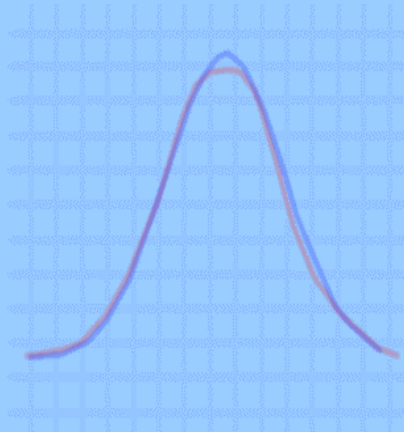




Universidade Federal da Bahia
Instituto de Matemática
Departamento de Estatística

Estatística IV (MAT027) e Introdução à Estatística (MAT050)



NOTAS DE AULA
UNIDADE III
INFERÊNCIA ESTATÍSTICA

1 INTRODUÇÃO

Até o presente momento, nós aprendemos a descrever uma amostra através das medidas de tendência central e de dispersão, que são parte da chamada “estatística descritiva”. Com a utilização da “inferência estatística”, desejamos inferir indutivamente propriedades de uma população (ou universo) com base nos resultados obtidos com a amostra (ou subconjunto do inverso) o que constitui uma ferramenta muito importante no desenvolvimento de uma disciplina científica. Toda a inferência na Estatística está baseada na “teoria das probabilidades”, que nós acabamos de ver.

Freqüentemente, devemos tomar decisões sobre populações com base em informações obtidas em amostras das mesmas. Tais decisões chamam-se decisões estatísticas. Por exemplo, com base em resultados amostrais, podemos querer decidir se determinada droga é eficiente na cura de determinada doença, se um processo educacional é melhor do que outro, se um determinado número de caixas de um banco é suficiente para um atendimento rápido ao cliente, etc. Para a tomada de tais decisões utilizaremos a inferência estatística.

Nesta etapa do nosso curso, iniciaremos a discussão falando sobre aspectos fundamentais da amostragem, após a qual introduziremos as noções sobre distribuição amostral da média e da proporção, intervalos de confiança e testes de hipóteses para médias e proporções, finalizando com um método para avaliação de relações entre variáveis qualitativas, que é o teste qui-quadrado de independência.

2 AMOSTRAGEM

Amostragem é o ato de obter uma amostra de uma população, podendo-se definir população como um conjunto de elementos, cada um deles apresentando uma ou mais características em comum. Amostra é, simplesmente, uma parte da população.

O levantamento por amostragem, quando comparado com o levantamento total, apresenta certas vantagens:

- custo menor;
- resultado em menor tempo;
- objetivos mais amplos;
- dados mais fidedignos.

Há situações em que a amostragem se impõe. Assim, pode-se ter o caso de a população de estudo ser muito grande, sendo impraticável o levantamento total. Em casos em que o processo de investigação das características de cada elemento for destrutivo (teste de resistência de materiais, por exemplo), só tem sentido trabalhar-se com amostras. Há também os problemas de ordem ética: novas drogas, vacinas, técnicas cirúrgicas devem ser testadas inicialmente em amostras, antes de seu uso amplo na população.

2.1 ETAPAS DE UM LEVANTAMENTO POR AMOSTRAGEM

Quer a amostragem seja feita dentro de um laboratório (por exemplo, para selecionar ratos que serão usados em um experimento sobre agentes cancerígenos), quer a amostragem seja feita sobre a população geral (por exemplo, para obter informações sobre aspectos de fertilidade de mulheres moradoras em uma grande cidade, como São Paulo), existem etapas que devem ser seguidas, intimamente ligadas aos princípios de metodologia de pesquisa científica. Tais itens serão comentados com linguagem mais dirigida a levantamentos objetivando estimar parâmetros de populações reais de seres humanos. Tais comentários, todavia, são aplicáveis a qualquer situação em que se requeira amostragem, bastando para isso pequenos ajustes, basicamente de forma e não de conteúdo.

As principais etapas de um levantamento por amostragem são:

1. Explicitação dos objetivos com bastante clareza, a fim de evitar dúvidas posteriores ou mesmo esquecimentos, devendo ficar bem definida qual a unidade elementar (elemento) ou unidade de análise a ser trabalhada.
2. Definição da população a ser amostrada.
 - (a) Em certas situações isto pode ser relativamente fácil, como no caso de se desejar tomar uma amostra de uma população de crianças que estejam matriculadas e frequentando certo grupo escolar. Ter-se-ia uma situação mais complexa para se estudar gestantes que procuram centros de saúde para fazer pré-natal.
3. Escolha das variáveis a serem observadas em cada unidade de análise.
 - (a) Deve ser verificado se todos os dados que vão ser levantados são relevantes para a pesquisa e se nenhum dado relevante foi omitido. Existe uma tendência, particularmente ao se trabalhar com populações humanas, usando questionário, de se fazer muitas perguntas, um grande número das quais nunca são analisadas. Questionários longos, em geral, levam a diminuir a qualidade da resposta.
4. Especificação do grau de precisão desejado. Os resultados de levantamento por amostragem são sujeitos a incerteza, devido a erros de medida e ao fato de apenas parte da população ser examinada. O grau de incerteza pode ser diminuído tomando-se amostras maiores e empregando-se melhores técnicas ou aparelhos de medida.
5. Escolha dos instrumentos de medida e da forma de abordagem.
 - (a) Em caso de inquéritos sobre nutrição, por exemplo, poderá haver escolha entre observação única ou observação durante sete dias; em estudos antropométricos será decidido o tipo de balança, calibrador e outros aparelhos a serem usados. Questionários podem ser preenchidos pelo próprio indivíduo ou serem aplicados a cada indivíduo por um entrevistador treinado.
6. Escolha da unidade amostral, que é definida como a menor parte distinta e identificável da população, para fins de enumeração e sorteio da amostra.
 - (a) Uma unidade amostral pode ser o próprio elemento de estudo (criança, cobaia, cidade, trecho da estrada) ou um conjunto de elementos (classe de escola, ninhada, Estado, conjunto de trechos continuados de estrada). As unidades amostrais devem cobrir toda a população e não podem apresentar transvariação, ou seja, um elemento de estudo não pode pertencer ao mesmo tempo a mais de uma unidade amostral. À relação, lista ou mapa contendo todas as unidades amostrais dá-se o nome de sistema de referência ou fundamentos da pesquisa.
7. Execução de prova experimental, prova-piloto ou pré-teste.
 - (a) Nesta etapa é feito um verdadeiro ensaio do trabalho a ser desenvolvido, sendo testados os instrumentos de medida, questionário, pessoal de campo, a sistemática proposta, a reação da população. Orienta os reajustes necessários e pode dar informações valiosas sobre possível duração e custo da pesquisa e indicações sobre a variabilidade do fenômeno pesquisado, o que permite calcular melhor o tamanho da amostra.
8. Seleção da amostra após decidido qual deve ser o respectivo tamanho.

Esta seleção deve ser feita, de preferência, por meio de sorteio do tipo lotérico.

2.2 TIPOS DE AMOSTRAGEM

A amostragem é probabilística quando cada unidade amostral na população tem uma probabilidade conhecida e diferente de zero de pertencer à amostra. De outra forma, a amostragem é dita não-probabilística.

Admita-se, por exemplo, que seja definida uma população de dez grupos escolares, cada qual com certo número de alunos, desejando-se uma amostra de tamanho igual a cinco grupos escolares. Se o pesquisador decidiu simplesmente escolher os grupos escolares A, B, C, I, J, ter-se-ia uma amostragem não-probabilística. É possível, no entanto (e mesmo desejável), que o investigador obtenha o número de alunos de cada grupo escolar e faça um sorteio para obtenção das cinco escolas, cada escola tendo uma probabilidade de ser sorteada proporcionalmente ao seu número de alunos; será uma amostragem probabilística.

A amostragem não-probabilística pode prejudicar sensivelmente a validade externa de um estudo, pois muitos fatores podem influir na escolha de uma unidade amostral para pertencer à amostra, prejudicando sua representatividade em relação à população. Mesmo assim, existem situações em que ela é usada, havendo então interesse em se conhecer algumas formas de amostragem não-probabilística:

- por voluntários, que é bastante usada em ensaios clínicos para teste de novos medicamentos;
- intencional, quando as unidades que compõe a amostra são escolhidas pelo pesquisador; é usada na verificação de poluição de praias;
- por acesso mais fácil, em que as unidades são escolhidas por estarem em melhores condições de acessibilidade. Veja-se, por exemplo, o caso em que se defina para estudo todo o conjunto de habitantes de uma área rural, tendo o domicílio como unidade amostral; se o entrevistador escolheu os dez primeiros domicílios do seu caminho, teremos este tipo de amostragem, tendo sido desconsideradas as outras unidades amostrais que também pertencem à população.

2.3 AMOSTRAGEM PROBABILÍSTICA

2.3.1 AMOSTRAGEM CASUAL SIMPLES

Também conhecida por amostragem ocasional, acidental, casual, randômica, etc. A amostragem simples ao acaso destaca-se por ser um processo de seleção bastante fácil e muito usado. Neste processo, todos os elementos da população tem igual probabilidade de serem escolhidos, não só antes de ser iniciado, como também até completar-se o processo de coleta. Eis o procedimento para seu uso:

1. Devemos numerar todos os elementos da população. Se, por exemplo, nossa população tem 5.000 elementos, devemos numerá-los de 0000 a 4999 ou, como acontece geralmente, usamos um número que já identifica o elemento.
2. Devemos efetuar sucessivos sorteios com reposição (ou não) até completar o tamanho da amostra (n).

Para realizar este sorteio, podemos usar as “tábuas de números aleatórios” ou ainda preferencialmente, o uso de programas computacionais próprios para estes fins.

