula conceitual) ou } \sigma = \sqrt{\frac{\sum_{i=1}^{N} X_i^2}{N} - \mu^2} \text{ (fórmula operacional)}$$

No exemplo: O empregado B tem desvio padrão:

$$\sigma = \sqrt{\sigma^2} = \sqrt{95{,}52} = 9{,}77 \text{ formulários}$$

**Coeficiente de Variação*: É uma medida relativa de concentração dos dados em torno da média para a comparação de grupos distintos com médias diferentes ou unidades diferentes. Quanto menor o coeficiente de variação, mais homogêneo será o grupo de dados.

$$\gamma = \frac{\sigma}{\mu}$$

No exemplo: Empregado A: $\gamma = 6{,}32 / 800 = 0{,}0079$

Empregado B: $\gamma = 9{,}77 / 810 = 0{,}1142$

* *PROPORÇÃO:* É a fração ou percentagem de itens de determinado grupo ou classe em relação ao total observado.

$$\pi = \frac{x}{N}$$

onde "x" é o número de itens que apresentam certa característica e
"N" é o número total de observações.

Exemplo: Em um grupo de 40 pessoas 10 são fumantes dizemos que a proporção de fumantes é 10/40 = 0,25 ou 25%.

1.3. Distribuição de Freqüências

Representam as séries de dados agrupados onde o tempo, o espaço e a espécie do fenômeno permanecem constantes e as variações do fenômeno são agrupadas em subintervalos ou pontos dos dados observados. São divididas em:

* *DISTRIBUIÇÃO DE FREQÜÊNCIAS POR PONTO*: É uma tabela que contém para cada valor observado o número de vezes que ele ocorre(freqüência).

Exemplo: Em uma empresa com 20 funcionários foi realizado um estudo sobre o número de cafezinhos tomados durante o dia.

Suponha que os valores observados foram: 3, 2, 2, 0, 2, 1, 4, 0, 1, 1, 2, 3, 2, 2, 1, 0, 2, 2, 1, 2

Agrupando numa tabela de distribuição de freqüências por ponto , temos:

Classe de índice i	número de cafezinhos	f_i	f_{ri}	F_i	F_{ri}
1	0	3	3/20=0,15	3	0,15
2	1	5	5/20=0,25	8	0,40
3	2	9	9/20=0,45	17	0,85
4	3	2	2/20=0,10	19	0,95
5	4	1	1/20=0,05	20	1,00
Total	-	20	1,0	-	-

Sendo que:

1. **Freqüência Absoluta** (f_i) - é o número de observações ocorridas na classe i.
2. **Freqüência Acumulada** (F_i) - é a soma das freqüência absolutas até a classe i.
3. **Freqüência Relativa** (f_{ri}) - é a freqüência absoluta da classe i em relação ao total observado.
4. **Freqüência Relativa Acumulada** (F_{ri}) - é a soma das freqüências relativas até a classe i.

* *DISTRIBUIÇÃO DE FREQÜÊNCIAS POR INTERVALO*: É uma tabela que contém divisões da variável em estudo (intervalos) onde é observado o número de vezes que ocorrem os valores contidos nesses intervalos (freqüência).

Exemplo: Em um grupo com 40 pessoas foi realizado um levantamento das idades.

15	45	21	28	47	30	39	22	36	34
25	35	42	26	29	30	27	23	49	43
31	40	18	46	39	17	22	41	35	27
38	48	35	32	24	20	44	34	28	17

Agrupando os dados em uma tabela de distribuição de freqüências, temos:

Idades	f_i
15 |- 20	4
20 |- 25	6
25 |- 30	7
30 |- 35	6
35 |- 40	7
40 |- 45	5
45 |- 50	5
total	40

**Uma maneira de como montar a tabela*

1) Encontrar o maior valor dado observado e o menor valor observado.

Xmáx = 49 Xmín = 15

2) Calcular a amplitude total H

H = Xmáx - Xmín = 49 -15 = 34

3) Calcular o número de classes k

$k = \sqrt{N}$ (valor aproximado)

$k = \sqrt{40} = 6{,}32 \cong 7$ teremos então 7 classes

4) Calcular a amplitude das classes h

h = H / k

h = 34 / 7 = 4,85 $\cong$ 5 , isto é amplitude de 5 anos.

5) Estabelecer os limites de classes

Onde : l_i = limite inferior da classe i

L_i = limite superior da classe i

1ª classe: Podemos estabelecer o valor mínimo observado com o limite inferior da classe ou um valor inferior que melhor represente os dados.

$l_1 = 15$ somando a amplitude h, temos o limite superior da classe

$L_1 = 15 + 5 = 20$

2ª classe : $l_2 = 20$

$L_2 = 20 + 5 = 25$

e assim sucessivamente ...

classe i	Idades	f_i	f_{ri}	F_i	F_{ri}	Pto médio
1	15 |- 20	4	0,100	4	0,100	17,5
2	20 |- 25	6	0,150	10	0,250	22,5
3	25 |- 30	7	0,175	17	0,425	27,5
4	30 |- 35	6	0,150	23	0,575	32,5
5	35 |- 40	7	0,175	30	0,750	37,5
6	40 |- 45	5	0,125	35	0,875	42,5
7	45 |- 50	5	0,125	40	1,000	47,5
	total	40	1,000	-	-	

Sendo que: Ponto médio é o valor que representa a classe

Ponto médio = $(l_i + L_i)$ /2

Ponto médio da 1ª classe = (15+20)/2 = 17,5

Interpretação: $f_6 = 5$ => 5 alunos tem entre 40 e 45 anos

$F_2 = 10$ => 10 alunos tem de 15 a 25 anos

$f_{r5} = 0,175$ => 17,5 % dos alunos tem entre 35 e 40 anos

$F_{r3} = 0,425$ => 42,5% dos alunos tem de 15 a 30 anos

* *Medidas de posição e variabilidade*

* *Média aritmética*: Os valores são multiplicados por suas respectivas freqüências e para dados agrupados em distribuições de freqüência por intervalos, x_i são representados pelo pontos médios dos intervalos correspondentes.

$$\mu = \frac{\sum f_i x_i}{N} \text{ onde } N = \sum f_i$$

Como exemplo tomemos a velocidade de 70 motocicletas

classe i	velocidade	f_i (num de motos)	Ponto médio (x_i)	$f_i x_i$
1	50 |- 60	6	55	330
2	60 |- 70	9	65	585
3	70 |- 80	11	75	825
4	80 |- 90	22	85	1870
5	90 |- 100	16	95	1520
6	100 |- 110	4	105	420
7	110 |- 120	2	115	230
	total	70	-	5780

$$\mu = \frac{5780}{70} = 82,57$$

* *Mediana*: Para calcular a mediana nesse caso, devemos seguir os passos:

1) Encontrar a classe mediana
- Achar a posição da medida => P = N/ 2
- Calcular as freqüências acumuladas F_i

2)Calcular o valor da mediana

$$Md = l_i + h_i\left(\frac{N/2 - F_{i-1}}{f_i}\right)$$

onde: l_i : limite inferior da classe mediana
F_{i-1}: freqüência acumulada da classe anterior à classe mediana
f_i: freqüência da classe mediana
h_i: amplitude da classe mediana

No exemplo da velocidade das motocicletas:

Posição da mediana: P = 70/2 = 35

Calculando as freqüências acumuladas

classe i	velocidade	f_i (num de motos)	F_i	
1	50 |- 60	6	6	
2	60 |- 70	9	15	
3	70 |- 80	11	26	
4	80 |- 90	22	48	=> classe mediana
5	90 |- 100	16	64	
6	100 |- 110	4	68	
7	110 |- 120	2	70	
	total	70	-	

Como a mediana se encontra posicionada no 35° elemento e, este se encontra na 4ª classe, temos então a classe mediana.

Calculando o valor da mediana:

$$Md = 80 + 10\left(\frac{70/2 - 26}{22}\right) = 80 + 10.\frac{9}{22} = 80 + 4{,}09 = 84{,}09$$

* *Moda*: Para calcular a moda nesse caso, devemos seguir os passos:

1) Encontrar a classe modal, isto é, a classe com maior freqüência

2)Calcular o valor da moda: (pela fórmula de Czuber)

$$Mo = l_i + h_i\left(\frac{f_i - f_{i-1}}{2f_i - f_{i-1} - f_{i+1}}\right)$$

onde: l_i : limite inferior da classe modal
h_i: amplitude da classe modal
f_i: freqüência da classe modal
f_{i-1}: freqüência da classe anterior a classe modal
f_{i+1}: freqüência da classe posterior a classe modal

No exemplo da velocidade das motocicletas:

classe i	velocidade	f_i	
1	50 |- 60	6	
2	60 |- 70	9	
3	70 |- 80	11	
4	80 |- 90	22	=> classe modal
5	90 |- 100	16	
6	100 |- 110	4	
7	110 |- 120	2	
	total	70	

Calculando o valor da moda, temos:

$$Mo = 80 + 10\left(\frac{22-11}{2x22-11-16}\right) = 80 + 10.\frac{11}{17} = 80 + 6{,}47 = 86{,}47$$

**Variância absoluta*: É uma medida estatística que leva em consideração todas as informações do conjunto em análise, fazendo uso da soma de quadrados dos desvios em torno de μ. Denotada pelo símbolo σ^2 (na população).

$$\sigma^2 = \frac{\Sigma f_i (X_i - \mu)^2}{N} \text{ (fórmula conceitual) ou } \sigma^2 = \frac{\Sigma f_i x_i^2}{N} - \mu^2 \text{ (fórmula operacional)}$$

classe i	velocidade	f_i	Pto médio (x_i)	$(x_i-\mu)$	$(x_i-\mu)^2$	$f_i(x_i-\mu)^2$
1	50 |- 60	6	55	-27,57	760,105	4560,6294
2	60 |- 70	9	65	-17,57	308,705	2778,3441
3	70 |- 80	11	75	-7,57	57,305	630,3539
4	80 |- 90	22	85	2,43	5,905	129,9078
5	90 |- 100	16	95	12,43	154,505	2472,0784
6	100 |- 110	4	105	22,43	503,105	2012,4196
7	110 |- 120	2	115	32,43	1051,705	2103,4098
	total	70	-			14687,143

$$\sigma^2 = \frac{\Sigma f_i (X_i - \mu)^2}{N} = \frac{14687{,}143}{70} = 209{,}82$$

1.4. Exercícios

1) Classifique as variáveis abaixo como qualitativa (QL), quantitativa discreta (QTD) ou quantitativa contínua (QTC):

( ) QI funcional
( ) tempo para realizar uma prova
( ) número de acertos em um exame
( ) tamanho de camisa
( ) preço de um automóvel
( ) número da camisa dos jogadores

2) Cinco pessoas que estão tomando cafezinho na lancheria tem idade média 23 anos. Chegou uma sexta pessoa e a idade média passou a ser 26 anos. Qual a idade desta sexta pessoa?

3) A seguir temos o valor do aluguel (em R$ 1.000) de 20 fábricas situadas em certo distrito industrial.

8	9	8	10	7	12	10	12	8	9
12	10	10	7	8	7	9	9	15	7

a) Calcule o valor do aluguel médio e interprete.
b) Determine e interprete o valor mediano.
c) Determine e interprete o valor modal.
d) Calcule e interprete o desvio padrão.

4) Os dados abaixo representam as temperaturas em duas cidades. Qual a cidade que tem temperaturas mais homogêneas?

cidade A:	23	25	24	28	21
cidade B:	32	29	27	32	26

5) Ensaios em quarenta corpos de prova de concreto forneceram as seguintes resistências à ruptura :

64	61	65	43	45	54	51	74
30	100	91	75	78	68	80	69
72	27	40	93	99	94	78	72
59	78	95	62	42	96	100	95
81	84	78	103	98	60	84	91

a) Monte uma distribuição de distribuição de freqüências.
b) Calcule e interprete F_4, f_5, fr_2,Fr_3
c) Calcule e interprete a média e o desvio padrão.

2. PROBABILIDADE:

Independente da aplicação, a utilização da probabilidade indica que existe um elemento ao acaso (ou incerteza) quanto à ocorrência ou não de um evento futuro. Assim em muitos casos, pode ser impossível afirmar o que ocorrerá, mas é possível dizer o que pode ocorrer.

Esta área da estatística visa estabelecer um modelo matemático do fenômeno aleatório. O problema pode ser colocado como segue: dado um sistema que é completamente conhecido, tal como um baralho ou os componentes em uma mistura química, como pode o resultado de certos procedimentos ser descrito? Este tipo de questão deve ser respondida antes de técnicas estatísticas serem desenvolvidas ou utilizadas para a análise dos dados. Desta forma, o modelo matemático de teoria de probabilidade serve como base para técnicas estatísticas.

2.1. Definições Iniciais:

**EXPERIMENTO*: Qualquer procedimento que pode ser repetido e que, em cada uma das repetições produz um resultado (não necessariamente um valor, pode ser um vetor ou uma função).

**Experimento Determinístico*: As condições sob as quais um experimento é executado determina o resultado do experimento. Sob condições idênticas, os resultados são sempre os mesmos, qualquer que seja o número de ocorrência dos mesmos.

Exemplo: Ao aquecermos um determinado sólido, sabemos que a certa temperatura haverá a passagem para o estado líquido.

**Experimento Não-Determinístico ou Aleatório:* Apesar de repetirmos o experimento nas mesmas condições, não podemos afirmar que resultado particular ocorrerá. Quando vamos realizar um experimento aleatório, não podemos predizer, com certeza, qual o resultado ocorrerá, pois existe mais de um resultado possível, isto é, há uma variabilidade nos resultados das realizações do experimento.

Exemplos: Precipitação de chuva que cairá em determinada localidade. Com todas as informações precisas (pressão, velocidade do vento, altitudes, etc) não torna possível predizer "quanto" de chuva irá cair.

Em uma linha de produção, que está sujeita a alterações nos equipamentos e ao desempenho dos operadores entre fatores, fabricar peças em série e contar o número de peças defeituosas produzidas em um período de 24h.

**ESPAÇO AMOSTRAL*: Quando estamos diante de um fenômeno aleatório podemos descrever o conjunto de todos resultados possíveis. A este conjunto chamamos de *espaço amostral*. Denotado pelo conjunto S.

Quando descrevemos um espaço amostral associado a um experimento devemos ter idéia bastante clara do que estamos mensurando ou observando. Por isso, devemos falar de um espaço amostral associado a um experimento não "o" espaço amostral.

Exemplos: Lançar uma moeda até que ocorra cara. Contar o número de lançamentos

$$S = \{ 1,2,3, \ldots ,\infty\}$$

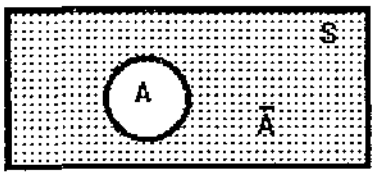
Lançar um dado e observar a face voltada para cima.

$$S = \{ 1,2,3,4,5,6\}$$

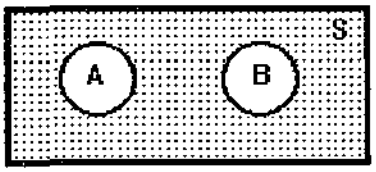
**EVENTO*: É um subconjunto de um espaço amostral S, isto é, um particular resultado dentre os existentes no espaço amostral. O conjunto vazio também constitui um evento.

Exemplo: A_1: A ocorrência de face par no lançamento de um dado. A_1:{ 2,4,6 }
A_2: Mais que dois rebites sejam defeituosos. A_2: { 3,4, ... , m }

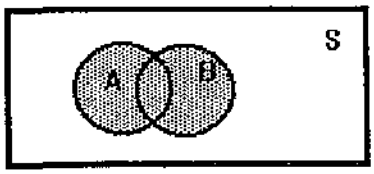
**Evento complementar*: $\overline{A}$ = A complementar. É formado por todos os pontos que pertencem ao espaço amostral S mas não pertencem a A.



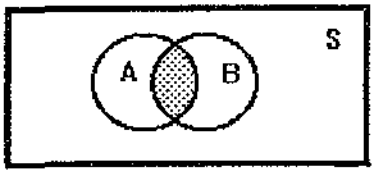
**Eventos Mutuamente Exclusivos*: Dois ou mais eventos são mutuamente exclusivos quando a ocorrência de um exclui a ocorrência de outro, isto é interseção de A e B é vazia (o conjunto vazio) .



**União de Eventos*: O evento A ∪ B ocorre se somente se A ocorre ou B ocorre ou ambos ocorrem. É formado pelos pontos que pertencem a pelo menos um dos eventos.



**Interseção de Eventos*: O evento A ∩ B ocorre se e somente se A e B ocorrem.

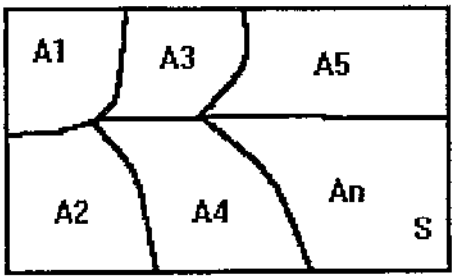


**PARTIÇÃO DO ESPAÇO AMOSTRAL*: Dizemos que os eventos $A_1, A_2, \ldots, A_n$ tornam um espaço a partição do espaço amostral S se:

i) $P(A_i) > 0$

ii) $A_i \cap A_j = \varnothing$ para $i \neq j$ ou seja os eventos A_i são mutuamente exclusivos.

iii) $\bigcup_{i=1}^{n} A_i = S$ => a união dos A_i é o espaço amostral



2.2. Conceitos de Probabilidade:

As probabilidades são utilizadas para exprimir a chance de ocorrência de um determinado evento.

**CONCEITO CLÁSSICO*: Supõe-se que todos os possíveis resultados de um experimento aleatório são igualmente prováveis.

Existem "n" resultados possíveis dos quais "a" são favoráveis a ocorrência do evento A. A probabilidade do evento A ocorrer é dada por:

$$P(A) = \frac{a}{n}$$

Exemplo: Qual a probabilidade de retirar uma dama de um baralho?
P(A) = 4 / 52

**CONCEITO FREQUENCIAL*: Baseia-se em dados históricos. É a relação do número de observações de um evento e o total observado.

Quanto maior o número de observações realizadas mais o valor da freqüência observada tenderá ao verdadeiro valor da probabilidade.

Exemplo: Se lançamos um dado não viciado 100 vezes e a fase 3 ocorreu 18 vezes. Temos que a face 3 tem 18 chances em 100. Lançando-se mais 200 vezes foi obtida 32 vezes a fase 3. Aumentando o número de lançamentos chegaremos cada vez mais próximos do valor que corresponderá à probabilidade de ser face 3 em um único lançamento.

**CONCEITO SUBJETIVO*: É o grau de crença do indivíduo de que o evento irá ocorrer, baseado em alguma evidência disponível.

Exemplo: Qual a probabilidade de tirar uma nota boa em estatística, antes de iniciar a prova?

2.3. Teoremas de Probabilidades. Teorema de Bayes.

**PROBABILIDADE AXIOMÁTICA*: Seja um espaço amostral e P uma função real definida em S, dizemos que P(A) é a probabilidade de A ocorrer, sendo:

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

$P(S) = 1$

Se A e B são mutuamente exclusivos, então:

$P(A_1 \cup A_2 \cup A_3 \cup ... \cup A_n) = P(A_1) + P(A_2) + P(A_3) + ... P(A_n)$

Exemplo: Uma fábrica que tem 100 funcionários, foi realizada uma pesquisa a respeito de uso de drogas e problemas de alcoolismo.

tem \ sexo problemas \	M	F	totais
Não	35	30	65
Sim	25	10	35
totais	60	40	100

Qual a probabilidade de selecionar um funcionário que seja do sexo feminino?

Qual a probabilidade de selecionar um funcionário que não seja do sexo feminino?

Qual a probabilidade de selecionar um funcionário do sexo masculino que não tenha problemas?

Qual a probabilidade de selecionar uma máquina da empresa do sexo feminino e que tenha problemas?

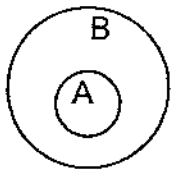
**TEOREMA 1*: Teorema da Soma. Se A e B são dois eventos quaisquer então:

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

Corolário: Para quaisquer eventos A, B, C temos:

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

* *TEOREMA 2*: Se $A \subset B$ então $P(A) \leq P(B)$.



**TEOREMA 3*: Probabilidade Condicional. Se A e B são dois eventos definitos no espaço amostral, a probabilidade de A ocorrer uma vez que B tenha ocorrido sendo denotado por P(A/B) é obtido por:

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

**TEOREMA 4*: Teorema do Produto. Através da definição de probabilidade condicional, temos que:

$$P(A \cap B) = P(B) \, . \, P(A/B)$$

**EVENTOS INDEPENDENTES*: Se A e B são independentes se P(A/B) = P(A) ou P(B/A) = P(B). A ocorrência de um não afeta (ou influencia) a ocorrência do outro.

