Estatística Aplicada à Administração II

Marcelo Menezes Reis

Catalogação na publicação por: Onélia Silva Guimarães CRB-14/071

Sumário

Apresentação

Caro estudante!

Você já cursou com aproveitamento a disciplina de Estatística Aplicada à Administração I. Todos os conceitos lá estudados serão importantes para Estatística Aplicada à Administração II, especialmente os da Unidade 6 – Probabilidade.

Conforme mencionado anteriormente os métodos estatísticos são ferramentas primordiais para o administrador de qualquer organização, pois possibilitam obter informações confiáveis, sem as quais a tomada de decisões seria mais difícil ou mesmo impossível. E, não se esqueça, a essência de administrar é tomar decisões. Por este motivo, esta disciplina faz parte do currículo do curso de Administração.

Nesta disciplina você aprenderá como, a partir de dados confiáveis (conceitos de planejamento de pesquisa estatística e amostragem), resumidos e organizados pelas técnicas de análise exploratória de dados vistas na primeira disciplina,aplicar técnicas apropriadas (probabilidade aplicada e inferência estatística) para generalizar os resultados encontrados, que por sua vez serão usados para tomar decisões. Procurei apresentar exemplos concretos de aplicação, usando ferramentas computacionais simples (como as planilhas eletrônicas, com as quais você teve um primeiro contato na disciplina de Informática Básica). O domínio dos métodos estatísticos dará a você um grande diferencial, pois permitirá tomar melhores decisões, o que, em essência, é o objetivo primordial de qualquer organização.

Sucesso em sua caminhada.

Prof. Marcelo Menezes Reis

Unidade 1

Variáveis aleatórias

Objetivo

Nesta Unidade você vai compreender o conceito de variável aleatória e seu relacionamento com os modelos probabilísticos. Vai aprender também que os modelos probabilísticos podem ser construídos para as variáveis aleatórias.

1.1 -Definição de variável aleatória: discreta e contínua.

Caro estudante!

Uma pergunta que é normalmente feita a todos que trabalham com ciências exatas: "por que a obsessão em reduzir tudo a números"? Vimos em Análise Exploratória de Dados que uma variável quantitativa geralmente, porque nem tudo pode ser reduzido a números, como a inteligência e criatividade, apresenta mais informação que uma variável qualitativa, pode ser resumida não somente através de tabelas e gráficos mas também através de medidas de síntese.

Nos exemplos sobre probabilidade apresentados na Unidade 5 os eventos foram geralmente definidos de forma verbal: bolas da mesma cor, 2 bolas vermelhas, soma das faces menor ou igual a 5, etc. Não haveria problema em definir os eventos através de números. Bastaria associar aos resultados do Espaço Amostral números, através de uma função.

Esta função é chamada de Variável Aleatória. Os modelos probabilísticos podem então ser construídos para as variáveis aleatórias. O administrador precisa conhecer estes conceitos porque eles proporcionam maior objetividade na obtenção das probabilidades, o que torna o processo de tomada de decisões mais seguro. Vamos conhecer esses conceitos nesta Unidade?

Uma definição inicial de Variável Aleatória poderia ser: trata-se de uma "variável quantitativa, cujo resultado (valor) depende de fatores aleatórios".

Formalmente, **Variável Aleatória** é uma função matemática que associa números reais (contradomínio da função) aos resultados de um Espaço AmostralGLOSSÁRIO - Espaço Amostral é o conjunto de todos os resultados possíveis de um experimento aleatório. Fonte: Barbetta, Reis e Bornia, 2010. Fim GLOSSÁRIO(domínio da função), por sua vez vinculado a um Experimento Aleatório. Se o Espaço Amostral for finito ou infinito

numerável a variável aleatória é dita discreta. Se o Espaço Amostral for infinito a variável aleatória é dita contínua.

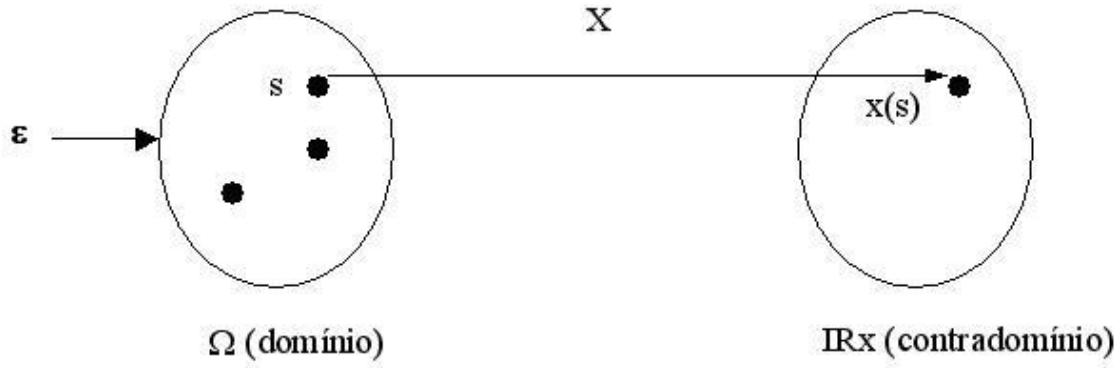


Figura 1 - Variável aleatória

Fonte: elaborada pelo autor

Por exemplo, imaginemos o Experimento AleatórioGLOSSÁRIO Experimento Aleatório é um processo de obtenção de um resultado ou medida que apresenta as seguintes características: não se pode afirmar, antes de realizar o experimento, qual será o resultado de uma realização, mas é possível determinar o conjunto de resultados possíveis; quando é realizado um grande número de vezes (replicado) apresentará uma regularidade que permitirá construir um modelo probabilístico para analisar o experimento. Fonte: adaptado pelo autor de Lopes, 1999. Fim GLOSSÁRIOjogar uma moeda honesta duas vezes e observar a face voltada para cima. O Espaço Amostral seria finito:

$$\Omega = \{CaraCara; CaraCoroa; CoroaCara; CoroaCoroa\}$$

Se houvesse interesse no número de caras obtidas, poderia ser definida uma variável aleatória discreta X, onde X = Número de caras em dois lançamentos. Os valores possíveis de X seriam:

$$X = \{0, 1, 2\}$$

O valor 0 é associado ao evento CoroaCoroa, o valor 1 é associado aos eventos CaraCoroa e CoroaCara, e o valor 2 é associado ao evento CaraCara.

Quando o Espaço Amostral é infinito muitas vezes já está definido de forma numérica, pela própria natureza quantitativa do fenômeno analisado, facilitando a definição da variável aleatória.

Os Modelos Probabilísticos podem ser construídos para as variáveis aleatórias: assim haverá Modelos Probabilísticos Discretos e Modelos Probabilísticos Contínuos. Para construir um modelo probabilístico para uma variável aleatória é necessário definir os seus possíveis valores (contradomínio), e como a probabilidade total (do Espaço Amostral, que vale 1) distribui-se entre eles: é preciso então definir a distribuição de probabilidades. GLOSSÁRIO Distribuição de probabilidades: função que relaciona os valores possíveis que uma variável aleatória pode assumir com as respectivas probabilidades, em suma é o próprio modelo probabilístico da variável aleatória. Fonte: Barbetta, Reis e Bornia, 2010. GLOSSÁRIO

Veja que dependendo do tipo de variável aleatória haverá diferenças na construção da distribuição.

1.2 – Distribuições de probabilidades para variáveis aleatórias discretas

Podemos ver alguns exemplos de variáveis aleatórias discretas:

a) número de coroas obtido no lançamento de 2 moedas;

b) número de itens defeituosos em uma amostra retirada aleatoriamente de um lote;

c) número de defeitos em um azulejo numa fábrica de revestimentos cerâmicos;

d) número de pessoas que visitam um determinado site num certo período de tempo;

Quando uma variável aleatória X é discreta, a obtenção da distribuição de probabilidades consiste em definir o conjunto de pares $[x_i, p(x_i)]$, onde x_i é o i-ésimo valor da variável X, e $p(x_i)$ é a probabilidade de ocorrência de x_i, como no Quadro 1:

$X = x_i$	$p(X = x_i)$
x_1	$p(x_1)$
x_2	$p(x_2)$
...	...

x_n	$p(x_n)$

Quadro 1-Distribuição de Probabilidades para uma Variável Aleatória Discreta

Fonte: elaborado pelo autor

Onde $p(xi) \geq 0$, n é o número de valores que X pode assumir, e $\sum_{i=1}^{n} p(x_i) = 1,0$

Ao obter a distribuição de probabilidades para uma variável aleatória discreta, se você quiser conferir os resultados, some as probabilidades, se elas não somarem 1, há algo errado. Vamos ao primeiro exemplo.

Exemplo 1 - O jogador Ruinzinho está treinando cobranças de pênaltis. Dados históricos mostram que: a probabilidade de ele acertar uma cobrança, supondo que ele acertou a anterior é de 60%. Mas, se ele tiver errado a anterior a probabilidade de ele acertar uma cobrança cai para 30%. Construa a distribuição de probabilidades do número de acertos em 3 tentativas de cobrança.

A variável aleatória X, número de acertos em três tentativas, é uma variável aleatória discreta: o seu contradomínio é finito, o jogador pode acertar 0, 1, 2 ou 3 vezes. Mas, para calcular as probabilidades associadas a esses valores é preciso estabelecer todos os eventos possíveis, pois mais de um evento contribui para as probabilidades de 1 e 2 acertos. Observando a árvore de eventos abaixo (onde A é acertar a cobrança e E significa errar).

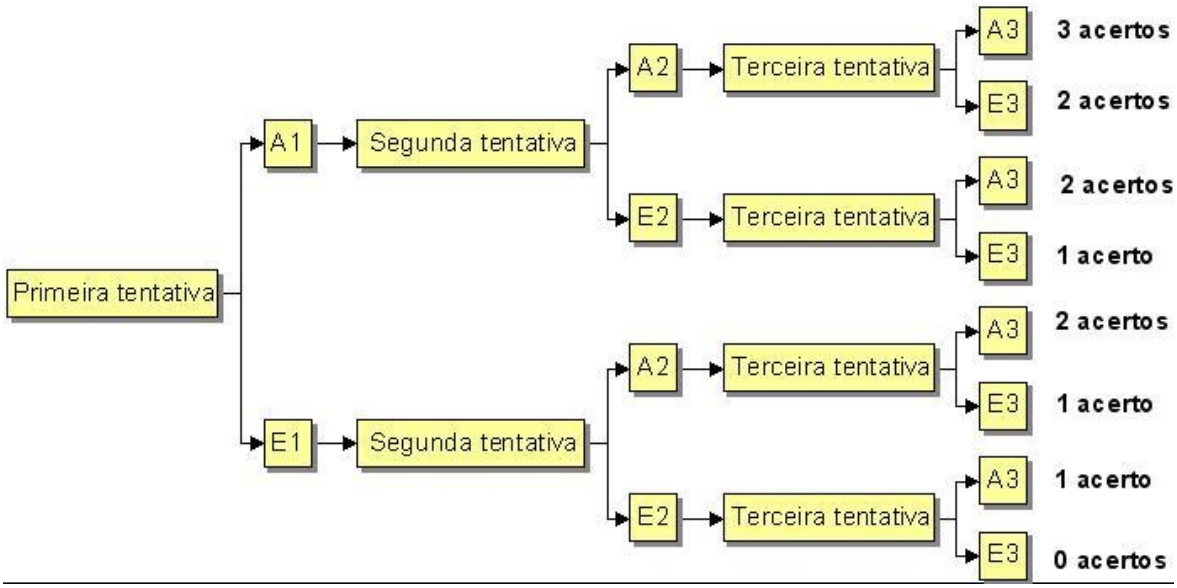


Figura 2 - Árvore de eventos

Fonte: elaborada pelo autor

Observe que todos os eventos são mutuamente exclusivos, o jogador não pode, na mesma seqüência de 3 cobranças, errar e acertar a primeira. É preciso explicitar os valores da variável, e os eventos em termos de teoria dos conjuntos.

Valores possíveis = {0, 1, 2, 3} acertos. A equivalência entre os valores da variável e os eventos é estabelecida abaixo:

$X = 0 \Leftrightarrow [E_1 \cap E_2 \cap E_3]$

$X = 1 \Leftrightarrow [(A_1 \cap E_2 \cap E_3) \cup (E_1 \cap A_2 \cap E_3) \cup (E_1 \cap E_2 \cap A_3)]$

$X = 2 \Leftrightarrow [(A_1 \cap A_2 \cap E_3) \cup (E_1 \cap A_2 \cap A_3) \cup (A_1 \cap E_2 \cap A_3)]$

$X = 3 \Leftrightarrow [A_1 \cap A_2 \cap A_3]$

Então:

$P(X=0) = P[E_1 \cap E_2 \cap E_3]$

$P(X=1) = P[(A_1 \cap E_2 \cap E_3) \cup (E_1 \cap A_2 \cap E_3) \cup (E_1 \cap E_2 \cap A_3)]$

$P(X=2) = P[(A_1 \cap A_2 \cap E_3) \cup (E_1 \cap A_2 \cap A_3) \cup (A_1 \cap E_2 \cap A_3)]$

$P(X=3) = P[A_1 \cap A_2 \cap A_3]$

Assume-se que na primeira tentativa o jogador tem 50% de chance de acertar,LINK E_1, errar a primeira cobrança, é o evento complementar de A_1, acertar a primeira cobrança LINK

então: $P(A_1) = 0,5$ e $P(E_1) = 0,5$

Além disso estabeleceu-se que quando o jogador acertou a cobrança na tentativa anterior a probabilidade de acertar a próxima é de 0,6, e caso tenha errado na anterior a probabilidade de acertar na próxima é de apenas 0,3. Tratam-se de duas probabilidades condicionais, estabelecidas em função de eventos já ocorridos.

Se o jogador acertou na tentativa i (qualqueruma), as probabilidades de acertar e errar na próxima tentativa serão:

$P(A_{i+1}|A_i) = 0,6$ Pelo complementar obtém-se $P(E_{i+1}|A_i) = 0,4$

Se o jogador errou na tentativa i, as probabilidades de acertar e errar na próxima tentativa serão:

$P(A_{i+1}|E_i) = 0,3$ Pelo complementar obtém-se $P(E_{i+1}|E_i) = 0,7$

Com estas probabilidades estabelecidas, lembrando da regra do produto, e considerando o fato de que os eventos são mutuamente exclusivos é possível calcular as probabilidades de ocorrência de cada valor da variável aleatória X.

$$P(X=0) = P[E_1 \cap E_2 \cap E_3] = P(E_1) \times P(E_2|\, E_1) \times P(E_3|\, E_1 \cap E_2)$$

Como os resultados em uma tentativa só dependem daqueles obtidos na imediatamente anterior, o terceiro termo da expressão acima pode ser simplificado para $P(E_3|\, E_2)$, e a probabilidade será:

$$P(X=0) = P(E_1) \text{ x } P(E_2|\, E_1) \text{ x } P(E_3|\, E_2) = 0,5 \times 0,7 \times 0,7 = 0,245 \ (24,5\%)$$

Estendendo o procedimento acima para os outros valores:

$P(X=1) = P[(A_1 \cap E_2 \cap E_3) \cup (E_1 \cap A_2 \cap E_3) \cup (E_1 \cap E_2 \cap A_3)]$

$P(X=2) = P[(A_1 \cap A_2 \cap E_3) \cup (E_1 \cap A_2 \cap A_3) \cup (A_1 \cap E_2 \cap A_3)]$

$P(X=3) = P[A_1 \cap A_2 \cap A_3]$

Como os eventos são mutuamente exclusivos:

$P(X=1) = P(A_1 \cap E_2 \cap E_3) + P(E_1 \cap A_2 \cap E_3) + P(E_1 \cap E_2 \cap A_3)$

$P(X=1) = P(A_1)\times P(E_2|A_1)\times P(E_3|E_2)+P(E_1)\times P(A_2|E_1)\times P(E_3|A_2)+P(E_1)\times P(E_2|E_1)\times P(A_3|E_2)$

$P(X=1) = 0,5\times 0,4 \times 0,7 + 0,5\times 0,3 \times 0,4+ 0,5\times 0,7 \times 0,3 = 0,305$

$P(X=2) = P(A_1\cap A_2\cap E_3)+ P(E_1\cap A_2\cap A_3) + P(A_1\cap E_2\cap A_3)$

$P(X=2) =P(A_1)\times P(A_2|A_1)\times P(E_3|A_2)+P(E_1)\times P(A_2|E_1)\times P(A_3|A_2)+P(A_1)\times P(E_2|A_1)\times P(A_3|E_2)$

$P(X = 2) = 0,5 \times 0,6 \times 0,4 +0,5 \times 0,3 \times 0,6 + 0,5 \times 0,4 \times 0,3 = 0,27$ (27%)

$P(X=3) = P[A_1\cap A_2\cap A_3] = P(A_1) \times P(A_2|A_1) \times P(A_3|A_2) = 0,5 \times 0,6 \times 0,6 = 0,18$ (18%)

Com os valores calculados acima é possível construir o Quadro 2com os pares valores-probabilidades.

X	$p(X = x_i)$
0	0,245
1	0,305
2	0,270
3	0,180
Total	1,0

Quadro 2 - Distribuição de probabilidades: número de acertos em 3 cobranças

Fonte: elaborado pelo autor

Ao longo dos séculos, matemáticos e estatísticos deduziram modelos matemáticos para tornar mais simples a obtenção de distribuição de probabilidades para uma variável aleatória discreta. Alguns destes modelos serão vistos na Unidade 2.

Vamos agora passar para a análise das variáveis aleatórias contínuas.

1.3 – Distribuições de probabilidades para variáveis aleatórias contínuas

Podemos ver alguns exemplos de variáveis aleatórias contínuas:

- volume de água perdido em um sistema de abastecimento;
- renda familiar em salários mínimos de pessoas selecionadas por amostragem aleatória para responder uma pesquisa;
- a demanda por um produto em um mês;
- tempo de vida de uma lâmpada incandescente;

Uma variável aleatória contínua está associada a um Espaço Amostral infinito. Assim, a probabilidade de que a variável assuma exatamente um valor x_i é zero, não havendo mais sentido em representar a distribuição pelos pares $x_i - p(x_i)$. Igualment sem sentido fica a distinção entre > e ≥ existente nas variáveis aleatórias discretas. Utiliza-se então uma função não negativa, a função densidade de probabilidades, definida para todos os valores possíveis da variável aleatória.

Uma função densidade de probabilidades poderia ser apresentada graficamente da seguinte forma:

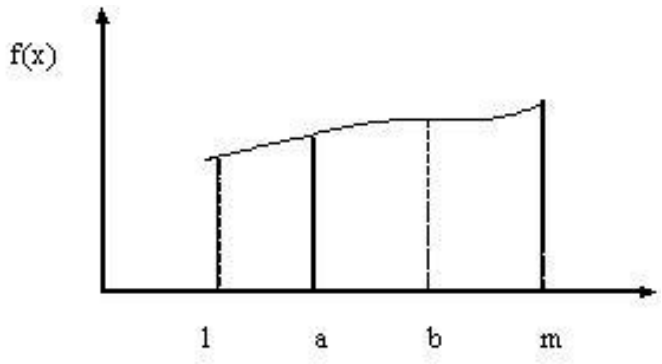


Figura 3 - Função densidade de probabilidades

Fonte: elaborada pelo autor

Para calcular a probabilidade de uma variável aleatória contínua assumir valores entre **a** e **b** (dois valores quaisquer), basta calcular a área abaixo da curva entre a e b. Se a área for calculada entre l e m (limites da função) tem que dar 1, que é a probabilidade total. Usualmente isso é feito calculando a integral da função no intervalo de interesse. Em muitas situações de nosso interesse tais probabilidades podem ser calculadas através de fórmulas matemáticas relativamente simples, ou foram dispostas em tabelas, que são encontradas em praticamente todos os livros de estatística, e que serão vistas na Unidade 7.

Agora vamos ver alguns conceitos muito importantes como valor esperado e variância de uma variável aleatória.

1.4 – Valor Esperado e Variância

Todos os modelos probabilísticos apresentam duas medidas (dois momentos) que permitem caracterizar a variável aleatória para a qual eles foram construídos: o Valor Esperado e a Variância da variável aleatória. O Valor Esperado (simbolizado por E(X)) nada mais é do que a média aritmética simples vista em Análise Exploratória de Dados (Unidade 3 de Estatística Aplicada à Administração I), utilizando probabilidades ao invés de frequências no cálculo. Analogamente, a Variância (simbolizada por V(X)) é a variância vista anteriormente, utilizando probabilidades. Da mesma forma que em Análise Exploratória de Dados é também comum trabalhar com o Desvio Padrão, raiz quadrada positiva da Variância (que aqui será simbolizado por $\sigma(X)$, "sigma de X"). A interpretação dos resultados obtidos pode ser feita de forma semelhante à Análise Exploratória de Dados, apenas recordando que se trata de uma variável aleatória, e estão sendo usadas probabilidades e não frequências.

Para uma variável aleatória discreta o valor esperado e a variância podem ser calculados da seguinte forma:

$$E(X) = \sum_{i=1}^{n} x_i \times p(x_i) \qquad V(X) = E\left(X^2\right) - [E(X)]^2 \quad \text{onde } E\left(X^2\right) = \sum_{i=1}^{n} x_i^2 \times p(x_i)$$

Para uma variável aleatória contínua a obtenção do valor esperado e da variância exige o cálculo de integrais das funções de densidade de probabilidades. Para as distribuições mais importantes as equações encontram-se disponíveis nos livros de estatística, em função dos parâmetros da distribuição, e algumas serão vistas na Unidade 2.

Uma das principais utilidades do valor esperado é na comparação de propostas. Suponha que os valores de uma variável aleatória sejam lucros, ou prejuízos, advindos de

decisões tomadas, por exemplo, decidir por uma proposta de compra do cliente A, ou do cliente B. Associados aos valores há probabilidades, como decidir qual é a mais vantajosa? O cálculo do valor esperado possibilita uma comparação objetiva: decidiríamos pela que apresentasse o lucro esperado mais elevado. Há um campo de conhecimento que se ocupa especificamente de fornecer as ferramentas necessárias para tais tomadas de decisão: a teoria estatística da decisão ou análise estatística da decisão.