Se, durante o sorteio, unidades amostrais já sorteadas poderem ser novamente sorteadas, sendo representadas uma, duas ou mais vezes na amostra, ter-se-á a chamada amostragem casual simples com reposição.

Em geral, dar-se preferência ao tipo de amostragem casual simples sem reposição, principalmente quando se trata de populações com reduzido número de unidades amostrais.

2.3.2 AMOSTRAGEM SISTEMÁTICA

Trata-se de uma variação da amostragem simples ao acaso, muito conveniente quando a população está naturalmente ordenada, como fichas em um fichário, listas telefônicas, etc.

Procedimento:

Seja N o tamanho da população e n o tamanho da amostra. Então, calcula-se o intervalo de amostragem $\frac{N}{n}$ ou o inteiro mais próximo que chamaremos de “ a ”. Sorteia-se um número entre 1 e “ a ” e seja x esse número. Formamos, assim, a amostra dos elementos correspondentes aos números: $x; (x + a); (x + 2a); \dots; [x + (n - 1)a]$.

Exemplos:

1. Seja $N = 500$ e $n = 50$. Então, $\frac{500}{50} = 10$, ou $a = 10$.

Sorteia-se um número de 1 a 10. Seja 3 ($x = 3$) o número sorteado. Logo, os elementos numerados por 3; 13; 23; 33; ... serão os componentes da amostra.

2. Para uma população de tamanho $N = 32$, numerada seqüencialmente de 1 a 32, e amostra de tamanho $n = 8$, tem-se $a = 4$.

O início casual “ x ” deve ser sorteado entre 1, 2, 3 e 4; admita-se que tenha sido $x = 3$.

A amostra fica constituída das unidades amostrais de número (ou ordem):

3; 7; 11; 15; 19; 23; 27; 31.

Pode ser visto que o resultado é obtido somando-se sucessivamente o intervalo de amostragem $a = 4$ a cada número obtido imediatamente antes.

2.3.3 AMOSTRAGEM ESTRATIFICADA

No caso de população heterogênea, na qual podemos distinguir subpopulações mais ou menos homogêneas denominadas estratos, podemos usar a amostragem estratificada.

Estratificar uma população é dividi-la em L subpopulações denominadas estratos, tais que $n_1 + n_2 + \dots + n_L = n$, onde os estratos são mutuamente exclusivos.

Após a determinação dos estratos, seleciona-se uma amostra aleatória de cada subpopulação.

Muitas vezes uma população é composta de subpopulações (ou estratos) bem definidos, havendo maior homogeneidade entre as unidades amostrais dentro de cada estrato do que entre as unidades amostrais de estratos diferentes. Sexo, idade, condição sócio-econômica, são exemplos típicos. Nestas condições, tais estratos devem ser levados em consideração e o sorteio da amostra deve ser feito em cada um deles independentemente; daí o nome de amostragem estratificada.

Um caso muito importante da amostragem estratificada é aquele em que o pesquisador deseja que as subpopulações sejam representadas na amostra com a mesma proporcionalidade com que compõe a população total. Trata-se da situação denominada amostragem casual simples estratificada com partilha proporcional ou simplesmente amostragem estratificada proporcional.

3 COMPARAÇÃO ENTRE ESTATÍSTICAS E PARÂMETROS

Quando estamos trabalhando com dados provenientes de uma amostra, poderemos calcular algumas estatísticas, que são caracterizações da amostra. Por exemplo, pode-se calcular a média ou a variância desses dados. Logo, estatística é uma função dos dados da amostra X_1, X_2, \dots, X_n :

$$T = f(X_1, X_2, \dots, X_n)$$

Temos então que:

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n} \implies \text{a média da amostra}$$

$$S^2 = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n - 1} \implies \text{a variância da amostra}$$

que são funções de X_1, X_2, \dots, X_n .

\bar{X} e S^2 são *as estatísticas* mais comuns.

Por sua vez, *os parâmetros* caracterizam a população.

Os símbolos mais comuns utilizados para distinguir se estamos trabalhando com amostra ou população são dados na tabela a seguir:

Discriminação	Estatística	Parâmetros
Média	\bar{X}	μ
Variância	S^2	σ^2
Número de elementos	n	N
Proporção	\hat{p}	p

4 DISTRIBUIÇÕES AMOSTRAIS

O problema da Inferência Estatística é fazer uma afirmação sobre parâmetros da população através da amostra.

4.1 DISTRIBUIÇÃO AMOSTRAL DA MÉDIA

Suponha uma população identificada pela v.a. X , cujos parâmetros média populacional $\mu = E(X)$ e variância populacional $\sigma^2 = Var(X)$ são supostamente conhecidos. Vamos retirar todas as possíveis amostras ao acaso simples, de tamanho n dessa população, e para cada uma calcular a média \bar{X} .

Vamos supor a seguinte população:

A população $\{2, 3, 4, 5\}$ tem média $\mu = 3,5$ e variância $\sigma^2 = 1,25$.

Vamos relacionar todas as amostras possíveis de tamanho 2 dessa população:

(2, 2) (2, 3) (2, 4) (2, 5)
 (3, 2) (3, 3) (3, 4) (3, 5)
 (4, 2) (4, 3) (4, 4) (4, 5)
 (5, 2) (5, 3) (5, 4) (5, 5)

Agora vamos calcular a média de cada amostra acima relacionada. Então teremos:

2,0 2,5 3,0 3,5
 2,5 3,0 3,5 4,0
 3,0 3,5 4,0 4,5
 3,5 4,0 4,5 5,0

Por fim, vamos calcular a média das médias, ou seja,

$$E(\bar{X}) = \frac{2,0 + 2,5 + 3,0 + 3,5 + \dots + 4,5 + 5,0}{16} = 3,5$$

De modo análogo, vamos calcular a variância da média \bar{X} por:

$$Var(\bar{X}) = \frac{1}{n} \underbrace{\sum_{i=1}^n (x_i - \bar{x})^2}_A$$

Para o cálculo da parte denotada por A , teremos que:

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x})^2 &= (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \dots + (x_n - \bar{x})^2 \\ &= (2,0 - 3,5)^2 + (2,5 - 3,5)^2 + \dots + (5,0 - 3,5)^2 = 10 \end{aligned}$$

Substituindo na fórmula da variância, teremos:

$$Var(\bar{X}) = \frac{1}{16}(10) = 0,625 = \frac{Var(X)}{n} = \frac{1,25}{2}$$

Teorema: Para amostras casuais simples (X_1, X_2, \dots, X_n) , retiradas de uma população com média μ e variância σ^2 , a distribuição amostral da média $= \frac{(X_1 + X_2 + \dots + X_n)}{n}$ aproxima-se de uma Distribuição Normal com média μ e variância $\frac{\sigma^2}{n}$, quando n tende ao infinito.

Desta forma:

$$E(\bar{X}) = \mu$$

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

Se $X \sim N(\mu, \sigma^2) \implies \bar{X} \sim N(\mu, \frac{\sigma^2}{n}), \forall n > 1$

Para padronizarmos a v.a. \bar{X} vamos usar a mesma transformação em que subtrai-se a média e divide-se pelo desvio-padrão: $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \implies Z \sim N(0, 1)$

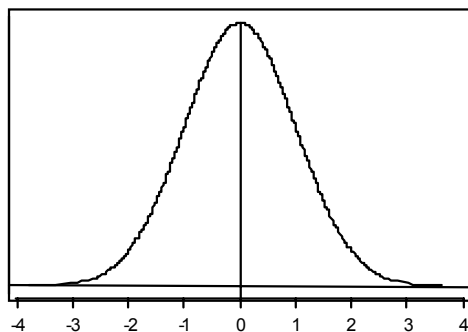
Exemplo: Uma variável aleatória X tem distribuição normal, com média 100 e desvio-padrão 10. Se \bar{X} é a média de uma amostra de 16 elementos retirados desta população, calcule?

1. $P(90 < \bar{X} < 110)$;
2. Que deveria ser o tamanho da amostra de modo a garantir a que $P(90 < \bar{X} < 110) = 95\%$?

Resolução:

- $X \sim N(100, 100) \implies \bar{X} \sim N(100; 6, 25)$

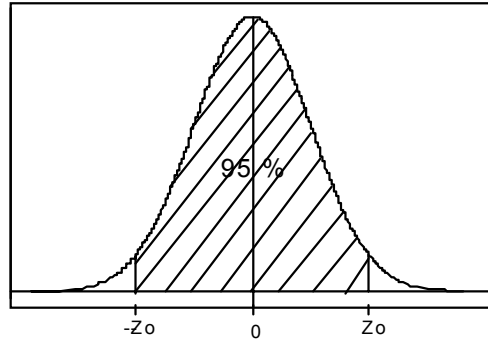
$$P(90 < \bar{X} < 110) = P\left(\frac{90 - \mu}{\frac{\sigma}{\sqrt{n}}} < Z < \frac{110 - \mu}{\frac{\sigma}{\sqrt{n}}}\right) = P\left(\frac{-10}{2,5} < Z < \frac{10}{2,5}\right) = P(-4 < Z < 4) = 1,0$$



- Calculando o valor de n para a condição dada:

$$P(90 < \bar{X} < 110) = P\left(\frac{-10}{\frac{10}{\sqrt{n}}} < Z < \frac{10}{\frac{10}{\sqrt{n}}}\right) = 95\%$$

Como o n é fixo, então teremos Z entre dois pontos simétricos e queremos saber que pontos são estes que deixam uma área de 95% entre eles.