$P(A \cap B) = P(B) \, . \, P(A/B)$ Como $P(A/B) = P(A)$, então:

$$P(A \cap B) = P(B) \, . \, P(A)$$

No exemplo:

O especialista de recursos humanos escolhe um funcionário do sexo masculino. Qual a probabilidade de ele não ter problemas?

Qual a probabilidade do funcionário ser do sexo feminino e ter problemas? (usando o teor.4)

**TEOREMA 5:* Teorema da Probabilidade Total: Sejam $A_1, A_2, \ldots, A_n$ eventos que formam um partição do espaço amostral. Seja B um evento desse espaço. Então:

$$P(B) = P(A_1). \, P(B/A_1) + P(A_2). \, P(B/A_2) + \ldots + P(A_n). \, P(B/A_n)$$

* Os eventos $(B \cap Ai)$ e $(B \cap Aj)$ para $i \neq j$ são mutuamente exclusivos.

* Através do teorema do produto, o evento B pode ser decomposto:

$$P(B) = P(A_1 \cap B) \cup P(A_2 \cap B) \cup P(A_3 \cap B) \cup ... \cup P(A_n \cap B)$$

Exemplo: Três máquinas A, B, C produzem 50%, 30%, 20% respectivamente do total de peças de uma fábrica. As percentagens de produção defeituosos destas máquinas são 3%, 4%, 5% respectivamente. Se uma peça é selecionada aleatoriamente, encontre a probabilidade de ela ser defeituosa.

**TEOREMA DE BAYES.* Sejam A_1, A_2, ... , A_n eventos que formam uma partição no espaço amostral. Seja B um evento deste espaço. Sejam conhecidos $P(A_i)$ e $P(B / A_i)$ onde i = 1,2,..., n então:

$$P(Aj / B) = \frac{P(Ai).P(B/Aj)}{\sum_{i=1}^{n} P(Ai).P(B/Ai)}$$

$$P(Aj / B) = \frac{P(Aj \cap B)}{P(B)}$$

No exemplo: Selecionamos uma peça e constatamos que era defeituosa. Determine a probabilidade dela ter sido fabricada pela máquina A.

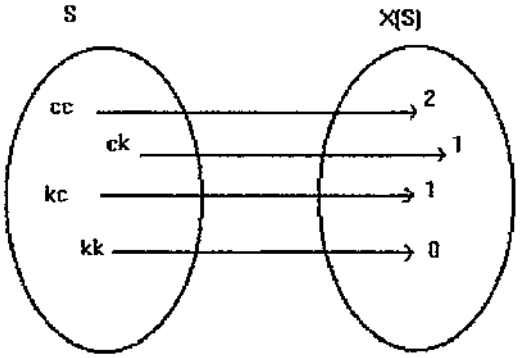
2.4. Distribuições de probabilidade de variáveis aleatórias discretas

**VARIÁVEIS ALEATÓRIAS*: Sejam E um experimento e S o espaço associado ao experimento. Uma função X, que associe cada elemento em S a um número real X(S) é denominado variável aleatória (v.a.).

Exemplo: Número de caras que ocorrem em dois lançamentos de um dados.

S: {CC, CK, KC, KK}

X(S) : {0,1,2}



**VALOR ESPERADO E VARIÂNCIA DE UMA V.A. DISCRETA*: Se X é uma variável que pode assumir os valores $x_1, x_2, \ldots, x_n$ e cada um desses valores estiver associado a uma e só uma probabilidade $P(x_1), P(x_2), \ldots, P(x_n)$.

O valor esperado de X é: $E[X] = P(x_1)x_1 + P(x_2)x_2, \ldots + P(x_n)x_n = \sum_{i=1}^{n} P(xi)xi$

A variância de X é: $VAR[X] = \sum_{i=1}^{n} Xi^2.P(Xi) - (E(X))^2$

Exemplo: Um empreiteiro faz as seguintes estimativas:

prazo de execução em dias (X)	probabilidade P(X)
10 dias	0,30
15 dias	0,20
22 dias	0,50

O prazo esperado para execução da obra é E[X] = 10x0,3+15x0,2+22x0,5 = 17 dias

A variação é dada por $VAR[X] = 317 - (17)^2 = 28$

**DISTRIBUIÇÃO DE PROBABILIDADE DE UMA V.A. DISCRETA*: São todos resultados de uma v.a. discreta e suas respectivas probabilidades.

A função $f(X)$ é a função de probabilidades de uma v.a.d. se, para cada possível resultado X temos:

1) $f(X) > 0$

2) $\sum_{i=1}^{n} f(x_i) = 1$

3) $P(X=x) = f(x)$

**FUNÇÃO DISTRIBUIÇÃO ACUMULADA (F(X))*: É a probabilidade acumulada de uma v.a.d. somando todas as probabilidades até um ponto X.

$$F(X) = P(X \leq x) = \sum_{i=1}^{x} f(x_i)$$

2.5. Distribuição de Probabilidade Binomial:

Usa-se o termo binomial para designar situações em que os resultados de uma v.a. podem ser agrupados em duas categorias, "sucesso" e "fracasso" que são mutuamente exclusivas. A distribuição binomial é útil para determinar a probabilidade de certo número de sucessos num conjunto de observações.

**CARACTERÍSTICAS*:

- O experimento consiste em n tentativas em iguais condições.
- Cada tentativa tem um resultado, entre dois possíveis: sucesso ou fracasso.
- As probabilidades de sucesso p e de fracasso $q = (1-p)$ permanecem constantes em todas as tentativas.
- Os resultados são independentes uns dos outros.

**CÁLCULO* : Para calcular uma probabilidade binomial, precisamos especificar:

n : número de tentativas

p : probabilidade de sucesso em cada tentativa

e é necessário observar:

x : número de sucessos (em n tentativas)

Em n tentativas, temos x sucessos com probabilidade p e n-x fracassos com probabilidade q. Como nessas n tentativas, não tem relevância a ordem de ocorrência dos x sucessos e n-x fracassos. Essa combinação é dada por:

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

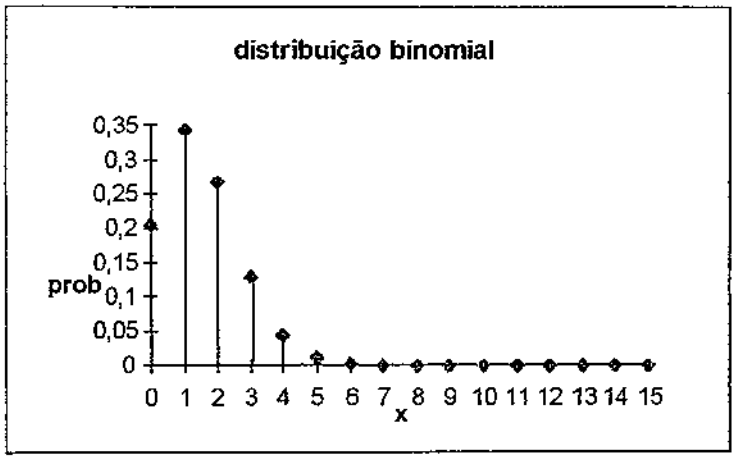
De modo que:

$$P(X=x) = C_x^n \; p^x (1-p)^{n-x}$$

Exemplo: Seja p=0,1 a probabilidade de encontrar um item defeituoso. Em 15 peças que tomamos aleatoriamente de uma linha produtiva, temos a probabilidade de obter x = 1, dada por:

$$P(X=1) = C_1^{15} 0{,}1^1 (1-0{,}1)^{15-1} = \frac{15!}{(15-1)!1!}.0{,}1.0{,}9^{14} = 0{,}3432$$

Para cada valor de X em {0,1,2, ..., 15} temos uma probabilidade, a figura abaixo mostra essas probabilidades graficamente.



**PARÂMETRO DA DISTRIBUIÇÃO:* A distribuição a binomial tem por parâmetro p (a probabilidade de sucesso). Seja X é o número de sucessos, então em função deste parâmetro podemos calcular:

$$E[X] = n\,p \qquad \text{e} \qquad VAR[X] = n\,p\,(1 - p)$$

No exemplo: $E[X] = n\,p = 0{,}10 \times 15 = 1{,}5$, logo temos em média 1,5 itens defeituosos nesta linha produtiva.

$$VAR[X] = n\,p\,(1 - p) = 0{,}1 \times 15 \times (1 - 0{,}1) = 1{,}35$$

2.6. Distribuições de probabilidades de variáveis aleatórias contínuas

**VARIÁVEL CONTÍNUA*: Quando uma variável pode tomar qualquer valor em determinado intervalo.

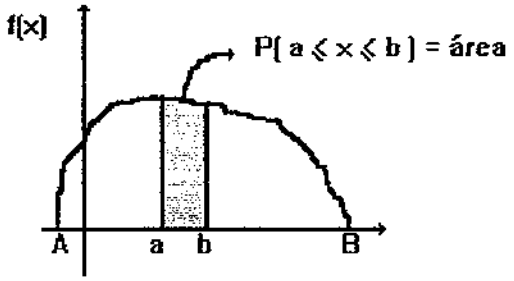
Exemplo: Concavidade de uma lente de contato.

**VALOR ESPERADO E VARIÂNCIA DE UMA VARIÁVEL ALEATÓRIA CONTÍNUA*:

$$E[X] = \int_{-\infty}^{+\infty} x\,f(x)\,dx$$

$$VAR[X] = E[X^2] - (E[X])^2 \quad \text{e} \quad E[X^2] = \int_{-\infty}^{+\infty} x^2 f(x)\,dx$$

**DISTRIBUIÇÃO DE PROBABILIDADE DE UMA V.A. CONTÍNUA*: Uma variável aleatória contínua pode assumir qualquer valor dentro de um intervalo definido, onde não podemos listar todos os valores com suas respectivas probabilidades. A solução é construir uma função densidade de probabilidade (f.d.p.), baseada na função f(x) correspondente.



* *FUNÇÃO DENSIDADE DE PROBABILIDADE*: Seja X uma v.a. contínua, a função f(x) é uma função densidade de probabilidade se satisfaz as seguintes condições:

1) $f(x) > 0$

2) $\int_{-\infty}^{+\infty} f(x)\,dx = 1$

3) $P(X \in A) = \int_{A} f(x)dx$, $A \subset \Re$

Observações;

- A probabilidade de X ser exatamente igual a um certo valor especificado x é igual a zero, isto é, P(X=x) = 0.
- Se x for uma v.a. contínua então $P(a \leq x \leq b) = P(a < x < b)$.
- A área abaixo da curva fornece a probabilidade, não a $f(x)$.

* *FUNÇÃO DISTRIBUIÇÃO ACUMULADA F(x):* Se X é uma v.a. contínua com função densidade de probabilidade f(X), então sua acumulada é:

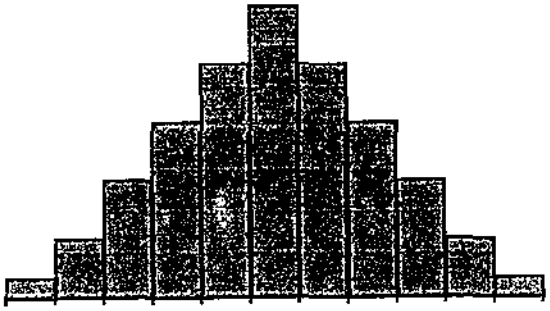
$$F(x) = \int_{-\infty}^{x} f(x)\,dx \qquad \text{assim} \qquad f(x) = \frac{dF(x)}{dx}$$

Assim sendo X uma v.a. contínua e F(x) sua função distribuição acumulada (f.d.a.) para dois pontos a e b quaisquer teremos:

$$P(a < x < b) = F(b) - F(a)$$

2.7. Distribuição de Probabilidade Normal:

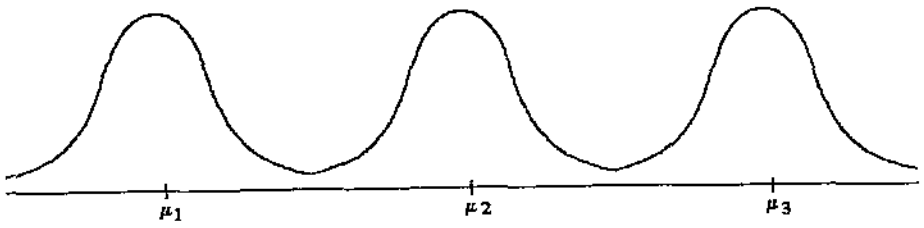
A curva normal é conhecida também por curva de Gauss, pois foi ele quem contribuiu para a sua teorização. A curva normal está associada a histogramas similares ao que vemos na figura abaixo, onde temos uma grande concentração em torno de um valor central e a medida que nos afastamos desse valor (para ambos os lados) a freqüência (ou probabilidade) ocorrência do fenômeno vai diminuindo.



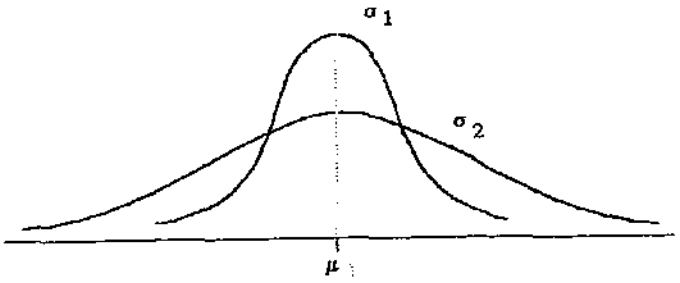
**CARACTERÍSTICAS*:

- A curva normal tem forma de sino.
- É simétrica em relação à média.
- Prolonga-se de $-\infty$ até $+\infty$.
- Cada distribuição normal é especificada por seus parâmetros média (μ) que varia de $[-\infty,+\infty]$ e o desvio padrão (σ) que varia entre $[0, +\infty]$. Existe uma curva normal distinta para cada combinação de μ,σ.

Curvas com médias diferentes (e mesmo desvio padrão):



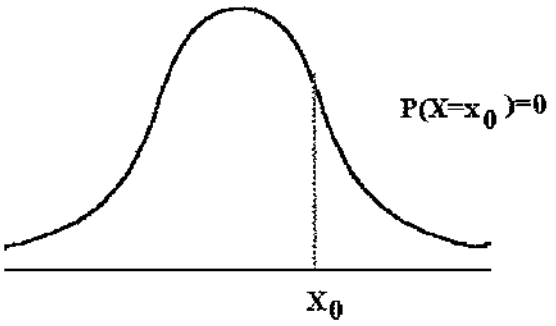
Curvas com desvios padrões diferentes (e mesma média):



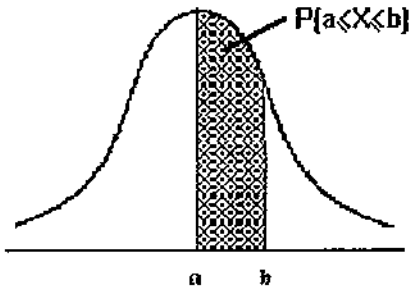
- A área total abaixo da curva é considerada como 100%. Isto é,

$$P(-\infty < x < +\infty)=1$$

- Como há um número infinito de valores entre $-\infty$ e $+\infty$, a probabilidade de uma variável aleatória normalmente distribuida assumir exatamente um valor Xo é zero.



- A área sob a curva entre dois pontos é a probabilidade de uma variável normalmente distribuída assumir um valor entre dois pontos.



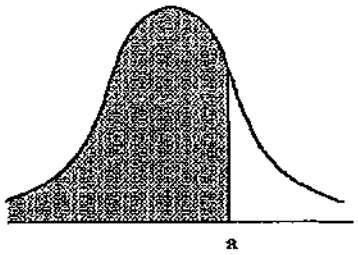
Para podermos calcular $P(a \le x \le b) = \int_a^b f(x)\,dx$, precisamos conhecer f(x) ou f.d.p. da normal.

* *FUNÇÃO DISTRIBUIÇÃO DE PROBABILIDADE*:

$$f(x)= \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \qquad -\infty < x < +\infty$$

* *FUNÇÃO DISTRIBUIÇÃO ACUMULADA*: A distribuição normal acumulada é definida como a probabilidade que a variável normal X é menor ou igual a algum valor "a", ou

$$P(X < a) = F(a) = \int_{-\infty}^{a} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$



Como a distribuição normal varia de local e forma para cada μ e σ, uma padronização e tabulação foi realizada para a curva normal com $\mu = 0$ e $\sigma = 1$.

* *DISTRIBUIÇÃO NORMAL PADRÃO*: As áreas correspondentes as probabilidades da distribuição normal padrão estão tabeladas. A unidade da distribuição normal padrão é chamada *escala z* que significa o número de desvios a contar da média.

As distribuições com $\mu \neq 0$ e/ou $\sigma \neq 1$, podem ser convertidas para a escala Z usando:

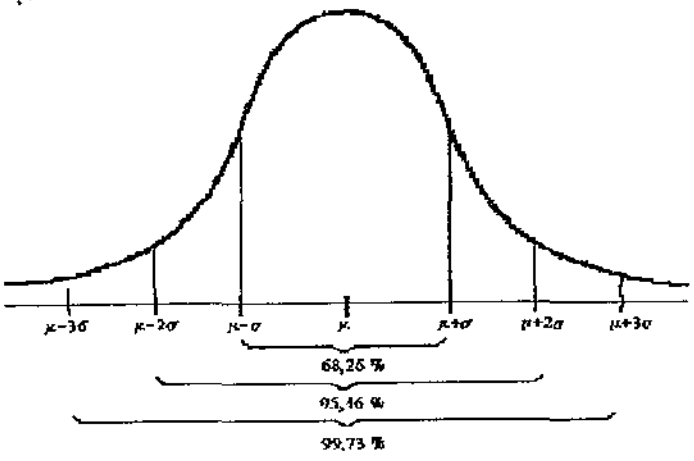
$$z = \frac{x - \mu}{\sigma}$$

Como z expressa a localização de unidades relativo a média usando o desvio padrão. Obtemos então:

$$P(X < a) = P\left(z < \frac{a-\mu}{\sigma}\right) = \phi\left(\frac{a-\mu}{\sigma}\right)$$

onde ϕ (.) é a distribuição normal acumulada e está tabelada conforme a tabela anexa.

Note que 68,26% dos valores estão entre os limites definidos por $\mu \pm \sigma$; 95,46% dos valores estão entre os limites definidos por $\mu \pm 2\sigma$; e 99,73% dos valores estão entre os limites definidos por $\mu \pm 3\sigma$.



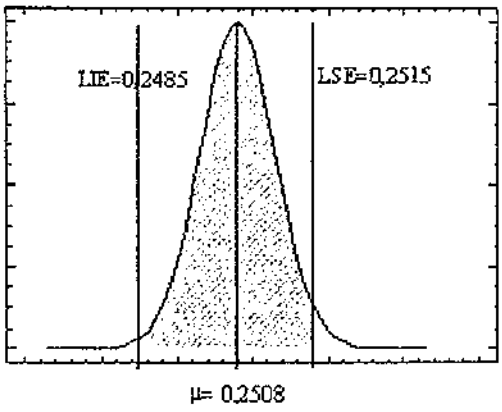
Exemplo: O diâmetro das hastes de metal de um disk drive é normalmente distribuído com $\mu = 0{,}2508$ polegadas e $\sigma = 0{,}0005$ polegadas. As especificações da haste forma estabelecidas como sendo $0{,}25 \pm 0{,}0015$ polegadas. Desejamos determinar a fração de hastes produzidas conforme as especificações.

$$P(0{,}2485 < x < 0{,}2515) = P(x < 0{,}2515) - P(x < 0{,}2485)$$

$$= \phi\left(\frac{0{,}2515 - 0{,}2508}{0{,}0005}\right) - \phi\left(\frac{0{,}2485 - 0{,}2508}{0{,}0005}\right)$$

$$= \phi(1{,}40) - \phi(-4{,}60) = 0{,}9192 - 0$$

$$= 0{,}9192$$



2.8 - Aproximação da Binomial pela Normal:

Muitas situações podem ser convenientemente descritas pela distribuição binomial. O que ocorre é que muitas vezes temos um grande número de observações (n grande), tornando os cálculos muito trabalhosos.

O uso da normal para aproximar a binomial apresenta dificuldade conceitual. A distribuição normal é contínua e, enquanto a binomial é discreta. A transição do caso discreto para o contínuo envolve a consideração de valores não-inteiros associados às variáveis contínuas, mas não a variáveis discretas.

O problema se resolve atribuindo intervalos da distribuição contínua para representar valores inteiros comuns as variáveis discretas. Por exemplo: os valores contínuos de 6,5 e 7,5 se associam ao inteiro 7. Assim para determinar a probabilidade binomial de exatamente 7 sucessos, deveríamos usar uma aproximação normal baseada na probabilidade (área abaixo da curva) entre 6,5 e 7,5.