O valor esperado (média) e a variância apresentam algumas propriedades, tanto para variáveis aleatórias discretas quanto contínuas. O seu conhecimento facilitará muito a obtenção das medidas em problemas mais sofisticados.

Para o valor esperado E(X), sendo k uma constante:

a) $E(k) = k$ A média de uma constante é a própria constante.

b) $E(k \pm X) = k \pm E(X)$ A média de uma constante somada a uma variável aleatória é a própria constante somada à média da variável aleatória.

c) $E(k \times X) = k \times E(X)$ A média de uma constante multiplicada por uma variável aleatória é a própria constante multiplicada pela média da variável aleatória.

d) $E(X \pm Y) = E(X) \pm E(Y)$ A média da soma de duas variáveis aleatórias é igual à soma das médias das duas variáveis aleatórias.

e) Sejam X e Y duas variáveis aleatórias independentes $E(X \times Y) = E(X) \times E(Y)$ A média do produto de duas variáveis aleatórias independentes é igual ao produto das médias das duas variáveis aleatórias.

Para a variância V(X), sendo k uma constante:

a) $V(k) = 0$ Uma constante não varia, portanto sua variância é igual a zero.

b) $V(k \pm X) = V(X)$ A variância de uma constante somada a uma variável aleatória é igual apenas à variância da variável aleatória.

c) $V(k \times X) = k^2 \times V(X)$ A variância de uma constante multiplicada a uma variável aleatória é igual ao quadrado da constante multiplicada pela variância da variável aleatória.

d) Sejam X e Y duas variáveis aleatórias independentes $V(X \pm Y) = V(X) + V(Y)$ A variância da soma ou subtração de duas variáveis aleatórias independentes será igual à soma das variâncias das duas variáveis aleatórias.

Agora vamos ver um exemplo.

Exemplo 2 - Calcular o valor esperado e a variância da distribuição do Exemplo 1.
Para uma variável aleatória discreta é aconselhável acrescentar mais uma coluna ao Quadro 2 com os valores e probabilidades, para poder calcular o valor de $E(X^2)$:

X	$p(X = x_i)$	$x_i \times p(X = x_i)$	$x_i^2 \times p(X = x_i)$
0	0,245	0	0
1	0,305	0,305	0,305
2	0,270	0,540	1,08
3	0,180	0,540	1,62
Total	1,0	1,385	3,005

Quadro 3 - Distribuição de probabilidades do Exemplo 1 (com coluna $x_i^2 \times p(X = x_i)$)
Fonte: elaborado pelo autor.

Substituindo nas expressões de valor esperado e variância:

$$\mathbf{E(X) = \sum_{i=1}^{n} x_i \times p(x_i) = 1{,}385}\ \text{acertos}$$

$$\mathbf{V(X) = \sum_{i=1}^{n} x_i^2 \times p(x_i) - \left[\sum_{i=1}^{n} x_i \times p(x_i)\right]^2 = 3{,}005 - (1{,}385)^2 = 1{,}087}\ \text{acertos}^2$$

$$\mathbf{\sigma(X) = \sqrt{V(X)} = \sqrt{1{,}087} = 1{,}042}\ \text{acertos}$$

Observe que o valor esperado (1,385 acertos) é um valor que a variável aleatória não pode assumir! Não é o "valor mais provável", é o ponto de equilíbrio do conjunto. Repare que a unidade da variância dificulta sua comparação com o valor esperado, mas ao se utilizar o desvio padrão é possível verificar que a dispersão dos resultados é quase do valor da média (valor esperado).

Tô afim de saber:

- Sobre Variáveis Aleatórias, BARBETTA, P.A., REIS, M.M., BORNIA, A.C. **Estatística para Cursos de Engenharia e Informática**. 3ª ed. - São Paulo: Atlas, 2010, capítulos 5 e 6.
- Sobre as propriedades de valor esperado e variância, BARBETTA, P.A., REIS, M.M., BORNIA, A.C. **Estatística para Cursos de Engenharia e Informática**. 3ª ed. - São Paulo: Atlas, 2010., capítulos 5 e 6.
- Também sobre variáveis aleatórias, STEVENSON, Willian J. **Estatística Aplicada à Administração**. São Paulo: Ed. Harbra, 2001, capítulos 5 e 6.
- Sobre teoria estatística da decisão: BEKMAN, O. R., COSTA NETO, P. O. **Análise Estatística da Decisão**. São Paulo: Edgard Blücher, 1980, 4ª reimpressão, 2006.

Atividades de aprendizagem

1) Três alunos estão tentando independentemente resolver um problema. A probabilidade de que o aluno A resolva o problema é de 4/5, de B resolver é de 2/3 e de C resolver é de 3/7. Seja X o número de soluções corretas apresentadas para este problema.
a) Construa a distribuição de probabilidades de X. (R.: 0,038; 0,257; 0,476; 0,228)
b) Calcule E(X) e V(X). (R.: 1,893; 0,630)

2) Um prédio possui 3 vigias dispostos em vários pontos de onde têm visão do portão de entrada. Se alguém não autorizado entrar, o vigia que o vê faz soar um alarme. Suponha que os vigias trabalham independentemente entre si, e que a probabilidade de que cada um deles veja uma pessoa entrar é 0,8. Seja X o número de alarmes que soam ao entrar uma

pessoa não autorizada. Encontre a distribuição de probabilidades de X. (R.: 0,008; 0,096; 0,384; 0,512)

3) Uma companhia petrolífera obteve a concessão de explorar uma certa região. Estudos anteriores estimam que a probabilidade de existir petróleo nessa região é 0,2. A companhia pode optar por um novo teste que custa $ 50, sendo que, se realmente existe petróleo, esse teste dirá com probabilidade 0,8 que existe, e se realmente não existe petróleo, dirá com probabilidade 0,7 que não existe. Considerando que o custo de perfuração será de $ 300 e se for encontrado petróleo, a companhia lucrará $1500 (lucro bruto), qual o valor esperado do lucro da companhia, se essa tomar as melhores decisões (perfurar quando o teste indicar que há petróleo e não perfurar quando o teste não indicar)? (R.: $70)

Resumo

O resumo desta Unidade está demonstrado na Figura4:

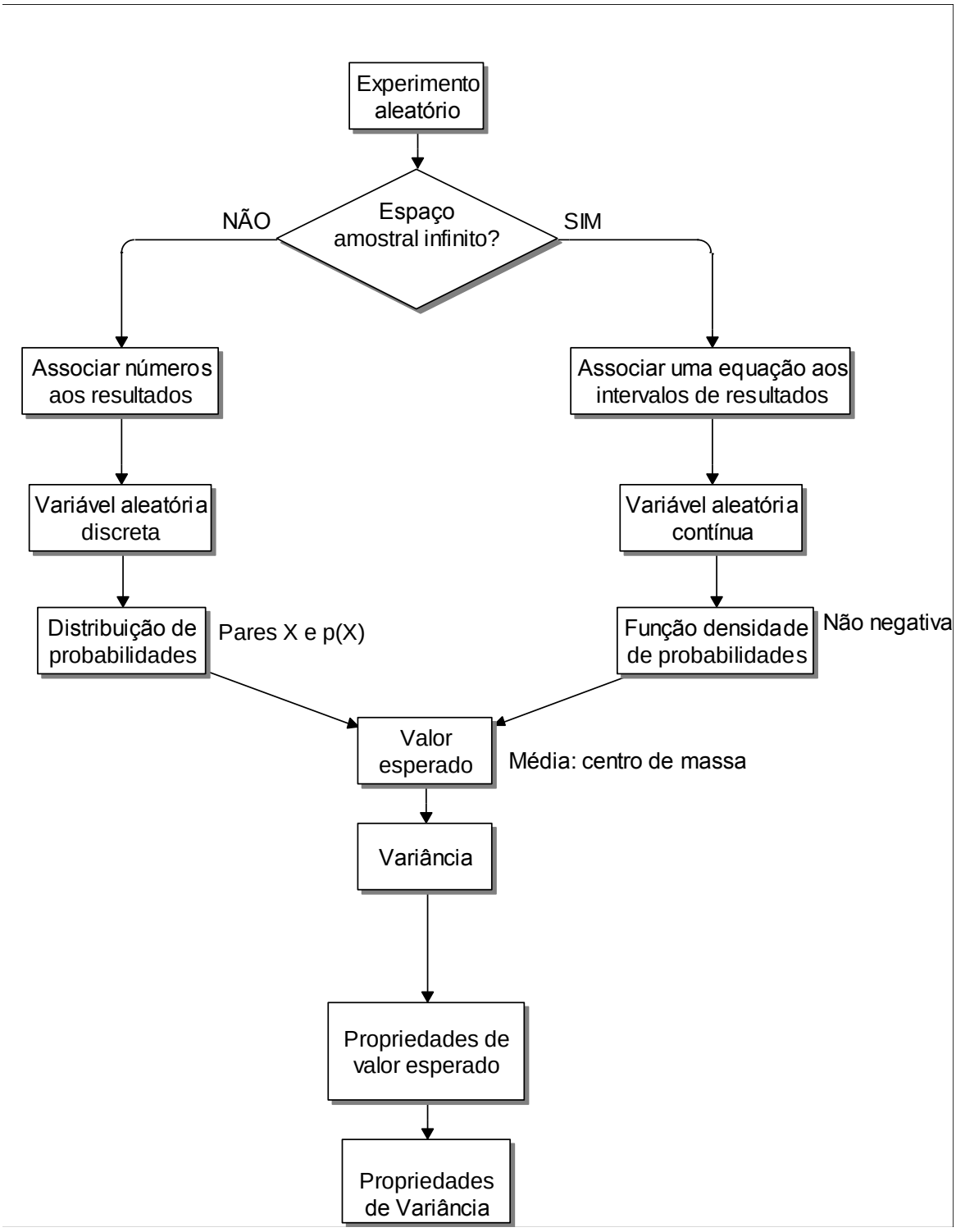


Figura 4 - Resumo da Unidade 1

Fonte: elaborado pelo autor

Chegamos ao final de mais uma Unidade. Veremos mais sobre os temas abordados na Unidade 2 quando estudaremos várias distribuições de probabilidade (modelos probabilísticos) que são extremamente úteis para modelar muitas situações práticas, auxiliando na tomada de decisões. Estes conhecimentos serão depois aplicados nas Unidades 4 e 5.

Unidade 2

Modelos probabilísticos mais comuns

Objetivo

Nesta Unidade você vai conhecer os modelos probabilísticos mais importantes para variáveis aleatórias discretas e contínuas. Você aprenderá a identificar as situações reais em que podem ser usados para o cálculo de probabilidades e a importância disso para o administrador.

2.1– Modelos Probabilísticos para Variáveis Aleatórias Discretas

Na Unidade 6 de Estatística Aplicada à Administração I e na Unidade 1 deste livro vimos os conceitos gerais de Probabilidade e Variáveis Aleatórias: podemos construir um modelo probabilístico do zero para um problema de administração, a partir de dados históricos ou experimentais.

Embora plenamente possível, o processo de construção de um modelo probabilístico do zero pode ser bastante longo: é preciso coletar os dados, fazer a análise exploratória deles, obter as probabilidades e validar o modelo. Mesmo tomando todos os cuidados, muitas vezes iremos reinventar a roda, e correndo o risco de ela sair quadrada...

Por que não usar os conhecimentos prévios desenvolvidos ao longo de centenas de anos de pesquisa e experimentação? Vamos procurar, dentre os vários modelos probabilísticos existentes aquele mais apropriado para o fenômeno que estamos estudando, que é materializado através de variáveis aleatórias.

Através da análise exploratória de dados podemos avaliar qual modelo é mais apropriado para os nossos dados. Contudo, para fazer isso precisamos conhecer tais modelos.

Nesta Unidade vamos estudar os modelos mais usados para variáveis aleatórias discretas (binomial e Poisson), e para variáveis aleatórias contínuas (uniforme, normal, t e qui-quadrado).

Aqui é importante avaliar com cuidado a variável aleatória discreta.GLOSSÁRIO Variável aleatória: é uma função matemática que associa números reais aos resultados de um Espaço Amostral, por sua vez vinculado a um Experimento Aleatório. Fonte: Barbetta, Reis e Bornia, 2010. FimGLOSSÁRIO

É preciso identificar se o **Espaço Amostral** é **finito** ou **infinito numerável**:GLOSSÁRIO Espaço Amostral finito é aquele formado por um número limitado de resultados possíveis.Fonte: Barbetta, Reis e Bornia, 2010. Fim GLOSSÁRIO.
- Espaço Amostral infinito numerável é aquele formado por um número infinito de resultados, mas que podem ser listados. Fonte: Barbetta, Reis e Bornia, 2010. Fim GLOSSÁRIOalguns modelos são apropriados para um caso e não para o outro.

Vamos ver os dois modelos mais importantes para variáveis aleatórias discretas: binomial e Poisson.

2.1.1 Modelo binomial

Seja um Experimento AleatórioGLOSSÁRIO Experimento Aleatório é um processo de obtenção de um resultado ou medida que apresenta as seguintes características: não se pode afirmar, antes de realizar o experimento, qual será o resultado de uma realização, mas é possível determinar o conjunto de resultados possíveis; quando é realizado um grande número de vezes (replicado) apresentará uma regularidade que permitirá construir um modelo probabilístico para analisar o experimento. Fonte: adpatado pelo autor de Lopes, 1999. Fim GLOSSÁRIOqualquer que apresenta as seguintes características:

- consiste na realização de um número finito e conhecido **n** de ensaios (ou repetições);
- cada um dos ensaios tem apenas dois resultados possíveis: "sucesso" ou "fracasso" (estão entre aspas porque a definição de sucesso não quer necessariamente algo "positivo", e também porque poderá incluir significar um grupo de resultados); e
- os ensaios são independentes entre si, apresentando probabilidades de "sucesso" (**p**) e de "fracasso" (**1-p**) constantes.

Neste caso estamos interessados no número de "sucessos" obtidos nos **n** ensaios: como o Espaço Amostral é finito (vai de 0 a **n**) uma variável aleatória associada seria discreta. Este tipo de experimento é chamado de Binomial.

Então, a variável aleatória discretaGLOSSÁRIO Variável aleatória discreta: o Espaço Amostral ao qual ela está associada é finito ou infinito numerável. Fonte: Barbetta, Reis e Bornia, 2010. FimGLOSSÁRIO X, número de "sucessos" nos **n** ensaios, apresenta uma distribuição (modelo) binomial com os seguintes parâmetros:

n = número de ensaios **p =** probabilidade de "sucesso"

Com esses dois parâmetros é possível calcular as probabilidades de um determinado número de sucessos, bem como obter o Valor Esperado e a Variância da variável X:

$E(X) = n \times p$ $V(X) = n \times p \times (1-p)$

Exemplo 1 - Experimentos binomiais:

a) Observar o número de caras em 3 lançamentos imparciais de uma moeda honesta: n=3; p=0,5

b) Observar o número de meninos nascidos em 3 partos de uma família: n=3; p = x

c) Observar o número de componentes defeituosos em uma amostra de 10 componentes de um grande número de peças que apresentaram anteriormente 10% de defeituosos: n = 10; p= 0,1.

Vamos ver com maiores detalhes o caso do número de meninos (e meninas) nascidos em uma família. Chamando menino de evento H, será o "sucesso", e menina de evento M, e sabendo pela história da família que P(H) = 0,52 e P(M) = 0,48 (então **p** = 0,52 e **1- p** = 0,48), quais serão as probabilidades obtidas para a variável aleatória número de meninos em 3 nascimentos? Vamos obter a distribuição de probabilidades.

Resolvendo usando os conceitos gerais de probabilidade é preciso primeiramente determinar o Espaço Amostral, como poderão ser os sexos das 3 crianças:

Ω = {H∩H∩H, H∩H∩M, H∩M∩H, M∩H∩H, H∩M∩M, M∩H∩M, M∩M∩H, M∩M∩M}

Supondo que os nascimentos sejam independentes, podemos calcular as probabilidades de cada intersecção simplesmente multiplicando as probabilidades individuais de seus componentes:

$P\{H \cap H \cap H\} = P(H) \times P(H) \times P(H) = p \times p \times p = p^3$

$P\{H \cap H \cap M\} = P(H) \times P(H) \times P(M) = p \times p \times (1- p) = p^2 \times (1 - p)$

$P\{H \cap M \cap H\} = P(H) \times P(M) \times P(H) = p \times (1 - p) \times p = p^2 \times (1 - p)$

$P\{M \cap H \cap H\} = P(M) \times P(H) \times P(H) = (1 - p) \times p \times p = p^2 \times (1 - p)$

$P\{H \cap M \cap M\} = P(H) \times P(M) \times P(M) = p \times (1 - p) \times (1 - p) = p \times (1 - p)^2$

$P\{M \cap H \cap M\} = P(M) \times P(H) \times P(M) = (1 - p) \times p \times (1 - p) = p \times (1 - p)^2$

$P\{M \cap M \cap H\} = P(M) \times P(M) \times P(H) = (1 - p) \times (1 - p) \times p = p \times (1 - p)^2$

$P\{M \cap M \cap M\} = P(M) \times P(M) \times P(M) = (1- p) \times (1 - p) \times (1 - p) = (1 - p)^3$

Observe que:

$P\{H \cap H \cap M\} = P\{H \cap M \cap H\} = P\{M \cap H \cap H\} = p^2 \times (1 - p)$ = Probabilidade de 2 "sucessos"

$P\{H \cap M \cap M\} = P\{M \cap H \cap M\} = P\{M \cap M \cap H\} = p \times (1 - p)^2$ = Probabilidade de 1 "sucesso"

Importa apenas a "natureza" dos sucessos, não a ordem em que ocorrem: com a utilização de **combinações**LINK em qualquer livro de matemática do ensino médio é possível encontrar a definição e exemplos de combinações. Fim LINK. é possível obter o número de resultados iguais para cada número de sucessos. Supondo que o número de ensaios **n** é o número de "objetos" disponíveis, e que o número de "sucessos" em que estamos interessados (doravante chamado **k**) é o número de "espaços" onde colocar os objetos (um objeto por espaço), o número de resultados iguais será:

$$C_{n,k} = \frac{n!}{k! \times (n - k)!}$$

Para o caso acima, em que há 3 ensaios (**n** =3):

- para 2 sucessos (**k** =2) $C_{3,2} = \frac{3!}{2! \times (3 - 2)!} = 3$ (o mesmo resultado obtido por enumeração)

- para 1 sucesso (**k** =1) $C_{3,1} = \frac{3!}{1! \times (3 - 1)!} = 3$ (o mesmo resultado obtido por enumeração)

O procedimento acima poderia ser feito para quaisquer valores de **n** e **k** (desde que **n** ≥ **k**), permitindo obter uma expressão geral para calcular a probabilidade associada a um resultado qualquer.

A probabilidade de uma variável aleatória discreta X, número de sucessos em **n** ensaios, com distribuição binomial de parâmetros **n** e **p**, assumir um certo valor **k** (0 ≤**k**≤**n**) será:

$$P(X = k) = C_{n,k} \times p^k \times (1-p)^{n-k} \qquad \text{onde:} \quad C_{n,k} = \frac{n!}{k! \times (n-k)!}$$

É importante lembrar que a probabilidade de ocorrer **k** sucessos é igual à probabilidade de ocorrer**n - k** fracassos, e que todos os axiomas e propriedades de probabilidade continuam válidos.

Exemplo 2 - Admitamos que a probabilidade de que companhia não entregue seus produtos no prazo é igual a 18%. Quais são as probabilidades de que em 3 entregas 1, 2 ou todas as 3 entregas sejam feitas no prazo. Calcular também valor esperado, variância e desvio padrão do número de entregas no prazo.

Para cada entrega ("ensaio") há apenas dois resultados: no prazo ou não. Há um número limitado de realizações, **n** = 3. Definindo "sucesso" como no prazo, e supondo as operações independentes, a variável aleatória X, número de entregas no prazo em 3 terá distribuição binomial com parâmetros

n = 3 e **p** = 0,82 (*e***1- p** = 0,18).

Então:

$$P(X = 0) = C_{3,0} \times 0{,}82^0 \times (0{,}18)^3 = \frac{3!}{0! \times (3-0)!} \times 0{,}82^0 \times (0{,}18)^3 = 0{,}006$$

$$P(X = 1) = C_{3,1} \times 0{,}82^1 \times (0{,}18)^2 = \frac{3!}{1! \times (3-1)!} \times 0{,}82^1 \times (0{,}18)^2 = 0{,}080$$

$$P(X = 2) = C_{3,2} \times 0{,}82^2 \times (0{,}18)^1 = \frac{3!}{2! \times (3-2)!} \times 0{,}82^2 \times (0{,}18)^1 = 0{,}363$$

$$P(X = 3) = C_{3,3} \times 0{,}82^3 \times (0{,}18)^0 = \frac{3!}{3! \times (3-3)!} \times 0{,}82^3 \times (0{,}18)^0 = 0{,}551$$

Somando todas as probabilidades o resultado é igual a 1, como teria que ser. LINK Lembre-se que a soma das probabilidades de todos os eventos que compõem o Espaço Amostral é igual a 1. E que 0! = 1, e que um número diferente de 0 elevado a zero é igual a 1. LINK O Valor Esperado, Variância e o Desvio Padrão serão:

$$E(X) = n \times p = 3 \times 0{,}82 = 2{,}46 \text{ entregas}$$

$$V(X) = n \times p \times (1-p) = 3 \times 0{,}82 \times 0{,}18 = 0{,}4428 \text{ entregas}^2.$$

$$\sigma(X) = \sqrt{V(X)} = \sqrt{0{,}4428} = 0{,}665 \text{ entregas}$$

A média é quase igual ao número de operações devido à alta probabilidade de sucesso.

Mas, e se o Espaço Amostral fosse infinito numerável? Teríamos que usar o modelo de Poisson. Você conhece este modelo? Sabe como tirar proveito de suas facilidades? Vamos estudar juntos para aprender ou para relembrar!

2.1.2 – Modelo de Poisson

Vamos supor um experimento binomial, com apenas dois resultados possíveis, mas com a seguinte característica: apesar da probabilidade **p** ser constante o valor de **n** teoricamente é infinito.