Olhando a tabela da distribuição normal, teremos que $z_0 = 1,96$. Então,

$$P\left(\frac{-10}{\sqrt{n}} < Z < \frac{10}{\sqrt{n}}\right) = 95\% \implies P(z_0 < Z < z_0) = 95\% \implies P(-1,96 < Z < 1,96) = 95\%$$

Logo,

$$\frac{-10}{\sqrt{n}} = -1,96 \longrightarrow 10\sqrt{n} = 19,6 \longrightarrow \sqrt{n} = 1,96 \longrightarrow n = 4,0$$

4.2 DISTRIBUIÇÃO AMOSTRAL DA PROPORÇÃO

Consideremos que numa população a proporção de elementos que possui determinada característica é p . Sabemos então que a proporção de indivíduos que não possuem esta característica será $(1 - p)$.

Seja X a v.a. que conta o número de pessoas que possuem a característica de interesse. Já vimos que esta v.a. tem Distribuição Binomial, com média $E(X) = np$ e variância $Var(X) = npq$; onde p é a proporção de indivíduos na **população** que têm a característica.

Definindo como \hat{p} a proporção de indivíduos portadores da característica na amostra, teremos que, para n suficientemente grande, a distribuição amostral da proporção \hat{p} será:

$$\hat{p} \sim N\left(p; \frac{p(1-p)}{n}\right)$$

Exemplo: Sabe-se que 20% das peças de um lote são defeituosas. Sorteiam-se 8 peças, com reposição, e calcula-se a proporção \hat{p} de peças defeituosas na amostra. Qual será a distribuição de \hat{p} ?

Resolução: Como foi visto $\hat{p} \sim N\left(p; \frac{p(1-p)}{n}\right)$, onde $p = 20\% = 0,2$ e $n = 8$.

Então, $E(\hat{p}) = 0,20$ e $Var(\hat{p}) = \frac{(0,20) \cdot (0,80)}{8} = 0,02$. Logo, $\hat{p} \sim N(0,20; 0,02)$.

5 INTERVALOS DE CONFIANÇA

Até o momento, nós calculamos uma estimativa do parâmetro desconhecido através de *estimadores pontuais* (as estatísticas), que especificam um único valor para o estimador. Por exemplo, quando calculamos a estatística \bar{X} para uma determinada amostra, o valor que estima μ é único e o chamaremos de estimativa da média populacional.

Este procedimento, porém, não nos permite julgar qual a possível magnitude do erro que estamos cometendo. Pode haver o interesse em se obter um indicador de precisão dessa estimativa. Daí surge a idéia de ser feita uma estimativa por intervalo, que consiste em estabelecer um intervalo de confiança para μ , por exemplo.

Vamos voltar ao exemplo do **item 4.1**.

De uma população de tamanho 4 foram retiradas todas as amostras possíveis de tamanho 2. Sabemos que a média populacional μ é 3,5 e para cada valor calculamos o erro entre a média da amostra e a média populacional.

Amostra	Elementos amostrais	Média da amostra - \bar{X}	Erro = $(\bar{X} - \mu)$
1	(2, 2)	2, 0	-1, 5
2	(2, 3)	2, 5	-1, 0
3	(2, 4)	3, 0	-0, 5
4	(2, 5)	3, 5	0, 0
5	(3, 2)	2, 5	-1, 0
6	(3, 3)	3, 0	-0, 5
7	(3, 4)	3, 5	0, 0
8	(3, 5)	4, 0	0, 5
9	(4, 2)	3, 0	-0, 5
10	(4, 3)	3, 5	0, 0
11	(4, 4)	4, 0	0, 5
12	(4, 5)	4, 5	1, 0
13	(5, 2)	3, 5	0, 0
14	(5, 3)	4, 0	0, 5
15	(5, 4)	4, 5	1, 0
16	(5, 5)	5, 0	1, 5

Verificamos na última coluna que, para algumas amostras, o erro entre \bar{X} e μ foi igual a zero. Entretanto, outras amostras apresentaram erros grandes de 1, 5 para mais ou para menos em relação à média populacional.

Porém, quando vamos elaborar uma estimativa para um parâmetro populacional, utilizamos apenas uma dessas possíveis amostras. Interessa-nos construir um intervalo de confiança que garanta que essa única amostra escolhida seja uma dentre as amostras que admitimos que o erro cometido não seja significativo para o estudo. No exemplo dado, admitindo que o erro amostral máximo desejável seja 0, 5, as amostras 1, 2, 5, 12, 15 e 16 nos levariam a estimativas com erro muito grande. Existe portanto uma probabilidade de $\frac{6}{16}$ (ou 37, 5%) de sortearmos uma amostra não desejável.

Nosso objetivo é construir um intervalo de confiança de tal maneira que tenhamos uma probabilidade pequena de sortearmos amostras com erros não desejáveis. Ou seja, admitindo que essa probabilidade seja 10%, teremos apenas 10% de amostras que gerarão estimativas cujo o erro amostral superará o valor que estamos admitindo como máximo. Existirão, portanto, 90% de amostras cujas estimativas serão consideradas “boas”, isto é, cujos intervalos gerados a partir dessas estimativas pontuais esperamos que contenham o verdadeiro parâmetro populacional. Chamamos a essa última probabilidade de nível de confiança da estimativa e simbolizamos por $(1 - \alpha)$.

Observe que $(1 - \alpha)$ pode ser igual a 99%, 95%, 90%, 80%, etc. Se estabelecermos que desejamos um intervalo de confiança $(1 - \alpha) = 95\%$, α será igual a 0, 05; o que significa que, ao estimarmos o parâmetro, poderemos estar utilizando uma daquelas amostras dentre as 5% que geram estimativas intervalares cujos erros amostrais encontram-se acima do desejável.

Resumindo, desejamos construir um intervalo, por exemplo para a média, de tal maneira que $P(|\text{média amostral} - \text{média populacional}| < \text{erro amostral}) = 95\% = 1 - \alpha$

5.1 INTERVALO DE CONFIANÇA para a MÉDIA POPULACIONAL (μ)

Quando a VARIÂNCIA POPULACIONAL (σ^2) é conhecida.

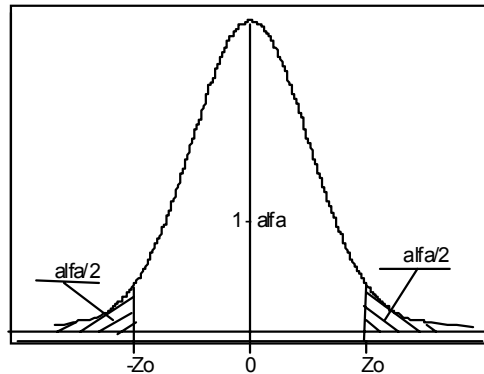
Seja $X \sim N(\mu, \sigma^2)$. Como já vimos, $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$. Padronizando a variável \bar{X} , teremos que:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \text{ tem distribuição } N(0, 1)$$

Vamos então considerar o intervalo de valores de Z, simétrico em torno da média, tal que a probabilidade de Z assumir valor neste intervalo seja igual a $(1 - \alpha)$, isto é:

$$P(-z_0 < Z < z_0) = 1 - \alpha$$

Conforme o gráfico da distribuição de Z, verifica-se:



Se o intervalo de confiança é definido pelo intervalo $(-z_0; z_0)$, a área em branco é a que fornece a probabilidade acima requerida. Logo, a área restante será igual a α . Como temos duas áreas iguais, cada uma delas será igual a $\frac{\alpha}{2}$.

Vamos chamar os pontos z_0 de $z_{\frac{\alpha}{2}}$ e $-z_0$ de $-z_{\frac{\alpha}{2}}$, pois eles nos fornecerão os pontos que deixam nas caudas da distribuição uma área de $\frac{\alpha}{2}$.

Então, rescrevendo a probabilidade, teríamos que: $P(-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}) = 1 - \alpha$

Vamos substituir: $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$. Logo, $P(-z_{\frac{\alpha}{2}} < Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{\frac{\alpha}{2}}) = 1 - \alpha$

Rearranjando a expressão entre parênteses, obtemos:

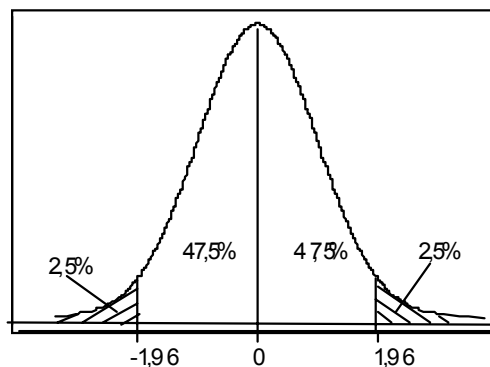
$$P(\bar{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

Por definição, a expressão $\bar{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ é o intervalo de confiança para o parâmetro μ , ao nível de confiança $1 - \alpha$, quando σ é conhecido e o erro amostral é definido por $z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ que chamaremos de ε .

Exemplo: Seja X a v.a. que representa a taxa de colesterol no plasma sanguíneo humano. Vamos supor que essa variável tem distribuição aproximadamente normal com média μ e desvio padrão $\sigma = 20\text{mg}/100\text{ml}$ de plasma.

Vejam como se obtém o intervalo de confiança para o parâmetro μ . Vamos estabelecer o nível de confiança $(1 - \alpha) = 0,95$; isto é, um nível de 95% de confiança.

Observe a seguinte figura:



Através de uma tabela da Distribuição Normal Zero-Um ou $X \sim N(0, 1)$, obtemos a probabilidade correspondente a área 0,475, o valor $z = 1,96$. Portanto, os valores de $z = \pm 1,96$ delimitam a área de probabilidade igual a 0,95 (0,475 + 0,475).