Exemplo: Numa linha produtiva a proporção de defeituosos é 0,4, em 20 itens que tomamos aleatoriamente da produção. A probabilidade de encontrarmos 3 itens defeituosos é:

$$P(X=3) = C_3^{20}\, 0{,}4^{3}\,(1-0{,}4)^{20-3} = 0{,}0124$$

Como a normal é expressa em função da média e desvio padrão, calculamos:

$$\mu = n.p = 20.\ 0{,}4 = 8 \qquad \text{e} \qquad \sigma = \sqrt{np(1-p)} = \sqrt{20.0{,}4.0{,}6} = 2{,}2$$

"exatamente 3" deve ser interpretado como o intervalo de 2,5 a 3,5 na curva normal.

$$\begin{aligned} P(2{,}5 < X < 3{,}5) &= P(X<3{,}5) - P(X<2{,}5) \\ &= P(\,Z < 3{,}5 - 8 / 2{,}2\,) - P\,(\,Z < 2{,}5 - 8 / 2{,}2) \\ &= \phi\,(\,-2{,}5) - \phi(\,-2{,}05) \\ &= 0{,}9938 - 0{,}9798 \\ &= 0{,}0140 \end{aligned}$$

2.9. EXERCÍCIOS

1) De 120 pessoas que solicitaram emprego, em uma empresa, 50 possuem experiência anterior e 30 possuem um certificado profissional especial, dos quais 13 possuem experiência anterior e o certificado. Qual a probabilidade de um candidato escolhido aleatoriamente :

a) ter exper. anterior ou certificado especial.

b) ter certificado, dado que tem experiência anterior.

2) Sabendo que 3 máquinas produzem 3 tipos de lente, obtivemos:

tipo lente\ máq	m1	m2	m3
lente A	3	4	2
lente B	1	3	3
lente C	5	2	3

Escolheu-se uma máquina ao acaso e uma lente ao acaso, verificando-se que é uma lente B. Qual a probabilidade da lente ter vindo da m1? e da m2?

3) Sejam P(A) = 0,5 P(B) = 0,4 e P(A∪B) = 0,7. Pergunta-se se A e B são:

(a) mutuamente exclusivos? (b) independentes?

4) O número de chamadas telefônicas recebidas por uma telefonista e suas probabilidades para um intervalo de 3 min são:

# chamadas	0	1	2	3	4	5
P(X)	0.6	0.2	0.1	0.04	0.03	0.03

Em média quantas chamadas podem ser esperadas num intervalo de 3 min?

5) Num lote que tem 2% de defeituosos, foram retiradas 40 peças, que será rejeitado se forem encontradas duas ou mais peças defeituosas. Qual a probabilidade de rejeitar o lote?

6)Os registros de uma pequena companhia indicam que 40% das faturas por ela emitidas são pagas após o vencimento. De 14 faturas expedidas, determine a probabilidade de:

a)nenhuma ser paga com atraso.

b)no máximo 2 serem pagas com atraso.

c)pelo menos 3 serem pagas com atraso.

d)uma ser paga em dia.

7) Uma amostra de 3 m de cabo foi retirada de uma bobina. O cabo tem em média uma falha por m. Qual a probabilidade de não encontrar falha na amostra?

8)Um banco recebe em média 3 cheques sem fundo por dia. Qual a probabilidade de receber 8 cheques sem fundo numa semana de compensação?

9) Determine a probabilidade para os seguintes valores de z, traçando a curva e sombreando a área desejada.

a) entre 0 e 2. b)a esquerda de -1,87

c)a direita de 2,33 d)a esquerda de 1,34

e) entre -0,56 e -0,20 f) a direita de -1,29

10)Determine os seguintes valores de z para as seguintes áreas:

a) 0,5517 - área à esquerda
b) 0,0228 - área à esquerda
c) 0,0228 - área à direita
d) 0,9750 - área à esquerda

11) A vida útil de lavadora de pratos automáticas é de 1,5 anos, com desvio padrão 0,3 anos. Se os defeitos se distribuem normalmente, qual é a probabilidade de uma lavadora necessitar conserto antes de expirar o período de 1 ano de garantia?

12) O tempo necessário, em uma oficina, para o conserto de transmissão para certo carro é normalmente distribuído com média 45 min e desvio padrão 8 min. O mecânico planeja começar o conserto do carro 10 min após o cliente deixa-lo na oficina, comunicando que o carro estará pronto em 1 h. Qual a probabilidade de que o cliente tenha que esperar caso o mecânico esteja enganado e o cliente fique esperando?

13) Sabe-se que o conteúdo de uma lata de cerveja é 350 ml e que tem distribuição aproximadamente normal com média 350 ml e desvio padrão 10 ml.

a) que % de latas tem menos que 345 ml de conteúdo?
b) que % de latas tem mais que 360 ml de conteúdo?

14) Uma fábrica de pneus fez um teste para medir o desgaste de seis pneus e verificou que ele seguia o comportamento de uma curva normal com média 48.000 km e desvio padrão de 2.000 km. Calcule a probabilidade de um pneu escolhido ao acaso:

a)Dure mais que 47.000 km?
b) dure entre 45,000 e 51,000 km?
c) até que quilometragem duram 90% dos pneus?

3. AMOSTRAGEM e DISTRIBUIÇÕES AMOSTRAIS:

3.1. Introdução:

Até o momento, tomamos o conhecimento de alguns modelos probabilísticos que procuram medir a variabilidade de fenômenos aleatórios de acordo com suas ocorrências que eram as distribuições de probabilidade de variáveis aleatórias.

Na prática, raramente sabemos qual distribuição representa a variável. Obter a distribuição exata de alguma variável é muito dispendioso e as vezes impraticável, pois teríamos de ter todos elementos da população.

Por exemplo, se quiséssemos saber a resistência média de uma marca de lâmpada, teríamos que testar todas as lâmpadas até queimarem.

Assim, a solução é selecionar parte dos elementos (amostra), analisá-los e tirar conclusões para o todo (população). Este é o objetivo da *Estatística Inferencial.*

Logo, *Estatística inferencial* é o ramo da estatística que se preocupa em obter informações sobre o todo a partir de parte deste todo, ou seja, tomar decisões com base em dados colhidos de uma amostra.

3.2. Amostragem:

Por falta de tempo e recursos econômicos raras vezes se estuda individualmente todos os sujeitos da população na qual se está interessado. Em lugar disso, o pesquisador estuda uma amostra para generalizar as conclusões para a população.

Para que as nossas conclusões sejam confiáveis, é necessário que as amostras sejam obtidas de processos adequados que garantam a sua representatividade, ou seja, que a amostra reproduza as mesmas características da população no que diz respeito as variáveis de interesse.

**AMOSTRA REPRESENTATIVA*: É aquela amostra que representa todas as características importantes para o estudo existentes na população.

**TÉCNICAS DE AMOSTRAGEM*:

**Probabilísticas:* São aquelas em que a seleção das unidades é aleatória de tal forma que cada elemento da população tem uma probabilidade de pertencer a amostra.

**Não Probabilísticas*: São aquelas que não envolvem aleatoriedade na seleção dos elementos. Por exemplo: amostras intencionais, em que o especialista escolhe deliberadamente os elementos da amostra ou amostra de voluntários.

Para que possamos utilizar as técnicas de inferência estatística é necessário que o processo de escolha da amostra seja probabilístico, pois somente neste caso podemos avaliar a probabilidade de erro.

**TÉCNICAS DE AMOSTRAGEM PROBABILÍSTICA:*

**Amostragem Aleatória Simples (a.a.s)*: Também chamada de casual ou randômica. A característica principal é que todos os elementos têm igual probabilidade de pertencer à amostra. Para garantir que seja aleatório pode-se utilizar a tábua de números aleatórios que é desprovida de qualquer lei de informação.

**Amostragrem Aleatória Sistemática:* Quando os elementos da população se apresentam naturalmente ordenados e a retirada dos elementos é feita periodicamente.

**Amostragem Aleatória Estratificada*: Pode ser utilizada quando existem subgrupos dentro da população estudada que são homogêneos, mas que tem apresentam certas diferenciações entre os subgrupos.

**Amostragem Aleatória por Conglomerados*: Pode ser utilizada quando é possível identificar dentro da população subgrupos que representam uma miniatura da população. Estes subgrupos são chamados de conglomerados, diferenciando-se dos estratos por não haver homogeneidade interna (dentro de cada conglomerado os elementos são tão distintos quanto dentro da população).

3.3. Distribuição Amostral:

A finalidade da amostragem é obter uma indicação do valor de um ou mais parâmetros de uma população, tais como média, variância da população ou proporção.

Quando extraímos aleatoriamente repetidas amostras de uma mesma população a estatística amostral varia de uma amostra para a outra, chamamos esta variação de variabilidade amostral.

O objetivo é saber o quão próximo está a estatística amostral do verdadeiro parâmetro. Para isso três fatores são importantes: O estudo da distribuição de probabilidade da estatística amostral; o tamanho da amostra (grandes amostras têm menor variabilidade entre as estatísticas do que em pequenas amostras) e ainda a variabilidade na população (populações com muita variabilidade produzem estatísticas amostrais com maior variabilidade).

A variabilidade amostral pode ser expressa em uma distribuição de probabilidade que associa aos possíveis resultados de uma estatística amostral suas respectivas probabilidades.

**PARÂMETROS e ESTATÍSTICAS:*

**Parâmetros* são medidas estatísticas obtidas através do censo para descrever uma característica da população.

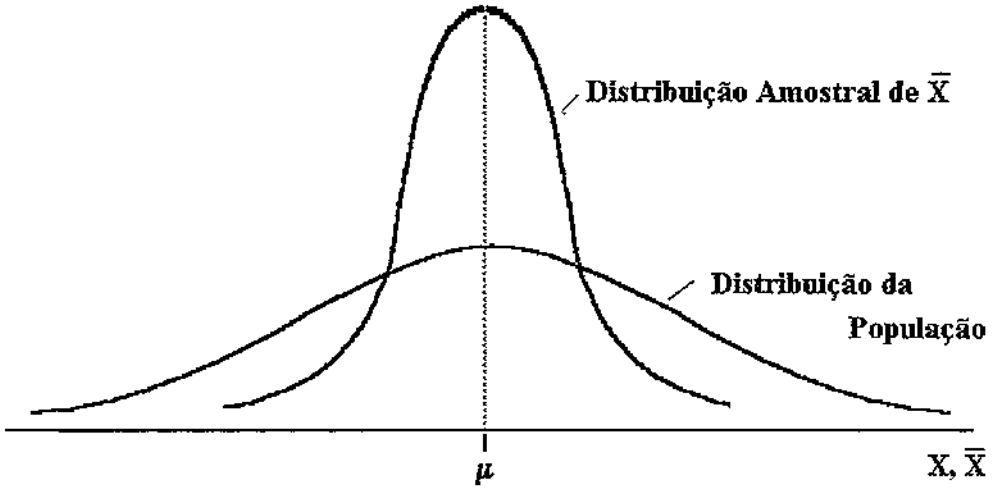
**Estatísticas* são medidas características obtidas através de uma amostra.

Medida	Parâmetro	Estatística
Média	μ	$\bar{x}$
Desvio padrão	σ	s
Variância	σ^2	s^2
Proporção	Π	p

3.4. Distribuição Amostral das Médias:

Uma distribuição amostral das médias indica a probabilidade de ocorrência de uma média amostral.

As médias amostrais tendem a agrupar-se em torno da média populacional.



Distribuição amostral de $\overline{X}$ - População Normal

A média das médias amostrais é igual a verdadeira média populacional.

$$E[\overline{X}] = \mu$$

E o desvio padrão da distribuição amostral das médias será dado por:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

** Teorema do Limite Central:*

- Se a população sob amostragem tem distribuição normal, a distribuição das médias amostrais também será normal.

- Mesmo que a população não seja considerada distribuição normal, a distribuição das médias amostrais será aproximadamente normal para grandes amostras.

$$\overline{x} \cong N(\mu, \sigma/\sqrt{n}) \quad \text{Logo,} \quad Z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$$

Exemplo: Uma var. aleatória X tem distribuição normal com média 100 e desvio padrão 10.

a) Qual a P(90 < X < 110)?

b) Se $\overline{x}$ é a média de uma amostra de 16 elementos retirados dessa população, calcule P(90< $\overline{x}$ < 110).

c) Que tamanho deveria ter a amostra para que P(90< $\overline{x}$ < 110) = 95%.

3.5.Distribuição Amostral das Médias quando σ é Desconhecido:

Quando desconhecemos o desvio-padrão populacional utilizamos como estimativa o valor de s. Desta forma, o desvio-padrão das médias (ou erro padrão) será dado por:

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}, \quad \text{onde} \quad s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

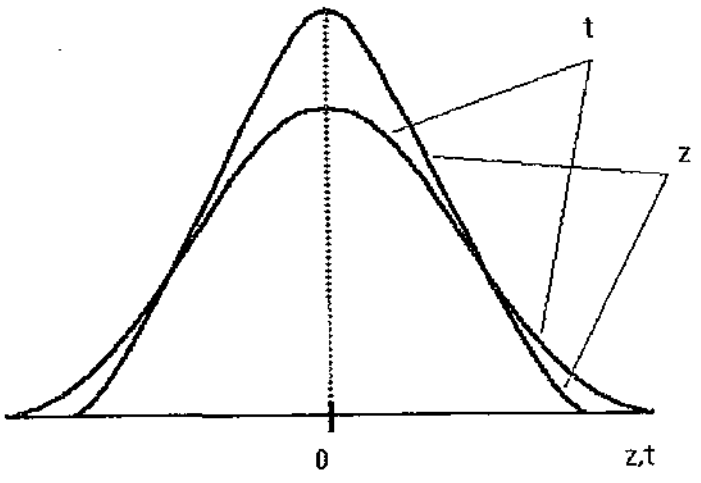
Para grandes amostras, podemos admitir que a variação dos valores observados na amostra seja semelhante a variação da população. Porém, para pequenas amostras isso pode não ser verdadeiro. Neste caso, a distribuição adequada é a distribuição t-student.

Assim, a estatística:

$$\frac{(\bar{x} - \mu)}{\frac{s}{\sqrt{n}}} \approx t - student \text{ com n-1 graus de liberdade.}$$

Esta distribuição é muito parecida com a distribuição normal, sendo simétrica em torno da média zero, porém tem maior dispersão comparado com a normal.

A forma da distribuição t-student depende do tamanho da amostra. Quanto menor o tamanho da amostra, menor serão os graus de liberdade e mais dispersa ("achatada") será a curva.



Distribuição t e Distribuição Normal Reduzida

3.6. Distribuição Amostral da Variância da Amostra(s^2):

A variância amostral é dada por:

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

Sabendo que a estatística:

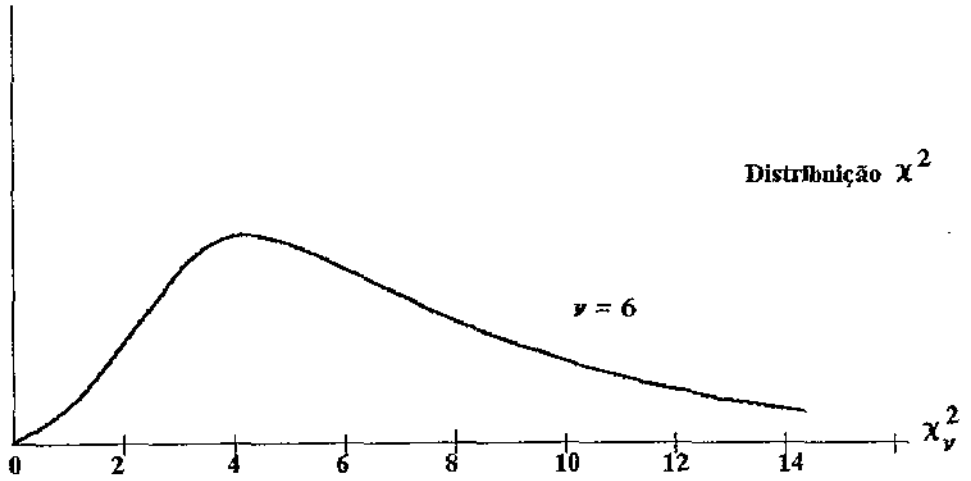
$$\chi^2_\nu = \sum_{i=1}^{\nu}\left(\frac{x_i - \mu}{\sigma}\right)^2 = \sum_{i=1}^{\nu} Z_i^2$$

onde x_i são valores aleatórios independentemente retirados de uma população normal de média μ e desvio-padrão σ tem distribuição χ^2 com n-1 graus de liberdade.

Podemos escrever:

$$\chi^2_{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma^2} = \frac{n-1}{\sigma^2} \cdot \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{(n-1).s^2}{\sigma^2}$$

Portanto, a variável $\frac{(n-1).s^2}{\sigma^2}$ tem uma distribuição χ^2 com n-1 graus de liberdade.

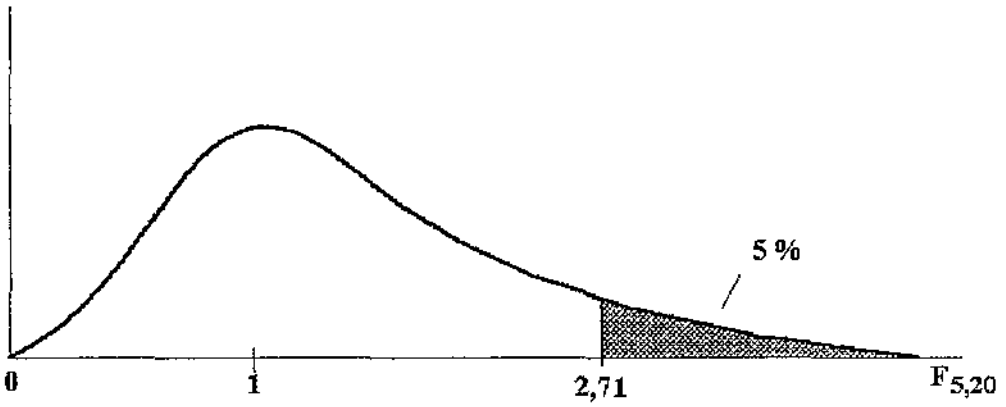


3.7.Distribuição Amostral do Quociente de Duas Variâncias Amostrais:

Suponhamos que duas amostras independentes retiradas de populações normais forneçam variâncias s_1^2 e s_2^2, o quociente S_1^2/S_2^2 terá distribuição F de snedecor com ν_1 graus de liberdade no numerador e ν_2 graus de liberdade no numerador.

Desta forma,

$$F_{v1,v2} = \frac{\chi^2_{v_1} / v_1}{\chi^2_{v_2} / v_2} = \frac{s_1^2}{s_2^2}$$



Distribuição F de Snedecor

3.8. Distribuição Amostral do Número de Sucessos na amostra e da Proporção da amostra:

Do cálculo de probabilidades temos que a distribuição amostral do número de sucessos *d* será uma distribuição binomial de parâmetros *n* e π e assim:

$$E[d] = n\pi \quad \text{e} \quad VAR[d] = n\pi(1-\pi)$$

A proporção *p*, que simplesmente é o quociente de *d* pelo tamanho da amostra *n*. Aplicando propriedades algébricas, temos que:

$$E[p] = \pi \text{ e } VAR[p] = \frac{\pi(1-\pi)}{n}$$

Se a amostra *n* for suficientemente grande, podemos aproximar as distribuições de *d* e *p* por distribuições normais com as respectivas médias e desvios padrões. Em termos práticos, em geral, podemos considerar que a amostra será suficientemente grande, para efeito dessa aproximação, se $np \geq 5$ e $n(1-p) \geq 5$.

3.9. EXERCÍCIOS:

1) Uma população (normalmente distribuída) consiste de cinco números: 2,3,6,8,11. Consideremos todas as amostras possíveis de 2 elementos que dela podemos retirar.

a) Determine a média e o desvio padrão da população.

b) Determine a média das médias amostrais e o desvio padrão das médias amostrais, para amostras com reposição.

c) Determine a média das médias amostrais e o desvio padrão das médias amostrais, para amostras sem reposição.

2)Certos amortecedores fabricados por uma empresa tem uma vida média de 800 dias e desvio padrão de 60 dias. Determine a probabilidade de que a média de uma amostra aleatória de 16 amortecedores:

a) esteja entre 770 e 830 dias

b) seja menor que 785 dias.

3) Os pesos de pacotes recebidos por um depósito tem uma média de 150 Kg e um desvio padrão de 25 Kg. Qual a probabilidade de 25 pacotes, retirados aleatoriamente e carregados em um elevador, não excedem o limite de segurança deste, que é de 4.100 Kg?