Na situação acima o modelo binomial não poderá ser utilizado. Nestes casos deve ser utilizado o modelo de Poisson.

Como seria a solução para o caso acima?

Como **n** é "infinito" deve-se fazer a análise das ocorrências em um período contínuo (de tempo, de espaço, entre outros) subdividido em um certo número de subintervalos (número tal que a probabilidade de existir mais de uma ocorrência em uma subdivisão é desprezível, e supondo ainda que as ocorrências em subdivisões diferentes são independentes); novamente é preciso trabalhar com uma quantidade constante que será chamada de **m** também:

$$m = \lambda \times t$$

onde **λ** é uma taxa de ocorrência do evento em um período contínuo (igual ou diferente do período sob análise), e **t** é justamente o período contínuo sob análise.LINK Apesar do símbolo t, o período contínuo não é necessariamente um intervalo de tempo. LINK

Como obter a taxa **λ**? Há duas opções: realizar um número suficiente de testes de laboratório para obter a taxa de ocorrência do evento a partir dos resultados, ou observar dados históricos e calcular a taxa.

Se uma variável aleatória discreta X, número de ocorrências de um evento, segue a distribuição de Poisson, a probabilidade de X assumir um valor **k** será:

$$P(X = k) = \frac{e^{-m} \times m^{k}}{k!}$$

Onde **e** é uma constante: **e**≅ 2,71. E $m = n \times p$ ou $m = \lambda \times t$.

Uma particularidade interessante da distribuição de Poisson é que o Valor Esperado e a Variância de uma variável aleatória que siga tal distribuição serão iguais:

$$E(X) = m = \lambda \times t$$

$$V(X) = m = \lambda \times t$$

O modelo de Poisson é muito utilizado para modelar fenômenos envolvendo filas: filas de banco, filas de mensagens em um servidor, filas de automóveis em um cruzamento.

Exemplo 3 - Alguns experimentos e fenômenos que seguem a distribuição de Poisson:

a) Número mensal de acidentes de tráfego em um cruzamento.

Observe que é uma variável aleatória discreta, pode assumir apenas valores inteiros (0, 1, 2, 3,...). Cada realização do "experimento" (acidente) pode ter apenas 2 resultados: ocorre o acidente ou não ocorre o acidente. Mas, o número máximo de realizações é desconhecido! Assim, a distribuição binomial não pode ser usada, e a análise do número de acidentes precisa ser feita em um período contínuo (no caso, período de tempo, 1 mês), exigindo o uso da distribuição de Poisson.

b) Número de itens defeituosos produzidos por hora em uma indústria.

Novamente, uma variável aleatória discreta (valores inteiros: 0,1, 2, 3, ...), cada realização só pode ter dois resultados possíveis (peça sem defeito ou peça defeituosa). Se o número máximo de realizações for conhecido, provavelmente a probabilidade de uma peça ser defeituosa será reduzida e apesar de ser possível a utilização da distribuição binomial o uso da distribuição de Poisson obterá resultados muitos próximos. Se o número máximo de realizações for desconhecido a distribuição binomial não pode ser usada, e a análise do número de acidentes precisa ser feita em um período contínuo (no caso, período de tempo, 1 hora), exigindo o uso da distribuição dePoisson.

c) Desintegração dos núcleos de substâncias radioativas: contagem do número de pulsações radioativas a intervalos de tempo fixos.

Situação semelhante a dos acidentes em um cruzamento, só que o "grau de aleatoriedade" deste experimento é muito maior. O número máximo de pulsações também é desconhecido, obrigando a realizar a análise em um período contínuo, utilizando a distribuição dePoisson.

Exemplo 4 - Uma telefonista recebe cerca de 0,20 chamadas por minuto (valor obtido de medições anteriores).

a) Qual é a probabilidade de receber exatamente 5 chamadas nos primeiros 10 minutos?

b) Qual é a probabilidade de receber até 2 chamadas nos primeiros 12 minutos?

c) Qual é o desvio padrão do número de chamadas em meia hora?

Há interesse no número de chamadas ocorridas em um período contínuo (de tempo no caso). Para cada "ensaio" há apenas dois resultados possíveis: a chamada ocorre ou não.

Observe que não há um limite para o número de chamadas no período (sabe-se apenas que o número mínimo pode ser 0): por esse motivo a utilização da binomial é inviável... Contudo há uma taxa de ocorrência (λ = 0,20 chamadas/minuto) e isso permite utilizar a distribuição de Poisson.

a) Neste caso o período **t** será igual a 10 minutos (**t** = 10 min.), P(X = 5)?

$$m = \lambda \times t = 0{,}20 \times 10 = 2 \text{ chamadas}$$

$$P(X = k) = \frac{e^{-m} \times m^{k}}{k!} = P(X = 5) = \frac{e^{-2} \times 2^{5}}{5!} = 0{,}0361$$

Então a probabilidade de que a telefonista receba exatamente 5 chamadas em 10 minutos é igual a 0,0361 (3,61%).

b) Neste caso o período **t** será igual a 12 minutos (**t** = 12 minutos). O evento de interesse é até 2 chamadas em 12 minutos ($X \leq 2$).

$$m = \lambda \times t = 0{,}20 \times 12 = 2{,}4 \text{ chamadas}$$

P(X ≤ 2) = P(X = 0) + P(X = 1) + P(X = 2)

$$P(X = 0) = \frac{e^{-2,4} \times 2{,}4^0}{0!} = 0{,}0907$$

$$P(X = 1) = \frac{e^{-2,4} \times 2{,}4^1}{1!} = 0{,}2177$$

$$P(X = 2) = \frac{e^{-2,4} \times 2{,}4^2}{2!} = 0{,}2613$$

P(X ≤ 2) = P(X = 0) + P(X = 1) + P(X = 2) = 0,0907 + 0,2177 + 0,2613 = 0,5697

Então a probabilidade de que a telefonista receba até 2 chamadas em 12 minutos é igual a 0,5697 (56,97%).

c) Neste caso o período **t** será igual a 30 minutos (**t** = 30 minutos). Primeiro calcula-se a variância:

$$V(X) = m = \lambda \times t = 0{,}2 \times 30 = 6 \text{ chamadas}^2$$

O Desvio Padrão é a raiz quadrada positiva da variância:

$$\sigma(X) = \sqrt{V(X)} = \sqrt{6} \cong 2{,}45 \text{ chamadas}$$

Há vários outros modelos para variáveis aleatórias discretas: hipergeométrico, geométrico, binomial negativo.

Na próxima seção vamos ver os principais modelos variáveis aleatórias contínuas.

2.2 – Modelos probabilísticos para Variáveis Aleatórias Contínuas

Nesta seção estudaremos os modelos uniforme, normal, t e qui-quadrado.

2.2.1 – Modelo uniforme

Quando o Espaço Amostral associado a um Experimento Aleatório é infinito torna-se necessário o uso de uma Variável Aleatória Contínuapara associar números reais aos resultados. Os modelos probabilísticos vistos anteriormente não podem ser empregados: a probabilidade de que uma variável aleatória contínua assuma exatamente um determinado valor é zero.

Para entender melhor a declaração acima, vamos relembrar a definição clássica de probabilidade: a probabilidade de ocorrência de um evento será igual ao quociente entre o número de resultados associados ao evento pelo número total de resultados possíveis. Ora, se o número total de resultados é infinito, ou tende ao infinito para ser mais exato, a probabilidade de ocorrência de um valor específico é igual a zero. Por esse motivo, quando se lida com Variáveis Aleatórias Contínuas calcula-se a probabilidade de ocorrência de eventos formados por intervalos de valores, através de uma função densidade de probabilidades (ver Unidade 1). Outra conseqüência disso é que os símbolos > e ≥ (< e ≤ também) são equivalentes para variáveis aleatórias contínuas.

O modelo mais simples para variáveis aleatórias contínuas é o modelo uniforme.

Seja uma variável aleatória contínua qualquer X que possa assumir valores entre A e B. Todos os valores entre A e B têm a mesma probabilidade de ocorrer, resultando no gráfico apresentado na Figura 5:

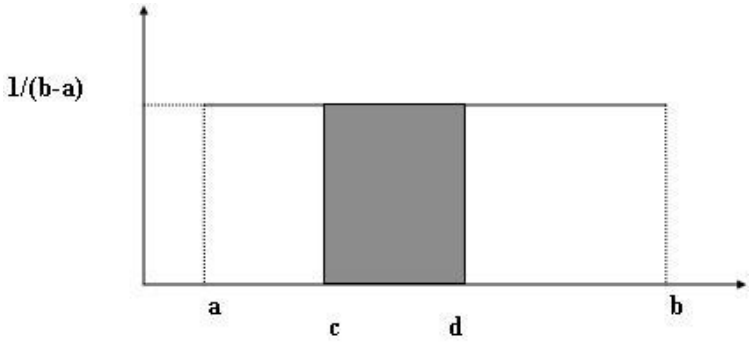


Figura 5 - Modelo uniforme

Fonte: elaborada pelo autor

Para que a área entre a e b seja igual a 1 o valor da ordenada precisa ser igual a **1/(b - a)**, constante, portanto, para todo o intervalo. A área escura representa a probabilidade da variável X assumir valores no intervalo **c - d**. Trata-se do modelo uniforme.

Dois intervalos de valores da variável aleatória contínua, que tenham o mesmo tamanho, têm a mesma probabilidade de ocorrer (desde que dentro da faixa de valores para os quais a função de densidade de probabilidades não é nula). Formalmente, uma variável aleatória contínua X tem distribuição uniforme, com parâmetros **a** e **b** reais (sendo **a** menor do que **b**), se sua função densidade de probabilidades for tal como a da Figuras49.

A probabilidade de que a variável assuma valores entre c e d (sendo $a < c < d < b$), é a área compreendida entre c e d:

$$P(c < X < d) = (d - c) \times \frac{1}{(b - a)}$$

Seu valor esperado e variância são:

$$E(X) = \frac{a + b}{2} \qquad V(X) = \frac{(b - a)^2}{12}$$

Intuitivamente podemos supor que muitas variáveis aleatórias contínuas terão um comportamento diferente do caso acima: em algumas delas haverá maior probabilidade de ocorrências de valores próximos ao limite inferior ou superior: para cada caso deverá ser ajustado um modelo probabilístico contínuo adequado.

O modelo uniforme é bastante usado para gerar números pseudo-aleatórios em processos de amostragem probabilística. LINK No ambiente virtual temos um exemplo resolvido de modelo uniforme, adaptado de BUSSAB, W.O., MORETTIN, P. A. **Estatística Básica**. 4ª ed. – São Paulo: Atual, 1987. LINK

Agora vamos passar ao modelo mais importante para variáveis aleatórias contínuas.

2.2.2 – Modelo normal

Há casos em que há maior probabilidade de ocorrência de valores situados em intervalos centrais da função densidade de probabilidades da variável aleatória contínua, e esta probabilidade diminui a medida que os valores se afastam deste centro (para valores menores ou maiores) o modelo probabilístico contínuo mais adequado seja o modelo Normal ou gaussiano. LINK O matemático alemão Gauss utilizou amplamente este modelo no tratamento de erros experimentais, embora não tenha sido o seu "descobridor". LINK Isso é especialmente encontrado em variáveis biométricas, resultantes de medidas corpóreas em seres vivos.

O Modelo Normal é adequado para medidas numéricas em geral, descrevendo vários fenômenos, e permitindo fazer aproximações de modelos discretos. É extremamente importante também para a Estatística Indutiva. O gráfico da função densidade de probabilidades de uma variável aleatória contínua que siga o modelo Normal (distribuição Normal) será como a Figura 6:

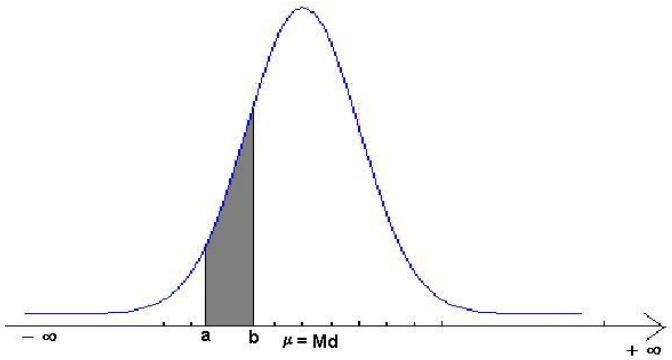


Figura 6 - Distribuição normal

Fonte: elaborada pelo autor a partir de Statgraphics®

Características:

- a curva apresenta forma de sino, há maior probabilidade da variável assumir valores próximos do centro.
- os valores de média (**μ**) e de mediana (**Md**) são iguais, significando que a curva é simétrica em relação à média.
- teoricamente a curva prolonga-se de - ∞ a + ∞ (menos infinito a mais infinito), então a área total sob a curva é igual a 1 (100%).

- qualquer distribuição normal é perfeitamente especificada por seus parâmetros média (**μ**) e variância (σ^2) =>**X**: **N** (**μ** , σ^2) LINK É comum a utilização de letras do alfabeto grego para representar algumas medidas. Não se esqueça que o desvio padrão (**σ**) é a raiz quadrada positiva da variância. LINK significa que a variável X tem distribuição normal com média **μ** e variância σ^2.

- a área escura na Figura6 é a probabilidade de uma variável que siga a distribuição normal assumir valores entre **a** e **b**: esta área é calculada através da integral da função normal de **a** a**b**.

- cada combinação (**μ** , σ^2) resulta em uma distribuição Normal diferente, portanto há uma família infinita de distribuições.

- a função normal citada acima tem a seguinte (e aterradora...) fórmula para sua função densidade de probabilidade:

$$f(x) = \frac{1}{\sqrt{2 \times \pi \times \sigma^2}} \times e^{\left(\frac{-1}{2} \times \left[\frac{x-\mu}{\sigma} \right]^2 \right)} \quad -\infty < x < +\infty$$

Saiba que não existe solução analítica para uma integral da expressão acima: qualquer integral precisa ser resolvida usando métodos numéricos de integração, que são extremamente trabalhosos quando implementados manualmente (somente viáveis se usarem meios computacionais). De Moivre, Laplace e Gauss desenvolveram seus trabalhos entre a metade do século XVIII e início do século XIX, e os computadores começaram a se popularizar a partir da década de 1960, do século XX. LINK Gauss, e todas as outras pessoas que usavam a distribuição Normal para calcular probabilidades até recentemente, resolviam as integrais usando métodos numéricos manualmente. LINK

Todas as distribuições normais apresentam algumas características em comum, porém, independentemente de seus valores de média e de variância:

- 68% dos dados estão situados entre a média menos um desvio padrão (**μ - σ**) e a média mais um desvio padrão (**μ + σ**);

- 95,5% dos dados estão situados entre a média menos dois desvios padrões (**μ - 2σ**) e a média mais dois desvios padrões (**μ + 2σ**);

- 99,7% dos dados estão situados entre a média menos três desvios padrões (**μ - 3σ**) e a média mais três desvios padrões (**μ + 3σ**).

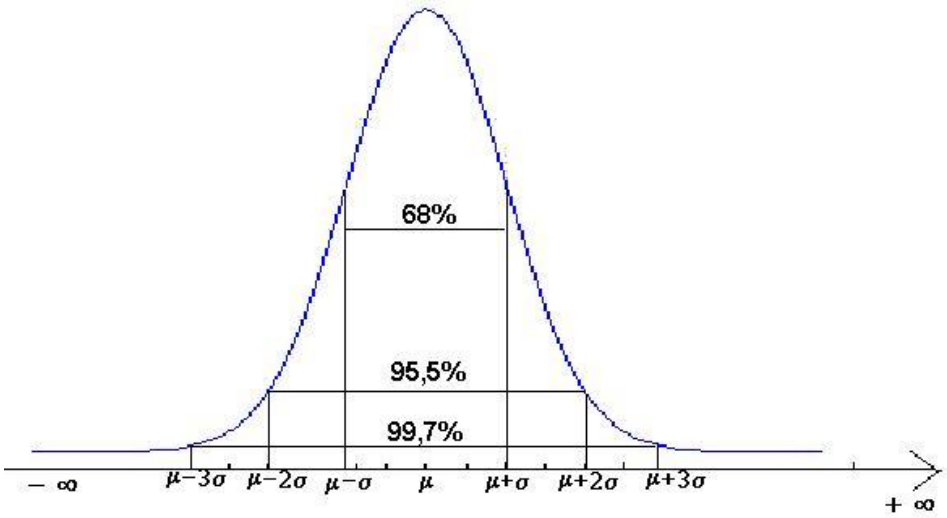


Figura 7 - Percentuais de dados e número de desvios padrões

Fonte: elaborada pelo autor a partir de Statgraphics®

Por causa dessas características alguém teve a idéia de criar um modelo normal padrão: uma variável Z com distribuição normal de média igual a zero e desvio padrão igual a 1 [Z: **N**(0, 1)]. As probabilidades foram calculadas para esta distribuição padrão e registradas em uma tabela. Através de uma transformação de variáveis chamada padronização é possível converter os valores de qualquer distribuição Normal em valores da distribuição Normal padrão e assim obter suas probabilidades - calcular o número de desvios padrões, a contar da média a que está um valor da variável, através da seguinte expressão:

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

Z - número de desvios padrões a partir da média **x** - valor de interesse

μ - média da distribuição normal de interesse **σ** - desvio padrão da distribuição normal

Z é um valor relativo: será negativo para valores de **x** menores do que a média, e será positivo para valores de **x** maiores do que a média. Pela transformação uma

distribuição Normal qualquer **X**: **N** (μ ,σ^2) passa a ser equivalente à distribuição Normal padrão Z: **N**(0 ,1), um valor de interesse x pode ser convertido em um valor z.

As probabilidades de uma variável com distribuição normal podem ser representadas por áreas sob a curva da distribuição normal padrão. No ambiente virtual, apresentamos a Tabela, que relaciona valores positivos de *z*, com áreas sob a cauda superior da curva. Os valores de *z* são apresentados com duas decimais. A primeira decimal fica na coluna da esquerda e a segunda decimal na linha do topo da tabela. A Figura 8 mostra como podemos usar essa Tabela para encontrar, por exemplo, a área sob a cauda superior da curva, além de *z* = 0,21.

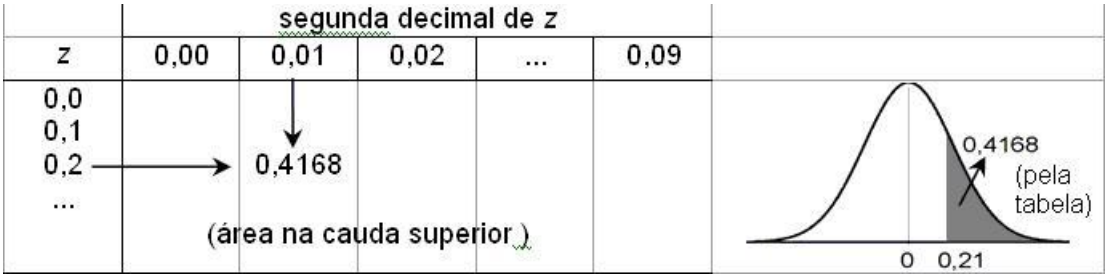


Figura 8 - Ilustração do uso da tabela da distribuição normal padrão (Tabela III do apêndice) para encontrar a área na cauda superior relativa ao valor de z = 0,21.

Fonte: Barbetta, Reis, Bornia(2010).

Exemplo 5 - Suponha uma variável aleatória contínua X, que tenha uma distribuição normal com média 50 e desvio padrão 10. Há interesse em calcular as probabilidades dos seguintes eventos:

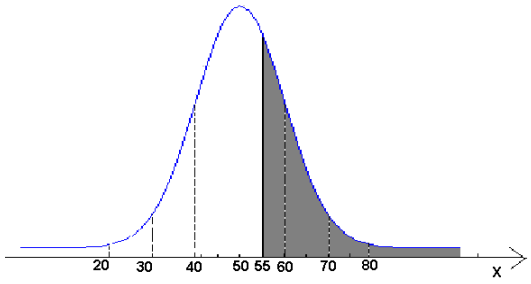
a) $X > 55$

b) $X < 50$

c) $X > 35$

d) $48 < X < 56$

a) Primeiramente, calculamos o valor de Z correspondente a 55. Z = (55 - 50)/ 10 = + 0,5.Pelas Figuras9 e 10pode-se ver a correspondência entre as duas distribuições:



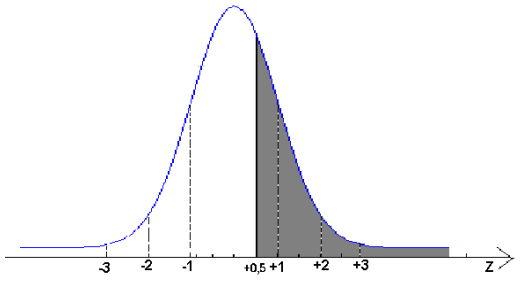
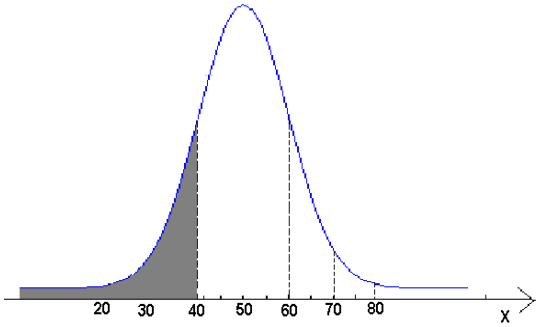


Figura 9–P(X>55) Figura 10–P(Z > 0,5)

Fonte: elaboradas pelo autor a partir de Stagraphics ®

O evento P (X>55) é equivalente ao evento P (Z> 0,5). Este valor pode ser obtido na tabela da distribuição normal padrão (ver ambiente virtual). Os valores de Z são apresentados com dois decimais: o primeiro na coluna da extrema esquerda e o segundo na linha do topo da tabela. Observe pelas Figurass que estão no alto da tabela que as probabilidades são para eventos do tipo do da Figuras acima [P(Z> z_1)]. Assim, poderíamos procurar a probabilidade do evento (Z > 0,5): fazendo o cruzamento do valor 0,5 (na coluna) com o valor 0,00 (na linha do topo) encontramos o valor 0,3085 (30,85%). Portanto, P(X>55) é igual a 0,3085. Observe a coerência entre o valor encontrado e as áreas na Figuras: a área é menor do que a metade da Figuras (metade da Figuras significaria 50%), e a probabilidade encontrada vale 30,85%.

b) Precisamos calcular o valor de Z correspondente a 40. Z = (40 - 50)/ 10 = -1,00.Pelas Figuras11 e 12 podemos ver a correspondência entre as duas distribuições:



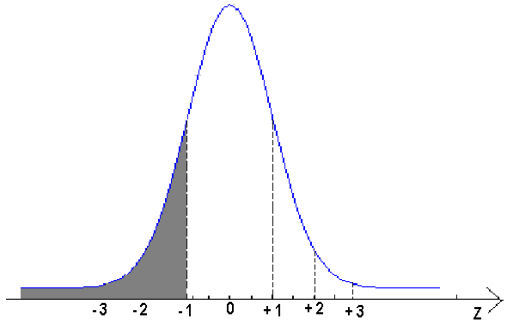


Figura 11–P(X<40) Figura 12–P(Z < -1,00)

Fonte: elaboradas pelo autor a partir de Statgraphics ®

O evento P (X<40) é equivalente ao evento P (Z < -1,00). Repare, porém, que queremos encontrar P (Z < -1,00), e a tabela nos apresenta valores apenas para P (Z > 1,00).