Para encontrarmos o intervalo de confiança para μ , supondo amostras de tamanho $n = 25$, podemos escrever:

$$P\left(\bar{X} - 1,96 \cdot \frac{20}{25} < \mu < \bar{X} + 1,96 \cdot \frac{20}{25}\right) = 0,95 \implies P(\bar{X} - 7,84 < \bar{X} + 7,84) = 0,95.$$

Portanto, a expressão:

$$\bar{X} - 7,84 < \mu < \bar{X} + 7,84$$

é o intervalo de confiança ao nível de 95% para a taxa média de colesterol no plasma sangüíneo humano, obtido com base em uma amostra qualquer de tamanho 25. Portanto, o erro amostral nesse exemplo é de 7,84 mg/100ml de plasma.

Para o cálculo dos valores entre os quais se encontra o verdadeiro valor do parâmetro, teríamos que calcular \bar{X} e substituir na expressão anterior. Vamos supor que a média da amostra é $\bar{X} = 198 \text{mg}/100 \text{ml}$ de plasma. Então, o intervalo de 95% de confiança, determinado com base nessa amostra é

$$190,16 < \mu < 205,84$$

Na prática, esperamos que a verdadeira média populacional esteja contida neste intervalo com 95% de confiança, ou seja, que a amostra utilizada para a estimativa seja uma das 95% das amostras possíveis da população cujo erro amostral seja menor ou igual a erro máximo admitido (7,84 mg/100ml de plasma). É importante frisar que o intervalo de confiança para μ é um indicador da precisão da estimativa de \bar{X} .

5.2 INTERVALO DE CONFIANÇA para a MÉDIA POPULACIONAL (μ)

Quando a VARIÂNCIA POPULACIONAL (σ^2) é desconhecida.

Neste caso, como não conhecemos σ^2 , precisaremos calcular a estimativa S^2 a partir de uma amostra. Devemos lembrar que:

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

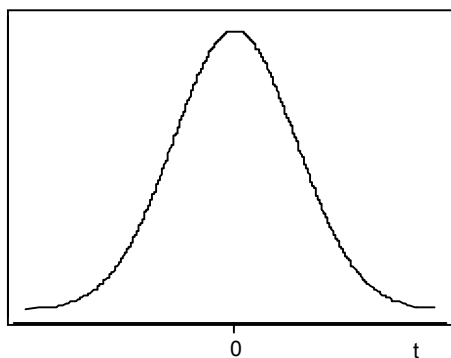
Sabemos que $X \sim N(\mu, \sigma^2) \implies \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ e $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ tem distribuição $N(0, 1)$

Agora, vamos considerar a variável aleatória $T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$.

Esta v.a. T tem uma distribuição conhecida como “*t*” de Student com $(n-1)$ graus de liberdade. O gráfico da distribuição “*t*” de Student é simétrico e tem forma similar à curva da Normal, só que menos “achatada”.

Os graus de liberdade de uma distribuição correspondem ao número de variáveis independentes que estão sendo somadas.

Sua representação gráfica seria a seguinte:



As probabilidades para uma v.a. T com distribuição “ t de Student também encontram-se tabeladas.

Vejamos agora como pode ser obtido um intervalo de confiança para μ quando o valor de σ é desconhecido. Dispostos de uma estimativa de σ obtida com base em uma amostra casual simples de tamanho n e sabendo-se que T tem distribuição “ t de Student com $(n - 1)$ graus de liberdade, associados a S . O procedimento a ser adotado é parecido com o adotado para o uso da distribuição Normal.

Dada uma amostra de tamanho n , imediatamente obtemos o valor $(n - 1)$ dos graus de liberdade associados à estimativa S de σ . Então, definido o intervalo de confiança $(1 - \alpha)$, iremos procurar na tabela a seguinte probabilidade:

$$P(-t_c < T < t_c) = 1 - \alpha$$

Os pontos $-t_c$ e t_c vão ser encontrados da mesma forma verificada para a Distribuição Normal.

Já vimos que $T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$. Então, podemos escrever $P(-t_c < T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} < t_c) = 1 - \alpha$.

Rearranjando a expressão entre parênteses, obtemos:

$$P(\bar{X} - t_c \cdot \frac{S}{\sqrt{n}} < \mu < \bar{X} + t_c \cdot \frac{S}{\sqrt{n}}) = 1 - \alpha.$$

Por definição, a expressão $\bar{X} - t_c \cdot \frac{S}{\sqrt{n}} < \mu < \bar{X} + t_c \cdot \frac{S}{\sqrt{n}}$ é o intervalo de confiança para o parâmetro μ , ao nível de confiança $1 - \alpha$, quando σ é desconhecido.

Exemplo: Seja X a variável aleatória, que representa a taxa normal de colesterol no plasma sanguíneo humano. Suponhamos que, com base em uma amostra casual simples de 25 indivíduos normais, um pesquisador obteve a média $\bar{X} = 198 \text{mg}/100 \text{ml}$ de plasma e o desvio-padrão $S = 30 \text{mg}/100 \text{ml}$ de plasma.

Vamos obter, com base nessa amostra, o intervalo de 90% de confiança para μ . Na tabela “ t de Student procuraremos o valor t_c para $(n - 1) = 24$ graus de liberdade e $\alpha = 10\%$. Este valor é $t_c = 1,71$. Então, escrevemos:

$$198 - 1,71 \cdot \frac{30}{25} < \mu < 198 + 1,71 \cdot \frac{30}{25} \implies 187,74 < \mu < 208,26$$

é o intervalo de 90% de confiança para μ , obtido com base na amostra. Então, espera-se que o intervalo calculado seja um dos 90% dos intervalos, para amostras casuais simples de 25, conterão o parâmetro μ . Ou seja, a média da população se encontra entre 187,74 mg/100ml de plasma e 208,26 mg/100ml de plasma com 90% de confiança.

5.3 INTERVALO DE CONFIANÇA para PROPORÇÃO

Já verificamos que $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$.

Padronizando a variável \hat{p} obteremos $Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$, onde $Z \sim N(0, 1)$

De maneira análoga ao intervalo de confiança encontrado para a média populacional (μ) quando a variância (σ^2) populacional é conhecida, teremos que:

$$P(-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}) = 1 - \alpha$$

$$\text{Logo, } P\left(-z_{\frac{\alpha}{2}} < \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\frac{\alpha}{2}}\right) = 1 - \alpha.$$

Rearranjando a expressão entre parênteses teremos:

$$P\left(\hat{p} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha$$

Como não conhecemos o valor p (proporção populacional) na expressão acima, para definirmos o intervalo de confiança devemos substituí-lo por seu estimador \hat{p} , temos:

$$\hat{p} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

que define o intervalo de confiança para o parâmetro p , ao nível de confiança $1 - \alpha$.

Exemplo: Entre 500 pessoas inquiridas a respeito de suas preferências eleitorais, 260 mostraram-se favoráveis ao candidato Y. Calcular um intervalo de confiança ao nível de 90% para a porcentagem dos eleitores favoráveis a Y.

Resolução: Temos $n = 500$; $x = 260$; $1 - \alpha = 90\%$

$$p = \frac{x}{n} = \frac{260}{500} = 0,52$$

$$P\left(0,52 - 1,64 \cdot \sqrt{\frac{0,52(1-0,52)}{500}} < p < 0,52 + 1,64 \cdot \sqrt{\frac{0,52(1-0,52)}{500}}\right) = 90\%$$

$$P\left(\hat{p} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1 - \alpha$$

$$P(0,488 < p < 0,552) = 90\%$$

6 TESTES DE HIPÓTESES

Um dos problemas a serem resolvidos pela Inferência Estatística é o de testar uma hipótese. Isto é, feito uma determinada afirmação sobre uma população, usualmente sobre um parâmetro desta, desejamos saber se os resultados de uma amostra contrariam ou não tal afirmação. Muitas vezes esta afirmação sobre a população é derivada de teorias desenvolvidas no campo substantivo do conhecimento. A adequação ou não dessa teoria ao universo real pode ser verificada ou refutada pela amostra. O objetivo do teste estatístico é fornecer ferramentas que nos permitam validar ou refutar uma hipótese (estatística) através dos resultados da amostra.

Por exemplo, de acordo com determinada literatura consultada pelos pesquisadores, a probabilidade de um indivíduo do sexo feminino contrair rubéola é 9%. Quer-se saber se esta mesma proporção é verificada entre as mulheres de Salvador. Para chegar a alguma conclusão seria necessário fazer um *teste de hipóteses* cujo procedimento veremos a seguir.

6.1 HIPÓTESES ESTATÍSTICAS

Na tomada de decisões é útil formular hipóteses ou suposições sobre a população em estudo. Tais hipóteses, que podem ou não ser verdadeiras, chamam-se *hipóteses estatísticas* e, em geral, consistem em afirmações sobre o valor de um parâmetro populacional.

São exemplos de hipóteses estatísticas:

1. A proporção de crianças do sexo masculino nascidas em determinada localidade nos últimos 5 anos é 0,5, isto é, $p = 0,5$.
2. A altura média dos brasileiros é 1,68 m, isto é, $\mu = 1,68$ m.

Vamos designar por H_0 (hipótese nula) a hipótese estatística principal, ou seja aquela que primeiro tentaremos validar no teste, e por H_1 (hipótese alternativa) a hipótese que contraria a hipótese nula.

Assim teremos;

1. $H_0 : \mu = 1,75 \implies$ hipótese nula
 $H_1 : \mu \neq 1,75 \implies$ hipótese alternativa
2. $H_0 : p = 0,10 \implies$ hipótese nula
 $H_1 : p \neq 0,10 \implies$ hipótese alternativa
3. $H_0 : \mu = 210 \implies$ hipótese nula
 $H_1 : \mu \neq 210 \implies$ hipótese alternativa

Através dos testes de hipóteses poderemos aceitar ou rejeitar a hipótese nula. A rejeição de H_0 implicará na aceitação de H_1 .

Usando a terminologia própria, diremos que o pesquisador pretende testar, por exemplo, $H_0 : p = 0,7$ contra $H_1 : p \neq 0,7$, com base em uma amostra de tamanho n .