4)Calcular os valores de t para os quais a área da extremidade direita da distribuição t de Student é de 5%, quando o número de graus de liberdade for igual a:

a) 16 b) 27 c) 200

5) Se a variável X tem distribuição t de Student com 10 gl (graus de liberdade) , determinar a constante K de modo que:

a) $P(X > K) = 0,05$ c) $P(X \leq K) = 0,30$

b) $P(-K < X < K) = 0,20$ d) $P(X > K) = 0,90$

6) Determinar o valor de $\chi^2_{0,95}$ para os graus de liberdade:

a) 5 b)18

7) Para uma distribuição Qui-quadrado (χ^2) com 12 gl , determine o valor do χ^2 de modo que:

a) a área à direita desse ponto seja de 5%.

b) a área à esquerda desse ponto seja de 99%.

8) Para a distribuição F ache:

a) $f_{0,05}$ com $v_1=7$ e $v_2=15$

b) $f_{0,025}$ com $v_1=15$ e $v_2=7$

c) $f_{0,01}$ com $v_1=24$ e $v_2=19$

d) $f_{0,95}$ com $v_1=7$ e $v_2=24$

e) $f_{0,99}$ com $v_1=28$ e $v_2=12$

9) Se s_1^2 e s_2^2 representam as variâncias amostrais de duas variáveis independentes de tamanho $n_1=25$ e $n_2=31$ tiradas de populações normais, qual a probabilidade de $P(\frac{s_1^2}{s_2^2} > 2,47)$?

10) Uma pesquisa de opinião pública numa comunidade mostrou 46% das pessoas são favoráveis a um projeto de lei. Determinar a probabilidade de que a maioria das pessoas , de um conjunto amostral de 1000 pessoas, seja favorável a tal projeto.

11) Um fabricante faz a remessa de 1.000 lotes, de 100 parafusos cada um. Se 5% dos parafusos são defeituosos, em quantos lotes pode-se esperar que existam:

a) Menos que 90 parafusos perfeitos.

b) 98 ou mais parafusos perfeitos.

4. ESTIMAÇÃO:

O objetivo da inferência estatística é obter conclusões a respeito de populações através de uma amostra extraída dessa população. Uma variável aleatória é caracterizada por sua distribuição de probabilidade. Em alguns casos, no controle estatístico da qualidade, a distribuição de probabilidade é usada para descrever ou modelar alguma característica de qualidade, como por exemplo, uma dimensão crítica de um produto ou a proporção de defeituosos de um processo de manufatura. Assim, estamos interessados em fazer inferências a respeito dos parâmetros da distribuição de probabilidade. Como estes parâmetros quase sempre são desconhecidos, iremos estimá-los a partir dos dados de uma amostra.

A Estatística Inferencial compreende a Estimação e Teste de hipótese. A estimação é um processo que consiste em utilizar dados amostrais (retirados segundo amostragem probabilística) a fim de obter conclusões sobre os parâmetros da população que são desconhecidos.

É através da estatística da amostra representada por um *estimador* que fornece uma estimativa dos *parâmetros populacionais.*

* *DEFINIÇÕES*:

* *Estimar*: Fornecer um valor para algum parâmetro populacional desconhecido, através de dados amostrais.

* *Estimador*: É uma função matemática obtida a partir de elementos da amostra que será no processo de estimação o parâmetro desejado.

* *Estimativa*: É um valor numérico particular de um estimador, obtido a partir de dados de uma amostra.

Exemplo: Numa população de municípios do estado desejamos estimar a média de investimento da receita municipal na área industrial.

Se investigássemos todos municípios teríamos a média populacional (###). Ao retirarmos uma amostra aleatória, estaríamos estimando a média populacional pela média amostral $\bar{x}$. Logo, $\bar{x}$ é um estimador de μ. E uma estimativa seria o valor de $\bar{x}$ para esta amostra particular.

4.1. Propriedade dos Estimadores:

* *NÃO-TENDENCIOSO* (não-viciado, justo ou não viesado): Um estimador é não tendencioso se sua média for igual ao parâmetro.

Se extraímos todas as possíveis amostras de mesmo tamanho (n) de uma única população e calcularmos para cada uma delas os respectivos valores da estatística amostral e se a média aritmética destes valores coincidir com o parâmetro, estaremos diante de um estimador não-tendencioso.

Exemplo: $E(\bar{x}) = \mu$, isso significa dizer que $\bar{x}$ é um estimador não-tendencioso de μ.

* *EFICIENTE:* Quando comparamos dois estimadores, dizemos que é eficiente o que apresentar a menor variância.

* *SUFICIENTE:* Um estimador é suficiente se contém o máximo de informações com referência ao parâmetro por ele estimado, ou seja, quando consegue sumarizar, "condensar" a informação em uma amostra a respeito do parâmetro a ser estimado.

* *CONSISTENTE*: Entre dois estimadores para o mesmo parâmetro, será considerado consistente aquele que for não tendencioso e de variância mínima.

ESTIMAÇÃO POR PONTO E POR INTERVALO

A estimação pode ser *por ponto ou por intervalo*. A estimação por ponto é um valor obtido pelos cálculos sobre os valores observados de uma variável que serve como aproximação do parâmetro. A estimação por intervalo fornece um intervalo em torno da estimativa por ponto, de modo que este intervalo tenha uma probabilidade de conter o parâmetro.

4.2. Estimação por Ponto:

Consiste em fornecer a melhor estimativa possível para o parâmetro que será estimado através de um único valor.

Exemplos: 1) A melhor estimador da média populacional μ é $\bar{x}$, pois é um estimador não tendencioso, eficiente, suficiente e consistente.

2) Sabendo que a variância da população $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$, poderíamos estimá-la por: $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$, utilizando $\bar{x}$, pois não conhecemos μ. Porém este estimador é tendencioso para σ^2, pois a média dos valores desta estatística para cada amostra possível de tamanho **n** é diferente de σ^2 .

Para tornar este estimador não tendencioso é necessário multiplicá-lo por $\frac{n}{n-1}$.

Teremos, então: $s^2 = \frac{n}{n-1} \cdot \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

Exemplo: Um pesquisador está estudando a resistência de determinado material sob certas condições. Uma amostra aleatoriamente escolhida de 9 elementos forneceu os seguintes valores: 4,9 7,0 8,1 4,5 5,6 6,8 7,2 5,7 6,2.

Estime a média e o desvio-padrão da resistência deste material.

4.3. ESTIMAÇÃO POR INTERVALO:

A estimação por intervalo nos fornece um intervalo de valores centrados na estatística amostral, no qual julgamos estar o parâmetro com uma probabilidade conhecida de erro.

Vimos que para uma população podemos retirar K amostras diferentes para um determinado tamanho de amostra n. Cada amostra possível tem um valor como estimativa e cada estimativa fornecerá um intervalo diferente para o parâmetro.

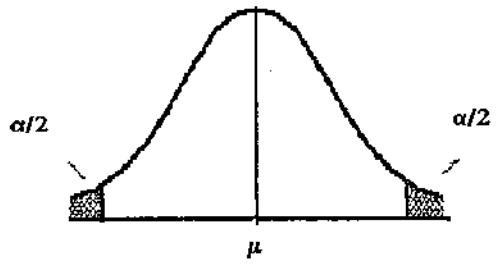
Assim, temos uma probabilidade (1-α) de que o valor do parâmetro esteja contido no intervalo estimado, chamado nível de confiança. Por esta razão, chamamos de intervalos de confiança.

O intervalo de confiança dependerá da distribuição amostral do estimador que foi utilizado para estimar o parâmetro.

4.4. Estimação por Intervalo para a Média Populacional:

Sabemos que as médias se distribuem segundo uma distribuição normal com média μ e desvio-padrão $\frac{\sigma}{\sqrt{n}}$. Quando retiramos uma amostra, a média $\overline{x}$ é uma das muitas médias possíveis de se obter de uma população.

Assim se adotarmos um nível de confiança de 95% , poderemos dizer que 95% das médias amostrais estarão dentro de 1,96 erros padrão.



**Erro Absoluto Máximo de Estimação:* O erro absoluto máximo de estimação diz respeito a diferença entre a média amostral e a média populacional.

$$\varepsilon = \left|\overline{x} - \mu\right|$$

Sabendo que o intervalo de confiança tem centro na média amostral, é determinado da seguinte maneira:

$$\left[\overline{x} \pm \varepsilon\right] \text{ onde } \varepsilon = z\frac{\sigma}{\sqrt{n}}$$

Quando n < 30 e σ desconhecido, usamos a distribuição t-student com n-1 graus de liberdade, sendo $\varepsilon = t.\frac{s}{\sqrt{n}}$

**CASO 1: INTERVALO DE CONFIANÇA PARA MÉDIA COM VARIÂNCIA POPULACIONAL CONHECIDA.*

Para uma variável aleatória X , com média desconhecida e variância conhecida σ^2, uma amostra aleatória é retirada e calcula-se $\bar{x}$. O intervalo de confiança com nível de confiança $1-\alpha$ é dado por:

$$\bar{x} - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Exemplo: Uma máquina enche pacotes de café com uma variância igual a 100 g^2. Ela estava regulada para enchê-los com 500g, em média. Agora ela se desregulou, e queremos saber qual a nova média μ. Uma amostra de 25 pacotes apresentou média igual a 485g. Estime a média por intervalo ao nível de 95% de confiança:

**CASO 2: INTERVALO DE CONFIANÇA PARA MÉDIA DE UMA DISTRIBUIÇÃO NORMAL COM VARIÂNCIA σ^2 DESCONHECIDA.*

Suponha que X seja uma variável aleatória de uma distribuição normal com média μ desconhecida e variância σ^2 desconhecida, retira-se um amostra aleatória e calcula-se a média amostral $\bar{x}$ e a variância amostral s^2. Utilizando a distribuição t-student:

$$\bar{x} - t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}$$

OBS: Quando n > 30, podemos utilizar a distribuição normal ou a distribuição t-student.

Exemplo: Um pesquisador está estudando a resistência de um determinado material sob determinadas condições. Ele sabe que essa variável é normalmente distribuída. Foi retirado uma amostra de 9 unidades 4,9 ; 7,0 ; 8,1; 4,5 ; 5,6; 6,8 ; 7,2 ; 5,7; 6,2.

a) Determine um intervalo de 90% de confiança para a resistência média populacional.
b) Determine um intervalo de 95% de confiança. para a resistência média populacional.
c) Verifique os resultados de a) e b), e conclua a respeito do erro de estimação e o nível de confiança?

4.5. Estimação por Intervalo para a Proporção Populacional:

**Intervalo de confiança para a proporção:* A distribuição de proporções amostrais indica o quão provável é determinado conjunto de proporções amostrais.

Seja, Π: proporção populacional de determinada característica e

p: proporção amostral de dessa característica

então o parâmetro Π de uma distribuição binomial, por exemplo, a proporção de peças defeituosas, poderá ter em uma amostra de n elementos tomada, "x" observações são possuidoras de determinada característica, a proporção de defeituosos na amostra estimado por $p = \frac{x}{n}$.

Para n<30 utilizaríamos a distribuição Binomial, lembrando que a distribuição amostral das proporções segue uma distribuição binomial conforme discutido no capítulo 5.

Quando n>30 e p>0,1, poderemos usar a distribuição normal como aproximação da binomial resultando no intervalo de confiança:

$$p - Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq \Pi \leq p + Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

Exemplo: Numa pesquisa de mercado, 400 pessoas foram entrevistadas sobre determinado produto e 60% destas pessoas preferiam a marca A. Estime um intervalo de 95% de confiança para a proporção populacional das pessoas que preferem a marca A.

4.6.Tamanho mínimo da amostra:

**PARA ESTIMAR MÉDIA POPULACIONAL*: Para determinarmos o tamanho da amostra, dependemos dos seguintes fatores:

* O nível de confiança a ser utilizado na estimação;
* O valor da variância absoluta da variável;
* O erro absoluto máximo de estimação:
* O custo financeiro de pesquisa

Quando conhecemos a variância populacional, podemos usar a seguinte fórmula:

$$n = \frac{Z^2 \sigma^2}{\varepsilon^2}$$

Exemplo: Qual o tamanho da amostra necessário para estimar a média de uma população cujo desvio-padrão é aproximadamente 4 mm, com 98% de confiança e precisão de 0,5 mm?

Sem conhecimento da variabilidade populacional estimamos a variância populacional através de uma amostra piloto de tamanho arbitrário. Assim:

$$n = \frac{(t_{n'-1,\alpha/2.})^2 s^2}{\varepsilon^2}$$

Exemplo: Foram realizadas 20 medidas do tempo gasto (em minutos) para se fabricar um componente industrial como uma amostra piloto, com o objetivo de estimarmos o tempo médio de produção(populacional), obtendo-se:

13	15	12	14	17	15	16	15	14	16
17	14	16	15	15	13	14	15	16	15

Verifique se estes dados são suficientes para estimar a média populacional com 95 % de confiança e precisão de 30 seg. Caso não for suficiente, qual é o tamanho de amostra complementar?

**PARA ESTIMAR A PROPORÇÃO POPULACIONAL:* Analogamente ao caso da média têm-se:

$$n = \frac{Z_{\alpha/2}^2}{\varepsilon^2} . \hat{p} . (1 - \hat{p})$$

onde, $\hat{p}$ é a proporção populacional ou alguma idéia da mesma obtida em estudos anteriores similares. Caso não se saiba o valor de $\hat{p}$, podemos estimá-lo através de uma amostra piloto n' ou usar p = 0,5.

Exemplo 1: Qual o tamanho de amostra suficiente para se estimar a proporção de defeituosos fornecidos por uma máquina, com precisão de 0,02 e 95% de confiança, sabendo que a proporção não é superior a 20%?

4.7. EXERCÍCIOS

1)A distribuição dos diâmetros de parafusos produzidos por uma certa máquina é normal, com desvio padrão igual a 0,17 mm. Uma amostra de seis parafusos retiradas ao acaso da produção apresentou os seguintes diâmetros (em mm): 25,4 ; 25,2 ; 25,6 ; 25,3 ; 25,0 ; 25,4. Estime a média da população e interprete.Construa o intervalo de 95 % de confiança para a média.

2) A empresa ABC enviou um questionário a uma amostra aleatória de clientes perguntando qual seria sua presumível necessidade de um certo produto no semestre seguinte. A partir dos dados estime (a) a média (b) o desvio padrão (c) a proporção de clientes que necessitam mais que 12 unidades deste produto.

unidades	Empresas
5	10
6	14
7	16
8	15
9	12
10	7
11	6
12	6
13	6
14	8

3) Solicitou-se a 100 estudantes de um colégio que anotassem suas despesas com alimentação e bebidas no período de uma semana. O resultado foi uma despesa média de R$ 40 e desvio-padrão R$ 10. Construa o intervalo de 98% de confinça para a média.

4) Em quatro leituras experimentais de um comercial de 30 segundos, um locutor levou em média 29,2 segundos com VARIÂNCIA de 5,76 segundos^2. Construa o intervalo de confiança para a média com 90% de confiança.

5) Uma amostra de 300 habitantes de uma cidade mostrou que 180 desejavam água florada. Encontre os limites de confiança 90% e 95% para a proporção da população favorável a fluoração.

6) Uma amostra de 50 bicicletas, de um estoque de 400, acusou 7 bicicletas com pneus vazios.
a) Estime o número de bicicletas com pneus vazios no estoque.
b) Construa o intervalo de confiança de 99% para a proporção de bicicletas com pneus vazios.
c) Se o reparo de cada bicicleta com pneu vazio leva 15 minutos, qual seria o tempo esperado de reparo dos pneus vazios do estoque?

7) Numa pesquisa de mercado, 57 das 150 pessoas entrevistadas afirmam que seriam compradoras de certo produto a ser lançado. Essa amostra é suficiente para estimar a proporção real de futuros compradores, com uma precisão de 0,08 e confiança 95%?

5.TESTE DE HIPÓTESE PARAMÉTRICO:

• A estimação é feita com base um uma variável convenientemente escolhida, função dos elementos da amostra, a qual denominamos estimadores.

• A segunda aplicação da teoria de amostragem consiste em verificar uma declaração feita sobre um parâmetro populacional.

Vamos supor que existe uma hipótese que será testada com base nos resultados amostrais, sendo aceita ou rejeitada.

5.1. Hipótese Estatística:

Hipótese em estatística é uma suposição formulada a respeito dos parâmetros de uma distribuição de probabilidade de uma ou mais populações.

Esta hipótese será testada com base em resultados amostrais, sendo aceita ou rejeitada. Ela somente será rejeitada se o resultado da amostra for claramente improvável de ocorrer quando a hipótese for verdadeira.

Consideremos Ho a hipótese existente, a ser testada e H1 a hipótese alternativa, complementar de Ho. O teste pode levar a aceitação ou rejeição de Ho que corresponde, respectivamente à negação ou afirmação de H1.

Exemplo: Suponhamos que uma indústria compre de certo fabricante parafusos cuja a carga média de ruptura por tração é especificada em 50 Kg, o desvio-padrão das cargas de ruptura é suposto ser igual a 4 Kg. O comprador deseja verificar se um grande lote de parafusos recebidos deve ser considerado satisfatório, no entanto existe alguma razão para se temer que a carga média de ruptura seja eventualmente inferior à 50 Kg. Se for superior não preocupa o comprador pois neste caso os parafusos seriam de melhor qualidade que a especificada. Neste exemplo, a hipótese do comprador é que a carga média da ruptura é inferior a 50 Kg.

O comprador pode ter o seguinte critério para decidir se compra ou não o lote: Resolve tomar uma amostra aleatória simples de 25 parafusos e submetê-los ao ensaio de ruptura.

Se a carga média de ruptura observada nesta amostra for maior que 48 Kg ele comprará o lote, caso contrário se recusará a comprar.

5.2. Passos para realizar um Teste de Hipótese:

1. HIPÓTESES:

Hipótese Nula (H_0): É um valor suposto para um parâmetro. No exemplo acima, Ho: μ=50.

Hipótese Alternativa(H_1) : É uma hipótese que contraria a hipótese nula, complementar de Ho, no exemplo, H1: μ <50.

ou seja,

Ho: $\mu = 50$
H1: $\mu < 50$

Supondo H_0 verdadeira, $\bar{x}$ da amostra aleatória de 25 valores será uma v.a com média também de 50 Kg e desvio padrão $\frac{\sigma}{\sqrt{n}}$.

No exemplo, $\sigma_{\bar{x}} = \frac{4}{\sqrt{25}} = 0{,}8$

Sabemos que $\bar{x}$ é aproximadamente normal, então podemos calcular a probabilidade de obtermos um valor inferior a 48.

$$P(\bar{x} < 48) = P(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < \frac{48-50}{0{,}8}) = (P(Z < -2{,}5) = 0{,}0062$$

Existe pois uma probabilidade de 0,0062 de que, mesmo sendo a hipótese H_0 verdadeira, $\bar{x}$ assuma um valor na região que leva à rejeição de H_0, conforme critério adotado anteriormente.

⇨ Nivel de significância de um Teste:

É a probabilidade máxima de rejeitar Ho. Se, por exemplo, utilizarmos o nível de significância de 5% a hipótese nula (Ho) será rejeitada, somente se o resultado da amostra for tão diferente do valor suposto que uma diferença igual ou maior ocorreria com uma probabilidade máxima de 0,05.

Na prática, o valor de α é fixo. (Normalmente $\alpha = 0{,}01$ ou 0,05 ou 0,10.)

No exemplo, fixado $\alpha = 0{,}05$, levaria à rejeição de Ho, pois $0{,}0062 < 0{,}05$.

• Uma outra maneira de tomar-se uma decisão é comparar o valor tabelado com a **estatística do teste.**

2. ESTATÍSTICA DO TESTE:

É o valor calculado a partir da amostra que será usado na tomada de decisão.

No exemplo, $Z_{calc} = -2{,}5$.

$$Z_{calc} = \frac{\text{valor da estimativa - valor alegado para o parâmetro}}{\text{desvio-padrão do estimador}}$$

3. REGIÃO CRÍTICA:

Região onde os valores da estatística dos teste levam à rejeição da hipótese nula. A sua área é igual ao nível de significância, e sua direção é a mesma da hipótese alternativa.

Unilateral à esquerda: H_0: $\mu = 50$
H_1: $\mu < 50$

Unilateral à direita: H_0: $\mu = 50$
H_1: $\mu > 50$

Bilateral: H_0: $\mu = 50$
H_1: $\mu \neq 50$

4. REGRA DE DECISÃO:

Se o valor da estatística do teste cair dentro da região crítica, rejeita-se H_0. Ao rejeitar a hipótese nula (H_0) existe uma forte evidência de sua falsidade.

Ao contrário, quando aceitamos, dizemos que não houve evidência amostral significativa no sentido de permitir a rejeição de Ho.

5. CONCLUSÃO:

O que significa, na situação de pesquisa, aceitar ou rejeitar Ho.

5.3. Tipos de erros:

Pelo fato de estarmos usando resultados amostrais para fazermos inferência sobre a população, estamos sujeito a erros.