Contudo, se rebatermos a Figura 12(da distribuição normal padrão com Z < -1,00) para a direita teremos o seguinte resultado (Figura 13):

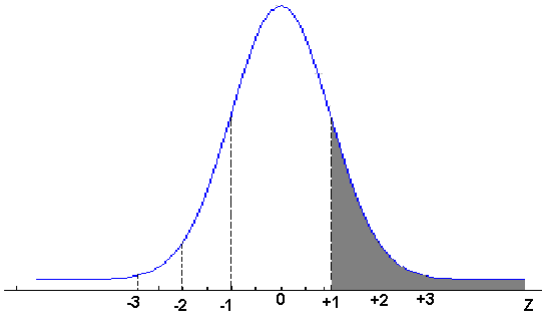
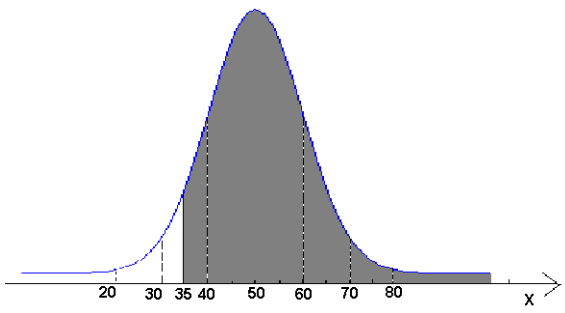


Figura 13–P (Z > 1,00) (rebatimento de P(Z< -1,00)

Fonte: elaborada pelo autor a partir de Stagraphics ®

Ou seja, a área P(Z < -1) = P(Z > 1). Esta probabilidade nós podemos encontrar diretamente pela tabela, fazendo o cruzamento do valor 1,0 (na coluna) com o valor 0,00 (na linha do topo) encontramos o valor 0,1587 (15,87%). Portanto,
P(X<40) = P(Z<-1) = P(Z>1), que é igual a 0,1587.

c) Agora há interesse em calcular a probabilidade de que X seja maior do que 35. É preciso calcular o valor de Z correspondente a 35. Z = (35 - 50)/ 10 = -1,50.Pelas Figuras14 e 15 se pode ver a correspondência entre as duas distribuições:



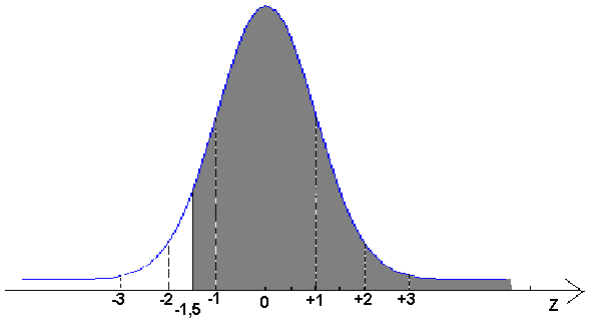


Figura 14–P(X > 35) Figura 15–P(Z > -1,50)

Fonte: elaboradas pelo autor a partir de Statgraphics ®

Não podemos obter a probabilidade P(Z>-1,50) diretamente, pois a tabela do ambiente virtual apresenta apenas resultados para valores positivos de Z. Sabemos que a

probabilidade total vale 1,0, podemos então considerar que P(Z > -1,50) = 1 - P(Z < -1,50). Usando o raciocínio descrito na letra b (rebatendo a Figura 15 para a direita), vamos obter: P(Z<-1,50) = P(Z>1,50). Esta última probabilidade pode ser facilmente encontrada na tabela da distribuição normal padrão: P(Z>1,50) = P(Z<-1,50) = 0,0668. Basta substituir na expressão: P(Z > -1,50) = 1 - P(Z < -1,50) = 1 - 0,0668 = 0,9332 (93,32%). Observe novamente a coerência entre as áreas da Figuras acima e o valor da probabilidade: a área na Figuras compreende mais do que 50% da probabilidade total, aproximando-se do extremo inferior da distribuição, perto de 100%, e a probabilidade encontrada realmente é próxima de 100%.

d) Agora há interesse em calcular a probabilidade de que X assuma valores entre 48 e 56. Calcular P (48 < X < 56), veja a Figura49 abaixo:

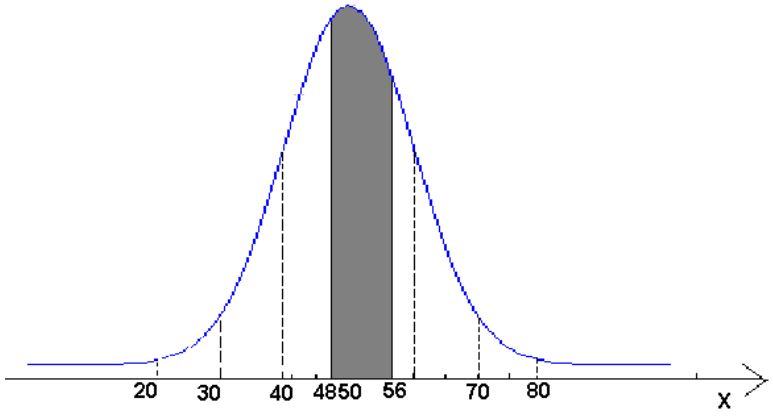


Figura 16–P(48 < X < 56)

Fonte: elaborada pelo autor a partir de Statgraphics ®

Novamente precisamos calcular os valores de Z correspondentes a 48 e a 56.

$Z_1 = (48 - 50)/ 10 = - 0,20$ $Z_2 = (56 - 50)/ 10 = 0,60$

Então: P (48 < X < 56) = P (-0,20<Z<0,60)

Repare que a área entre 48 e 56 é igual à área de 48 até +∞ MENOS a área de 56 até +∞:

P(48 < X < 56) = P(X > 48) - P(X > 56)= P(-0,20< Z <0,60) = P(Z > - 0,20) - P(Z > 0,60)

E os valores acima podem ser obtidos na tabela da distribuição normal padrão:

P(Z > 0,60) = 0,2743 P(Z > -0,20) = 1- P(Z > 0,20) = 1- 0,4207 = 0,5793

P(48 < X < 56)= P(-0,20< Z <0,60) = P(Z > -0,20) - P(Z > 0,60) = 0,5793 - 0,2743 = 0,3050

Então a probabilidade da variável X assumir valores entre 48 e 56 é igual a 0,305 (30,5%).

A distribuição Normal também pode ser utilizada para encontrar valores da variável de interesse correspondentes a uma probabilidade fixada.

Exemplo 6 - Supondo a mesma variável aleatória X com média 50 e desvio padrão 10. Encontre os valores de X, situados à mesma distância abaixo e acima da média, que contém 95% dos valores da variável.

Como a distribuição Normal é simétrica em relação à média, e como neste problema os valores de interesse estão situados à mesma distância da média "sobram" 5% dos valores, 2,5% na cauda inferior e 2,5% na superior, como na Figura17:

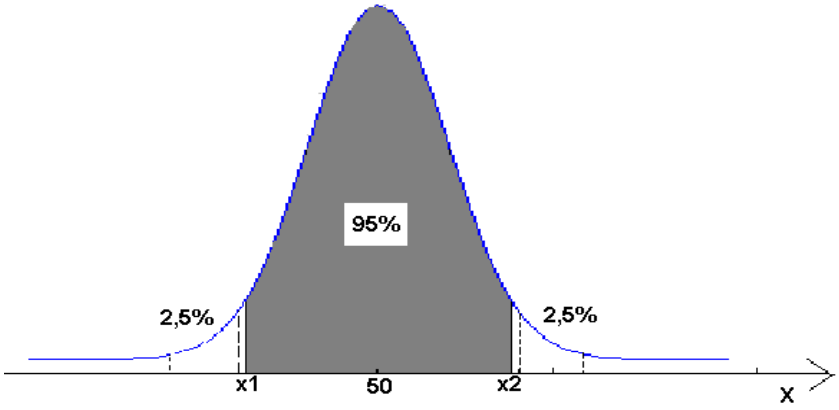


Figura 17–P(x_1< X < x_2) = 0,95

Fonte: elaborada pelo autor a partir de Statgraphics ®

É preciso encontrar os valores de Z (na tabela da distribuição Normal padrão) correspondentes às probabilidades da Figura acima, e a partir daí obter os valores de $\mathbf{x_1}$ e $\mathbf{x_2}$. Passando para a distribuição Normal padrão $\mathbf{x_1}$ corresponderá a um valor $\mathbf{z_1}$, e $\mathbf{x_2}$ a um valor $\mathbf{z_2}$, como na Figura18:

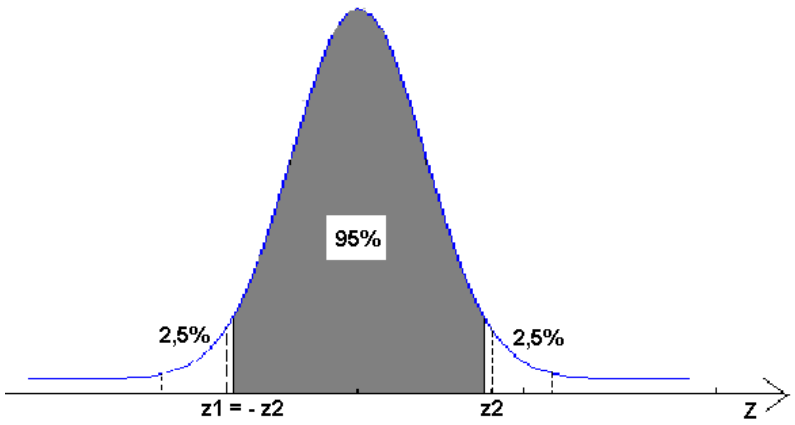


Figura 18–P(-z_2< Z < z_2) = 0,95

Fonte: elaborada pelo autor a partir de Statgraphics ®

Repare que a média da distribuição Normal padrão é igual a zero, fazendo com que $\mathbf{z_1}$ e $\mathbf{z_2}$ sejam iguais em módulo. Podemos encontrar $\mathbf{z_2}$, já que P(Z >$\mathbf{z_2}$) = 0,025

É necessário encontrar o valor da probabilidade na tabela da distribuição Normal padrão (ou o valor mais próximo) e obter o valor de Z associado.Para o caso de $\mathbf{z_2}$, ao procurar pela probabilidade 0,025 encontramos o valor exato 0,025, e, por conseguinte, o valor de $\mathbf{z_2}$ que é igual a 1,96: P (Z > 1,96) = 0,025.

Como $\mathbf{z_1 = -z_2}$, encontramos facilmente o valor de z1: $\mathbf{z_1}$ = -1,96. P (Z < -1,96) = 0,025.Observe que os valores são iguais em módulo, mas corresponderão a valores diferentes da variável X. A expressão usada para obter o valor de Z, em função do valor da variável X, pode ser usada para o inverso:

$$Z = \frac{x - \mu}{\sigma} \quad => \quad x = \mu + Z \times \sigma$$

E assim obteremos os valores de $\mathbf{x_1}$ e $\mathbf{x_2}$,LINK É muito importante que se preste atenção no sinal do valor de **z** ao obter o valor de **x**. LINK Observe se o resultado obtido faz sentido que correspondem a $\mathbf{z_1}$ e $\mathbf{z_2}$, respectivamente:

$$\mathbf{x_1 = \mu + (z_1 \times \sigma) = 50 + [(-1,96) \times 10] = 30,4}$$

$$\mathbf{x_2 = \mu + (z_2 \times \sigma) = 50 + (1,96 \times 10) = 69,6}$$

Observe que os resultados obtidos são coerentes: 30,4 está abaixo da média (1,96 desvios padrões) e 69,6 acima (também 1,96 desvios padrões). O intervalo definido por estes dois valores compreende 95% dos resultados da variável X.

Todo este trabalho poderia ter sido poupado se houvesse um programa computacional que fizesse esses cálculos. Há vários softwares disponíveis no mercado, alguns deles de domínio público, que calculam as probabilidades associadas a determinados eventos, como também os valores associados a determinadas probabilidades.

Uma das características mais importantes do modelo normal é a sua capacidade de aproximar outros modelos, permitindo muitas vezes simplificar os cálculos de probabilidade. Na próxima seção vamos ver como o modelo normal pode ser usado para aproximar o binomial. GLOSSÁRIO Modelo binomial: modelo probabilístico para variáveis aleatórias discretas que descreve o número de sucessos em n experimentos independentes (sendo n finito e conhecido), sendo que os experimentos podem ter apenas dois resultados possíveis, e a probabilidade de sucesso permanece constante durante os n experimentos. Fonte: Barbetta, Reis e Bornia, 2010; Lopes, 1999. FimGLOSSÁRIO

2.2.3 – Modelo normal como aproximação do binomial

O modelo Binomial (discreto) pode ser aproximado pelo modelo Normal (contínuo) se certas condições forem satisfeitas:

- quando o valor de **n** (número de ensaios) for tal que os cálculos binomiais trabalhosos demais. LINK Para os que pensam que o advento dos computadores eliminou este problema um alerta: em alguns casos os números envolvidos são tão grandes que sobrepujam suas capacidades. LINK
- quando o produto **n × p** (o valor esperado do modelo Binomial) e o produto **n × (1 - p)** forem ambos maiores ou iguais a 5.

Se isso ocorrer, uma binomial, de parâmetros **n** e **p,** pode ser aproximada por uma normal com:

média = μ = n × p (valor esperado do modelo Binomial)

variância = σ^2 = n × p × (1- p) (variância do modelo Binomial)

Usando o modelo Normal (contínuo) para aproximar o Binomial (discreto) é necessário fazer uma correção de continuidade: associar um intervalo ao valor discreto, para que o valor da probabilidade calculada pelo modelo contínuo seja mensurável. Este intervalo deve ser centrado no valor discreto, e deve ter uma amplitude igual à diferença entre dois valores consecutivos da variável discreta: se por exemplo a diferença for igual a 1 (a variável somente pode assumir valores inteiros) o intervalo deve ter amplitude igual a 1, 0,5 abaixo do valor e 0,5 acima**. Esta correção de continuidade precisa ser feita para garantir a coerência da aproximação.**

Seja uma variável aleatória X com distribuição Binomial.

1) Há interesse em calcular a probabilidade de X assumir um valor **k** genérico, P(X = **k**), ao fazer a aproximação pela Normal será: P(**k - 0,5**< X <**k + 0,5**).

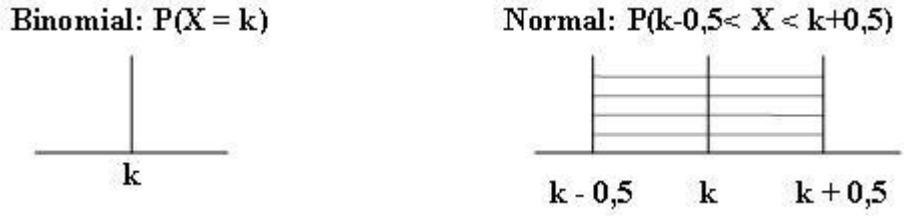


Figura 18 - Correção de continuidade da aproximação do modelo Binomial pelo Normal – 1º caso.

Fonte: elaborada pelo autor

2) Há interesse em calcular a probabilidade de X assumir valores menores ou iguais a um valor **k** genérico, P(X ≤**k**), ao fazer a aproximação pela Normal será: P(X <**k + 0,5**), todo o intervalo referente a k será incluído.

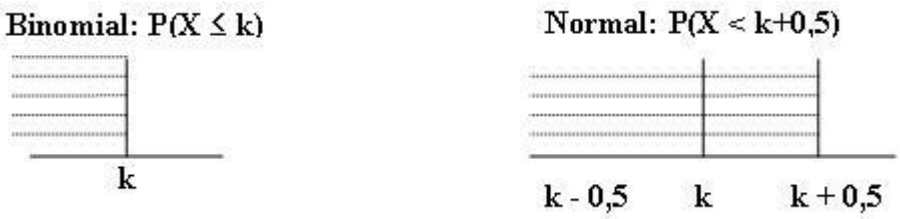


Figura 19 - Correção de continuidade da aproximação do modelo Binomial pelo Normal – 2º caso.

Fonte: elaborada pelo autor

3) Há interesse em calcular a probabilidade de X assumir valores maiores ou iguais a um valor **k** genérico, P(X ≥**k**), ao fazer a aproximação pela Normal será: P(X >**k - 0,5**), todo o intervalo referente a k será incluído.

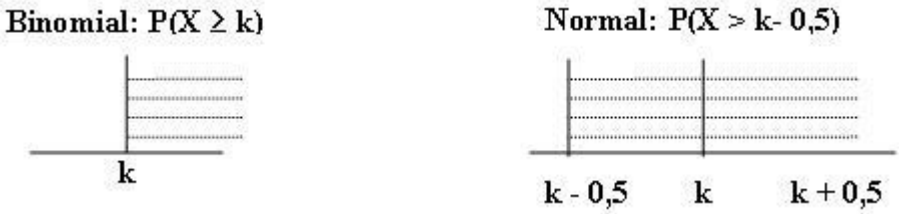


Figura 20 - Correção de continuidade da aproximação do modelo Binomial pelo Normal – 3º caso.

Fonte: elaborada pelo autor

4) Há interesse em calcular a probabilidade de X assumir valores menores do que um valor **k** genérico, P(X <**k**), ao fazer a aproximação pela Normal será: P(X <**k - 0,5**), todo o intervalo referente a k será excluído.

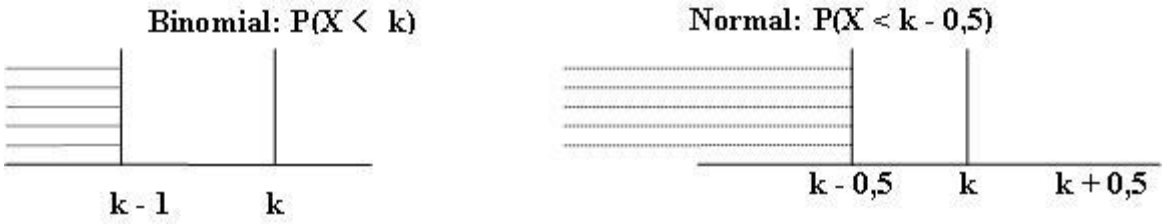


Figura 21 - Correção de continuidade da aproximação do modelo Binomial pelo Normal – 4º caso.

Fonte: elaborada pelo autor

5) Há interesse em calcular a probabilidade de X assumir valores maiores do que um valor **k** genérico, P(X >**k**), ao fazer a aproximação pela Normal será: P(X >**k + 0,5**), todo o intervalo referente a k será excluído.

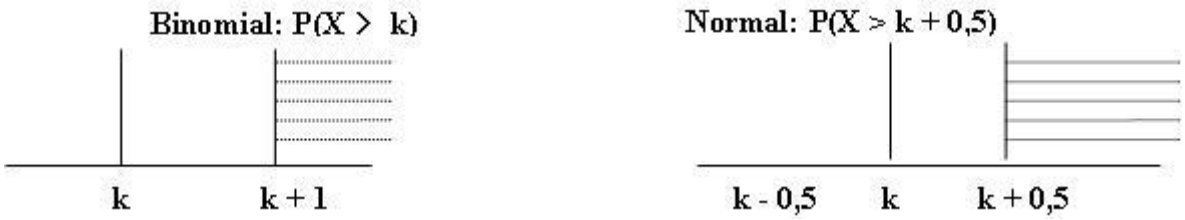


Figura 22 - Correção de continuidade da aproximação do modelo Binomial pelo Normal – 5º caso.

Fonte: elaborada pelo autor

Exemplo 7 - Um município tem 40000 eleitores. Para uma pesquisa de opinião eleitoral uma amostra aleatória de 1500 pessoas foi selecionada.Qual é a probabilidade de que pelo menos 500 dos eleitores seja menor de 25 anos se 35% dos 40000 são menores do que 25 anos?

Este problema poderia ser resolvido usando o modelo Binomial. Há apenas dois resultados possíveis para cada eleitor: menor de 25 anos (“sucesso”) e maior ou igual a 25 anos (“fracasso”). Existe um limite superior de realizações, no caso os 1500 eleitores da amostra, e há independência entre as retiradas, pois a amostra foi retirada de forma aleatória (e a amostra representa menos de 5% dos 40000 eleitores).

Então: “sucesso” = menor de 25 anos **p** = 0,35 **1 - p** = 0,65 **n** = 1500

A variável aleatória discreta X, número de eleitores menores de 25 anos em 1500, terá distribuição binomial com parâmetros **n** = 1500 e **p** = 0,35.

O evento “pelo menos 500 menores de 25 anos” seria definido como 500 ou mais eleitores:

$$P(X \geq 500) = P(X = 500) + P(X = 501) + ...+ P(X = 1500)$$

Há cerca de 1000 expressões binomiais.

Vamos ver se é possível aproximar pelo modelo Normal.