6.2 ERROS

Qualquer que seja a decisão a ser tomada estamos sujeitos a cometer erros. Para facilitar a linguagem necessitamos das seguintes definições:

6.2.1 ERRO do TIPO I

Esse erro ocorre quando rejeitamos H_0 (hipótese nula) quando esta é verdadeira. Chamamos de α a probabilidade de cometer este erro, isto é:

$$\alpha = P(\text{erro do tipo I}) = P(\text{rejeitar } H_0 / H_0 \text{ é verdadeira})$$

α é denominado nível de significância do teste. É claro que esperamos um valor pequeno de α . Na prática, costuma-se adotar um nível de significância de 0,05 ou de 0,01, embora outros valores possam ser usados. Se, por exemplo, ao delinear um teste, escolhermos um nível de significância de 0,05 (ou 5%), isto significa que rejeitaríamos a hipótese nula em 5 chances dentre 100 quando ela deveria ser aceita, ou seja, haveria 95% de confiança que teríamos tomado a decisão correta. Em tal caso dizemos que a hipótese foi rejeitada ao nível de 5% de significância, o que significa que podemos ter errado com uma probabilidade de 5%.

6.2.2 ERRO do TIPO II

Definimos esse erro quando aceitamos H_0 (hipótese nula) quando esta é falsa. A probabilidade de cometer este erro é indicada por β , Logo,

$$\beta = P(\text{erro do tipo II}) = P(\text{não rejeitar } H_0 / H_0 \text{ é falsa}).$$

A determinação do valor de β já é mais difícil, pois usualmente não se especificam valores fixos para o parâmetro na situação alternativa. Este procedimento não será visto no momento e estará relacionado com o *poder do teste*.

A tabela a seguir resume as situações em que cometermos os erros do tipo I e tipo II:

Realidade	Decisão	
	Aceitar H_0	Rejeitar H_0
H_0 é verdadeira	Decisão correta	Erro tipo I
H_0 é falsa	Erro tipo II	Decisão correta

6.3 TESTE DE HIPÓTESE para uma MÉDIA

Este teste é feito baseando-se nas distribuições da média amostral \bar{X} . Veremos dois testes para a média, que dependerão do conhecimento ou não do verdadeiro valor da variância.

6.3.1 Teste para a Média quando σ^2 é Conhecida (e quando a amostra é grande)

O procedimento para a realização desse teste de hipóteses pode ser resumido nos seguintes passos:

Assim teremos;

1. Enumerar as hipóteses:

$$H_0 : \mu = \mu_0$$

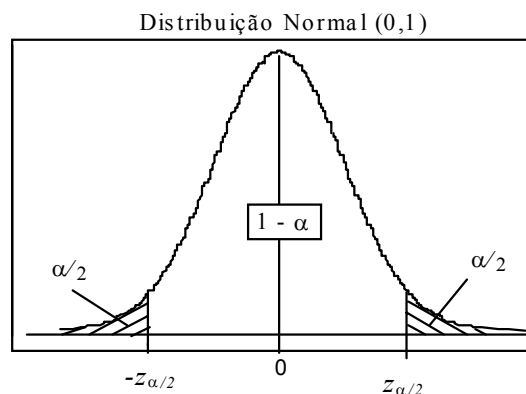
$$H_1 : \mu \neq \mu_0$$

2. Determinar a distribuição de \bar{X} . e, por conseguinte, a estatística de teste.

$$\text{Já vimos que } \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \Rightarrow \quad Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

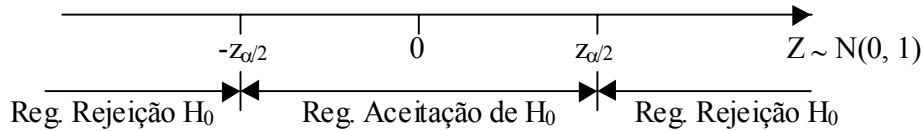
A estatística de teste é uma estatística amostral, usada para tomar uma decisão em relação à hipótese nula.

3. Fixar o nível de significância α .
4. Determinar a *Região de Aceitação* e a *Região de Rejeição* (crítica) que serão definidas de acordo com o α fixado:



A região crítica é o conjunto de todos os valores da estatística de teste que levam à rejeição da hipótese nula. Com base no gráfico acima, iremos formular a seguinte regra de decisão ou teste de hipótese ou de significância:

- Rejeitar a hipótese nula ao nível de significância α se o valor de Z for maior que o valor de $+z_{\frac{\alpha}{2}}$ ou menor que o valor $-z_{\frac{\alpha}{2}}$.
- Em caso contrário, não rejeitar a hipótese nula.



O valor $\pm z_{\frac{\alpha}{2}}$ é chamado valor crítico, separando a região crítica dos valores da estatística de teste que não levam à rejeição da hipótese nula.

Exemplo:

Em indivíduos sadios, a taxa de fósforo no sangue tem distribuição aproximadamente normal com média $\mu = 3mg/100cc$ e desvio-padrão $\sigma = 0,6mg/100cc$. Com o objetivo de saber se no artrismo essa taxa média era alterada, um pesquisador tomou uma amostra de 36 doentes e testou sua hipótese ao nível de 5% de significância; observou-se na amostra a média $\bar{X} = 3,12mg/100cc$.

Resolução:

Os hipóteses serão as seguintes:

$$H_0 : \mu = 3,0$$

$$1. H_1 : \mu \neq 3,0$$

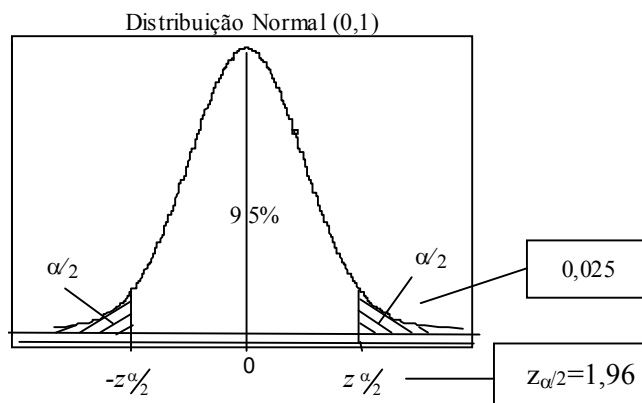
Usaremos $\alpha = 0,05$

$$X \sim N(3; (0,6)^2/36) \implies X \sim N(3; 0,01)$$

A estatística a ser utilizada será:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{3,12 - 3,0}{0,1} = 1,2$$

As regiões de aceitação e de rejeição serão as seguintes: Para $\alpha = 0,05$



Os pontos $-z_{\frac{\alpha}{2}}$ e $z_{\frac{\alpha}{2}}$ serão fornecidos pela tabela da faixa central de maneira idêntica à qual foi observado para construção de intervalos de confiança. Logo $z_{0,025} = 1,96$ e $-z_{0,025} = -1,96$. Estes pontos serão chamados de pontos críticos. A área *hachurada* corresponde à região de rejeição da hipótese nula.

Se a estatística $Z > z_{\frac{\alpha}{2}}$ ou $Z < -z_{\frac{\alpha}{2}}$, então rejeitaremos a hipótese nula. Caso contrário, aceitaremos-a.

Como $Z = 1,2 < z_{0,025} = 1,96 \implies$ não rejeitamos a hipótese de que a taxa de fósforo no sangue de pessoas com artrite seja igual a $3,0\text{mg}/100\text{cc}$.

Então não há evidência amostral suficiente para rejeitamos a hipótese ao nível de significância de 5%.

6.3.2 Teste para Média quando σ^2 é Desconhecida (e para pequenas amostras)

Neste caso, como não conhecemos σ^2 , precisamos calcular a estimativa S^2 a partir de uma amostra.

Seguindo os mesmos passos do teste anterior para a construção das hipóteses a serem analisadas; teremos como única diferença o fato de agora a estatística a ser utilizada ser:

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

T tem distribuição “t” de Student com $n - 1$ graus de liberdade.

1. De forma similar ao teste com base da distribuição Normal, agora teremos que:

- (a) Se $-t_{\frac{\alpha}{2}} < t < t_{\frac{\alpha}{2}}$, aceita-se H_0
- (b) Se $t > t_{\frac{\alpha}{2}}$, rejeita-se H_0
- (c) Se $t < -t_{\frac{\alpha}{2}}$, rejeita-se H_0 .

Exemplo: Os registros dos últimos anos de um colégio atestam para os calouros admitidos a nota média 115 (teste vocacional). Para testar a hipótese de que a média de uma nova turma é a mesma, tirou-se, ao acaso, uma amostra de vinte notas, obtendo-se média de 118 e desvio-padrão 20. Admitir que $\alpha = 0,05$ para efetuar o teste.

Resolução:

$$H_0 : \mu = 115$$

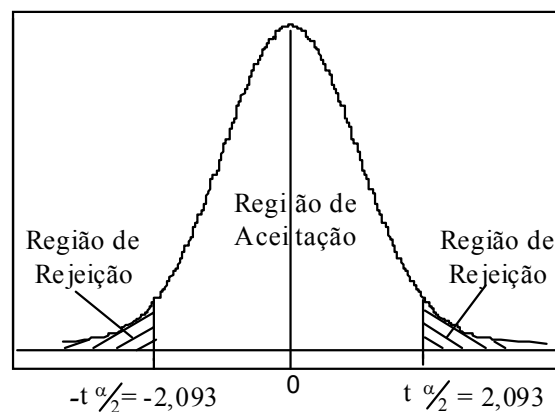
$$H_1 : \mu \neq 115$$

$$\alpha = 0,05; n = 20$$

Os graus de liberdade (g.l.) para o teste “t” de Student será obtido da seguinte forma

$$g.l. = n - 1 = 20 - 1 = 19.$$

Verificando-se na tabela da “t” de Student com 19 graus de liberdade e $\alpha = 0,05$, obteremos os seguintes pontos $-t_{\frac{\alpha}{2}} = -t_{0,025} = -2,093$ e $t_{\frac{\alpha}{2}} = t_{0,025} = 2,093$; que irão nos fornecer as seguintes áreas:



$$\text{A estatística } T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{118 - 115}{\frac{20}{\sqrt{20}}} = \frac{3}{1,4721} = 0,6708$$

Como $-2,093 < t < 2,093 \implies$ não rejeita-se a hipótese nula ao nível de significância de 5%.