Digamos que existe uma probabilidade α de que mesmo sendo Ho verdadeiro, $\bar{x}$ assuma um valor que leva Z_{calc} à rejeição de Ho.

Neste caso, no exemplo, o comprador iria cometer o erro do tipo I e a conseqüência seria de não comprar um lote satisfatório.

Porém se Ho fosse considerada verdadeira e na realidade $\mu < 50$, e $\bar{x}$ levasse à rejeição de H_0, o comprador cometerá o erro do tipo II, a qual consiste em aceitar Ho, sendo ela falsa.

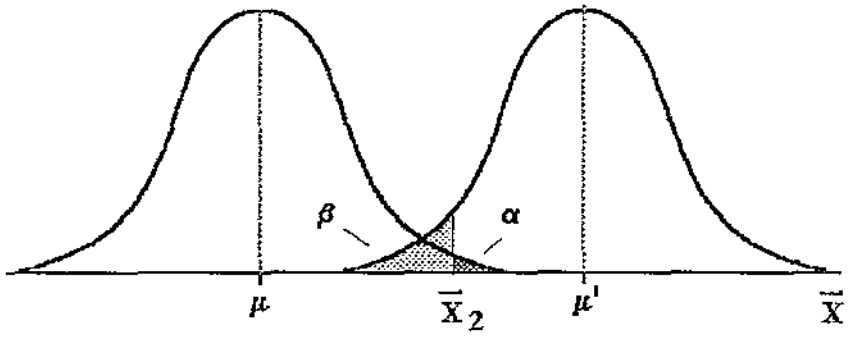
As probabilidades desses erros são chamadas α e β respectivamente.

$$\alpha = \textbf{P(erro tipo I)} = \textbf{P(rejeitar } H_0\textbf{/ } H_0 \textbf{ é verdadeiro)}$$

$$\beta = \textbf{P(erro tipo II)} = \textbf{P(aceitar } H_0\textbf{/ } H_0 \textbf{ é falso)}$$

REALIDADE

DECISÃO	H_0 verdadeira	H_0 falsa
Aceitar H_0	Decisão Correta $(1-\alpha)$	Erro do tipo II β
Rejeitar H_0	Erro do tipo I α	Decisão Correta $(1-\beta)$



Erros tipo I e tipo II

A probalidade de erro tipo I é determinada pelo pesquisador, mas para determinar a probabilidade de erro tipo II, devemos considerar a hipótese nula como falsa e, então determinar qual a verdadeira distribuição da característica em estudo.

Exemplo: O peso médio de litros de leite enchidas em uma linha de produção está sendo estudado. O padrão prevê um conteúdo médio de 1000 ml por embalagem. Sabe-se que o desvio padrão é de 10 ml.

Para encontrar a probabilidade de erro tipo II, quando testamos a média não ser igual a 1000 ml ao nível de 5% de significância com 4 unidades amostrais, e sendo o real conteúdo médio da embalagem de 1012 ml, temos:

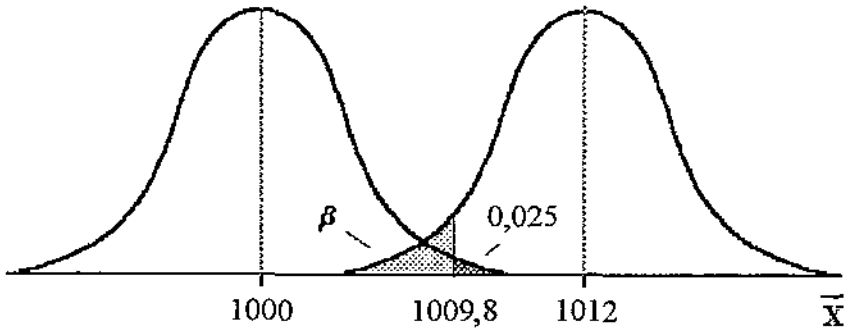
H_0: $\mu = 1000$

H_1: $\mu \neq 1000$

P (erro tipo II) = P (aceitar H_0/ H_0 é falsa) = ?

$Z_{\alpha/2} = Z_{0,025} = 1,96$

$$1{,}96 = \frac{\bar{x} - 1000}{10/\sqrt{4}} \Leftrightarrow \bar{x} = 1009{,}8$$



Erros tipo I e tipo II

P (aceitar H_0/ H_0 é falsa) = P ($\bar{x}$ < 1009,8 / μ = 1012)

$$= P\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < \frac{1009{,}8 - 1012}{10/\sqrt{4}}\right)$$

$$= P\,(Z < -0{,}44) = 0{,}33$$

Ou seja, a probabilidade de não rejeitarmos Ho, quando a média real da embalagem é de 1012 ml é de 0,33. A partir dessa informação podemos obter o poder do teste é de $1-\beta = 1 - 0{,}33 = 0{,}67$.

5.4. Teste de Hipótese para uma Média:

**ESTATÍSTICA DO TESTE:*

Tamanho de Amostra:	σ conhecido:	σ desconhecido:
n > 30	$Z_{cal} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$	$Z_{cal} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$
n < 30	$Z_{cal} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$	$t_{cal} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$

Comparamos com um t tabelado da distribuição t-student com n-1 graus de liberdade e nível de significância α.

Exemplo 1:

A resistência à tração do aço inoxidável produzido numa certa usina permanecia estável, com uma resistência média de 72 Kg/mm^2 e um desvio padrão de 2,0 Kg/mm^2. Recentemente, a máquina foi ajustada. A fim de determinar o efeito do ajuste, 10 amostras foram testadas. As resistências médias são apresentadas a seguir.

X: (Kg/mm^2): 76,2 78,3 76,4 74,7 72,6 78,4 75,7 70,2 73,3 74,2.

Presuma que o desvio padrão seja o mesmo que antes do ajuste. Podemos concluir que o ajuste mudou a resistência à tração de aço? (Adote 5% de significância)

5.5.Teste de Hipótese para comparação de duas médias (Independentes):

**SUPOSIÇÕES:*

- σ_1^2 e σ_2^2 desconhecidos e
- $\sigma_1^2 = \sigma_2^2$.

**HIPÓTESES:*

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 \neq \mu_2$ ou $H_1: \mu_1 > \mu_2$ ou $H_1: \mu_1 < \mu_2$

**ESTATÍSTICA DO TESTE:*

$$t_{cal} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_P^2 \cdot (1/n_1 + 1/n_2)}} \quad \text{onde,} \quad S_P^2 = \frac{(n1-1)S_1^2 + (n2-1)S_2^2}{n1+n2-2}$$

S_P^2 é a variância ponderada das variâncias amostrais.

**TOMADA DE DECISÃO:*

Comparar o valor da estatística do teste t_{calc} com o valor tabelado T_{tab} com n_1+n_2-2 graus de liberdade.

Exemplo: Sejam as amostras obtidas aleatoriamente de dois tipos de cabo de aço em relação à carga de ruptura. Ao nível de 2,5% de significância, pode-se concluir que o cabo do tipo I sejam mais resistentes que o do tipo II?

Carga de ruptura Kgf: Tipo I: 760, 755, 758, 761, 755
Tipo II: 758, 748, 757, 753, 755

Sabendo que as variâncias amostrais são 7,7 e 15,7 respectivamente e assumidas como iguais.

5.6. Teste de Hipótese para uma Variância Populacional:

**HIPÓTESES:*

$H_0: \sigma^2 = \sigma_0^2$

$H_1: \sigma^2 > \sigma_0^2$

**ESTATÍSTICA DO TESTE:*

Sendo Normal a distribuição da população, a estatística do teste será dada por:

$$\chi^2_{n-1} = \frac{(n-1).s^2}{\sigma_o^2}$$

**TOMADA DE DECISÃO:*

Rejeitamos Ho quando: $\chi^2_{calc} > \chi^2_{tab}$

Se o *teste for unilateral inferior:*

$H_0: \sigma^2 = \sigma_0^2$
$H_1: \sigma^2 < \sigma_0^2$
Rejeitamos H_0 se: $\chi^2_{calc} < \chi^2_{n-1;1-\alpha}$

Se o *teste for bilateral:*

$H_0: \sigma^2 = \sigma_0^2$
$H_1: \sigma^2 \neq \sigma_0^2$
Rejeitamos H_0 se: $\chi^2_{calc} < \chi^2_{n-1;1-\alpha/2}$ *ou* $\chi^2_{calc} > \chi^2_{n-1;\alpha/2}$

Exemplo:
Uma amostra de 10 elementos extraída de uma população suposta normal forneceu variância igual a 12,4. O resultado é suficiente para se concluir, ao nível de 5% de significância que a variância desta população é inferior a 25?

5.7. Teste de Hipótese para duas Variâncias Populacionais:

**HIPÓTESES:*

H_0: $\sigma_1^2 = \sigma_2^2$
H_1: $\sigma_1^2 < \sigma_2^2$

Sabendo que a distribuição amostral do quociente de duas variâncias s_1^2 e s_2^2 é uma F-snedecor, então, supondo H_0 verdadeira:

$$F_{n_1-1,n_2-1} = \frac{(\sigma^2/(n_1-1)).\chi^2_{n_1-1}}{(\sigma^2/(n_2-1)).\chi^2_{n_2-1}} = \frac{s_1^2}{s_2^2}$$

Sendo H_0 verdadeiro, devemos esperar que o valor de s_1^2 esteja próximo de s_2^2 e o quociente estará próximo de 1. Desta forma, rejeitamos H_0 se $\frac{s_1^2}{s_2^2}$ for significantemente superior a 1.

**ESTATÍSTICA DO TESTE:*

A estatística do teste será o quociente das estimativas de s_1^2 e s_2^2. ($s_1^2 > s_2^2$)

$$F_{calc} = \frac{s_1^2}{s_2^2}$$

**TOMADA DE DECISÃO:*

Compara-se o valor da estatística do teste com $F_{tab} \approx F(n_1 - 1, n_2 - 1; 1-\alpha)$

Rejeitamos H_0 se: $F_{calc} < F_{tab}$.

Se o *teste for bilateral:*

H_0: $\sigma_1^2 = \sigma_2^2$
H_1: $\sigma_1^2 \neq \sigma_2^2$

Rejeitamos H_0 se: $F_{calc} < F_{n1-1,n2-1;1-\alpha/2}$ ou $F_{calc} > F_{n1-1,n2-1;\alpha/2}$ ou

Se o teste for *unilateral à direita:*

$$H_0: \sigma_1^2 = \sigma_2^2$$
$$H_1: \sigma_1^2 > \sigma_2^2$$

Rejeitamos H_0 se: $F_{cal} > F_{n1-1,n2-1;\alpha}$

Exemplo 1:

Dois programas de treinamento de funcionários foram efetuados. Os 21 funcionários treinados no programa antigo apresentaram uma variância 146 em suas taxas de erro. No novo programa, 13 funcionários apresentaram uma variância de 200. Fixando um nível de significância α= 0,05, pode-se concluir que a variância é diferente para os dois programas?

Exemplo 2:

Uma empresa está estudando duas marcas de pneus A e B. testou 11 pneus de cada marca, quanto a durabilidade, e constatou: para a marca A uma média de 23.600 Km e um desvio-padrão de 3.200 Km, e, para a marca B, uma média de 24.800 Km e um desvio-padrão de 3.700 Km. Ao nível de 5%, testar a hipótese de igualdade das variâncias populacionais, contra a alternativa da variância de A ser menor que a variância de B.

5.8. Teste de Hipótese para uma Proporção Populacional:

Consideramos o problema de testar a hipótese que a proporção de sucessos de algum experimento binomial seja igual a um certo valor.

**HIPÓTESES:*

H_0: $\Pi = \Pi_0$

H_1: $\Pi \neq \Pi_0$ (ou unilateral)

**ESTATÍSTICA DO TESTE:*

Uma estatística apropriada a qual basearemos nosso critério de decisão é $p = \frac{f_i}{n}$ onde fi é o número de elementos portadores de determinada característica e n é o número de elementos da amostra.

A aproximação normal é usada para "n grande", sendo a estatística do teste:

$$Z_{calc} = \frac{p - \Pi_0}{\sqrt{\frac{(\Pi_0(1-\Pi_0)}{n}}}$$

Exemplo: Um comprador, ao receber de um fornecedor um grande lote de peças, decidiu inspecionar 200 delas. Decidiu também que o lote será aceito se ficar convencido ao nível 5% de significância que a proporção de defeituosos seja no máximo 4%. Qual será sua decisão (aceitar ou rejeitar o lote) se, na amostra foram encontradas 11 peças defeituosas?

5.9. Exercícios:

1. Defina sumariamente:

a. Erro tipo I
b. nível de significância
c. valor crítico
d. estatística do teste

2. Explique a relação existente entre:

a. Amostragem aleatória e distribuição amostral
b. A probabilidade do erro do tipo I e a região crítica

3. O que significa rejeitar a hipótese nula?

4. A aceitação da hipótese nula significa que ela esteja correta?

5. Estabeleça a hipótese nula a hipótese alternativa para as seguintes situações:

a) Um fornecedor afirma que o tempo de vida da marca de bateria que ele comercializa é maior que 3 meses.
b) Um engenheiro desconfia que um torno eletrônico está fora do ajuste produzindo eixos com diâmetro diferente do especificado que é de 2,54.
c) Um fabricante acha que o consumo de um certo modelo de eletrodoméstico é inferior a 20 watts.

6. A resistência dos cabos fabricados por determinada companhia acusam média de 1800 libras e desvio-padrão de 100 libras. Adotando-se uma nova técnica de fabricação espera-se aumentar essa resistência. Para testar tal hipótese selecionou-se uma amostra de 50 cabos fabricados pelo novo processo, obtendo-se uma resistência média de 1850 libras. Pode-se aceitar a hipótese ao nível de significância de 0,01?

7. Um fabricante de conservas anuncia que o conteúdo líquido de uma lata de seu produto é de 200 gramas com um desvio padrão de 40 gramas. A fiscalização de pesos e medidas investigou uma amostra aleatória de 64 latas, verificando que $\sum x = 127.360$. Fixado o nível de significância de 0,05, deverá o fabricante ser multado por não efetuar a venda do produto conforme anuncia?

8. Numa amostra de 10 lâmpadas elétricas produzidas por uma empresa verificou-se que seu tempo médio de duração foi calculado em 490h e desvio-padrão de 12h. Fixado o nível de significância de 0,05, realize um teste para verificar se o tempo médio é diferente de 500 horas?

9. Certa organização médica afirma que uma nova vacina é de qualidade superior a até então existente, que é 80% eficaz para curar certa enfermidade num determinado prazo. Examinada uma amostra de 100 pessoas que sofriam da referida doença, 86 ficaram curadas com a nova vacina . Fixado o nível de significância de 5% , verifique a aceitabilidade da afirmativa daquela organização.

10. O produtor de certa marca de cigarro afirma que a quantidade média de nicotina por cigarro é de 23 mg. Um interessado resolveu selecionar uma amostra aleatória de 6 cigarros desta marca, obtendo quantidade média de 25 mg e desvio padrão de 2,19 mg. Diante de tal pesquisa é possível que o produtor seja denunciado por falsa publicidade de nível teórico? Utilize um nível de 5% e suponha normalidade da população.

11. Os produtores de um programa de televisão acham que devem modificá-lo caso sua assistência regular seja inferior a um quarto de possuidores de aparelhos receptores. Uma pesquisa feita em 400 domicílios mostrou que em 80 deles o programa era assistido. Qual deve ser a decisão dos produtores se estão dispostos a correr um risco de 5% de modificar o programa sem que, diante da premissa inicial, isso seja necessário?

13. Para verificar a eficácia de uma nova droga injetadas em 72 ratos, obteve-se a seguinte tabela:

	Tamanho da amostra	variância
Machos	41	43,2
Fêmeas	31	29,5

Testar a igualdade de variâncias considerando nível de significância de 10%.

12. Uma fábrica de embalagens para produtos químicos precisa escolher entre suas técnicas de combate à corrosão de suas latas especiais. Uma amostra de 15 latas tratadas com a técnica "A" resultou em corrosão média de 48 com desvio-padrão 10. Outra amostra de 12 latas submetidas à técnica "B" produziu corrosão média 52 com desvio-padrão 15. Há significativa diferença entre as duas técnicas? Utilize 5%.

13. Um empresário acredita que há diferença significativa no tempo que homens e mulheres gastam para realizar determinada tarefa. Selecionou uma amostra de cada grupo e anotou o tempo gasto, em minutos, conforme abaixo. Supõe-se distribuição normal para o tempo:

Homens: 5 - 15 - 10 - 20 - 7 - 15
Mulheres: 10 - 15 - 22 - 20 - 10 - 7

14. Para uma amostra de 10 lâmpadas , a vida útil média foi de 4.000 horas com desvio padrão de 200 horas. Para outra marca, uma amostra de 8 lâmpadas acusou média de 4300 com desvio padrão de 250. Supõe-se que a vida útil esteja normalmente distribuída. Realize um teste para comparar as duas marcas com 1% de significância.

6. ANÁLISE DE VARIÂNCIA

É a técnica usada para verificar estatisticamente se duas ou mais médias são iguais, isto é, se provem de populações com mesma média. É uma técnica muito poderosa para poder identificar diferenças entre médias populacionais devidas à várias causas atuando simultaneamente sobre os elementos da população. Vamos abordar apenas o caso quando pode haver uma possível causa, ou seja apenas uma fonte de variação (caso ANOVA).

A análise de variância é uma extensão natural do teste de hipóteses onde passamos a verificar se a média de *k* amostras podem ser ou não consideradas iguais.

Por exemplo, os dados abaixo se referem a um teste realizado para determinar se a quilometragem é a mesma em quatro marcas de gasolina.

	marca da gasolina			
observação	1	2	3	4
1	15,1	14,9	15,4	15,6
2	15,0	15,2	15,2	15,5
3	14,9	14,9	16,1	15,8
4	15,7	14,8	15,3	15,3
5	15,4	14,9	15,2	15,7
6	15,1	15,3	15,2	15,7
total das amostras	91,2	90,0	92,4	93,6
médias amostrais	15,2	15,0	15,4	15,6
variâncias amostras	0,088	0,040	0,124	0,032

Note que não há duas médias *amostrais* iguais. A análise de variância pode ser utilizada para determinar se as médias amostrais sugerem diferenças efetivas entre as quilometragens, ou se tais diferenças decorrem apenas da variabilidade amostral.

Podemos então formular as hipóteses nula e alternativa:

H_0: as médias das populações são todas iguais, (H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$)

H_1: as médias das populações não são iguais.(no mínimo uma é diferente).

O teste se baseia em uma amostra extraída de cada população (marca da gasolina, no exemplo). Se o teste (análise de variância) nos levar a:

- Aceitar a hipótese de nulidade, concluiremos que as diferenças observadas entre as médias amostrais são devidas a variações aleatórias nas amostras (e assim, que as médias populacionais das quatro marcas são iguais).
- Rejeitar a hipótese de nulidade, concluiremos que as diferenças observadas são demasiadamente grandes para serem devidas apenas ao acaso (e assim, que as médias das populações não são iguais).

6.1.Suposições

Para aplicar a análise de variâncias as suposições que seguem devem ser satisfeitas:

1. As variâncias populacionais são iguais: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = ... = \sigma_k^2$
2. Cada população tem distribuição normal
3. As amostras devem ser aleatórias e independentes.

6.2.Cálculos iniciais da Análise de Variância

Vamos usar a notação segundo a qual x_{ij} (i = 1,2,...,k; j = 1,2,...,n) é o j-ésimo valor da i-ésima amostra de n elementos.

elementos\amostras	1	2	...	k
	x_{11}	x_{21}	...	x_{k1}
	x_{12}	x_{22}	...	x_{k2}
	...	...	...	...
j	x_{1j}	x_{2j}	...	x_{kj}
	...	...	...	...
	x_{1n}	x_{2n}	...	x_{kn}
Σ	T_1	T_2	...	T_k

Sendo que:

$T_i = \sum_{j=1}^{n} x_{ij}$ = soma dos valores da amostra i

$Q_i = \sum_{j=1}^{n} x_{ij}^2$ = soma dos quadrados dos valores da amostra i

$T = \sum_{i=1}^{n} T_i = \sum_{i=1}^{k} \sum_{j=1}^{n} x_{ij}$ = soma total dos valores

$Q = \sum_{i=1}^{n} Q_i = \sum_{i=1}^{k} \sum_{j=1}^{n} x_{ij}^2$ = soma total dos quadrados dos valores

$\overline{x_i} = T_i/n$ = média da amostra i

$\overline{\overline{x}} = T/nk$ = média de todos os valores

No exemplo temos que:

T_1= soma dos valores da amostra 1 é 91,2

T_2= soma dos valores da amostra 2 é _____

T_3= soma dos valores da amostra 3 é _____

T_4= soma dos valores da amostra 4 é _____

Q_1= soma dos quadrados dos valores da amostra 1 é:

$$(15,1)^2+(15,0)^2+(14,9)^2+(15,7)^2+(15,4)^2+(15,1)^2=1386,68$$

Q_2= soma dos quadrados dos valores da amostra 2 é:

_____

Q_3= soma dos quadrados dos valores da amostra 3 é:

_____

Q_4= soma dos quadrados dos valores da amostra 4 é:

_____

$T = T_1 + T_2 + T_3 + T_4 = 91,2 +$ _____ + _____ + _____ = _____

$Q = Q_1 + Q_2 + Q_3 + Q_4 = 1386,68 +$ _____ + _____ + _____ = _____

$\overline{x_1} = T_1/n$ = média da amostra 1 = 91,2 / 6 =15,2

$\overline{x_2} = T_2/n$ = média da amostra 2 = _____

$\overline{x_3} = T_3/n$ = média da amostra 3 = _____

$\overline{x_4} = T_4/n$ = média da amostra 4 = _____

$\overline{\overline{x}} = T/nk$ = média de todos os valores = ___ / 6.4 = ____ / 24 = _____

6.3.Decomposição das Variações

A análise de variância como o próprio nome diz , é um teste que analisa as variações entre as médias utilizando as variâncias. Para fazer isto, decompõem-se a variação total em variação entre as amostras (variações explicadas) e as variações entre as amostras (variações aleatórias).