O valor de **n** é grande:**n**×**p** = 1500 × 0,35 = 525 > 5e **n**×(**1 - p**) = 1500 × 0,65 = 975 > 5.

Como as condições foram satisfeitas é possível aproximar por um modelo Normal:

média = **μ** = **n** × **p** = **1500 x 0,35 = 525**

desvio padrão = **σ** = $\sqrt{n \times p \times (1-p)} = \sqrt{1500 \times 0{,}35 \times 0{,}65} = 18{,}47$

Pelo modelo Binomial: P (X ≥ 500). Pelo modelo Normal será: P (X ≥ 499,5).

P(X ≥ 499,5) = P(Z > z_1)z_1 = (499,5 - 525)/18,47 = -1,38 P(Z > -1,38) = 1 - P(Z >1,38)

Procurando na tabela da distribuição Normal padrão: P (Z > 1,38) = 0,0838

Então: P (X ≥ 500) ≅ P(X ≥ 499,5) = P (Z > -1,38) = 1 - P(Z >1,38) = 1 - 0,0838 = 0,9162.
A probabilidade de que pelo menos 500 dos eleitores da amostra sejam menores de 25 anos é igual a 0,9162 (91,62%).

Nas próximas duas seções vamos ver modelos probabilísticos derivados do modelo normal, usados predominantemente em processos de inferência estatística. Vamos introduzi-los agora para facilitar nosso trabalho quando chegarmos às Unidades 5 e 6.

2.2.4 – Modelo (distribuição) t de Student

Havia um matemático inglês, William Gosset, que trabalhava para a cervejaria Guiness, em Dublin, Irlanda, no início do século XX. Ele atuava no controle da qualidade do cultivo de ingredientes para a fabricação de cerveja.

Nesta época alguns estatísticos usavam a distribuição normal no estabelecimento de intervalos de confiança para a média a partir de pequenas amostras (veremos isso na Unidade 5). Calculavam média aritmética simples e variância da amostra e generalizavam os resultados através do modelo normal, como fizemos no Exemplo 7.

Gosset descobriu que o modelo normal não funcionava direito para pequenas amostras, e desenvolveu um novo modelo probabilístico, derivado do normal, introduzindo uma correção para levar em conta justamente o tamanho de amostra. Ele aplicou suas descobertas em seu trabalho, e quis publicá-las, mas a Guiness apenas permitiu após ele adotar o pseudônimo "Student". Por isso, o seu modelo é conhecido como t de Student para n – 1 graus de liberdade.

O valor **n -1** (tamanho da amostra menos 1) é chamado de número de **graus de liberdade** da estatística. Quando a variância amostral é calculada supõe-se que a média já seja conhecida, assim apenas um determinado número de elementos da amostra poderá ter seus valores variando livremente, este número será igual a **n - 1**, porque um dos valores não poderá variar livremente, pois terá que ter um valor tal que a média permaneça a mesma calculada anteriormente. Assim, a estatística terá **n - 1** graus de liberdade.

Trata-se de uma distribuição de probabilidades que apresenta média igual a zero (como a normal padrão), é simétrica em relação à média, mas apresenta uma variância igual a n / (n - 2), ou seja seus valores dependem do tamanho da amostra, apresentando maior variância para menores valores de amostra. LINK Esta é a correção propriamente dita, pois ao usar pequenas amostras o risco de que a variância amostral da variável seja diferente da variância populacional é maior, podendo levar a intervalos de confiança que não correspondem à realidade. A não utilização desta correção foi a fonte de muitos erros no passado, e, infelizmente, de ainda alguns erros no presente. LINK Quanto maior o tamanho da amostra mais a variância de **t** aproxima-se de 1,00 (variância da normal padrão). LINK Para tamanhos de amostra maiores do que 30 supõe-se que a variância de **t** é igual a 1: por isso a aproximação do item b.1. LINK A distribuição t de Student está na Figura23 para vários graus de liberdade:

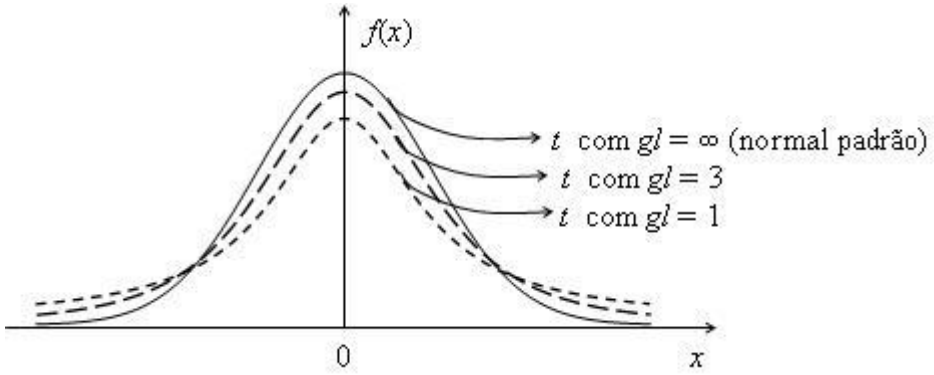


Figura 23 - Distribuição t de Student para vários graus de liberdade

Fonte: Barbetta, Reis, Bornia (2010)

Observe que tal como a distribuição normal padrão a distribuição **t** de Student é **simétrica** em relação à média (que é igual a zero).

A tabela da distribuição **t** de Student encontra-se no ambiente virtual, para vários graus de liberdade e valores de probabilidade.

Exemplo 8. Imagine a situação do Exemplo 7, obter os valores de t simétricos em relação à média que contêm 95% dos dados, supondo uma amostra de 10 elementos.

Temos que encontrar os valores $\mathbf{t_1}$ e $\mathbf{t_2}$, simétricos em relação à média que definem o intervalo que contém 95% dos dados. Como supomos uma amostra de 10 elementos a distribuição t de Student terá 10 – 1 = 9 graus de liberdade. Repare que a média da distribuição t de Student é igual a zero, fazendo com que $\mathbf{t_1}$ e $\mathbf{t_2}$ sejam iguais em módulo. Podemos encontrar $\mathbf{t_2}$, já que $P(t > \mathbf{t_2}) = 0{,}025$. Veja a Figura24:

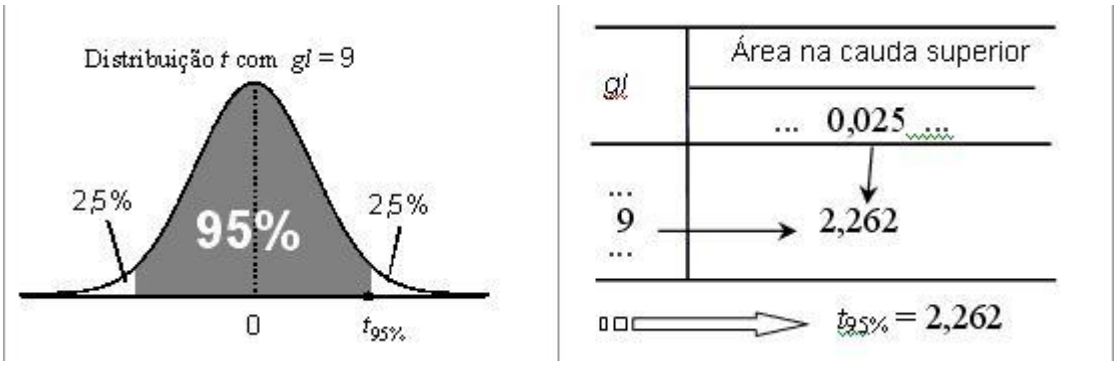


Figura 24 - Uso da tabela da distribuição t de Student. Ilustração com gl = 9 e área na cauda superior de 2,5%

Fonte: Barbetta, Reis, Bornia (2010)

Vamos utilizar bastante a distribuição t de Student nas Unidades 5 e 6.

2.2.5 – Modelo quiquadrado

Trata-se de mais um modelo derivado da distribuição normal, embora não vamos discutir como se dá esta derivação aqui.

Na Unidade 2 de Estatística Aplicada à Administração I estudamos como descrever os relacionamentos entre duas variáveis qualitativas, geralmente expresso através de uma tabela de contingências. No Quadro 4daquela Unidade analisamos o relacionamento entre modelo e opinião geral sobre os veículos da Toyord. Havíamos concluído que havia relacionamento, pois os modelos mais baratos apresentavam maiores percentuais de insatisfeitos do que os mais caros.

Na Unidade 6 vamos aprender a calcular uma estatística que relacionará as frequências observadas de cada cruzamento entre os valores de duas variáveis qualitativas, expressas em uma tabela de contingências, com as frequências esperadas desses mesmos cruzamentos, se as duas variáveis não tivessem qualquer relacionamento entre si. Esta estatística é chamada de quiquadrado, $\chi 2$, e caso a hipótese de que as variáveis não se relacionem ela seguirá o modelo quiquadrado com um certo número de graus de liberdade.

O número de graus de liberdade dependerá das condições da tabela: para o caso que será visto na Unidade 10 será o produto do número de linhas da tabela – 1 pelo número de colunas da tabela – 1. É uma distribuição assimétrica, sempre positiva, que tem valores diferentes dependendo do seu número de graus de liberdade. Sua média é igual ao número de graus de liberdade, e a variância é igual a duas vezes o número de graus de liberdade.

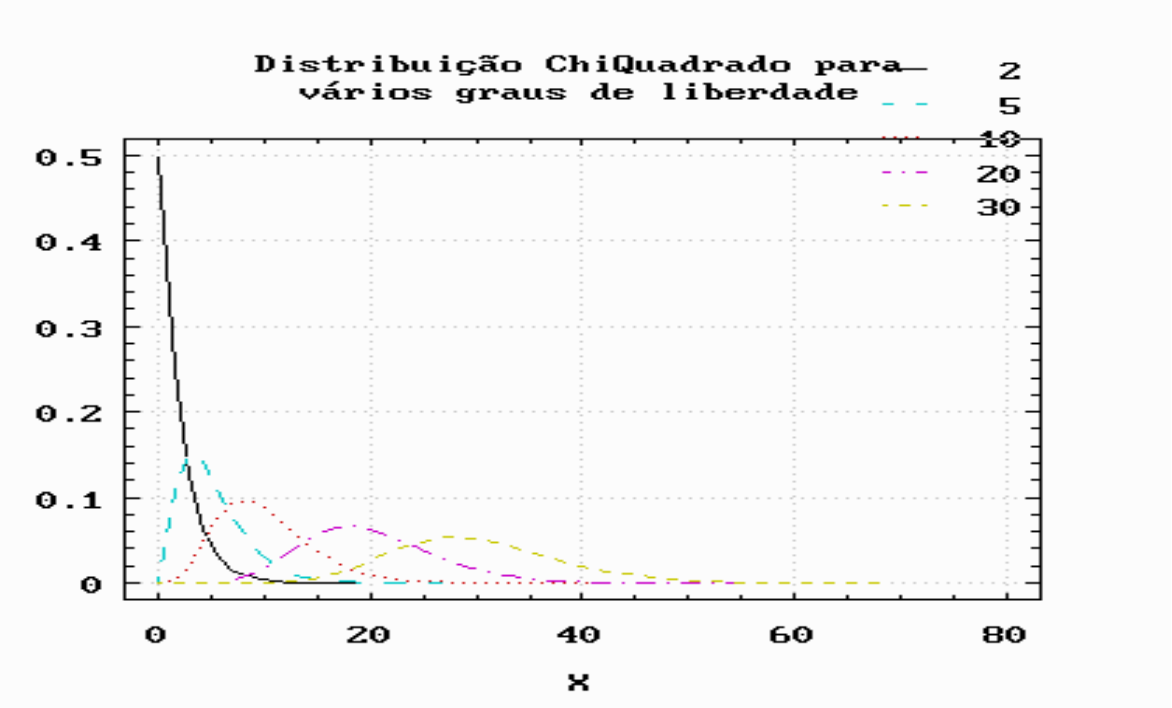


Figura 25 - Modelo quiquadrado com 2, 5, 10, 20 e 30graus de liberdade

Fonte: adaptada pelo autor de Stagraphics ®

A Figura 25 mostra as curvas do modelo (distribuição) quiquadrado para 2, 5, 10, 20 e 30 graus de liberdade. Observe como variam de forma dependendo do número de graus de liberdade da estatística.

A tabela da distribuição quiquadradoencontra-se no Ambiente Virtual de Ensino-Aprendizagem, para vários graus de liberdade e valores de probabilidade. Vamos ver um exemplo.

Exemplo 9 - Imagine que queremos encontrar o valor da estatística quiquadrado, para 3 graus de liberdade, deixando uma área na cauda superior de 5%.

O valor da estatística quiquadrado que define uma área na cauda superior de 5% pode ser encontrado através da Tabela, cruzando a linha de 3 graus de liberdade com a coluna de área na cauda superior igual a 0,05. Veja a Figura 26 a seguir:

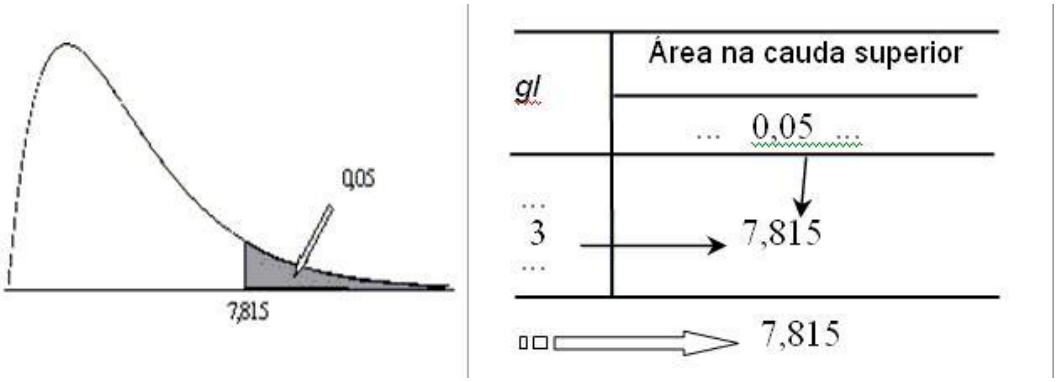


Figura 26 - Uso da tabela da distribuição quiquadrado. Ilustração com gl = 3 e área na cauda superior de 5%

Fonte: adaptado pelo autor de Barbetta, Reis, Bornia (2010)

2.3 – Modelos probabilísticosem planilhas eletrônicas

Atualmente todas as planilhas eletrônicas têm os principais modelos probabilísticos disponíveis, permitindo realizar cálculos de probabilidades ou obtenção de escores com facilidade, e praticamente eliminando a necessidade de aproximações ou tabelas.

Para os modelos binomial, Poisson, normal, t de Student e quiquadrado a planilha eletrônica Microsoft Excel ® dispõe de várias funções que permitem realizar os cálculos apresentados nos exemplos desta unidade. A seguir serão apresentadas as principais funções com os argumentos necessários para realizar os cálculos. Elas podem ser usadas mesmo nas versões mais antigas do Excel ®, embora nas mais recentes haja outras com uma sintaxe um pouco diferente.

Para uma variável aleatória X que siga um modelo binomial de parâmetros n e p, supondo um valor xi qualquer[1]:

$P(X = x_i)$ = DISTRBINOM(x_i;n;p;FALSO)

A função acima permitirá calcular a probabilidade de X ser exatamente igual a x_i. Se quisermos a probabilidade acumulada até x_i, basta fazer uma pequena modificação:

$P(X \leq x_i)$ = DISTRBINOM(x_i;n;p;VERDADEIRO)

Exemplo 10 - Estudos anteriores mostraram que há 73% de chance de consumidoras apresentarem uma reação positiva a anúncios publicitários com crianças. Uma agência apresentou um novo anúncio para 5 consumidoras. Qual é a probabilidade de que pelo menos 3 das 5 consumidoras apresentem reação positiva?

Para cada consumidora ("ensaio") há apenas 2 resultados: reação positiva ou não. Há um número finito de realizações, n = 5. Definindo "sucesso" como reação positiva, e considerando as consumidoras "independentes", a variável aleatória X, número de consumidoras com reação positiva em 5 que assistiram o novo anúncio terá distribuição binomial com parâmetros n = 5 e p = 0,73 (e 1- p = 0,27).
Evento de interesse: $X \geq 3$

$P(X \geq 3) = P(X=3) + P(X=4) + P(X=5)$
Pela fórmula binomial:

$$P(X = 3) = C_{5,3} \times 0{,}73^3 \times (0{,}27)^2 = \frac{5!}{3! \times (5-3)!} \times 0{,}73^3 \times (0{,}27)^2 = 0{,}284$$

$$P(X = 4) = C_{5,4} \times 0{,}73^4 \times (0{,}27)^1 = \frac{5!}{4! \times (5-4)!} \times 0{,}73^4 \times (0{,}27)^1 = 0{,}383$$

$$P(X = 5) = C_{5,5} \times 0{,}73^5 \times (0{,}27)^0 = \frac{5!}{5! \times (5-5)!} \times 0{,}73^5 \times (0{,}27)^0 = 0{,}207$$

[1] Para inserir qualquer fórmula no Excel é preciso selecionar uma célula e digitar =, seguido da fórmula/função desejada. Maiores detalhes em https://www.youtube.com/watch?v=gVH1VxpZ5iQ

$P(X \geq 3) = P(X = 3) + P(X = 4) + P(X = 5) = 0{,}284 + 0{,}383 + 0{,}207 = 0{,}874$

Pelo Excel ®, lembrando da propriedade do evento complementar:

$P(X \geq 3) = 1 - P(X < 3) = 1 - P(X \leq 2)$

Então: P(X ≥ 3) = 1 - DISTRBINOM(2;5;0,73;VERDADEIRO) = 1- 0,126 = 0,874

Para uma variável aleatória X que siga um modelo de Poisson com parâmetro $m = \lambda \times t$, supondo um valor x_i qualquer:

$P(X = x_i)$ = POISSON(x_i;m;FALSO)

A função acima permitirá calcular a probabilidade de X ser exatamente igual a x_i. Se quisermos a probabilidade acumulada até x_i, basta novamente fazer uma pequena modificação:

$P(X \leq x_i)$ = POISSON(x_i;m;VERDADEIRO)

Exemplo 11 - Em um porto estudos históricos mostram que chegam em média 2 navios por dia, de acordo com a distribuição de Poisson. Sabendo que o porto pode atender apenas 2 navios por dia, calcule a probabilidade de navios que chegarem em um determinado dia não serem atendidos.

A variável discreta número de navios que chegam em um dia ao porto segue uma distribuição de Poisson com $\lambda = 2$ navios/dia. O período de análise (para cálculo de probabilidade é um dia – determinado dia) t é igual a 1. Então $m = \lambda \times t = 2 \times 1 = 2$. Se *mais de 2* navios chegarem em um dia eles não serão atendidos, porque o porto pode atender apenas 2. Então procura-se:

$P(X > 2) = P(X = 3) + P(X = 4) + ...$ (não há limite superior)

Tal como está o problema não pode ser resolvido, temos que usar a propriedade do evento complementar:

$P(X > 2) = 1 - P(X \leq 2) = 1 - P(X = 0) - P(X = 1) - P(X = 2)$

Pela formula de Poisson:

$$P(X = 0) = \frac{e^{-2\times1} \times (2)^0}{0!} = 0{,}1353$$

$$P(X = 1) = \frac{e^{-2\times1} \times (2)^1}{1!} = 0{,}2707$$

$$P(X = 2) = \frac{e^{-2\times1} \times (2)^2}{2!} = 0{,}2707$$

$P(X>2)=1 - 0{,}1353 - 0{,}2707 - 0{,}2707 = 0{,}3233$

Pelo Excel ®, lembrando da propriedade do evento complementar:

P(X > 2) = 1 – POISSON(2;2;VERDADEIRO) = 0,3233

Para uma variável aleatória X que siga um modelo normal com média μ e desvio padrão σ, e para dois valores quaisquer x_1 e x_2 (sendo $x_2 > x_1$):

P(X ≤ x_1) = DISTNORM(x_1; μ; σ; VERDADEIRO)

P(X ≤ x_2) = DISTNORM(x_2; μ; σ; VERDADEIRO)

P(x_1≤X≤x_2)=DISTNORM(x_2;μ;σ;VERDADEIRO) - DISTNORM(x_1;μ;σ;VERDADEIRO)

Lembrando do Exemplo 5, item d, em que se procurava P(48 < X < 56). Pelo Excel, basta obter a probabilidade acumulada até 56 e subtrair a acumulada até 48:

P(48<X<56) =

DISTNORM(56;50;10;VERDADEIRO) –DISTNORM(48;50;10;VERDADEIRO) =

0,3050

Para uma variável aleatória X que siga um modelo normal com média μ e desvio padrão σ, se quisermos encontrar o valor de x_i correspondente a uma determinada probabilidade acumulada α:

x_i = INV.NORM(α; μ; σ)

Lembrando do Exemplo 6, em que supondo a mesma variável aleatória X com média 50 e desvio padrão 10. Encontre os valores de X (x_1 e x_2), situados à mesma distância abaixo e acima da média, que contém 95% dos valores da variável. Se entre os valores há 95%, e estão á mesma distância da média, então abaixo do primeiro valor há 2,5% ((100% - 95%)/2), e abaixo do segundo também há 2,5% + 95% totalizando 97,5%.

$P(x_1<X<x_2)=0,95$, $P(X<x_1)= 0,025$, $P(X<x_2)= 0,975$

Através do Excel:

x_1 = INV.NORM(0,025;50;10) = 30,4

x_2 = INV.NORM(0,975;50;10) = 69,6

Para uma variável aleatória X que siga um modelo t de Student com gl graus de liberdade, se quisermos encontrar a probabilidade de X ser *maior* do que x_i:

$P(X > x_i)$ =DISTT(x_i;gl;caudas)

Caso haja interesse apenas uma das "caudas" da distribuição t, usar 1 em caudas. Caso haja interesse nas duas caudas, usar 2 em caudas.