6.4 TESTE DE HIPÓTESE para a PROPORÇÃO

Este tipo de teste será realizado quando temos uma população e uma hipótese sobre a proporção de indivíduos portadores de uma certa característica. Esta hipótese afirma que essa proporção é igual a um certo número p_0 . Então, nossas hipóteses poderão ser descritas como:

$$H_0 : p = p_0$$

$$H_1 : p \neq p_0$$

Como já vimos, a estatística \hat{p} , proporção da amostra, tem uma distribuição normal, isto é: $\hat{p} \sim N \left(p, \frac{p(1-p)}{n} \right)$.

Fixado um valor α , devemos construir a região crítica para p na suposição de que os parâmetros definidos em H_0 sejam verdadeiros.

A estatística a ser usada neste teste será:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}, \text{ onde } p \text{ será a proporção na população e } \hat{p} \text{ a proporção na amostra.}$$

As nossas regiões de aceitação (RA) e rejeição (RR) serão dadas de maneira similar às definidas para a média.

1. Daí, poderemos concluir mais uma vez que:

- (a) Se $-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}$, aceita-se H_0
- (b) Se $Z < -z_{\frac{\alpha}{2}}$, rejeita-se H_0 .
- (c) Se $Z > z_{\frac{\alpha}{2}}$, rejeita-se H_0

Exemplo: As condições de mortalidade de uma região são tais que a proporção de nascidos que sobrevivem até 60 anos é de 0,6. Testar essa hipótese ao nível de 5% se em 1000 nascimentos amostrados aleatoriamente, verificou-se 530 sobreviventes até 60 anos.

Resolução:

Queremos testar:

$$H_0 : p = 0,6$$

$$H_1 : p \neq 0,6$$

$\alpha = 0,05$ será nosso nível de significância.

$$n = 1000; \quad p = \frac{530}{1000} = 0,53$$

$$\text{Como sabemos, } \hat{p} \sim N \left(p, \frac{p(1-p)}{n} \right) \implies \hat{p} \sim N(0,6; 0,00024)$$

Se $\alpha = 0,05$, os pontos críticos $-z_{\frac{\alpha}{2}}$ e $z_{\frac{\alpha}{2}}$ serão os pontos $\pm 1,96$.

Calculando a estatística do teste, teremos:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0,53 - 0,60}{\sqrt{0,00024}} = -4,52$$

Como $Z < -1,96$, rejeita-se H_0 ao nível de 5% de significância. Então, poderemos concluir que há evidências de que a proporção de nascidos vivos que sobrevivem até os 60 anos de idade é diferente de 0,60.

6.5 Teste de Hipótese para Comparação entre Médias de duas Populações.

É um dos testes mais utilizados por pesquisadores da área biomédica.

São bastante comuns as ocasiões em que se têm duas populações e se deseja comparar suas médias. Este tipo de teste é aplicado para verificação de existência de *diferença significativa* entre as duas médias populacionais. E nestas circunstâncias iremos utilizar os testes descritos a seguir.

6.5.1 Comparação entre Médias de duas Populações

Quando as variâncias populacionais σ_1^2 e σ_2^2 são Conhecidas.

Vamos supor que temos uma amostra X_1, X_2, \dots, X_{n_1} da população $N(\mu_1, \sigma_1^2)$ e uma amostra Y_1, Y_2, \dots, Y_{n_2} da população $N(\mu_2, \sigma_2^2)$, independentes. Queremos saber se suas médias diferem significativamente. Neste caso, as hipóteses estatísticas são:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Como σ_1^2 e σ_2^2 são conhecidos, têm-se que a distribuição amostral de $(\bar{X} - \bar{Y})$ é dada por:

$$(\bar{X} - \bar{Y}) \sim N\left(\mu_1 - \mu_2; \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right);$$

Se fixarmos α como o nível de significância do teste, então poderemos obter os pontos críticos $-z_{\frac{\alpha}{2}}$ e $z_{\frac{\alpha}{2}}$, respectivamente; definindo desta forma as áreas de aceitação e rejeição da hipótese nula.

Para procedermos com o teste, devemos padronizar a variável aleatória $(\bar{X} - \bar{Y})$, que resultará na seguinte estatística:

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

Sob a hipótese nula $\mu_1 = \mu_2$, logo:

$$Z_{H_0} = \frac{(\bar{X} - \bar{Y})}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

irá nos fornecer o valor que nos permitirá aceitar ou rejeitar a hipótese nula de forma similar aos testes anteriores.

Exemplo: Um estudo comparativo do tempo de vida médio de um doente após ter sido detectado o vírus Y no organismo humano foi realizado entre homens e mulheres, para os quais temos as seguintes distribuições, $N(3, 2; 0, 64)$ e $N(3, 7; 0, 81)$. Deseja-se saber se o tempo de vida médio de pacientes com este tipo de doença é o mesmo entre homens e mulheres. Para isto, estudou-se uma amostra de 50 homens e 50 mulheres, obtendo-se um tempo médio 3,0 e 3,5 anos, respectivamente.

Resolução: Queremos testar:

$$H_0 : \mu_1 = \mu_2 = 0$$

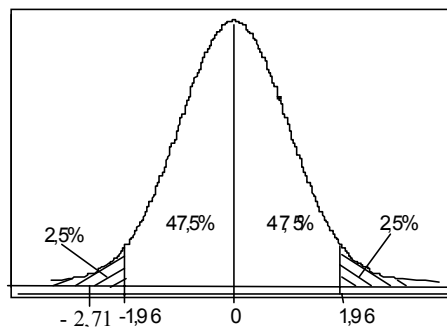
$$H_1 : \mu_1 \neq \mu_2 \neq 0$$

Vamos fixar $\alpha = 0,05$, logo os pontos críticos serão 1,96 e $-1,96$, respectivamente.

Temos a seguinte distribuição amostral para $(\bar{X}_1 - \bar{X}_2)$:

$$(\bar{X}_1 - \bar{X}_2) \sim N(-0,5; 0,034)$$

Calculando o valor da estatística sob H_0 , tem-se: $Z = \frac{-0,5}{0,034} = -2,7116$



Como $Z_{calculado} = -2,71 < -1,96 = Z_{tabelado}$, conforme figura acima, então rejeitamos H_0 , ao nível de 5% de significância. Logo, poderemos afirmar que existem indícios que o tempo de vida médio entre homens e mulheres que contraíram o vírus Y não é igual. Ou seja, os tempos médios de vida diferem significativamente entre sexo, ao nível de 5%.

6.5.2 Comparação entre Médias de duas Populações

Quando as variâncias populacionais σ_1^2 e σ_2^2 Desconhecidas.

Este teste de hipóteses é conhecido como teste “*t*” de Student para diferença de médias.

Quando as variâncias das populações não são conhecidas, caso muito comum na prática, há necessidade de estimá-las a partir dos dados amostrais. Neste caso, devemos substituir σ_1^2 e σ_2^2 pelos seus estimadores S_1^2 e S_2^2 , obtendo a seguinte estatística:

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_c \sqrt{\frac{n_1 + n_2}{n_1 \cdot n_2}}}$$

$$\text{onde } S_c = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

Dizemos que T tem distribuição “*t*” de Student com $n_1 + n_2 - 2$ graus de liberdade.

O teste de hipóteses deve ser conduzido de maneira análoga, lembrando-se da modificação dos valores críticos de *t*, que vão ter outro número de graus de liberdade.

Exemplo: Desejamos testar se dois tipos de ensino profissional são igualmente eficazes. Para isso, sortearam-se duas amostras de operários; a cada uma, deu-se um dos tipos de treinamento e, no final, submeteram-se os dois grupos a um mesmo teste. Que tipo de conclusão você poderia tirar, baseando-se nos resultados abaixo.

Amostra	N. de elementos	Média	Desvio-Padrão
Tipo I	12	75	5
Tipo II	10	74	10

Resolução:

As hipóteses serão:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Vamos fixar $\alpha = 0,05$.

Calcularemos a estatística:

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_c \sqrt{\frac{n_1 + n_2}{n_1 \cdot n_2}}}$$

$$\text{onde } S_c = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

Sob $H_0 \Rightarrow \mu_1 = \mu_2$

Então:

$$S_c = 7,88954 \quad \text{e} \quad T = \frac{(75 - 74)}{7,88954 \cdot 0,026846} = \frac{1}{1,2927}$$

$$T = 0,7735$$

Iremos definir os pontos críticos $-t_{\frac{\alpha}{2}}$ e $t_{\frac{\alpha}{2}}$, que serão, respectivamente, $-2,20$ e $2,20$.