Para realizar a análise de variância baseia-se que, *sendo a hipótese nula H_0 verdadeira*, essa três variações podem ser utilizadas para estimar σ^2.

* *VARIAÇÃO TOTAL:* Levando em conta que a suposição de que as variâncias populacionais são iguais e as médias são iguais se H_0 é verdadeira, então podemos estimar a variância fundindo as k amostras em uma só. Sendo que:

$$s_t^2 = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij}-\overline{\overline{x}}\right)^2}{nk-1} = \frac{Q-[T^2/nk]}{nk-1}$$

O numerador da expressão acima é conhecido como SQT = soma de quadrados total.

Para o exemplo das marcas de gasolina temos como estimativa da variação total:

$$s_t^2 = \frac{Q - [T^2/nk]}{nk-1} = \text{----------} =$$

* *VARIAÇÃO ENTRE AMOSTRAS*: Vimos acima que, sendo verdadeira a hipótese H0, podemos considerar todos os valore x_{ij} como provenientes de uma única população. nas mesmas condições podemos considerar as médias $\overline{x_i}$ das k amostras como uma amostra de k valores retirados da população dos possíveis valores de $\overline{x}$. Ora, sabemos da teoria da distribuição amostral que a população de valores de $\overline{x}$ é normalmente distribuída com variância σ^2/n. Logo, a variância da amostra formada pelos k valores $\overline{x_i}$ estima σ^2/n. temos pois, a segunda estimativa de σ^2, que será n vezes a variância dessa amostra, ou seja,

$$s_e^2 = \text{n.}\frac{\sum_{i=1}^{k}\left(\overline{x_i} - \overline{\overline{x}}\right)^2}{k-1} = \frac{\left[\sum_{i=1}^{k}(T_i^2/n)\right] - (T^2/nk)}{k-1}$$

O numerador da expressão acima é conhecido como SQE = soma de quadrados entre amostras.

Para o exemplo das marcas de gasolina temos como estimativa da variação entre as amostras:

$$s_e^2 = \frac{\left[\sum_{i=1}^{k}(T_i^2/n)\right] - (T^2/nk)}{k-1} = \text{----------} =$$

* *VARIAÇÃO RESIDUAL*: Evidentemente a variância σ^2 pode ser também estimada individualmente a partir dos elementos de cada uma das k amostras disponíveis, ou seja, dentro de cada amostra. Teríamos portanto, k estimativas individuais de σ^2, todas válidas, independente da veracidade ou não de H_0. Através de uma estimativa ponderada podemos construir uma estimativa única de σ^2 combinando as k estimativas. Cada amostra individual fornecerá uma estimativa, dada por:

$$s_i^2 = \frac{\sum_{j=1}^{n}\left(x_{ij} - \overline{x_i}\right)^2}{n-1} = \frac{Q_i - \left(T^2/n\right)}{n-1}$$

Sendo as amostras de mesmo tamanho, a estimativa resultante para o conjunto de amostras será a média aritmética das k estimativas individuais, ou seja,

$$s_r^2 = \frac{\sum_{i=1}^{k} s_i^2}{k} = \frac{Q - \sum_{i=1}^{k} \left(T_i^2/n\right)}{k(n-1)}$$

O numerador da expressão acima é conhecido como SQR = soma de quadrados dos resíduos.

Para o exemplo das marcas de gasolina temos como estimativa da variação residual:

$s_r^2 =$ ---------- =

Obs.: Uma vez que a variação total é a composição entre a variação entre as amostras e a variação residual, podemos escrever SQR = SQT - SQE.

6.4.Tomada de Decisão: a Tabela F

A estimativa de s_e^2 será uma estimativa não viciada de σ^2 apenas se H_0 for verdadeira, pois se isso não ocorrer, os desvios esperados $\left(\bar{x}_i - \bar{\bar{x}}\right)$ serão maiores que os desvios $\left(\bar{x}_i - \mu\right)$ superestimando σ^2.

Sendo assim podemos comparar as duas estimativas da variância através do teste *F*, que é a razão entre a "variação entre" e a "variação dentro". Em outras palavras a estatistica do teste consiste em:

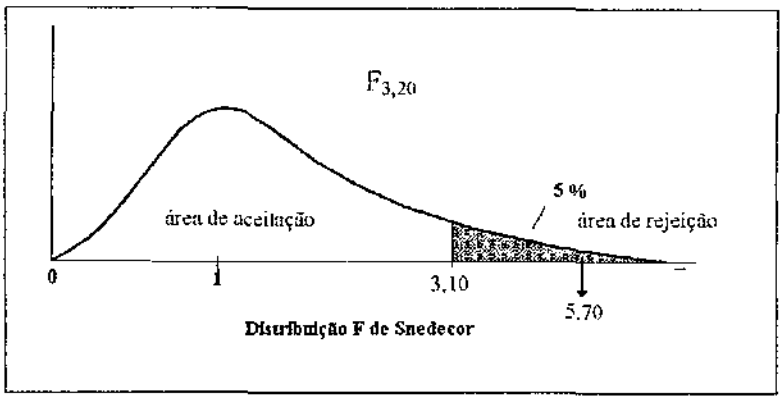
$$F_{calc} = \frac{s_e^2}{s_r^2}$$

Se H_0 for verdadeira, ambas as estimativas de σ^2 (entre amostras e residual) serão não viciadas e o valor do quociente entre elas será próximo de 1. Por outro lado, se o valor de F for elevado, poderemos concluir que s_e^2 superestima σ^2 e podemos rejeitar H_0.

Em outras palavras, o teste F será conduzido com k-1 graus de liberdade no numerador e k(n-1) graus de liberdade no denominador ou seja, H_0 será rejeitada se $F_{calc} > F_{(k-1);\, k(n-1),\alpha}$, onde α é o nível de significância escolhido para o teste.

Como conhecemos $s_e^2 = 0{,}402$ e $s_r^2 = 0{,}071$ então temos que $F_{calc} = 5{,}70$.

Fixando α=0,05 olhamos na tabela o valor de $F_{(4-1);\, 4(6-1);\, 0{,}05} = F_{3;20;0{,}5} = 3{,}10$



Decisão: Como $F_{calc} > F_{tab}$, isto é 5,70 > 3,10 então rejeitamos H0 ao nível de 5% de significância.

Conclusão: A quilometragem média desenvolvida por pelo menos uma das marcas de gasolina não é igual as demais ao nível de 5% de significância.

6.5.Tabela da Análise de Variância

Ao se fazer Análise de Variância é usual e recomendável dispor os cálculos segundo a tabela de análise de variância:

Fontes de variação	Soma de Quadrados (SQ)	Graus de Liberdade (GL)	Quadrado Médio (QM)	F_{calc}	F_{tab}
Entre amostras	$SQE=\left[\sum_{i=1}^{k}(T_i^2/n)\right]-(T^2/nk)$	k-1	$SQE/k-1$	QME/QMR	$F_{(k-1);k(n-1),\alpha}$
Residual	$SQR = Q-\sum_{i=1}^{k}\left(T_i^2/n\right)$	k(n-1)	$SQR/k(n-1)$		
Total	$SQT = Q-[T^2/nk]$	nk-1			

6.6. Exercícios

1) Use a análise de variância para testar a eficiência de quatro planos de dieta. Vinte e quatro pessoas foram aleatoriamente submetidas aos planos - seis pessoas para cada plano. Os dados abaixo fornecem a perda média de peso e a variância para cada grupo. Faça o teste ao nível de 5%.

Plano	Perda média de Peso, em Kg	Variância
Frutas	10,5	3,8
Somente líquidos	12	3,6
enlatados	9	2,0
chás naturais	15	4,6

2) Uma associação de consumidores está interessada numa comparação de preços de vendas de carros novos, tomou uma amostra aleatória de cinco capitais. Em cada capital anotou-se o preço médio de 10 carros do mesmo modelo, com os mesmos acessórios. Use o nível de 1% para verificar se os preços médios de vendas diferem significativamente entre as cinco capitais.

Capital	preço médio (em mil $)	variância
A	42,5	6
B	44,0	5
C	48,0	7
D	46,0	4
E	44,5	8

3) Os dados abaixo dão a vida observada dos pneus de quatro caminhões distribuidores de sorvete, conforme a posição. Supondo comparáveis os caminhões e os motoristas, poderemos afirmar que a duração média é independente da posição do pneu no veículo? (use nível de 1%)

posição do pneu			
dianteiro direito	dianteiro esquerdo	traseiro direito	traseiro esquerdo
17	25	22	26
19	27	21	24
20	18	19	30
24	22	26	28

4) Três pilotos de corrida de automóveis estão treinando para a próxima corrida do campeonato. Cada piloto faz cinco de troca dos quatro pneus nos carros. Faça uma análise de variância ao nível de 5% para verificar se as equipes de troca tem o mesmo desempenho.

Equipe	tempo em min				
piloto A	0,8	1,0	0,8	0,7	0,7
piloto B	0,8	0,6	0,6	0,5	0,5
piloto C	0,7	0,6	0,5	0,5	0,8

7. ANÁLISE DE CORRELAÇÃO E REGRESSÃO LINEAR

• Compreende a análise de dados amostrais para saber se e como duas ou mais variáveis estão relacionadas uma com a outra na população.

• Portanto, correlação e regressão envolvem uma forma de estimação, a diferença é que essas técnicas se referem à estimação de uma relação que possa existir na população.

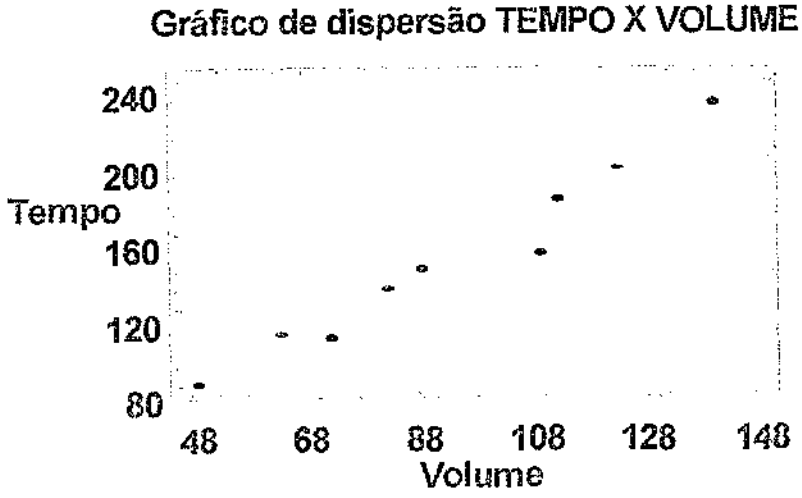
• A análise de correlação nos resume o grau de relacionamento entre duas ou mais variáveis enquanto que na regressão tem como resultado uma equação matemática que descreve o relacionamento.

• A análise de correlação linear simples diz respeito ao relacionamento de duas variáveis: uma variável dependente e uma variável independente que possuem uma relação linear entre elas.

7.1. Diagrama de Dispersão:

É um gráfico no qual cada ponto representa um par observado de valores onde podemos visualizar intuitivamente a relação entre as variáveis.

A dispersão entre os pontos do diagrama indicam a possibilidade de relacionamento entre as variáveis.



7.2. Análise de Correlação Linear:

Uma outra maneira de avaliar a correlação é através de um coeficiente que mede a intensidade da associação existente entre duas variáveis quantitativas independente da unidade de medida de cada variável.

** SUPOSIÇÕES PARA ANÁLISE DE CORRELAÇÃO:*

- ✓ Ambas variáveis são aleatórias (X e Y);
- ✓ Tanto X quanto Y tem distribuição normal;
- ✓ A variação dos valores de X para cada valor fixo Y é sempre a mesma, isto é, o valor de σ é sempre o mesmo para cada valor dado Y (homocedasticidade);
- ✓ A variação dos valores de Y para cada valor fixo de X é sempre a mesma (homocedasticidade);

** COEFICIENTE DE CORRELAÇÃO LINEAR DE PEARSON:*

- População: ρ
- Amostra: r

**INTERPRETAÇÃO DO COEFICIENTE DE CORRELAÇÃO LINEAR:*

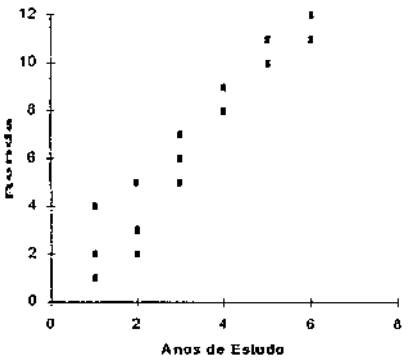
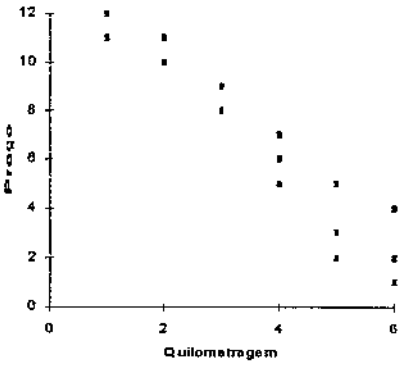
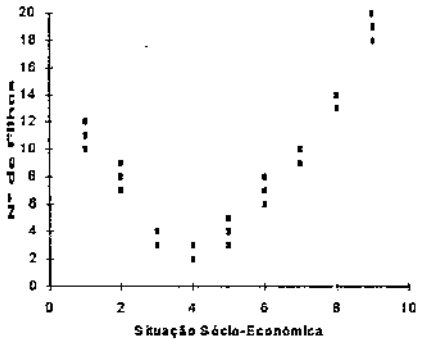
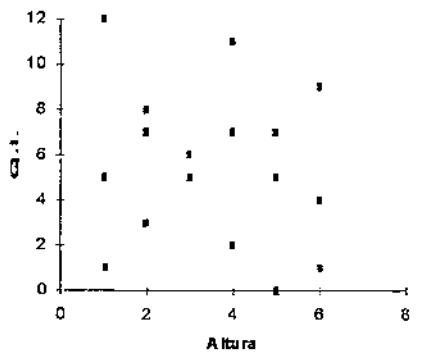
- ♦ Coeficiente de correlação linear é uma medida adimensional e varia de -1 a +1.
- ♦ Na população: $-1 \leq \rho \leq +1$ ou na amostra: $-1 \leq r \leq +1$
- ♦ O coeficiente de correlação fornece informação através do sinal:

* Se ρ for positivo, existe uma relação direta entre as variáveis (valores altos de uma variável correspondem a valores altos de outra variável).

* Se ρ for negativo a relação é inversa (valores altos de uma variável correspondem a valores baixos de outra variável).

* Se ρ for nulo, significa que não existe correlação linear.

Exemplos:

Valor de ρ	Descrição do relacionamento	Gráfico de dispersão
$\rho \cong 0,8$	Correlação linear direta entre renda e anos de estudo. Valores altos de renda correspondem a valores altos de anos de estudo.	
$\rho \cong -0,92$	Correlação linear inversa. Valores altos de quilometragem correspondem a valores baixos de preços.	
$\rho \cong 0$	Não há correlação linear (pode haver uma correlação curvilínea).	
$\rho \cong 0$	Não há correlação linear.	

** CÁLCULO DO COEFICIENTE DE CORRELAÇÃO LINEAR:*

Como freqüentemente trabalhamos com amostra, calculamos o coeficiente de correlação amostral denotado por r . Portanto r é uma estimativa de ρ . O coeficiente de correlação linear é dado pela divisão da covariação de X e Y e pelo produto do desvio padrão de X e o desvio padrão de Y.

$$r = \frac{Cov(x,y)}{Sx.Sy} \quad \text{ou} \quad r = \frac{n\sum xy - \sum x.\sum y}{\sqrt{\left(n\sum x^2 - (\sum x)^2 \cdot n\sum y^2 - (\sum y)^2\right)}}$$

7.3. Teste de Hipótese para o Coeficiente de Correlação Linear:

Quando calculamos "r" numa amostra temos que ter presente o fato de que estamos na realidade estimando a associação verdadeira entre X e Y que ocorre na população. Por esta razão realizamos um teste de hipótese.

Os possíveis valores de "r" obtidos em amostras do mesmo tamanho se distribuem segundo a distribuição t-student, quando $\rho = 0$.

**ETAPAS DO TESTE DE HIPÓTESE:*

⇨ Hipóteses:

H_0: $\rho = 0$
H_1: $\rho \neq 0$

⇨ Região Crítica:

Compara a estatística do teste com t_{tab} com n-2 graus de liberdade.

⇨ Estatística do teste:

$$t_{calc} = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

⇨ Decisão:

Se t_{calc} não pertence a Região Crítica, aceita-se Ho.

⇨ Conclusão:

O que significa aceitar ou rejeitar Ho no contexto, dependendo do problema estudado.

Exemplo:

Para cinco carros do mesmo modelo/ano e estado de conservação, foram verificadas a sua quilometragem e os respectivos preços de venda:

Quilometragem (em mil Km)	75	60	65	80	50
preço (em mil $)	10	12	10	9	13

7.4. Análise de Regressão Linear:

O estudo da regressão se aplica àquelas situações em que há suspeita da relação entre duas variáveis quantitativas e se deseja expressar matematicamente esta relação.

Os objetivos do estudo de regressão são: reconhecer a existência da dependência de Y em relação a X e expressar por meio de uma equação esta relação.

O *gráfico de dispersão* nos dá uma idéia da existência ou não da regressão.

**SUPOSIÇÕES SOBRE A ANÁLISE DE REGRESSÃO:*

✓ A variável dependente é uma v.a (os valores da variável independente podem ser fixados, os da dependente devem ser obtidos através de um processo de amostragem);

✓ Na regressão Linear as variáveis independente e dependente devem estar associadas linearmente;

✓ As variâncias das distribuições condicionais da variável dependente dados diferentes valores da variável independente são todos iguais (homocedasticidade).

**EQUAÇÃO DA REGRESSÃO LINEAR:*

A regressão linear fornece uma equação linear através do qual, pode-se determinar os valores da variável independente.

O modelo linear será dado por:

$$y_c = \alpha + \beta x + u$$

onde: α: coeficiente linear
β: coeficiente angular
u: erro aleatório

Quando usamos dados amostrais a equação da reta é dada por:

$$Y = a + bx$$

Dado um valor de x, este será usado para predizer o de Y. Como o valor de x é conhecido, resta-nos saber quem são os coeficientes da reta.

Onde "a" e "b" podem ser determinados pelo sistema de equações da reta, que por sua vez foi obtido pelo "***MÉTODO DOS MÍNIMOS QUADRADOS***".

Esse método é o mais usado para estimar os parâmetros a e b. Os valores de a e b que minimizam a soma dos quadrados dos desvios são dados pelo sistema de equações abaixo:

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2$$

Para que $\sum_{i=1}^{n}(Yi - Yc)^2$ seja mínima, o valor de "a" e "b", encontrados pelo sistema de equações da reta, resultou em:

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$a = \bar{y} - b\bar{x}$$

onde: $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ e $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ (n = número de pares de observações).

⇨ **INTERPRETAÇÃO DOS COEFICIENTES a e b:**

a: representa o intercepto, o valor que Y assume quando X=0.
b: indica a variação de Y por unidade da variação de X.

Exemplo: Para os dados dos automóveis, estime a equação da reta para o preço em função da quilometragem.

**ESTIMATIVA DO ERRO PADRÃO DA REGRESSÃO:*

A dispersão na população significa que para cada valor de X haverá muitos valores de Y, dependendo da equação que for estimada em função do conjunto de dados.