Para uma variável aleatória X que siga um modelo t de Student com gl graus de liberdade, se quisermos encontrar o valor de t que corresponde à soma das probabilidades das caudas (mesma probabilidade para cada lado):

t = INVT(probabilidade;gl)

Lembrando do Exemplo 8, obter os valores de t simétricos em relação à média que contêm 95% dos dados, supondo uma amostra de 10 elementos. Como a amostra tem 10 elementos a distribuição t terá 10 – 1 = 9 graus de liberdade.Se há 95% dentro do intervalo há 5% fora. Através do Excel:

t = INVT(0,05;9) = 2,262

Para uma variável aleatória X que siga um modelo quiquadrado com gl graus de 2liberdade, se quisermos encontrar a probabilidade de que X ser menor do que x_i:

$P(X < x_i)$ = DIST.QUIQUA(x_i;gl)

Para uma variável aleatória X que siga um modelo quiquadrado com gl graus de liberdade, se quisermos encontrar o valor de quiquadrado que corresponde a uma probabilidade na cauda superior:

$\chi 2$ = INV.QUI(probabilidade;gl)

Lembrando do Exemplo 9, queremos encontrar o valor da estatística quiquadrado, para 3 graus de liberdade, deixando uma área na cauda superior de 5%. Pelo Excel:

$\chi 2$ = INV.QUI(0,05;3) = 7,815

Com este tópico terminamos a Unidade 2. Na Unidade 3 você estudará os conceitos e técnicas de amostragem e na Unidade 4 você verá o importante conceito de distribuição amostral.Ambas são indispensáveis para o processo de generalização (inferência) estatística que será estudado nas Unidades 5 e 6.

Tô afim de saber:

- Sobre modelos probabilísticos para variáveis aleatórias discretas -

BARBETTA,P. A. **Estatística Aplicada às Ciências Sociais**. 7ª. ed. – Florianópolis: Ed. da UFSC, 2007, capítulo 7;

BARBETTA, P.A., REIS, M.M., BORNIA, A.C. **Estatística para Cursos de Engenharia e Informática**. 3ª ed. - São Paulo: Atlas, 2010, capítulo 5;

STEVENSON, Willian J. **Estatística Aplicada à Administração**. São Paulo: Ed. Harbra, 2001, capítulo 4.

- Sobre modelos probabilísticos para variáveis aleatórias contínuas -

BARBETTA,P. A. **Estatística Aplicada às Ciências Sociais**. 7ª. ed. – Florianópolis: Ed. da UFSC, 2007. capítulo 8;

BARBETTA, P.A., REIS, M.M., BORNIA, A.C. **Estatística para Cursos de Engenharia e Informática**. 2ª ed. - São Paulo: Atlas, 2010, capítulo 6;

STEVENSON, Willian J. **Estatística Aplicada à Administração**. São Paulo: Ed. Harbra, 2001, capítulo 5.

- Sobre a utilização do Microsoft Excel ® para cálculo de probabilidades para os principais modelos probabilísticos veja LEVINE, D. M., STEPHAN, D., KREHBIEL, T. C., BERENSON, M. L. **Estatística:** Teoria e Aplicações - Usando Microsoft Excel em Português. 5ª ed. – Rio de Janeiro: LTC, 200, capítulos 4 e 5.
- Sobre o uso do Microsoft Excel ® para cálculo de probabilidades para o modelo binomial assistir https://www.youtube.com/watch?v=wddGzOrwup8.
- Sobre o uso do Microsoft Excel ® para cálculo de probabilidades para o modelo normal assistir https://www.youtube.com/watch?v=pR8Yd0ZAXOA.

Resumo

O resumo desta Unidade está mostrado na Figura27:

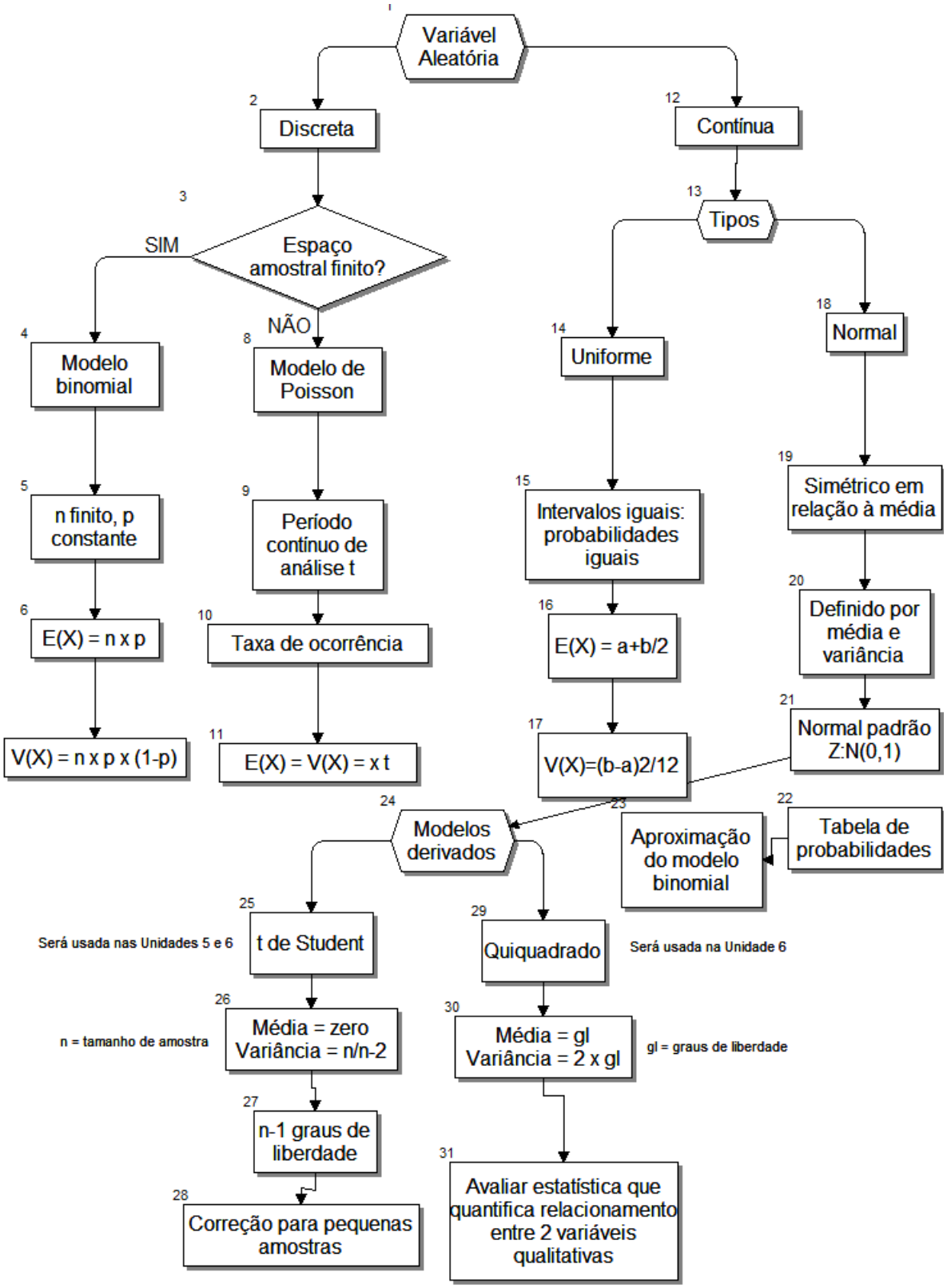


Figura 27 - Resumo da Unidade 2

Fonte: elaborado pelo autor

Atividades de aprendizagem

1) Em um sistema de transmissão de dados existe uma probabilidade igual a 0,05 de um dado ser transmitido erroneamente. Ao se realizar um teste para analisar a confiabilidade do sistema foram transmitidos 20 dados.

a) Qual é o modelo teórico mais adequado para este caso? Por quê?

b) Qual é a probabilidade de que tenha havido erro na transmissão? (R.: 0,6415)

c) Você acha a probabilidade encontrada no item b um valor aceitável? JUSTIFIQUE.

d) Qual é o número esperado de erros no teste realizado? (R.: 1 erro)

2) Suponha que você vai fazer uma prova de TGA com 10 questões do tipo verdadeiro-falso. Você nada sabe sobre o assunto e vai responder as questões por adivinhação.

a) Qual é o modelo probabilístico mais adequado para calcular as probabilidades de acertar um número X de questões dentre as 10? Por quê?

b) Qual é a probabilidade de acertar pelo menos 8 questões? (R.: 0,05468)

Adaptado de DOWNING, D. e CLARK, J.. Estatística Aplicada, São Paulo: Saraiva, 2000.

3) Um revendedor de automóveis novos constatou que 80% dos carros vendidos são devolvidos ao departamento mecânico para corrigir defeitos de fabricação, nos primeiros 25 dias após a venda. De 11 carros vendidos há interesse em calcular as probabilidades de que o número de automóveis que retornam para reparo seja 0, 1, 2, etc.

a) Qual é o modelo teórico mais adequado para este caso? Por quê?

b) Qual é a probabilidade de que todos voltem dentro de 25 dias para reparo? (R.: 0,085899)

c) Qual é a probabilidade de que nenhum volte? (R.: 0,0000002)

d) Uma organização de consumidores pretende processar o revendedor, e a fábrica dos automóveis, se a probabilidade de que a maioria deles (dentre os 11 vendidos) retornar para reparo seja superior a 75%. O revendedor e fábrica devem se preocupar com o processo? JUSTIFIQUE (R.: 0,98834).

Adaptado de STEVENSON, W.J. Estatística Aplicada à Administração, São Paulo: Harper do Brasil, 2001.

4) Em um determinado processo de fabricação 10% das peças são defeituosas. As peças são acondicionadas em caixas com 5 unidades cada uma. As caixas só serão aceitas se apresentarem no máximo uma peça defeituosa. Pergunta-se:

a) Qual é o modelo teórico mais adequado para este caso? Por quê?

b) Qual é a probabilidade de uma caixa ser aceita? (R.: 0,9185)

c) Você considera a probabilidade obtida no item b um valor apropriado? JUSTIFIQUE.

d) Qual é a probabilidade de que em um lote de 10 caixas pelo menos 8 sejam aceitas? (R.: 0,9579)

5) Em uma fábrica 3% dos artigos produzidos são defeituosos. O fabricante pretende vender 4000 peças recebendo 2 propostas:

Proposta 1: o comprador A examina uma amostra de 80 peças e pagará $60 por peça, se houver 3 ou menos defeituosas, caso contrário pagará $30 por peça apenas.

Proposta 2: o comprador examina 40 peças e está disposto a pagar $65 por peça, se todas forem perfeitas, porém pagará $20 por peça se houver alguma peça defeituosa.

Qual é a melhor proposta? JUSTIFIQUE. (R.: proposta 1).

6) Uma comissão responsável pelo recebimento de equipamentos em uma empresa faz testes em equipamentos selecionados aleatoriamente dentre os que chegam. Para avaliar uma determinada marca de transformadores de pequeno porte, a comissão selecionou aleatoriamente 18 dentre os que chegaram e classificará a marca como satisfatória se não existir nenhum defeituoso nesta amostra. Sabe-se que a produção destes equipamentos apresenta um percentual de 6% de defeituosos.

a) Qual é a probabilidade de que a marca venha a ser considerada satisfatória? (R.: 0,328)

b) Você considera a probabilidade encontrada no item a apropriada? JUSTIFIQUE.

7) Em um estudo de reconhecimento de marca, 95% dos consumidores reconheceram o refrigerante “Guaranazinho”. Mas, dentre 15 consumidores selecionados ao acaso apenas 10 reconheceram a marca.

a) Determine a probabilidade de obter no máximo 10 consumidores que reconheceram “Guaranazinho” dentre os 15 selecionados. (R.: 0,0006146)

b) Você acha que o resultado possa ser conseqüência de mero acaso? JUSTIFIQUE.
c) Suponha que será realizada uma nova pesquisa com 1200 pessoas. Determine a média e o desvio padrão do número de consumidores que reconhecem “Guaranazinho”. (R.: 1140; 7,55)
Adaptado de TRIOLA, M. Introdução à Estatística, Rio de Janeiro: LTC, 1999.

8) Certo pequeno município de SC relata que em média nascem 2,25 crianças por dia. Argumentam que tal taxa justificaria a instalação de um hospital com maternidade no local. O governo do estado, com problemas de caixa declara que somente se a probabilidade de nascerem mais de 2 crianças por dia for superior a 50% o hospital será instalado. Calcule as probabilidades apropriadas e responda se o hospital deve ser instalado. JUSTIFIQUE. (R.: $P(X > 2) = 0{,}390660733$)
Adaptado de TRIOLA, M. Introdução à Estatística, Rio de Janeiro: LTC, 1999, página 109.

9) O sistema de atendimento utilizado por uma central telefônica possui telefonistas para atender às chamadas dos usuários. Certa telefonista recebe em média 1,75 chamadas por minuto, durante um turno de trabalho de 6 horas consecutivas. Qual é a probabilidade de que esta telefonista:
a) A telefonista queixou-se ao sindicato que está trabalhando demais, e que precisaria de uma auxiliar. O sindicato concordou em ajudar desde que a probabilidade de ela receber mais de 600 chamadas no turno fosse maior do que 50%. O sindicato deve ajudar a telefonista? JUSTIFIQUE.
b) Qual é a média de chamadas em uma hora e em um turno completo? (R.: 105 chamadas, 630 chamadas)

10) Uma operadora de pedágios está preocupada com o dimensionamento de uma de suas praças. Muitos motoristas estão reclamando das filas, pois há apenas duas gôndolas operando todo o tempo. Estudos mostraram que em média 4 carros chegam na praça de pedágio a cada 15 minutos.
a) Qual é a probabilidade de que mais de 2 carros cheguem à praça em 30 minutos? (R.: 0,9862)

b) Você recomenda que a empresa aumente o número de gôndolas? Por quê?

11) Trace uma curva normal e sombreie a área desejada, obtendo então as probabilidades

a) P(Z > 1,0) (R.: 0,1587) b) P(Z < 1,0) (R.:0,8413) c) P(Z > -0,34) (R.: 0,6331)

d) P(0 < Z < 1,5) (R.: 0,4332) e) P(-2,88 < Z < 0) (R.: 0,498)

f) P(-0,56 < Z < -0,20) (R.: 0,133) g) P(-0,49 < Z < 0,49) (R.: 0,3758)

h) P(2,5 < Z < 2,8) (R.: 0,0036) i) P(Z < -0,2) (R.: 0,4207) j) P(Z > -0,2) (R.:0,5793)

k) P(-0,2 < Z < 0) (R.: 0,0793) l) P(-0,2 < Z < 0,4) (R.: 0,2347)

12) Determine os valores de z_1 que correspondem às seguintes probabilidades:

a) $P(Z > z_1) = 0,0505$ (R.: 1,64) b) $P(Z > z_1) = 0,0228$ (R.: 2) c) $P(Z < z_1) = 0,0228$ (R.: -2)

d) $P(0 < Z < z_1) = 0,4772$ (R.: 2) e) $P(-z_1 < Z < z_1) = 0,95$ (R.: 1,96)

f) $P(Z < z_1) = 0,0110$ (R.: -2,29) g) $P(Z < z_1) = 0,0505$ (R.: -1,64) h) $P(Z < z_1) = 0,5$ (R.: 0)

i) $P(-z_1 < Z < z_1) = 0,6825$ (R.: 1,0) j) $P(-z_1 < Z < z_1) = 0,9544$ (R.: 2,0)

Adaptado de STEVENSON, W.J. Estatística Aplicada à Administração, São Paulo: Harper do Brasil, 2001.

13) Suponha que o escore dos estudantes no vestibular seja uma variável aleatória com distribuição normal com média 550 e variância 900. Se a admissão em certo curso exige um escore mínimo de 575, qual é a probabilidade de um estudante ser admitido? E se o escore mínimo for 540? (R.: 0,2033; 0,6293)

Adaptado de DOWNING, D. e CLARK, J.. Estatística Aplicada, São Paulo: Saraiva, 2000, página 172.

14) Você pode escolher entre 2 empregos. Em uma indústria seus ganhos mensais terão distribuição normal com média de $4000 e desvio padrão de $500. Como vendedor de uma firma seus ganhos mensais terão distribuição normal com média de $3200 e desvio padrão de $2600.

a) Você ganha atualmente (salário fixo) $3500. Qual é a probabilidade de ganhar mais nos dois possíveis empregos? (R.: 0,8413; 0,4562)
b) Com base no resultado do item a, qual dos dois empregos você escolheria? JUSTIFIQUE.
Adaptado de DOWNING, D. e CLARK, J.. Estatística Aplicada, São Paulo: Saraiva, 2000.

15) Existe um processo para fabricação de eixos que apresenta comportamento praticamente normal com média de 3,062 mm e variância de 0,0001 mm^2.
a) Qual é o percentual de eixos produzidos com diâmetro superior a 3,05 mm? (R.: 0,8849)
b) Se o diâmetro deverá ter no mínimo 3,04 mm e no máximo 3,08 mm, e se o custo por eixo é de $1,2 e é vendido por $5, e que eixos produzidos ou muito largos ou muito estreitos são perdidos, qual é o lucro esperado numa produção de 100 eixos? (R.: $355,1)

16) Sabe-se que a precipitação anual de chuva em certa localidade, cuja altura é medida em cm, é uma variável aleatória normalmente distribuída com altura média igual a 29,5 cm e desvio padrão de 2,5 cm de chuva. Se em mais de 45% das vezes a altura de chuva ultrapassar 32 cm torna-se viável a instalação de um sistema para coleta e armazenamento de água da chuva (como complemento à atual malha de abastecimento). É viável instalar o sistema na localidade? JUSTIFIQUE.

17) Um professor aplica um teste e obtém resultados distribuídos normalmente com média 50 e desvio padrão 10. Se as notas são atribuídas segundo o esquema a seguir, determine os limites numéricos para cada conceito:
A: 10% superiores; (R.: 62,8) B: notas acima dos 70% inferiores e abaixo dos 10% superiores; (R.: 55,2)
C: notas acima dos 30% inferiores e abaixo dos 30% superiores; (R.: 44,8)
D: notas acima dos 10% inferiores e abaixo dos 70% superiores; (R.: 37,2) E: 10% inferiores
Sugestão: faça um desenho da distribuição normal com os percentuais (áreas).
Adaptado de TRIOLA, M. Introdução à Estatística, Rio de Janeiro: LTC, 1999.

18) O tempo de vida de um determinado componente eletrônico distribui-se normalmente com média de 250 horas e variância de 49 horas^2. Você adquire um destes componentes.

a) Qual é a probabilidade de que seu tempo de vida ultrapasse as 260 horas? (R.: 0,0778)

b) Qual deveria ser o prazo de garantia para estes componentes para que o serviço de reposição atendesse a somente 5% dos componentes adquiridos? (R.: 238,45 horas)

19) Imagine que a UFSC tivesse antecipado os resultados abaixo, referentes aos candidatos não eliminados, antes de divulgar a relação com as notas de todos os candidatos.

Pontuação Final Vestibular UFSC - 2002

	Economia	Administração
Média	50,92	55,11
Desvio padrão	9,09	8,22
Vagas/Candidatos	0,370	0,412

Admitindo que as notas são normalmente distribuídas:

a) O que você responderia para candidatos aos cursos de Economia e Administração que estimassem ter conseguido, respectivamente, 55 e 58 pontos? (R.: Ambos aprovados)

b) Imagine que você tenha que responder a dezenas de vestibulandos; para poupar trabalho, estime a nota mínima para classificação em cada curso. (R.: economia = 54; administração = 57)

20) Para os casos abaixo encontre a probabilidade pela distribuição binomial e pela aproximação pela normal. Identifique se o resultado da aproximação foi bom ou não, e explique por quê.

a) Com n = 14 e p = 0,50, determine P(X = 8). (R.: 0,1833; 0,1817)

b) Com n = 10 e p =0,40, determine P(X = 7). (R.: 0,0425; 0,0143)

c) Com n = 15 e p = 0,80, determine P(X ≥ 8). (R.: 0,9957; 0,9981)

d) Com n = 14 e p = 0,60, determine P(X < 9). (R.: 0,5141; 0,5199)

e) Com n = 20 e p = 0,20, determine P(X ≤ 2). (R.: 0,2061; 0,2005)

f) Com n = 20 e p = 0,35, determine P(15 < X ≤ 18). (R.: 0,517; 0,516)

21) Em um teste de múltipla escolha temos 200 questões, cada uma com 4 possíveis respostas, das quais apenas 1 é correta. Qual é a probabilidade de que um estudante acerte entre 25 e 30 questões de 80 dentre as 200 das quais ele não sabe nada? (R.: 0,1196)

Caro estudante,

Chegamos ao final da Unidade 2do nosso livro. Nela estudamos os modelos probabilísticos mais comuns. Essa Unidade foi repleta de Figuras, Quadros, representações e exemplos de utilização das técnicas e das diferentes formas de utilização destes modelos. Releia, caso necessário, todos os exemplos, leia as indicações do Saiba Mais e discuta com seus colegas. Responda a atividade de aprendizagem e visite o Ambiente Virtual de Ensino-Aprendizagem. Conte sempre com o acompanhamento da tutoria e das explicações do professor. Ótimos estudos!

Unidade 3

Técnicas de Amostragem

Objetivo

Nesta Unidade você vai compreender em detalhes o que é amostragem, quando deve usá-la, as suas principais técnicas, a definição do plano de amostragem, e aprenderá a utilizar uma fórmula simplificada para cálculo do tamanho mínimo de amostra.

Caro estudante,

Conforme vimos na Unidade 1 de Estatística Aplicada à Administração I, a amostragem é uma das formas de coleta de dados e observamos também que se trata de uma das subdivisões da Estatística, cujo conhecimento é indispensável para o administrador.Tenha em mente que estamos interessados em obter dados confiáveis para a tomada de decisões, e muitas vezes precisaremos realizar pesquisas para coletar tais dados. Convidamos você a conhecer um pouco mais sobre esta técnica de pesquisa e seus diferentes métodos de aplicação.

Há vários argumentos para justificar a utilização da amostragem, mas há casos em que seu uso pode não ser a melhor opção. O administrador precisa conhecer tais argumentos, para que, confrontando com os recursos disponíveis e os objetivos da pesquisa, possa tomar a melhor decisão sobre a forma de coleta dos dados.