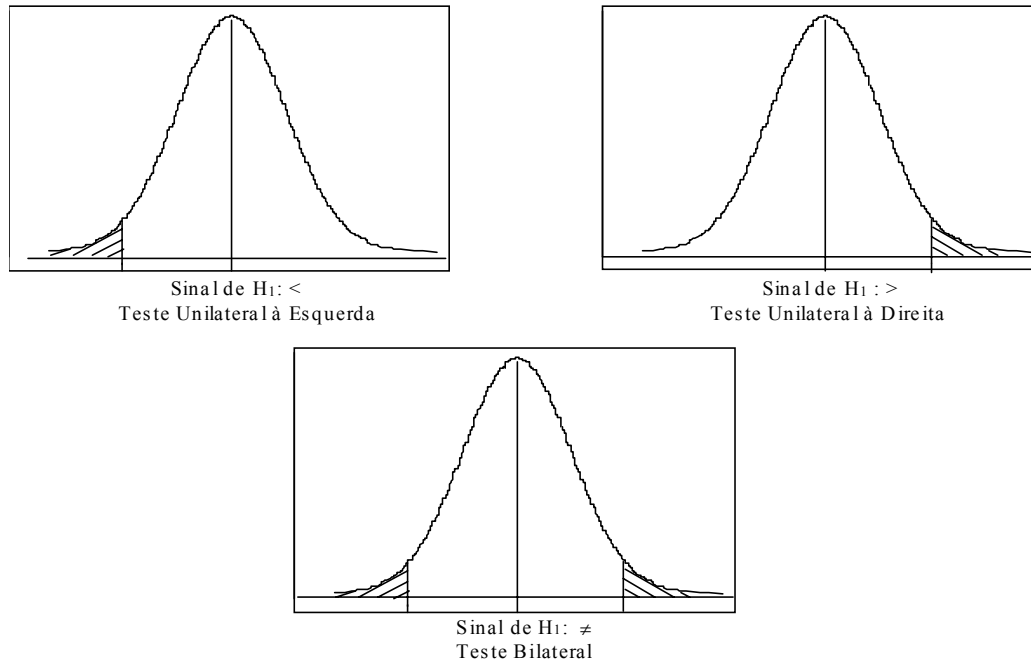
Como $-2,20 < T < 2,20$, aceitamos a hipótese nula. Com base no teste acima, pode-se concluir que os dois tipos de testes não diferem significativamente, ao nível de 5%. Logo, qualquer um dos dois tipos de ensino profissional são igualmente eficazes.

6.6 TESTES BILATERAIS e UNILATERAIS.

As caudas de uma distribuição são as regiões extremas delimitadas por valores críticos. Nos exemplos desenvolvidos nesta apostila os testes de hipótese envolveram testes bilaterais no sentido de que a região crítica está situada nas duas regiões extremas (*caudas*) sob a curva. Rejeitamos a hipótese nula se nossa estatística de teste está na região crítica porque isto indica uma discrepância significativa entre a hipótese nula e os dados amostrais.

Alguns testes são unilaterais, podendo ser unilaterais esquerdos ou direitos. Os testes unilaterais esquerdo têm a região crítica localizada na região extrema esquerda sob a curva, enquanto que os unilaterais direitos têm a região crítica localizada na região extrema direita sob a curva.

Nos *testes bilaterais*, o nível de significância é dividido igualmente entre as duas caudas que constituem a região crítica. Em *testes unilaterais*, a área crítica é igual a α . A figura a seguir esquematiza estas informações.



Exemplo: Determinação de valores críticos:

Muitos passageiros de navios de cruzeiro utilizam adesivos que fornecem dramamina ao corpo a fim de evitar o enjôo. Testa-se a informação sobre a quantidade da dosagem média, ao nível de significância de 0,05. As condições são tais que permitem a utilização da distribuição normal.

1. Determine o(s) valor(e)s crítico(s) de z se o teste é:
 - (a) bilateral;
 - (b) unilateral esquerdo;
 - (c) unilateral direito.

Resolução:

- (a) Em um *teste bilateral*, o nível de significância é dividido igualmente entre as duas caudas, o que determina uma área de 0,025 em cada cauda. Podemos encontrar os valores críticos pelo uso da tabela da normal padrão através dos valores que correspondem a áreas de 0,475 à direita ou à esquerda de média. Assim, obteremos os valores críticos $z = -1,96$ e $z = 1,96$.

- (b) Em um *teste unilateral esquerdo*, o nível de significância 0,05 é a área da região crítica à esquerda, de forma que o valor crítico corresponde a uma área de 0,45 à esquerda da média. Recorrendo à tabela da normal padrão, obtemos o valor crítico $z = -1,645$.
- (c) Em um *teste unilateral à direita*, o nível de significância 0,05 a área da região crítica à direita, de forma que o valor crítico corresponde a uma área de 0,45 à direita da média. Com o uso da tabela normal padrão, obtém-se o valor crítico $z = 1,645$.

6.7 NÍVEL DESCRITO (p-valor).

O procedimento descrito até o momento é conhecido como procedimento clássico de testes de hipóteses. Um outro procedimento que vem sendo muito adotado consiste em apresentar o nível descritivo (**ou p-valor**) do teste. A diferença básica entre esses dois procedimentos é que, neste último, não é necessário construir a região crítica. Em lugar disto, indica-se a probabilidade de ocorrerem valores da estatística mais extremos que o observado, supondo a hipótese H_0 verdadeira.

Regra de decisão

p-valor $> \alpha \implies$ aceitar H_0

p-valor $\leq \alpha \implies$ rejeitar H_0

Então, o p-valor é a probabilidade, sob a hipótese nula, de se observar um valor igual ou maior que o obtido. Ou seja, um p-valor (ou valor de probabilidade) é a probabilidade de obter um valor da estatística amostral de teste no *mínimo tão extremo* como o que resulta dos dados amostrais, na suposição de a hipótese nula ser verdadeira.

Este tipo de procedimento é usualmente o adotado quando se trabalha com *softwares estatísticos* para análise de dados, pois a informação a respeito do **p-valor** está disponível dentre os resultados apresentados pelo computador.

Com o objetivo de realizar o teste de hipóteses através desse procedimento, iremos apresentar dois exemplos. Para auxiliar nessa análise usaremos alguns resultados apresentados por um pacote estatístico para análise de dados denominado **Minitab**.

Exemplo 1: Uma companhia de cigarros anuncia que o índice médio de nicotina dos cigarros que fabrica, apresenta-se abaixo de 23 mg por cigarro. Um laboratório realiza 10 análises desse índice, obtendo: 27, 24, 21, 25, 26, 22, 23, 25, 27, 22. Sabe-se que o índice de nicotina se distribui normalmente. Pode-se aceitar, ao nível de 10%, a afirmação do fabricante?

$$H_0 : \mu = 23mg \quad \text{vs} \quad H_1 : \mu \neq 23mg \quad (\text{suponha um teste bilateral})$$

$$\text{A Estatística do teste é } T = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} = \frac{24,2 - 23}{\frac{2,1429}{\sqrt{10}}} = 1,77$$

O Valor da tabela que servirá de base para comparar com a estatística do teste é $-t_\alpha(n-1) = -t_{0,10,9} = -1,383$, logo não podemos rejeitar H_0 ao nível de significância de 10%.

O **Minitab** utiliza como critério o nível descritivo do teste:

Test of $\mu = 23.000$ vs $\mu \neq 23.000$						
Variable	N	Mean	StDev	SE Mean	T	P-Value
C1	10	24.200	2.150	0.680	1.77	0.94

Observa-se que o **p-valor** $> 0,10$ e, portanto, aceitamos a hipótese nula. Logo, os dados amostrais não possuem evidência suficiente para aceitarmos a afirmação do fabricante.

Exemplo 2: Uma rede de pizzarias deseja testar se o teor médio de gordura em peças de salame produzidas por determinada indústria de alimentos é igual a 15%. De um grande lote que adquiriu retirou uma amostra de 50 peças de salame e os resultados estão a seguir:

19.8 23.4 13.6 6.6 13.7 5.2 14.3
 13.3 12.2 14.3 8.5 15.8 16.0 18.3
 28.7 11.6 16.4 14.4 26.2 17.0 6.5
 10.0 24.5 34.9 19.1 6.9 19.5 11.0
 8.9 10.6 9.5 14.0 6.0 18.0 10.8
 16.7 18.4 10.1 12.3 6.5 25.4 15.3
 12.1 13.1 7.7 17.4 10.7 24.1 14.0
 21.4

As hipóteses a serem testadas são

$$H_0 : \mu = 15 \text{ versus } H_1 : \mu \neq 15$$

Test of $\mu = 15.000$ vs $\mu \neq 15.000$						
Variable	N	Mean	StDev	SE Mean	T	P-Value
C2	50	14.984	6.387	0.903	-0.12	0.91

Como **p-valor** $> 0,05$ então aceita-se H_0 , ou seja, não existe evidência de que o teor de gordura nas peças de salame produzidas pela indústria seja diferente de 15%

7 TESTE QUI-QUADRADO

Até o momento nós testamos hipóteses referentes a um parâmetro populacional ou mesmo à comparação de dois parâmetros. Vamos estudar agora um teste não-paramétrico, assim chamado por não depender dos parâmetros populacionais nem de suas respectivas estimativas.

Para a realização desses testes não-paramétricos nós utilizaremos uma nova distribuição estatística chamada Qui-quadrado, cuja medida baseia-se no grau de concordância entre as frequências observadas e as frequências esperadas de acordo com as regras de probabilidade.

Suponha um experimento realizado n vezes, onde se observou um conjunto A_1, A_2, \dots, A_k com as respectivas frequências O_1, O_2, \dots, O_k . No entanto, de acordo com as regras de probabilidade deveriam apresentar as seguintes frequências teóricas ou esperadas e_1, e_2, \dots, e_k .

Como ilustração, considere o seguinte experimento: Lançamento de um dado 60 vezes. Neste tipo de experimento temos 6 possíveis eventos (as faces do dado). E sabemos que se o dado é “honesto” a probabilidade de cada face ser sorteada é $\frac{1}{6}$ em cada jogada.

Os resultados deste experimento aparecem no quadro a seguir.