Pois bem,

se Y = a + bx é uma estimativa de $y_c = \alpha + \beta x + u$

então a e b são estimativas de α e β respectivamente.

A dispersão populacional é estimada com base nas observações amostrais em relação à reta de regressão calculada.

$$\sigma_u = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_c)^2}{n-2}}$$

OBSERVAÇÕES:

* A equação de regressão serve para predizer o valor da variável dependente, dado o valor da variável independente. Portanto, devemos observar que a estimação deve ser feita dentro do intervalo de variação dos valores da variável independente amostrada.

Ou seja, para construir um modelo de regressão, deve-se coletar os dados nos extremos do intervalo de X, numa região que se tem interesse prático de estudar e supõe-se válida a relação linear.

* A análise de regressão não indica que uma variável tende a "causar" os valores da outra, isto é, não acusa relação causa e efeito. Ela apenas indica que relação matemática existe entre as variáveis, se existir.

7.5.Coeficiente de Determinação ou Explicação:

O coeficiente de determinação (r^2) significa a variação explicada em relação a variação total (regressão).

r^2 é expresso em percentagem indicando quanto porcento da variação da variável "Y" está relacionada com a variação da variável "X".

O coeficiente de explicação nos indica se o modelo ajustado é adequado aos dados. Ele é dado pelo quociente entre a variação explicada pela regressão e a variação total.

$$r^2 = \frac{VE}{VT}$$

onde:

$$\underbrace{\sum_{i=1}^{n}(Y_i - \overline{Y})}_{VT} = \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y})}_{VR} + \underbrace{\sum_{i=1}^{n}(\hat{Y}i - \overline{Y})}_{VE}$$

VT = VR + VE

O coeficiente r^2 pode ser calculado pelo quadrado do coeficiente de correlação linear (r).

Exemplo: Para os dados do exemplo dos automóveis, interprete o coeficiente de determinação.

7.6. Teste de Hipótese para o Coeficiente Angular β :

A equação de regressão obtida depende dos valores da amostra, portanto é uma estimativa da reta verdadeira.

Mesmo quando há pouco ou nenhum relacionamento entre as variáveis na população é possível obter valores amostrais que façam as variáveis parecerem correlacionadas.

Como a dependência de Y em relação a X é representada pelo coeficiente angular β, então para sabermos se este coeficiente representa uma dependência real e não foi obtido casualmente devemos realizar um teste de hipótese sobre β.

**ETAPAS DO TESTE DE HIPÓTESE:*

⇨Hipóteses:

H_0: $\beta = 0$
H_1: $\beta \neq 0$ ou $\beta > 0$ ou $\beta < 0$

⇨ Estatística do Teste:

$$t_{calc} = \frac{b - \beta_0}{\sigma_b}$$

sendo $\sigma_b = \dfrac{\sigma_u}{\sqrt{(\sum_{i=1}^{n} x_i^2) - n\overline{X}^2}}$ onde σ_u é a estimativa do erro padrão da regressão.

O valor de t_{calc} será comparado com o valor tabelado da distribuição t-student. com n-2 graus de liberdade. Se n > 30 podemos usar o valor correspondente da distribuição Normal.

7.7. Estimação por Intervalo para o Coeficiente Angular β:

Estatística: b	Parâmetro: β

O intervalo de confiança para β será dado por:

$$b - t_{n-2} \cdot \sigma_b \leq \beta \leq b + t_{n-2} \cdot \sigma_b$$

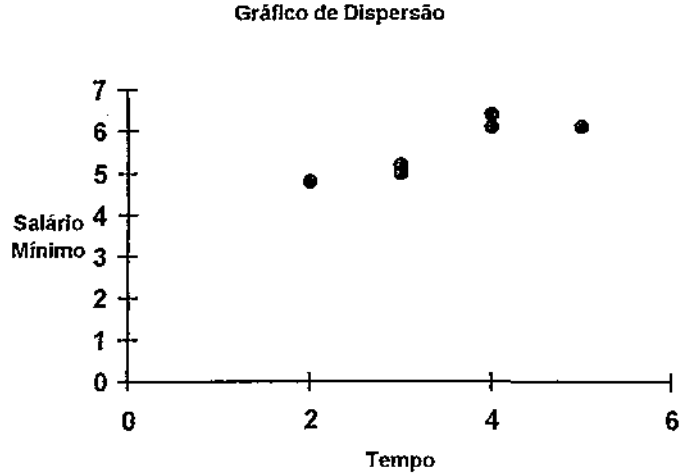
Para n > 30, podemos usar a distribuição normal.

O intervalo de confiança indica o intervalo provável em que o verdadeiro parâmetro pode estar. Mas, também serve para testar a significância β. Se o intervalo de confiança para inclui zero isso eqüivaleria dizer que a hipótese nula (Ho: β = 0) não pode ser rejeitada. Se Ho especifica algum valor diferente de zero, e se este estiver incluído no intervalo de confiança, então a alegação não pode ser rejeitada.

Exemplo 1: Uma amostra de funcionários de uma repartição pública foi selecionada aleatoriamente. Relacionou-se o tempo de serviço (em anos) com seu salário bruto mensal:

FUNCIONÁRIO	TEMPO (x)	SAL. MIN. (y)	x.y	x^2
A	3	5,2	15,6	9
B	4	6,1	24,4	16
C	3	5,0	15,0	9
D	2	4,8	9,6	4
E	5	6,1	30,5	25
F	4	6,4	25,6	16
TOTAL	21	33,6	120,7	79

Diagrama de dispersão:



$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{21}{6} = 3{,}5 \qquad \bar{y} = \frac{\sum_{j=1}^{n} y_j}{n} = \frac{33{,}6}{6} = 5{,}6$$

$$b = \frac{n\sum_{i,j=1}^{n} x_i y_j - \sum_{i=1}^{n} x_i \sum_{j=1}^{n} y_j}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{6.120{,}7 - 21.33{,}6}{6.79 - (21)^2} = \frac{724{,}2 - 705{,}6}{474 - 441} = 0{,}56$$

$a = \bar{y} - b\bar{x}$

$a = 5{,}6 - 0{,}56(3{,}5) = 3{,}64$

A estimativa da equação da regressão é

$$Yc = 3{,}64 + 0{,}56\ x$$

Interpretação: O salário médio esperado para um funcionário desta repartição é de 3,64 salários mínimos mais 0,56 salários mínimos para cada ano de serviço.

Se desejarmos saber o salário esperado de um funcionário que tenha 3,5 anos de serviço, basta substituir x por 3,5 na equação:

$$Yc = 3{,}64 + 0{,}56(3{,}5) = 5{,}6 \text{ sal. mín.}$$

7.8. Verificação da Validade do Modelo:

A adequação do modelo e as suposições para análise de regressão poderão ser feita pela análise dos resíduos.

Um gráfico é plotado com a relação entre X e os resíduos.

Os resíduos são calculados por:

$$Ri = \frac{Yi - (a + bx)}{S}$$

Um gráfico é plotado com a relação entre X e os resíduos. Para que o ajuste esteja adequado, os resíduos devem estar distribuídos aleatoriamente em torno de zero.

Uma suposição para o ajuste é que os resíduos deverão ter distribuição aproximadamente normal com variância constante σ_i , Isto é $\varepsilon_i \cong N(0, \sigma_i)$.

Para testar-se a normalidade dos resíduos pode utilizar o papel de probabilidade da distribuição Normal ou utilizar testes estatísticos adequados que verificam a hipótese de normalidade.

Dados atípicos: Alguns dados coletados podem ser resultado de fatores externos ao estudo ou podem ser digitados errados ou ainda proveniente de erros de leitura.

Quando há desconfiança da presença destes dados, deve-se verificar a procedência dos mesmos e caso sejam valores realmente atípicos, deverão ser retirados e uma nova regressão será feita.

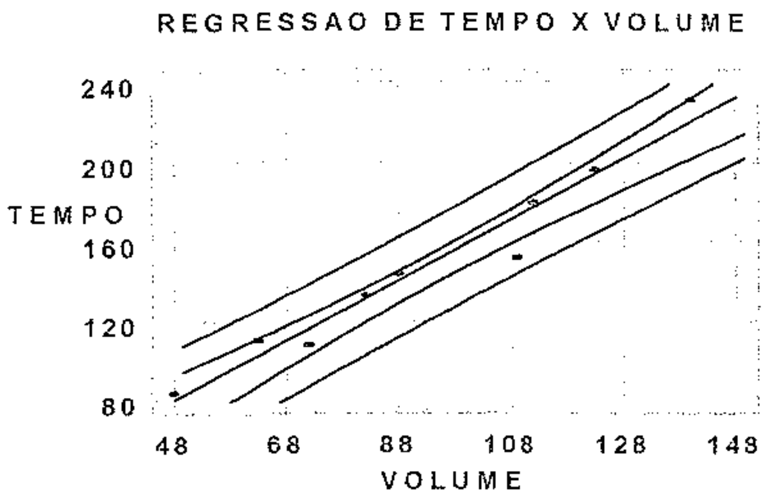
EXEMPLO:

Os dados abaixo referem-se a uma amostra de 9 pedidos de mercadoria. O objetivo do estudo é saber se existe relação entre o volume de uma carga e o tempo gasto para acondicioná-la. Por esta razão, sortearam-se os pedidos abaixo e mediu-se as duas variáveis de interesse.

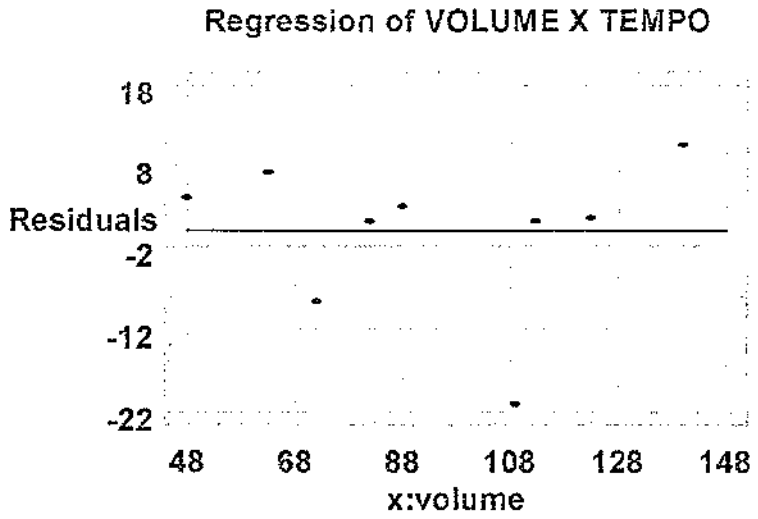
Tempo	84	108	110	133	144	152	180	196	231
Volume	48	72	63	82	88	109	112	123	140

Abaixo têm-se a saída do pacote estatístico Statgraphics:

```
    Regression Analysis - Linear model: Y = a+bX
---------------------------------------------------------------------------
Dependent variable: Tempo   Independent variable: Volume
---------------------------------------------------------------------------
                                  Standard            T            Prob.
  Parameter          Estimate       Error           Value          Level
---------------------------------------------------------------------------
  Intercept          6.58405       11.5575        0.569678         .58670
  Slope              1.52777       0.11887         12.8524         .00000
---------------------------------------------------------------------------
                        Analysis of Variance
---------------------------------------------------------------------------
Source              Sum of Squares    Df   Mean Square     F-Ratio  Prob. Level
Model                    16894.082     1     16894.082      165.18       .00000
Residual                 715.91821     7     102.27403
---------------------------------------------------------------------------
otal (Corr.)             17610.000     8
:orrelation Coefficient = 0.979462              R-squared =   95.93 percent
tnd. Error of Est. = 10.1131
```

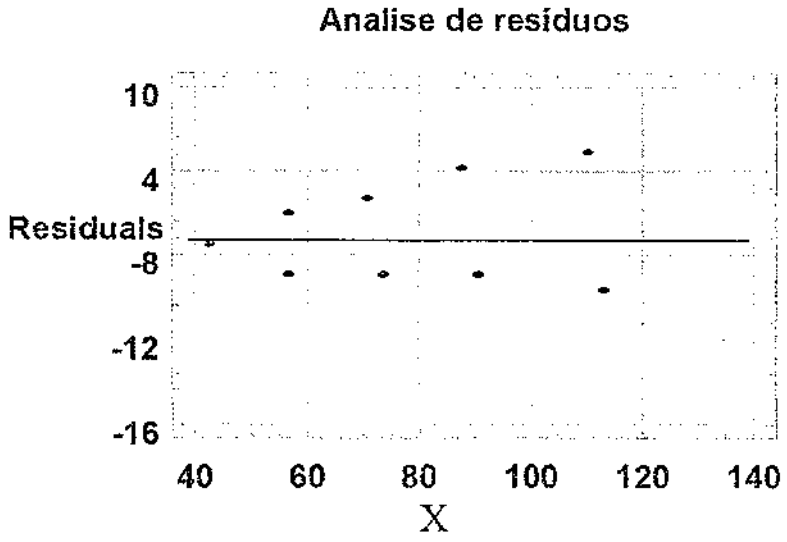


O gráfico acima mostra reta de regressão e as linhas indicam um intervalo de 95% de confiança para a resposta média dado um determinado valor de X.



O gráfico dos resíduos mostra que os pontos encontram-se aleatoriamente em torno de zero. Não há indícios de que haja um padrão não-aleatório.

Exemplo de um padrão não-aleatório:



O gráfico acima representa um modelo não adequado.

7.9.Exercícios:

1. Uma cadeia de supermercados financiou um estudo dos gastos realizados por famílias de 4 pessoas com renda mensal líquida entre 8 e 20 salários mínimos. A pesquisa levou à equação de regressão Y= -1,2+0,4X , onde Y representa a despesa mensal estimada e X a renda mensal líquida em salários mínimos.

a . Estime a despesa mensal de uma família com renda líquida de 15 s.m.
b. A equação parece sugerir que em uma família com renda mensal líquida de 3 s.m. nada gasta com mercadorias. O que você tem a dizer sobre isso?
c. A equação em questão serve para estimar a despesa mensal de uma família com 5 pessoas com renda mensal líquida de 12 s.m.? Justifique.

2. Uma população é composta por N=6 pontos (X, Y). São eles: (1,2) (5,6) (2,4) (2,3) (3,5) (5,10).

a. Determine a reta de regressão $Y=\alpha+\beta+u$.
b. Faça um gráfico que apareçam os pontos populacionais, a reta determinada no item anterior. Verifique se a $\sum u = 0$
c. Sorteie uma amostra de 4 pontos e use-os para estimar a reta de regressão determinada no item a. Desenhe no mesmo gráfico.

3. Uma amostra de fábricas de uma indústria levou **a:**

Custo total	**y**	**80**	**44**	**51**	**70**	**61**
Produção	**x**	**12**	**4**	**6**	**11**	**8**

a. Determine a reta de regressão.
b. Quais os significados econômicos de a e b?
c. Você diria a 10% que o custo marginal da indústria é superior a 4 ?

4. Uma amostra casual de 50 sujeitos com idade entre 35 e 54 anos foi investigada quanto à renda anual Y(dólares) e quanto à escolaridade X (anos) . A renda anual média resultante foi 11 anos, e ainda se verificou $\sum x^2 = 9000$. Os dados conduziram a Y=1200+ 800X, estimativa da reta de regressão.

a. Estime a renda anual de um sujeito dessa faixa etária que tenha 10 anos de escolaridade.
b. É válida a informação de que cada ano de escolaridade rende 800 dólares anuais para pessoas nesta faixa etária?

5. Abaixo, você encontra 3 afirmações. Indique, justificando, se concorda ou não com cada uma delas.

a. Se entre X e Y o coeficiente de correlação é 1, apenas uma dessas variáveis exerce influência sobre a outra, nenhuma outra participa de tal relação. Isso já não é verdade se o citado coeficiente for igual a -1.
b. Se o coeficiente angular da reta de regressão é nulo, o coeficiente de correlação entre as variáveis também o é.

6. Em certa população o coeficiente de correlação entre X e Y é -0,8.
a. O que isto significa?
b. Que percentual de variância de Y não é explicada por X?

7. Abaixo você encontra os tempos X de aquecimento de 5 iguais volumes de uma mesma solução e as respectivas temperaturas Y de ebulição.

X	20	22	19	23	17
Y	75	80	75	82	78

a. Calcule o coeficiente de correlação entre X e Y.
b. Interprete o coeficiente de determinação.
c. Teste, a 5%, se existe correlação linear populacional entre X e Y.

8. Calcule o coeficiente de correlação entre os escores de matemática e estatística de 7 estudantes.

escore em matemática	55	60	52	40	41	42,5	47,5
escore em estatístística	42	45	41	36	32	44	40

9. Um grupo de pesquisa estabeleceu uma escala de quociente de violência para programas de TV, classificou 10 programas, e coletou dados da % de pessoas que assistem ao programa.

programa	1	2	3	4	5	6	7	8	9	10
quoc. violência	10	20	30	40	40	50	55	65	70	70
% de assistência	15	16	20	24	25	30	30	35	35	35

a. Calcule o coeficiente de correlação, classifique e interprete.
b. Calcule e interprete o coeficiente de determinação.
c. Estabeleça a reta de regressão da assistência em termos do quociente de violência

10. A velocidade máxima de automóveis de fórmula 1 com motores de mesma potência é função, entre outras variáveis, do peso do veículo, no intervalo entre 700 e 800 Kg. Assim, verificou-se qual a velocidade máxima atingida em uma reta de 1.200 m. Os resultados foram:

Peso(Kg)	750	755	777	782	793
Veloc. Máx. (Km/h)	380	354	348	330	320

a. Estime a velocidade esperada para um veículo com 760 Kg?
b. Teste o coeficiente angular, com 10% de significância, para verificar uma tendência negativa.

11. Durante uma semana do verão, verificou-se o número de internações por causa de desitratação na praia em função da vendas de sorvete da única sorveteria da praia.

a. Ajuste a reta de regressão pelo método do mínimos quadrados.
b. Teste, com 5 % de significância, o coeficiente angular ser diferente de zero.
c. Interprete os resultados obtidos.
d. Estime o número de internações para uma venda de 85.

vendas de sorvete (unidades)	50	67	54	70	63	81	90
número de internações	5	7	3	8	8	10	12

8. TESTES NÃO-PARAMÉTRICOS

Todos os testes previamente estudados anteriormente impõem certas exigências, tais como igualdade de variâncias das populações, populações normalmente distribuídas, etc. Estudaremos agora um conjunto de testes, chamados testes *não-paramétricos*, ou testes *livres de distribuição*, que não exigem tais restrições.

A par da eliminação das suposições, os testes não-paramétricos são em geral fáceis de aplicar, servem para pequenas amostras, e são intuitivamente atraentes. Podem, pois, ser usados quando as suposições exigidas pelos testes paramétricos não são satisfeitas, ou quando não é possível verificar essa suposições, em razão do pequeno tamanho da amostra. Além disso, em muitas situações precisamos analisar dados qualitativos e os teste não-paramétricos são muito úteis nesse caso.

8.1. Testes de Aderência - Teste Qui-Quadrado

Uma importante classe de testes não-paramétricos é constituída pelos chamados *testes de aderência*, em que a hipótese testada refere-se à forma da distribuição da população. Nesses testes, admitimos, por hipóteses, que a distribuição da variável de interesse na população seja descrita por determinado modelo de distribuição de probabilidade e testamos esse modelo, ou seja, verificamos a boa ou má aderência dos dados da amostra ao modelo.

Se obtivermos uma boa aderência e a amostra for razoavelmente grande , poderemos, em princípio, admitir que o modelo fornece uma boa idealização da distribuição populacional. Inversamente, a rejeição de H_0 com certo nível de significância indica que o modelo testado é inadequado para representar a distribuição da população.

O teste χ^2 (qui-quadrado) é o teste de aderência mais utilizado, mas para tanto é necessário que uma suposição seja satisfeitas, essa suposição é que a freqüência esperada em cada categoria ou classe seja maior que um.

Como qualquer teste estatístico é necessário estabelecer em primeiro lugar as hipóteses. Como os testes de aderência tem por objetivo verificar se os dados observados modelam-se a alguma distribuição, temos que as hipóteses nula e alternativa devem necessariamente especificar um tipo de distribuição. Além disso, o teste para uma distribuição pode simplesmente focalizar o tipo (normal, por exemplo) ou o tipo mais seus parâmetros (normal com μ=5,2 e σ=2,4). Assim, uma hipótese nula típica poderia ser:

H_0: A distribuição da população é do tipo poisson

ou então,

H_0: A distribuição da população é poisson, com média 3,2.

Após estabelecer as hipóteses, passamos ao cálculo da estatística de teste:

$$\chi^2_{calc} = \sum_{i=1}^{k} \frac{(Oi - Ei)^2}{Oi}$$

onde: O_i = freqüência observada em certa classe
E_i = freqüência esperada em cada classe
k = número de classes

O fundamento do teste qui-quadrado é comparar as freqüências observadas com as freqüências esperadas para cada categoria ou classe i.