Se o administrador decidir por amostragem, é preciso delinear o plano de amostragem, indicando como ela será implementada, e qual será o seu tamanho, item crucial e que irá influenciar muito nos custos da pesquisa. Vamos ver isso em detalhes, nesta Unidade.

3.1 – O que é amostragem?

Amostragem é a subdivisão da Estatística que reúne os métodos necessários para coletar adequadamente amostras representativas e suficientes para que os resultados obtidos possam ser generalizados para a população de interesse. A pressuposição básica é que todas as etapas prévias do planejamento da pesquisa (veja na Unidade 1 de Estatística Aplicada À Administração I) já foram cumpridas, e que o administrador agora precisa decidir se coletará os dados por censoGlossário **Censo**: forma de coleta de dados em que a pesquisa é realizada com todos os elementos da população. Fonte: Barbetta, Reis e Bornia, 2010FimGlossário ou por amostragem.Glossário **Amostragem**: forma de coleta de dados em que apenas uma pequena parte, considerada representativa, da população é pesquisada.

Os resultados podem ser então generalizados, usualmente através de métodos estatísticos apropriados, para toda a população. Fonte: Barbetta, 2007. Fim Glossário.

O censo consiste simplesmente em estudar todos os elementos da **população**,Glossário **População**: é o conjunto de medidas da(s) característica(s) de interesse em todos os elementos que a(s) apresenta(m). Fonte: Andrade e Ogliari, 2007. Fim Glossárioe a amostragem pesquisa apenas uma pequena parte dela, suposta representativa do todo. Para realizar um estudo por amostragem, de maneira que seus resultados sejam válidos e possam generalizados para a população, algumas técnicas precisam ser empregadas. A essência deste processo é mostrada na Figura 28a seguir:

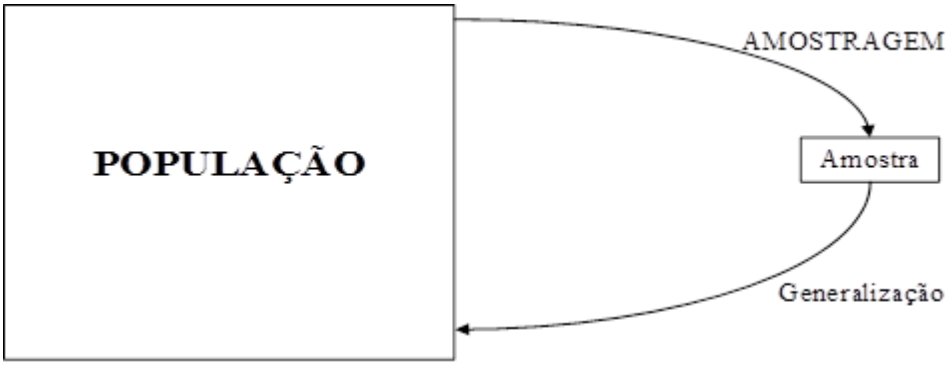


Figura 28 - Processo de Amostragem e Generalização

Fonte: elaborada pelo autor

É importante saber avaliar os argumentos a favor de cada forma de coleta.

3.2 –Condições e recomendações para uso

Podemos enumerar basicamente três motivos para usar amostragem em uma pesquisa: economia, rapidez de processamento e quando há a necessidade de testes destrutivos.Glossário **Testes destrutivos**: são ensaios realizados para avaliar a durabilidade, resistência, ou conformidade com as especificações de determinados produtos, que causam a sua inutilização, impedindo a sua comercialização. Muitos testes

destrutivos são previstos em legislação específica das mais diversas áreas. Fonte: elaborado pelo autor. Fonte: elaborado pelo autor. Fim Glossário

- **Economia**: é muito mais barato levantar as características de uma pequena parcela da população do que de todos os seus integrantes, especialmente para grandes populações. O custo do censo demográfico do IBGE é tão colossal que somente pode ser feito a cada dez anos.
- **Rapidez de processamento**:como a quantidade de dados coletada é muito menor do que a produzida em um censo, especialmente para grandes populações, o seu processamento é mais rápido. Os resultados ficam disponíveis em pouco tempo, permitindo tomar decisões em seguida. Tal característica é especialmente importante em pesquisas de opinião eleitoral, cujo resultado precisa ser conhecido rapidamente, para que candidatos e partidos possam reavaliar suas estratégias.
- **Testes destrutivos**: se para realizar a pesquisa precisamos realizar testes destrutivos (de resistência, tempo de vida útil, entre outros), o censo torna-se impraticável, exigindo a utilização de amostragem. Em muitos casos, como no caso de produtos alimentícios e farmacêuticos, há normas legais que precisam ser cumpridas rigorosamente quando da realização dos ensaios.

A Figura 29 sintetiza os motivos:

Economia Rapidez de processamento Testes destrutivos

Figura 29 – Motivos para usar amostragem

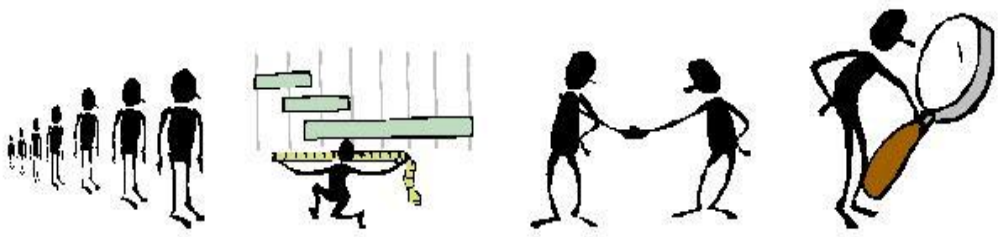
Fonte: adaptado pelo autor de Microsoft ®

Após reconhecer os motivos de se utilizar a amostragem, pense em algumas situações em que seria recomendável utilizar esta técnica.

Existem situações em que a utilização de amostragem pode não ser a melhor opção. Neste caso, podemos enumerar basicamente quatro motivos: população pequena, característica de fácil mensuração, necessidades políticas e necessidade de alta precisão.

- **População pequena**:quando é utilizada uma amostra probabilística (aleatória), e a população é pequena (digamos, menos de 100 elementos) o tamanho mínimo de amostra para obter bons resultados será quase igual ao próprio tamanho da população (veremos isso mais adiante, ainda nesta Unidade). Vale a pena então realizar um censo.
- **Característica de fácil mensuração**: a característica pode não precisar de mecanismos sofisticados de mensuração, simplesmente resume-se em uma opinião direta - a favor ou contra uma proposta. Neste caso a coleta dos dados seria bastante simples, possibilitando avaliar todos os elementos da população. Outro caso freqüente na indústria são os sistemas automatizados de medição, por exemplo, em uma fábrica de cubos de rodas de bicicletas, situada na zona franca de Manaus, os diâmetros de todos os cubos produzidos são medidos automaticamente por um sistema de telemetria a laser, dispensando a coleta por amostragem e um inspetor humano para realizar a medição.
- **Necessidades políticas**: muitas vezes uma proposta irá afetar dramaticamente todos os elementos da população, como a adoção de um regime ou forma de governo, por exemplo, o que pode ensejar a realização de um censo, para que todos manifestem sua opinião.
- **Necessidade de alta precisão**:por que o IBGE conduz um censo a cada dez anos? Porque as informações demográficas têm que ser precisas, para orientar políticas governamentais, e somente dessa maneira esse objetivo pode ser atingido.

A Figura30sintetiza os motivos:



População pequena Fácil mensuração Necessidades políticas Alta precisão

Figura 30 – Motivos para não usar amostragem

Fonte: adaptado pelo autor de Microsoft ®

Exercite a mente! Pense em algumas situações onde seja aconselhável usar um censo. Você deve se lembrar da pesquisa que esboçamos na Unidade 1 de Estatística Aplicada à Administração I: "o CRA de Santa Catarina está interessado em conhecer a opinião dos seus registrados sobre o curso em que se graduaram, desde que tal curso esteja situado em Santa Catarina". Além disso, vimos que o número de registrados no CRA, com graduação em Santa Catarina foi suposto igual a 9000. Além disso, há uma listagem com os registrados, para fins de cobrança de anuidade inclusive, que contém informações sobre endereço, curso em que se graduou, entre outras. Para conhecer a opinião das pessoas precisamos entrevistá-las (via correio, Internet, telefone ou pessoalmente). Com base no que foi dito até agora, você sabe responder se a pesquisa deve ser conduzida por censo ou por amostragem? Vamos ver juntos então!

3.2.1 – Aspectos necessários para o sucesso da amostragem

Há três aspectos necessários para que uma pesquisa realizada por amostragem gere resultados confiáveis: representatividade, suficiência e aleatoriedade da amostra.

A **representatividade** é o mais óbvio.Glossário Amostra representativa é aquela que representa na sua composição todas as subdivisões da população, procurando retratar da melhor maneira possível a sua variabilidade. Fonte: elaborado pelo autor. Fim GlossárioA amostra precisa retratar a variabilidade existente na população: ela precisa ser uma "cópia reduzida" da população. Sendo assim, todas as subdivisões da população

precisam ter representantes na amostra. A chave é avaliar se as subdivisões da população (por sexo, classe econômica, cidade, atividade profissional) podem influenciar nos resultados da pesquisa. Imagine uma pesquisa eleitoral para governador: devemos entrevistar eleitores em todas as regiões do Estado (assume-se que haja diferenças de opinião de região para região), pois se escolhermos apenas uma delas, e ela for a base política de um candidato, o resultado será distorcido.

A **suficiência** também é um aspecto relativamente óbvio. Glossário Amostra suficiente é aquela que tem um tamanho tal que permite representar adequadamente a variabilidade da população (por exemplo, além de ter representantes de cada subdivisão da população, a amostra precisa ter uma quantidade suficiente de elementos para retratar a variabilidade dentro de cada subdivisão). Fonte: elaborado pelo autor. Fim GlossárioÉ necessário que a amostra tenha um tamanho suficiente para representar a variabilidade existente na população. Quanto mais homogênea for a população (menor variabilidade), menor poderá ser o tamanho da amostra, e quanto mais heterogênea (maior variabilidade), maior terá que ser o tamanho da amostra para representá-la. LINK Vamos aprender ainda nesta Unidade uma fórmula simplificada para o cálculo do tamanho de amostra, e na Unidade 9 veremos uma expressão mais completa. Em ambos os casos, porém, veremos que o tamanho de amostra também dependerá da precisão que queremos para o nosso resultado. LINK

A **aleatoriedade** da amostra é o aspecto menos intuitivo, mas extremamente importante. Glossário Amostra aleatória, casual ou probabilística é a amostra retirada por meio de um sorteio não viciado, que garante que cada elemento da população terá uma probabilidade maior do que zero de pertencer à amostra. Fonte: Barbetta, Reis e Bornia, 2010. Fim GlossárioSignifica que os elementos da amostra serão selecionados da população por meio de sorteio não viciado: todos os elementos da população têm chance de pertencer à amostra. É necessária uma listagem com os elementos da população, permitindo a atribuição de números a cada um deles, e faz-se o sorteio. Idealmente, nós escreveríamos os números dos elementos da população em pequenos papéis, depositaríamos em uma urna, misturaríamos os papéis, e, de olhos vendados, escolheríamos os números, selecionando a

amostra. Para grandes populações esse procedimento é inviável, e com a disponibilidade de recursos computacionais, contraproducente.

O sorteio pode ser realizado através de **tabelas de números aleatórios** ou **algoritmos de geração de números pseudo-aleatórios**. Glossário Algoritmos de geração de números pseudo-aleatórios são programas computacionais que geram números aleatórios (pseudo-aleatórios, pois têm uma regra de formação), procurando simular os sorteios manuais de números de 0 a 9, procurando garantir que todo número com a mesma quantidade de algarismos tenha a mesma probabilidade de ocorrência. Fonte: elaborado pelo autor. Fim Glossário

As tabelas de números aleatórios são instrumentos usados para auxiliar na seleção de amostras aleatórias. São formadas por sucessivos sorteios de algarismos do conjunto {0, 1, 2, 3, 4, 5, 6, 7, 8, 9}, fazendo com que todo número com a mesma quantidade de algarismos tenha a mesma probabilidade de ocorrência. Quando o sorteio é realizado "manualmente", a tabela é realmente chamada de tabela de números aleatórios. LINK Muitos estatísticos realizaram tais sorteios, registraram os resultados e os publicaram em livros e periódicos para uso geral. LINK. Se, porém, os números são obtidos mediante simulação computacional, passamos a ter uma tabela de números pseudo-aleatórios, pois os números são provenientes da execução de um **algoritmo** matemático, que tem uma lógica e uma lei de formação dos resultados. LINK Neste caso há sempre o risco dos números se repetirem se a série for muito longa, descaracterizando a aleatoriedade. LINK Não obstante, tal problema, caso o algoritmo seja bom, somente ocorre após milhões ou bilhões de sorteios, quantidade muitíssimo superior àquela usada nas nossas pesquisas. Alguns estatísticos construíram tabelas de números pseudo-aleatórios e as deixaram disponíveis para o público em geral.

Nos dias de hoje, com todas as facilidades da informática, é cada vez mais comum bases de dados armazenadas em meio digital, desde uma simples planilha do Microsoft Excel ®, ou do Br.Office Calc ®até grandes bancos de dados.

Então pergunta-se: por que não realizar também o processo de amostragem, em meio digital, com os algoritmos citados no parágrafo anterior: os **algoritmos de geração de números pseudo-aleatórios**?

Trata-se de programas computacionais que procuram simular os sorteios reais de números. A grande vantagem do seu uso é a possibilidade de adaptar facilmente o sorteio ao tamanho da população envolvida, e, obviamente, a velocidade de processamento. Veja um exemplo de números aleatórios de 4 dígitos (de 0001 a 9000) gerados pelo Br.Office Calc ®LINK Na seção "Para saber mais" vamos disponibilizar um link que explica como gerar números pseudo-aleatórios com este aplicativo LINK:

3439	907	5369	8092	7962	8626	131	3667	7769	1248
2206	410	292	1478	1977	155	2566	3088	4983	3217
3347	3201	8193	4195	3836	2736	8781	7260	8921	2307

No caso da nossa pesquisa para o CRA de Santa Catarina, em que temos 9.000 registrados graduados em Santa Catarina, e há uma listagem da população, pense como seria o sorteio?

No caso mais simples de amostragem aleatória, o registrado de número 3.439 seria sorteado, seguido pelo 907, e pelo 5.369, e assim por diante, até completar o tamanho de amostra. Usualmente, cria-se automaticamente uma nova base de dados com os elementos sorteados.

Toda a teoria de inferência estatísticaLINK Veremos sobre a teoria da inferência estatística nas Unidades 4, 5 e 6 LINKpressupõe que a amostra, a partir da qual será feita a generalização (veja a Figura 28), foi retirada de forma aleatória.

Agora que já conhecemos os aspectos principais para o sucesso da amostragem podemos detalhar o plano de amostragem.

3.2.2 – Plano de Amostragem

Uma vez tendo decidido realizar a pesquisa selecionando uma amostra da população é preciso elaborar o **plano de amostragem**, que consiste em definir as unidades amostrais, o modo como a amostra será retirada (o tipo de amostragem), e o próprio tamanho da amostra.

As **unidades amostrais** são as unidades selecionadas para se chegar aos elementos da própria população. Podem ser os próprios elementos da população, quando há acesso direto a eles, ou qualquer outra unidade que possibilite chegar até eles: selecionar os domicílios como unidades de amostragem, para chegar até as famílias (que são os elementos da população); selecionar as turmas como unidades de amostragem, para chegar até os alunos (que são os elementos da população). No caso da pesquisa do CRA de Santa Catarina as unidades amostrais são os próprios elementos da população, uma vez que temos a sua listagem. No caso da Pesquisa Nacional por Amostragem de Domicílios do IBGE, as unidades amostrais são os domicílios, através dos quais chega-se às famílias.

O modo como a amostra será retirada é outra decisão importante, que precisa constar do plano de amostragem. Na Figura 31a seguir vemos o resumo dos diversos tipos de amostragem:

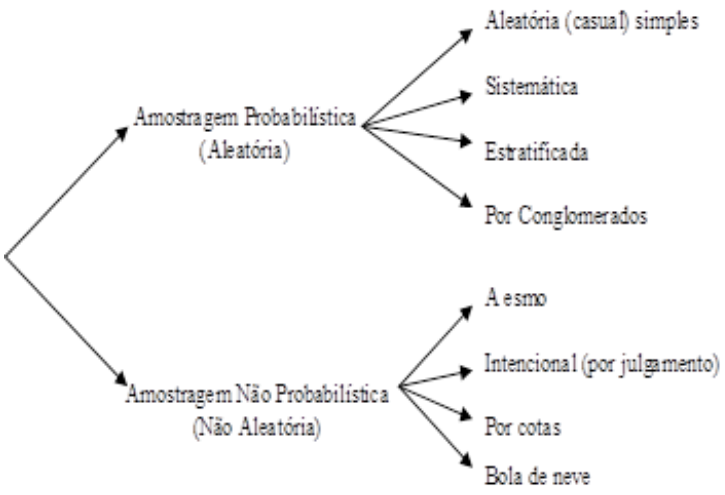


Figura 31 - Tipos de Amostragem

Fonte: elaborada pelo autor

3.3 - Amostragem probabilística ou aleatória: conceito, subtipos

Amostragem probabilística, aleatória ou casual é aquela que garante que cada elemento da população tenha probabilidade de pertencer à amostra. Para que isso ocorra é necessário que a amostra seja selecionada por sorteio não viciado, ou seja, exige-se aleatoriedade. A sua importância decorre do fato de que apenas os resultados provenientes de uma amostra probabilística podem ser generalizados estatisticamente para a população da pesquisa.

Você deve estar se perguntando, mas afinal o que significa estatisticamente? Significa que podemos associar aos resultados uma probabilidade de que estejam corretos, ou seja, uma medida da confiabilidade das conclusões obtidas. Se a amostra não for probabilística não há como saber se há 95% ou 0% de probabilidade de que os resultados sejam corretos, e as técnicas de inferência estatística, porventura utilizadas, terão validade questionável.

A condição primordial para uso da amostragem probabilística é que todos os elementos da população tenham uma probabilidade maior do que zero de pertencerem à amostra. Tal condição é materializada se:

1) Há acesso a toda a população. Ou seja, não há teoricamente problema em selecionar nenhum dos elementos, todos poderiam ser pesquisados. Concretamente, há uma lista da população, como no caso da pesquisa do CRA, que dispõe de uma lista com os 9.000 registrados que se graduaram em Santa Catarina.

2) Os elementos da amostra são selecionados através de alguma forma de sorteio não viciado: tabelas de números aleatórios, números pseudo-aleatórios gerados por computador. Com a utilização de sorteio elimina-se a ingerência do pesquisador na obtenção da amostra, e garante-se que todos os integrantes da população têm probabilidade de pertencer à amostra.

Agora vamos lhe apresentar os tipos de amostragem probabilística.

3.3.1 - Amostragem aleatória (casual) simples.

A amostragem aleatória simples Glossário - Amostragem aleatória simples é o processo de amostragem em que todos os elementos da população têm a mesma probabilidade de pertencer à amostra, e cada elemento é sorteado. Fonte: Barbetta, Reis e Bornia, 2010. Fim Glossário é o tipo de amostragem probabilística recomendável, somente, se a população for homogênea em relação aos objetivos da pesquisa, por exemplo, quando admite-se que todos os elementos da população têm características semelhantes em relação aos objetivos da pesquisa. Há uma listagem dos elementos da população, atribuem-se números a eles, e através de alguma espécie de sorteio não viciado, por meio de tabelas de números aleatóriosGlossário - Tabelas de números aleatóriossão instrumentos usados para auxiliar na seleção de amostras aleatórias, formadas por sucessivos sorteios de algarismos do conjunto {0, 1, 2, 3, 4, 5, 6, 7, 8, 9}, fazendo com que todo número com a mesma quantidade de algarismos tenha a mesma probabilidade de ocorrência. Fonte: Barbetta, 2007. Fim Glossárioou números pseudo-aleatórios gerados por computador, os integrantes da amostra são selecionados. Neste tipo de amostragem probabilística todos os elementos da população têm a mesma probabilidade de pertencer à amostra. Foi exatamente o que fizemos no final do tema “Aspectos necessários para o sucesso da amostragem” para a nossa pesquisa do CRA.

3.3.2 - Amostragem sistemática

Quando a lista de respondentes for muito grande a utilização de amostragem aleatória simples pode ser um processo moroso, ou se o tamanho de amostra for substancial, teremos que realizar um grande número de sorteios: caso estejamos utilizando números pseudo-aleatórios aumenta o risco de repetição dos números. Utiliza-se então uma variação, a amostragem sistemática,Glossário - Amostragem sistemática é a variação da amostragem aleatória simples em que os elementos da população são retirados a intervalos regulares, até compor o total da amostra, sendo o sorteio realizado apenas no ponto de partida. Fonte: Barbetta, 2007. Glossárioque também supõe que a população é homogênea em relação à variável de interesse, mas que consiste em retirar elementos da população a

intervalos regulares, até compor o total da amostra. A amostragem sistemática somente pode ser retirada se a ordenação da lista não tiver relação com a variável de interesse. Imagine que queremos obter uma amostra de idades de uma listagem justamente ordenada desta forma, neste caso a amostragem sistemática não seria apropriada, a não ser que reordenássemos a lista.

Veja a seguir o procedimento para a amostragem sistemática:

- obtém-se o tamanho da população (N);
- calcula-se o tamanho da amostra (n) – veremos isso mais adiante;
- encontra-se o intervalo de retirada k = N/n

  # se k for fracionário, deve-se aumentar n até tornar o resultado inteiro;

  # se N for um número primo, excluem-se *por sorteio* alguns elementos da população para tornar k inteiro.
- sorteia-se o ponto de partida (um dos k números do primeiro intervalo), usando uma tabela de números aleatórios, ou qualquer outro dispositivo (isso precisa ser feito para garantir que todos os elementos da população tenham chance de pertencer à amostra).
- a cada k elementos da população, retira-se um para fazer parte da amostra, até completar o valor de n.