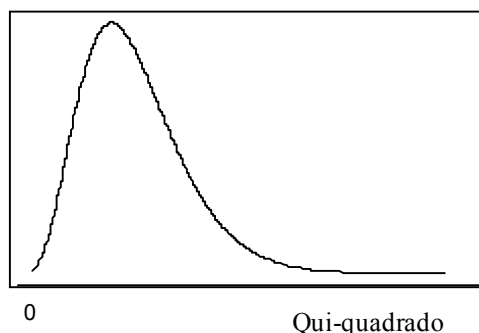
Eventos	1	2	3	4	5	6
Freq. Observada	15	7	4	11	6	17
Freq. esperada	10	10	10	10	10	10

O modelo que serve para determinar as discrepâncias existentes entre as frequências observadas e esperadas é dado pela estatística:

$$\chi^2 = \sum \frac{(O_j - e_j)^2}{e_j} = \frac{(O_1 - e_1)^2}{e_1} + \frac{(O_2 - e_2)^2}{e_2} + \dots + \frac{(O_k - e_k)^2}{e_k}$$

onde O_j é a frequência observada e e_j é a frequência esperada, sob a hipótese em estudo.

A distribuição amostral da distribuição Qui-quadrado χ^2 terá a seguinte forma aproximada:



A χ^2 , assim como a “t” de Student, terá graus de liberdade associados à sua distribuição, que serão definidos de acordo com o procedimento a ser adotado.

7.1 TESTE DE INDEPENDÊNCIA

Uma importante aplicação do teste χ^2 ocorre quando queremos estudar a relação entre duas ou mais variáveis de classificação. A representação das freqüências observadas, neste caso, pode ser feita por meio de uma tabela de contingência. Considerando-se dois critérios de classificação, teremos tabelas de dupla entrada de classificação $h \times k$, em que as freqüências observadas ocupam h linhas e k colunas.

Tabela de contingência (2 x 2)

Resultados observados			
Classificação	I	II	TOTAL
A	A_1	A_2	N_a
B	B_1	B_2	N_b
TOTAL	N_1	N_2	N

Na tabela acima consideramos duas variáveis. A primeira variável (variável 1) admitimos que seja classificada em duas categorias A e B; a variável também em duas categoria I e II.

Uma tabela de contingência é usualmente construída com o propósito de estudar a relação entre as duas variáveis de classificação. Por meio do teste χ^2 é possível testar a hipótese de que as duas variáveis são independentes.

No caso do teste de independência, em particular, o valor da estatística χ^2 será maior quanto mais afastado o valor observado for do valor esperado e maior será a dependência. Em relação ao número de graus de liberdade, este será definido por $(h - 1)(k - 1)$, onde h representa o número de modalidades da variável 1 e k o número de modalidades da variável 2.

Como o pesquisador está preocupado em testar a associação entre as variáveis, as hipóteses serão formuladas da seguinte maneira:

H_0 : as variáveis são independentes

H_1 : as variáveis não são independentes, ou seja, elas apresentam algum grau de associação entre si.

Exemplo: Deseja-se verificar a relação entre a classificação dada ao ajustamento matrimonial e ao nível de escolaridade do casal. Os dados encontram-se na tabela a seguir:

Educação	Ajustamento Matrimonial				TOTAL
	Muito baixo	Baixo	Alto	Muito alto	
3º grau	18 (27)	29 (39)	70 (64)	115 (102)	232
2º grau	17 (13)	28 (19)	30 (32)	41 (51)	116
1º grau	11 (6)	10 (9)	11 (14)	20 (23)	52
TOTAL	46	67	111	176	400

Neste caso queremos testar a hipótese de não haver relação entre o nível educacional do indivíduo e o êxito no seu casamento.

Para podermos calcular as frequências esperadas para as diversas caselas, sob a suposição de independência, estaremos esperando que não haja diferença entre as modalidades, logo poderemos usar os totais marginais: 232, 116 e 52. Assim, encontraremos a distribuição de escolaridade sem levar em consideração o ajustamento matrimonial, de forma que:

$$\frac{232}{400} = 0,58 \quad ; \quad \frac{116}{400} = 0,29 \quad ; \quad \frac{52}{400} = 0,13$$

Aplicaremos estas taxas em todas as caselas da tabela anterior, obtendo as frequências esperadas, que estão representadas dentro dos parênteses da tabela anterior.

Nosso teste procederá do seguinte modo:

H_0 : O ajustamento matrimonial é independente da educação

H_1 : O ajustamento matrimonial não é independente da educação

$\alpha = 5\%$

$\varphi =$ número de graus de liberdade $= (h - 1)(k - 1) = (3 - 1)(4 - 1) = 2 \times 3 = 6 \implies \chi^2 = 12,6$

A estatística do teste, sob H_0 , será:

$$\chi^2 = \sum \frac{(O_j - e_j)^2}{e_j} = \frac{(18 - 27)^2}{27} + \frac{(17 - 13)^2}{13} + \dots + \frac{(20 - 23)^2}{23} = 20,7$$

Como $\chi^2_{calculado} > 12,6$, rejeita-se a hipótese de independência.

Por inspeção da tabela parece que os indivíduos com alguma educação colegial parecem ajustar-se melhor ao casamento do que aqueles que não a possuem.

8 EXERCÍCIOS DE FIXAÇÃO

1) Um antropólogo mediu as alturas (em polegadas) de uma amostra aleatória de 100 homens de determinada população, encontrando a média amostral de 71,3. Se a variância da população for $\sigma^2 = 9$, determine um intervalo de 99% de confiança para a altura média de toda a população.

2) Para os tempos de reação de 30 motoristas selecionados aleatoriamente, encontrou-se uma média de 0,83 segundos e um desvio-padrão de 0,20 segundos. Determinar um intervalo de 95% de confiança para o tempo médio de reação de toda a população de motoristas.

3) Uma amostra aleatória de 400 domicílios mostra que 25% deles são casas de aluguel. Qual é o intervalo de confiança que podemos razoavelmente supor que seja o da proporção de casas de aluguel, com um nível de significância de 2%?

4) Certo tipo de semente cresce, em média, até a altura de 8,5 polegadas, com desvio-padrão de uma polegada. Semeiam-se 100 delas em um solo enriquecido, a fim de testar qual a média na melhora de crescimento. Se nesta amostra, encontrarmos uma altura média de 8,8 polegadas de altura, qual será nossa conclusão?

5) Estamos desconfiados de que a média das receitas municipais per capita das cidades pequenas (até 20.000 habitantes) é maior do que a das receitas do estado, que é de 1.229 unidades. Para comprovar ou não esta hipótese, sorteamos dez cidades pequenas e obtivemos os seguintes resultados: 1.230; 582; 576; 2.093; 2.621; 1.045; 1.439; 717; 1.838; 1.359. O relatório do software **Minitab** encontra-se a seguir.

Test of $\mu = 1229$ vs $\mu > 1229$						
Variable	N	Mean	StDev	SE Mean	T	P-Value
Renda	10	1350	676	214	0.57	0.29

Com base nestas informações, que conclusões podem ser feitas. Construa as regiões críticas para este teste de hipóteses e compare este procedimento com o do uso do p-valor.

6) Duas máquinas, A e B, são usadas para empacotar pó de café. A experiência passada garante que o desvio padrão de ambas é de 10g. Porém, suspeita-se de que elas têm médias diferentes. Para verificar, sortearam-se duas amostras: uma com 25 pacotes da máquina A e outra com 16 pacotes da máquina B. As médias foram, respectivamente, $\bar{X}_A = 502,74g$ e $\bar{X}_B = 490,60g$. Com estes números, e ao nível de 5%, qual seria a conclusão do teste $H_0 : \mu_A = \mu_B$?

7) Suponha que está sendo realizada uma pesquisa para avaliar diversas características físicas e comportamentais dos ursos polares. Alguns pesquisadores acreditam que o peso nesse grupo de animais não difere segundo o sexo. Com o objetivo de se verificar se esta afirmação é verdadeira, 99 ursos do sexo masculino e 44 do sexo feminino foram pesados. Os resultados obtidos encontram-se a seguir. Com base nesses resultados, que conclusões você pode obter?

Twosample T for Weight				
Sex	N	Mean	StDev	SE Mean
1	99	214	120	12
2	44	143.0	64.5	9.7
95% C.I. for $\mu_1 - \mu_2$: (40, 101.5)				
T - Test $\mu_1 = \mu_2$ (vs not =): T = 4.59 P = 0.0000 DF = 135				

8) Fez-se um estudo de 531 pessoas feridas em acidentes de bicicleta; os resultados de uma amostra aleatória constam da tabela abaixo. Com o nível de 0,05 de significância, teste a afirmação de que o uso do capacete não reduz a possibilidade de ferimentos no rosto. Com base nos resultados, acha que o uso do capacete ajuda a reduzir o risco de ferimentos no rosto?

Resultados	Com capacete	Sem capacete
Com ferimentos faciais	30	182
Todos os ferimentos não faciais	83	236

Fonte: “A Case-control study of the Effectiveness of bicycle safety helmets in preventing facial injury”, de Thompson, Thompson, Rivara e Wolf, American Journal of Public Health, vo. 80, no. 12.

9) Um estudo feito para determinar a taxa de fumantes entre pessoas de diferentes grupos etários originou os dados amostrais aleatórios resumidos na tabela a seguir. Com o nível de 0,05 de significância, teste a afirmação de que o fumo é independente do grupo etário. Com base nesses dados, tem sentido dirigir a propaganda de cigarros a grupos etários específicos?

Hábito de fumar	20-24 anos	25-34 anos	35-44 anos	45-64 anos
Fumante	18	15	17	15
Não fumante	32	35	33	35
TOTAL	50	50	50	50

Fonte: Baseado em dados do Centro Nacional de Estatísticas em Saúde (E.U.A)