Para finalizar estabelecemos a região crítica e tomamos uma decisão. Rejeitaremos Ho quando:

$$\chi^2_{calc} > \chi^2_{(k-1;\alpha)}$$

Obs: Para usar essa região crítica, em outras palavras o teste qui-quadrado é necessário que a freqüência esperada de cada categoria seja maior ou igual a 5.

Exemplo: Alega-se que uma máquina de encher e fechar garrafas de cerveja produz um enchimento médio de 1 litro, com desvio padrão de 0,2 litros e, que a distribuição da quantidade de cerveja por garrafa é normal. Examinam-se 250 garrafas, anotando-se o conteúdo de cerveja por garrafa. Teste a alegação ao nível de 5% de significância.

1) Estabelecer as hipóteses:
H_0: A distribuição é normal com $\mu = 1$ l e $\sigma = 0{,}2$ l.
H_1: A distribuição não é normal com $\mu = 1$ l e $\sigma = 0{,}2$ l.

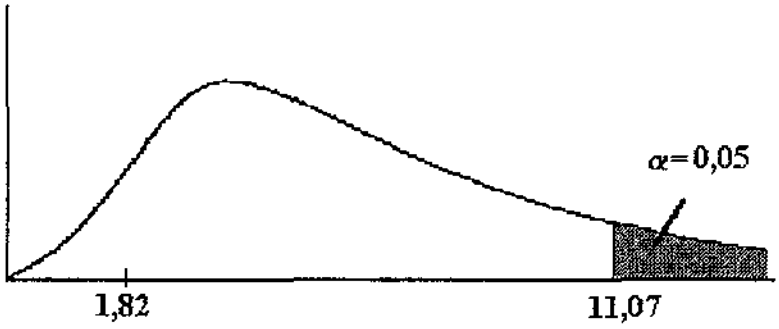
2) Calculando a estatística do teste: (com base nos dados da distribuição de freqüência que foram observados)

classe	freq obs (O)		E (x 250)	(O - E)	$(O-E)^2$	$\frac{(O-E)^2}{E}$
$\leq 0{,}96$	8	$\leq \mu - 2\sigma$	5,7 (*)	2,3	5,29	0,93
0,96 a < 0,98	36	$\mu-2\sigma$ a $< \mu-\sigma$	34,02	1,98	3,92	0,12
0,98 a < 1,00	84	$\mu-\sigma$ a $< \mu$	85,32	-1,32	1,74	0,02
1,00 a < 1,02	79	μ a $< \mu+\sigma$	85,32	-6,32	39,94	0,47
1,02 a < 1,04	37	$\mu+\sigma$ a $< \mu+2\sigma$	34,02	2,98	8,88	0,26
$\geq 1{,}04$	6	$\geq \mu + 2\sigma$	5,7	0,3	0,09	0,02
	250					1,82

* $P(X < \mu - 2\sigma) = P\left(Z < \frac{\mu - 2\sigma - \mu}{\sigma}\right) = P(Z < -2{,}00) = 0{,}0228$

3) Tomada de decisão:

$\chi^2_{calc} = 1{,}82$ e $\chi^2_{(k-1;\alpha)} = \chi^2_{(6-1;0,05)} = 11{,}07$ => Aceita-se Ho



8.2. Tabelas de contingência - Teste χ^2 de Independência

Quando existem duas ou mais variáveis qualitativas de interesse, a representação tabular das freqüências observadas pode ser feita através de uma tabela de contingência. No caso de duas variáveis apenas, essa representação torna-se muito cômoda, mediante uma tabela de duas entradas.

Seja, por exemplo, uma amostra de 500 pessoas , que foram entrevistadas quanto a suas preferências sobre o sabor de sorvete, tendo sido obtido os dados da tabela abaixo (Stevenson, 1986).

	Região			
Sabor do sorvete	Nordeste	Sul	Meio-Oeste	totais
baunilha	86	44	70	200
chocolate	45	30	50	125
morango	34	6	10	50
outros	85	20	20	125
totais	250	100	150	500

Temos uma tabela de contingência de dimensão 4 x 3, pois a variável sabor do sorvete apresenta 4 categorias possíveis no estudo, e a variável região apresentada três classificações no estudo. As freqüências registradas na parte interna da tabela indica que 86 pessoas do nordeste preferem sorvete de baunilha, 45 de chocolate, etc, no total de 500 pessoas entrevistadas. A linha e a coluna de totais fornecem a distribuição de freqüências marginais, isto é, as distribuições de cada variável qualitativa considerada individualmente, não importando a outra variável.

Podemos estar interessados em saber se as preferências de sabor variam conforme a região, isto é,

H_0: a preferência pelo sabor é *independente* da região

H_1: a preferência pelo sabor depende da região.

A hipótese nula pode ser interpretada como: as percentagens de cada população na categoria 1 são todas iguais; as percentagens de cada população na categoria 2 são todas iguais; e assim sucessivamente até a r-ésima linha. Isto é,

		população						
		1		2		....		k
	1	p_{11}	=	p_{12}	=	....	=	p_{1k}
categoria	2	p_{21}	=	p_{22}	=	....	=	p_{2k}
	...	...	=	...	=	....	=	...
	r	p_{r1}	=	p_{r2}	=	....	=	p_{rk}

Após estabelecer as hipóteses, passamos ao cálculo da estatística de teste:

$$\chi^2_{calc} = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(Oij - Eij)^2}{Eij}$$

onde: O_{ij} = freqüência observada na interseção da linha i com a coluna j
E_{ij} = freqüência esperada na interseção da linha i com a coluna j
k = número de colunas
r = número de linhas

A fim de realizar o teste precisamos conhecer o valor das freqüências esperadas em cada cela (encontro da linha e coluna) supondo a H_0 como sendo verdadeira.

As freqüências esperadas de cada cela da tabela são estimadas por:

$$E_{ij} = np_{ij}$$

onde: n = tamanho total da amostra
p_{ij} = proporção na cela ij

Ora havendo independência entre as variáveis (conforme H_0), temos que:

$$p_{ij} = p_{i.} \times p_{.j}$$

$$\text{onde } p_{i.} = \frac{\text{total da linha i}}{n}$$

$$p_{.j} = \frac{\text{total da coluna j}}{n}$$

Assim temos que:

$$E_{ij} = n\, p_{ij} = n\, p_{i.} \times p_{.j} = n \frac{\text{total da linha i}}{n} \frac{\text{total da coluna j}}{n}$$

$$\boxed{Eij = \frac{\text{total da linha i} \times \text{total da coluna j}}{n}}$$

Calculando as freqüências esperadas do exemplo temos:

	Região			
Sabor do sorvete	Nordeste	Sul	Meio-Oeste	totais
baunilha	$\frac{200x250}{500} = 100$	$\frac{200x100}{500} = 40$	$\frac{200x150}{500} = 60$	200
chocolate	$\frac{125x250}{500} = 62,5$	$\frac{125x100}{500} = 25$	$\frac{125x150}{500} = 37,5$	125
morango	25	10	15	50
outros	62,5	25	37,5	125
totais	250	100	150	500

Obs: Para usar o teste qui-quadrado, como no caso anterior, é necessário que a freqüência esperada de cada categoria seja maior ou igual a 5 ($E_{ij} \geq 5$).

Agora calculamos a estatística de teste e concluímos.

classe	freq obs (O_{ij})	freq esp (E_{ij})	(O_{ij} - E_{ij})	$(O_{ij}-E_{ij})^2$	$\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$
baunilha /nordeste	86	100	-14	196	1,96
baunilha/sul	44	40	4	16	0,4
baunilha/meio-oeste	70	60	10	100	1,67
chocolate/nordeste	45	62,5	-17,5	306,25	4,9
chocolate/sul	30	25	5	25	1
chocolate/meio-oeste	50	37,5	12,5	156,25	4,17
morango/nordeste	34	25	9	81	3,24
morango/sul	6	10	-4	16	1,6
morango/meio-oeste	10	15	-5	25	1,67
outros/nordeste	85	62,5	22,5	506,25	8,1
outros/sul	20	25	-5	25	1
outros/meio-oeste	20	37,5	17,5	306,25	8,17
				soma	37,88

Para finalizar estabelecemos a região crítica e tomamos uma decisão. Rejeitaremos Ho quando:

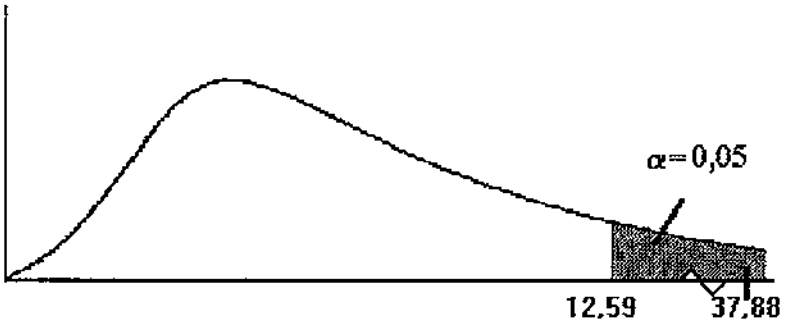
$$\chi^2_{calc} > \chi^2_{tab}$$

onde: $\chi^2_{tab} = \chi^2$ com (r - 1)(k - 1) gl para um nível de significância α.

No exemplo:

$$\chi^2_{calc} = 37{,}88$$
$$\chi^2_{[(r-1)(k-1);\alpha]} = \chi^2_{[(4-1)(3-1);0{,}05]} = \chi^2_{6;0{,}05} = 12{,}59$$

Como se verifica que $\chi^2_{calc} = 37{,}88 > \chi^2_{tab} = 12{,}59$ então rejeitamos H_0.



Conclusão: Como a estatística de teste está na região de rejeição, o estudo indica ao nível de 5% de significância que a preferência pelo sabor parece não depender da região.

8.3. Teste de Mann-Whitney

Há situações em que desejamos comparar duas populações, mas as suposições para realizar um teste paramétrico não são atendidas, assim independente da forma da distribuição de probabilidade apresentada pela variável em estudo. Um dos testes que podemos usar para comparar duas populações é o teste Mann-Whitney.

Este teste é uma alternativa para comparar duas populações, baseado na soma de postos dos valores observados. O posto de um valor em um conjunto de n valores é um número que indica sua posição no conjunto ordenado. Havendo valores iguais, considera-se um posto médio, de não afetar os postos seguintes.

Após estabelecidos os postos calculamos as seguintes estatísticas:

$$u_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - T_1$$
$$u_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - T_2$$

onde: n_1 e n_2 são os tamanhos das duas amostras e
T_1 e T_2 as respectivas somas de postos.

Tabelas para a realização do teste com base em u_1 e u_2 são disponíveis na literatura. Entretanto, para $n_1 > 7$ e $n_2 > 7$, o teste pode ser realizado por aproximação normal, sendo que para H_0 verdadeira, temos

$$\mu(u_1) = \mu(u_2) = \frac{n_1 n_2}{2} \quad \text{e} \quad \sigma(u_1) = \sigma(u_2) = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

Exemplo: Doze pneus selecionados aleatoriamente de cada um de dois fabricantes de pneus foram colocados à prova quanto à duração. Pode-se dizer que a vida média seja diferente ao nível de 5% de significância?

Fabricante 1	Fabricante 2	Postos Fab.1	Postos Fab.2
35.500	33.400	24	23
25.400	29.650	7	18
24.605	25.500	6	8
25.670	27.900	9	15
30.645	24.570	20	4,5
27.850	23.800	13	2
24.570	27.890	4,5	14
31.800	30.100	21	19
27.760	28.865	12	16
28.875	27.700	17	11
21.900	24.450	1	3
26.560	32.300	10	22
	total	144,5	155,5

$n_1 = 12$ $n_2 = 12$ $T_1 = 144,5$ $T_2 = 155,5$

$$\mu(u_1) = \mu(u_2) = \frac{n_1 n_2}{2} = \frac{12.12}{2} = 72$$

$$\sigma(u_1) = \sigma(u_2) = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{12.12(12 + 12 + 1)}{12}} = \sqrt{300} = 17,32$$

$$u_1 = n_1 n_2 + \frac{n_1 (n_1 + 1)}{2} - T_1 = 12.12 + \frac{12.13}{2} - 144,5 = 77,5$$

$$u_2 = n_1 n_2 + \frac{n_2 (n_2 + 1)}{2} - T_2 = 12.12 + \frac{12.13}{2} - 155,5 = 66,5$$

O teste pode ser feito com u_1 ou u_2, os quais são simétricos em relação à média 72. Assim,

$$z = \frac{u_1 - \mu(u_1)}{\sigma(u_1)} = \frac{72 - 77,5}{17,32} \cong -0,32$$

como $z > -z_{0,025} = -1,96$ aceitamos H_0, logo aparece que as médias são iguais.

8.4. O Coeficiente de correlação de Spearman

A correlação de postos de Spearman é uma técnica não-paramétrica para avaliar o grau de relacionamento entre observações emparelhadas de duas variáveis, quando os dados se dispões em postos.

Dados preferenciais são muito comuns em áreas como de teste de alimentos , eventos competitivos (concursos de beleza, competições atléticas)e estudo de atitudes. O objetivo de obter o coeficiente de correlação de Spearman nesses casos é determinar até que ponto dois conjunto de postos concordam ou discordam.

Podemos obter o valor do coeficiente de correlação de Spearman através da fórmula:

$$r_{sp} = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)}$$

onde: n é o número de observações

Σd^2 é a soma dos quadrados das diferenças entre os postos

Exemplo: Dois provadores devem julgar 12 vinhos. Cada um atribuirá postos denotando a preferência, desde 1 (mais alta) até 12 (mais baixa). Se os provadores estão de acordo, é de esperar que os postos atribuídos por eles aos vários tipos de vinhos sejam aproximadamente os mesmos.

	preferências		diferença	quadrado da diferença
vinho	Juiz 1	Juiz 2	d	d^2
1	1	3	+2	4
2	5	4	-1	1
3	2	1	-1	1
4	7	5	-2	4
5	4	2	-2	4
6	8	9	+1	1
7	3	7	+4	16
8	6	6	0	0
9	9	8	-1	1
10	12	10	-2	4
11	11	11	0	0
12	10	12	+2	4
			total	40

$$r_{sp} = 1 - \frac{6x40}{12(144-1)} = +0{,}86$$

O valor de +0,86 implica que os juizes são concordantes em seus julgamentos. O coeficiente de correlação por postos de Spearman pode variar entre -1 e +1. Quando o coeficiente está próximo de +1 isto indica que os dois conjuntos de postos são semelhantes , enquanto que se o coeficiente está próximo de -1, os conjuntos são diferentes. Se há acordo em alguns itens e discordância em outros, o coeficiente fica próximo de zero, o que sugere ausência de relacionamento entre os dois conjuntos.

Como utilizamos dados amostrais é necessário verificar a significância do coeficiente. Para amostras maiores que 10, a hipótese nula $r_{sp} = 0$ pode ser testada pela fórmula:

$$t = \frac{r_{sp} - 0}{\sqrt{\left(1 - 0{,}86^2\right)/(n-2)}}$$

com n-2 graus de liberdade.

8.5. Exercícios

1) Uma rotina de computador foi usada para gerar quarenta número supostos com distribuição χ^2 com dez graus de liberdade. Obtiveram os seguintes valores:

9,28	11,82	10,83	5,20	17,61	5,56	15,94	10,06	12,99	8,35
13,39	13,66	12,44	12,17	7,59	11,22	8,12	4,55	10,2	23,36
7,44	10,45	9,51	11,70	6,80	14,02	9,75	9,10	4,58	7,25
18,45	8,88	11,15	6,47	12,47	8,98	11,75	7,01	14,85	12,13

Teste, ao nível de 5% de significância, a adequabilidade da rotina usada para fim proposto. [*sugestão:* agrupe convenientemente...]

2) Uma amostra de duzentos adultos foi entrevistada a respeito de certo projeto de lei. Os resultados são os que seguem abaixo. Verifique ao nível de 1% de significância, se a opinião depende do sexo e/ou estado civil.

	favoráveis	contrários
Homens casados	56	24
Homens solteiros	15	25
Mulheres casadas	24	16
Mulheres solteiras	13	27

3) Compare as velocidade médias de dois grupos de alunos de um curso de digitação. O grupo I aprendeu a digitar por um método tradicional, enquanto o grupo II aprendeu pelo método "cego". Teste a afirmação de que o resultado dos alunos do método "cego" foi pior ao nível de 5% de significância.

Grupo I, palav/min	Grupo II, palav/min
26	23
29	28
25	20
27	18
32	24
26	23
23	22
33	

4) Solicita-se a dois gerentes que classifiquem 11 empregados de acordo com o potencial gerencial. Determine o grau de concordância, ou discordância, entre os dois gerentes e verifique ao nível de 5% se o relacionamento é ou não significativo.

Empregado	Gerente 1	Gerente 2
João	6	9
Pedro	7	10
Cláudia	5	8
Joana	4	7
Ana	9	11
Paulo	1	1
Roberto	8	6
Maria	2	2
Carla	3	4
Alberto	11	3
José	10	5

5) No decurso de um ano, determina empresa teve 50 acidentes. Um dos aspectos da investigação realizada pelo engenheiro de segurança diz respeito ao dia de ocorrência do acidente. Pelos dados que seguem abaixo, pode-se dizer que o dia da semana tenha alguma influência? Teste a hipótese de nula, de que os dias são igualmente prováveis. $\alpha = 0{,}01$.

dia	segunda	terça	quarta	quinta	sexta
num. acidentes	15	6	4	9	16

6) Uma amostra de cinqüenta peças produzidas por uma máquina forneceu distribuição de comprimentos das peças dada a seguir, valores em mm. A especificação de produção indica que as peças tem comprimento médio de 500 mm e que o comprimento se distribui normalmente em torno dessa média. Ao nível de 5% de significância, concordamos ou discordamos dessa especificação? As peças foram medidas com precisão de centésimos de milímetro.

Comprimentos	Freqüência
480 |- 485	1
485 |- 490	5
490 |- 495	11
495 |- 500	14
500 |- 505	9
505 |- 510	5
510 |- 515	4
515 |- 520	1
total	50

7) Use o teste de Mann-Whitney para determinar se um novo processo de treinamento resulta em uma redução do tempo médio de conserto (use 5% de significância)

Antigo	15,0	15,1	15,3	15,5	15,6	15,6	16,0	16,2
Novo	15,1	15,2	15,7	15,8	15,9			

8) Proprietários de certo modelo de automóvel foram entrevistados acerca do desempenho e do consumo de combustível de seus carros. O resultado da pesquisa de opiniões é resumido na seguinte tabela:

consumo\desempenho	mau	regular	bom
alto	133	125	179
baixo	21	34	58

Verifique ao nível de 5% de significância, se devemos considerar que, no consenso geral, o desempenho e consumo guardam relação entre si.

9) Recentemente foi realizada em um bairro uma pesquisa. Os resultados obtidos para 120 lares seguem abaixo. Determine se há correlação positiva ou negativa e o grau dessa relação.

	número de televisões por casa			
renda por casa	0	1	2	3
baixa	7	11	6	0
média	4	4	3	13
média alta	3	7	28	10
alta	1	3	8	12

9. REFERÊNCIAS BIBLIOGRÁFICAS

1. COSTA NETO, P.L. de O. **Estatística**. Edgard Blücher. São Paulo, 1977.

2. FONSECA, J.S. DA & MARTINS, G. de A. **Curso de Estatística**. Editora Atlas, 3ª edição, São Paulo, 1982.

3. GUERRA,M.J. & DONAIRE, D. **Estatística Indutiva. Teoria e Aplicações**. Livraria Ciência e Tecnologia Editora, 4ª ed., São Paulo, 1990.

4. HOEL, P.G. **Estatística Elementar**. Editora Atlas. São Paulo, 1977.

5. SOARES, J.F. et alli. **Introdução à Estatística**. Guanabara Koogan, 1991.

6. KAZMIER, L. **Estatística Aplicada à Economia e Administração**. Editora McGraw-Hill do Brasil, 1977.

7. MENDENHALL, W. **Probabilidade e Estatística. Vol.1**. Editora Campus. Rio de Janeiro, 1985.

8. MENDENHALL, W. **Probabilidade e Estatística. Vol.2.** Editora Campus. Rio de Janeiro, 1985.

9. PEREIRA, Rivadavia. **A Estatística e suas aplicações**. Grafosul. Porto Alegre, 1978.

10. SNEDECOR,G.W. & COCHRAN, W. **Statisitcal Methods**. Iowa State Press, 7ª edição, 1980.

11. SPIEGEL, M. **Probabilidade e Estatística.** McGraw-Hill do Brasil, 1977.

12. STEVENSON,W. **Estatística Aplicada à Administração.** Editora Harbra, São Paulo, 1986.