O resumo deste processo é retratado na Figura 32, veja:

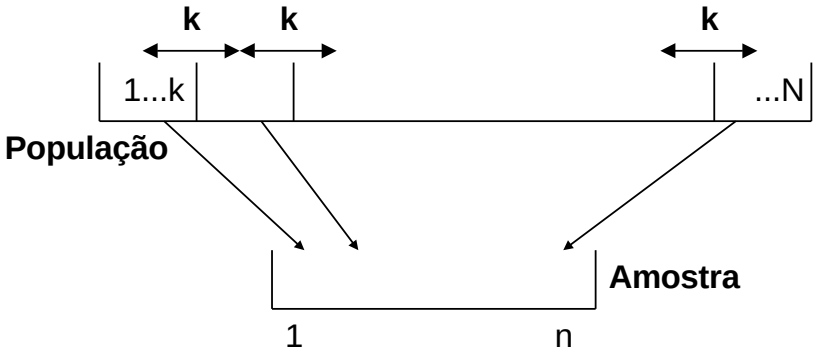


Figura 32 - Processo de amostragem sistemática

Fonte: elaborada pelo autor.

O exemplo a seguir ajudará você a entender melhor sobre o processo de amostragem sistemática. Leia com atenção!

Exemplo 1 - Uma operadora telefônica pretende saber a opinião de seus assinantes comerciais sobre seus serviços na cidade de Florianópolis. Supondo que há 25.037 assinantes comerciais, e a amostra precisa ter no mínimo 800 elementos, mostre como seria organizada uma amostragem sistemática para selecionar os respondentes.

A operadora dispõe de uma lista ordenada alfabeticamente com todos os seus assinantes, o intervalo de retirada será:**k = N/n = 25037/800 = 31,2965**

Como o valor de k é fracionário algo precisa ser feito. Aumentar o tamanho da amostra não resolverá o problema, porque 25.037 é um número primo. Como não podemos reduzir o tamanho de amostra, devendo permanecer igual a 800, se excluirmos por sorteio 237 elementos da população, e refizermos a lista teremos:**k = N/n = 24800/800 = 31**

A cada 31 assinantes um é retirado para fazer parte da amostra. Devemos sortear o ponto de partida: um número de 1 a 31 (do 1^{o} ao 31^{o} assinante). Imagine que o sorteio resultasse em 5, então a amostra seria (número de assinantes):{5, 36, 67, 98, ..., 24774}

3.3.3 - Amostragem estratificada

É bastante comum que a população de uma pesquisa seja heterogênea em relação aos objetivos da pesquisa. No caso de uma pesquisa eleitoral para governador, por exemplo, podemos esperar que a opinião deva ser diferente dependendo da região onde o eleitor mora, classe social e mesmo profissão dos entrevistados. Contudo, podemos supor que haja certa homogeneidade de opinião dentro de cada grupo. Então, supõe-se que haja heterogeneidade entre os estratos, mas homogeneidade dentro dos estratos, e que eles sejam mutuamente exclusivos (cada elemento da população pode pertencer a apenas um estrato). Para garantir que a amostra seja representativa da população Glossário Amostra

representativa: aquela que representa na sua composição todas as subdivisões da população, procurando retratar da melhor maneira possível a sua variabilidade. Fonte: elaborado pelo autor. Fim Glossárioprecisamos garantir que os diferentes estratos sejam nela representados: deve usar a amostragem estratificada, Glossário - Amostragem estratificada é a amostragem probabilística usada quando a população for heterogênea em relação aos objetivos da pesquisa (as opiniões tendem a variar muito de subgrupo para subgrupo), e amostra precisa conter elementos de cada subgrupo da população para representá-la adequadamente. Fonte: Barbetta, 2007. Fim Glossário como representa a Figura 33:

Figura 33 - Amostragem estratificada

Fonte: elaborada pelo autor

Veja que a seleção dos elementos de cada estrato pode ser feita usando amostragem aleatória simples ou sistemática.

A amostragem estratificada pode ser:

- proporcional, quando o número de elementos selecionados de cada estrato é proporcional ao seu tamanho na população (por exemplo, se o estrato representa 15% da população, 15% da amostra deverá ser retirada dele); e
- uniforme, quando os mesmos números de elementos são selecionados de cada estrato.

A amostragem estratificadaproporcional possibilita resultados melhores, mas exige um grande conhecimento da população (para saber quantos são e quais são os tamanhos dos estratos). A amostragem estratificada uniforme é mais utilizada em estudos comparativos.

No caso da pesquisa do CRA você acredita que a população é heterogênea em relação aos objetivos da pesquisa? Será que a região do Estado, o fato de ter estudado em faculdade pública ou particular pode influenciar as opiniões dos registrados sobre os cursos onde se graduaram?

3.3.4 - Amostragem por conglomerados

Teoricamente, a amostragem estratificada proporcional apresenta os melhores resultados possíveis. Sua grande dificuldade de uso deve-se ao grau de conhecimento necessário sobre a população, que geralmente não existe ou é impraticável de obter. Uma alternativa consiste no uso de conglomerados.Glossário - Amostragem por conglomerados é a amostragem probabilística em que a população é subdividida em grupos definidos por conveniência (usualmente geográfica), e alguns destes grupos são selecionados por sorteio, e elementos dos grupos sorteados podem também ser sorteados para compor a amostra. Fonte: Barbetta, 2007. Fim Glossário

Os conglomerados também são grupos mutuamente exclusivos de elementos da população, mas são definidos de forma mais arbitrária do que os estratos: é bastante comum definir os conglomerados geograficamente. Por exemplo, os bairros de uma cidade, que constituiriam conglomerados de domicílios.

O procedimento para a amostragem por conglomerados ocorre da seguinte forma:

- divide-se a população em conglomerados;
- sorteiam-se os conglomerados (usando tabela de números aleatórios ou qualquer outro método não viciado);
- pesquisam-se todos os elementos dos conglomerados sorteados, ou sorteiam-se elementos deles.

A utilização de amostragem por conglomerados permite uma redução substancial nos custos de obtenção da amostra, sem comprometer demasiadamente a precisão, sendo

que em alguns casos é a única alternativa possível. Veja a Figura34e entenda como ocorre essa amostragem:

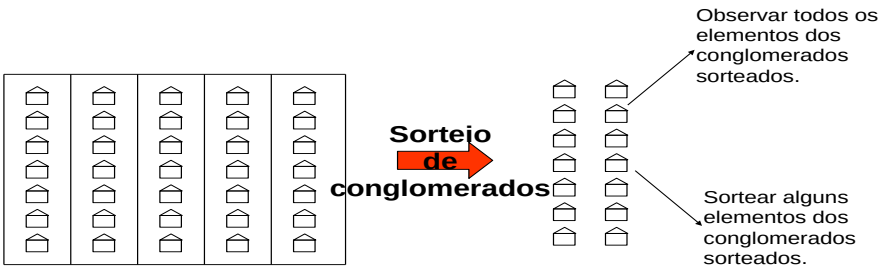


Figura 34 - Amostragem por conglomerados

Fonte: elaborada pelo autor

A Pesquisa Nacional por Amostra de Domicílios (PNAD) do IBGE, coleta informações demográficas e sócio-econômicas sobre a população brasileira. Utiliza amostragem por conglomerados em três estágios: LINK Mais informações em http://www.ibge.gov.br/home/estatistica/populacao/trabalhoerendimento/pnad98/saude/metodologia.shtm LINK

- **Primeiro estágio**: amostras de municípios (conglomerados) para cada uma das regiões geográficas do Brasil;
- **Segundo estágio**: setores censitários sorteados em cada município (conglomerado sorteado); e
- **Terceiro estágio**: domicílios sorteados em cada setor censitário.

Você deve estar se perguntando, e quando não for possível garantir a probabilidade de todo elemento da população pertencer à amostra? Então este é o momento de partirmos para a amostragem não probabilística.

3.4 – Amostragem não probabilística

A obtenção de uma amostra probabilística exige uma listagem com os elementos da população. Em suma, exige acesso a todos os elementos da população. Nem sempre é possível obter tal listagem na prática, o que teoricamente inviabilizaria a retirada de uma amostra probabilística. Então, pode-se recorrer à amostragem não probabilística. Glossário

Amostragem não probabilística é o processo de amostragem em que nem todos os elementos da população têm chance de pertencer à amostra, pois a seleção não é feita por sorteio não viciado. Fonte: Barbetta, 2007. Fim Glossário

Ao usar a amostragem não probabilística o pesquisador não sabe qual é a probabilidade de que um elemento da população tem de pertencer à amostra. Portanto, os resultados da amostra não podem ser estatisticamente generalizados para a população, porque não se pode estimar o erro amostral. Glossário Erro amostral é o valor máximo que o pesquisador admite errar na estimativa de uma característica da população a partir de uma amostra aleatória desta mesma população. Fonte: Barbetta, 2007. Fim Glossário

Alguns dos usos habituais da amostragem não probabilística são os seguintes:

- a etapa preliminar em projetos de pesquisa;
- em projetos de pesquisa qualitativa; e
- em casos onde a população de trabalho não pode ser enumerada.

Veja que existem ainda vários tipos de amostragem não probabilística e que serão descritos na seqüência.

3.4.1 - Amostragem a esmo

Na Amostragem a esmo, o pesquisador procura ser o mais aleatório possível, mas sem fazer um sorteio formal. Imagine um lote de 10.000 parafusos, do qual queremos tirar uma amostra de 100, se fôssemos realizar uma amostragem aleatória simples o processo talvez fosse trabalhoso demais. Então simplesmente retiramos os elementos a esmo. Este tipo de amostragem também pode ser utilizado quando a população for formada por material contínuo (gases, líquidos, minérios), bastando homogeneizar o material e retirar a amostra.

3.4.2 - Amostragem por julgamento (intencional).

Na amostragem por julgamento, o pesquisador deliberadamente escolhe alguns elementos para fazer parte da amostra, com base no seu julgamento de aqueles seriam representativos da população. Este tipo de amostragem é bastante usado em estudos qualitativos. Obviamente o risco de obter uma amostra viciada é grande, pois se baseia totalmente nas preferências do pesquisador, que pode se enganar (involuntária ou "voluntariamente").

3.4.3 - Amostragem por cotas

A Amostragem por cotas parece semelhante a uma amostragem estratificada proporcional, da qual se diferencia por não empregar sorteio na seleção dos elementos. A população é dividida em vários subgrupos, na realidade é comum dividir em um grande número para compensar a falta de aleatoriedade, e seleciona-se uma cota de cada subgrupo, proporcional ao seu tamanho.

Em uma pesquisa de opinião eleitoral, por exemplo, poderíamos dividir a população de eleitores por sexo, nível de instrução, faixas de renda entre outros aspectos, e obter cotas proporcionais ao tamanho dos grupos (que poderia ser obtido através das informações do IBGE). Na amostragem por cotas os elementos da amostra são escolhidos pelos entrevistadores (de acordo com os critérios), geralmente em pontos de grande movimento, o que sempre acarreta certa subjetividade (e impede que qualquer um que não esteja passando pelo local no exato momento da pesquisa possa ser selecionado).

Na prática, muitas pesquisas são realizadas utilizando amostragem por cotas, incluindo as polêmicas pesquisas eleitorais. LINK Leia um texto muito interessante sobre o tema que encontra-se disponível em: http://www.ime.unicamp.br/~dias/falaciaPesquisaEleitoral.pdf LINK

No exemplo apresentado no Quadro 4, imagine que queremos saber a opinião dos eleitores do bairro Goiaba sobre o governo municipal. Supõe-se que as principais variáveis que condicionariam as respostas seriam sexo, idade e classe social. O bairro apresenta a seguinte composição demográfica para as variáveis:

Sexo	Idade (faixa etária)	Classe social	% populacional
Masculino	18| -- 35	A	1%
Masculino	18| -- 35	B	4%
Masculino	18| -- 35	C	10%
Feminino	18| -- 35	A	1%
Feminino	18| -- 35	B	2%
Feminino	18| -- 35	C	9%
Masculino	35| -- 60	A	5%
Masculino	35| -- 60	B	8%
Masculino	35| -- 60	C	12%
Feminino	35| -- 60	A	4%
Feminino	35| -- 60	B	8%
Feminino	35| -- 60	C	10%
Masculino	Mais de 60	A	1%
Masculino	Mais de 60	B	9%
Masculino	Mais de 60	C	3%
Feminino	Mais de 60	A	3%
Feminino	Mais de 60	B	7%
Feminino	Mais de 60	C	3%

Quadro 4 - Esquema de amostragem por cotas.

Fonte: adaptado pelo autor de Marconi e Lakatos (2003)

Se, por exemplo, o tamanho de nossa amostra fosse igual a 200 (200 pessoas serão entrevistadas), o número de pessoas deveria ser dividido de forma proporcional: 1% do sexo masculino, com idade entre 18 e 25 anos, da classe A, totalizando 2 pessoas; 4% do sexo masculino, com idade entre 18 e 25 anos, da classe B, totalizando 8 pessoas, e assim por diante. Os entrevistadores receberiam suas cotas, e deveriam escolher pessoas, em pontos de movimento do referido bairro, que se aproximem dos critérios e entrevistá-las, recolhendo suas opiniões sobre o governo municipal. Usualmente os resultados são generalizados estatisticamente para a população, empregando as técnicas que serão vistas na Unidade 5 deste livro-texto, mas rigorosamente os resultados da amostragem por cotas

não têm validade estatística, visto que não contemplam o princípio de aleatoriedade na seleção da amostra.

3.4.4 - Amostragem "bola de neve".

A Amostragem "bola de neve" é particularmente importante quando é difícil identificar respondentes em potencial. A cada novo respondente que é identificado e entrevistado, pede-se que identifique outros que possam ser qualificados como respondentes. Pode levar a amostras compostas apenas por "amigos" dos primeiros entrevistados, o que pode causar viesamentos nos resultados finais.

Agora que você já conhece sobre o importante e interessante tema do cálculo do tamanho de amostra, passaremos para uma amostra probabilística.

3.5 – Cálculo do tamanho de uma amostra probabilística (aleatória) para estimar proporção.

A determinação do tamanho de amostra é um dos aspectos mais controversos da técnica de amostragem, e envolve uma série de conceitos (probabilidade, inferência estatística e a própria teoria da amostragem). Nesta seção apresentaremos uma visão simplificada para obter o tamanho mínimo de uma amostra aleatória simples que atenda aos seguintes requisitos:

- o interesse na proporção de ocorrência de um dos valores de uma variável qualitativa na população;
- a confiabilidade dos resultados da amostra deve ser aproximadamente igual a 95% (ou seja, há 95% de probabilidade de que a proporção populacional do valor da variável qualitativa esteja no intervalo definido pelos resultados da amostra);
- estamos fazendo uma estimativa exagerada do tamanho de amostra;
- não vamos nos preocupar com aspectos financeiros relacionados ao tamanho da amostra (embora, obviamente, seja uma consideração importante).

O primeiro passo para calcular o tamanho da amostra é definir o **erro amostral** tolerável, que será chamado de $\mathbf{E_0}$. Este erro é o valor máximo que o pesquisador admite errar na estimativa de uma característica da população.

Lembre-se das pesquisas de opinião eleitoral: "o candidato Fulano está com 18% de intenção de voto, a margem de erro da pesquisa é de 2% para mais ou para menos". O 2% é o valor do erro amostral tolerável, então o percentual de pessoas declarando o voto no candidato Fulano é igual a 18% ± 2%. Além disso, há uma probabilidade de que este intervalo não contenha o valor real do parâmetro, ou seja, o percentual de eleitores que declaram o voto no candidato, pelo fato de que estamos usando uma amostra, embora isso raramente seja dito na mídia, especialmente na televisão.

É razoável imaginar que quanto menor o erro amostral tolerável escolhido maior será o tamanho da amostra necessário para obtê-lo. Isso fica mais claro ao ver a fórmula para obtenção da primeira estimativa do tamanho de amostra:

$$n_0 = \frac{\mathbf{1}}{E_0^{\,2}}$$

Onde E_0 é o erro amostral tolerável, e n_0 é a primeira estimativa do tamanho de amostra. Se o tamanho da população, N, for conhecido podemos corrigir a primeira estimativa:

$$n = \frac{N \times n_0}{N + n_0}$$

Exemplo 2 – Calcule o tamanho mínimo de uma amostra aleatória simples para estimar uma proporção, admitindo com alto grau de confiança, um erro amostral máximo de 2%, supondo que a população tenha:

a) 200 elementos.

b) 200.000 elementos.

Observe a diferença entre os tamanhos das duas populações: a da letra b é mil vezes maior do que a da letra a. Como a primeira estimativa, n_0 não depende do tamanho da

população, e o erro amostral é 2% para ambas podemos calculá-lo apenas uma vez. Devemos dividir o 2% por 100 antes de substituir na fórmula:

$$n_0 = \frac{1}{E_0^{\,2}} = \frac{1}{(0{,}02)^2} = 2500$$

Então nossa primeira estimativa, para um erro amostral de 2%, é retirar uma amostra de 2.500 elementos.

a) Obviamente precisamos corrigir a primeira estimativa, pois a população conta com apenas 200 elementos. Então:

$$n = \frac{N \times n_0}{N + n_0} = \frac{200 \times 2500}{200 + 2500} = 185{,}185$$

Precisamos arredondar, sempre para cima, o tamanho mínimo da amostra. Então a amostra deverá ter pelo menos 186 elementos para garantir um erro amostral de 2%. Observe que a amostra representa 93% da população. Será que um censo não seria mais aconselhável neste caso?

b) Corrigindo a primeira estimativa com o tamanho da população:

$$n = \frac{N \times n_0}{N + n_0} = \frac{200000 \times 2500}{200000 + 2500} = 2469{,}136$$

Arredondando, a amostra deverá ter no mínimo 2.470 elementos para garantir um erro amostral de 2%. Observe que a amostra representa 1,235 % da população. Claríssimo caso em que a amostragem é a melhor opção de coleta.

Poderíamos ter usado diretamente a primeira estimativa, 2.500 elementos, pois a correção não causou grande mudança. Este exemplo prova que não precisamos de grandes amostras para obter uma boa precisão nos resultados.

A Figura 35 mostra um gráfico relacionando tamanhos de amostra para diferentes tamanhos de população, considerando um erro amostral tolerável igual a 2%.

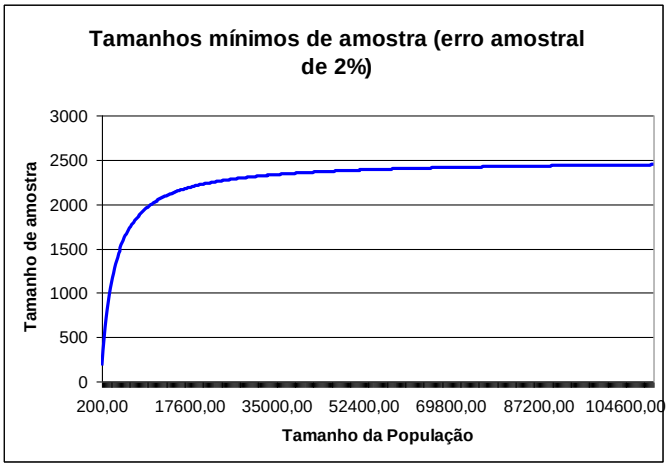


Figura 35 - Tamanho de amostra x tamanho da população (e_0 = 2%)

Fonte: elaborado pelo autor a partir de Microsoft ®

Observe que a partir de um determinado tamanho de população, para o mesmo erro amostral, o ritmo de crescimento do tamanho da amostra vai diminuindo, para 70.000 elementos ou mais praticamente não há mais aumento. Isso mostra que não há necessidade de retirar, por exemplo, 50% da população para ter uma boa amostra.

É importante alertar que ao calcular o tamanho de amostra para amostragem estratificada, deve-se fazê-lo para cada estrato, e o tamanho total será a soma dos valores. Se isso não for feito, não podemos garantir o erro amostral dentro de cada estrato: se calcularmos um valor geral e dividirmos o tamanho da amostra por estrato (mesmo proporcionalmente), a margem de erro dentro de cada estrato será maior do que a prevista.

Tô afim de saber:

- Sobre amostragem, consulte BARBETTA,P. A. *Estatística Aplicada às Ciências Sociais*. 7ª. ed. – Florianópolis: Ed. da UFSC, 2008, Capítulo 3.

- Sobre características de fácil mensuração consulte em LAGO NETO, J.C. *O Efeito da Autocorrelação em Gráficos de Controle para Variável Contínua*: Um Estudo de Caso. Florianópolis. 1999. Dissertação (Mestrado em Engenharia de Produção)- Programa de Pós-Graduação em Engenharia de Produção, UFSC.

- Sobre pesquisas eleitorais, consulte SOUZA, J. *Pesquisas Eleitorais*: Críticas e Técnicas; Brasília: Centro Gráfico do Senado Federal, 1990.

- Sobre como gerar números pseudo-aleatórios ou obter amostras aleatórias simples no Br.Office Calc ®, leia o texto *Como gerar uma amostra aleatória simples com o Br.Office Calc®*, no Ambiente Virtual de Ensino-Aprendizagem.

-Sobre Amostragem a esmo, leia COSTA NETO, P.L. da O. *Estatística*. 2ª ed, São Paulo: Edgard Blücher, 2002.

Atividades de Aprendizagem

O que você acha de testar seus conhecimentos com relação ao estudo da Unidade 3? Para tanto, faça as atividades propostas a seguir e encaminhe-as para seu tutor através do Ambiente Virtual de Ensino-Aprendizagem. Não hesite em buscar o auxílio do seu tutor se encontrar dificuldades.

1) Analise os planos de amostragens apresentados abaixo. Você concorda com a maneira como foram elaborados? Justifique. Apresente as soluções que você julgar necessárias.

a) Para ser conhecida a opinião dos estudantes da UFSC sobre o Jornal Universitário, foram colhidas as opiniões de 40 estudantes da última fase do curso de Jornalismo daquela instituição.

b) Há interesse em medir o índice de luminosidade das salas de aula da UFSC. A coleta de dados será feita em todos os centros da UFSC, durante os períodos diurno e noturno, nas salas que estiveram desocupadas no momento da pesquisa. Cada centro será visitado apenas uma vez.

c) As constantes reclamações dos usuários motivaram a direção da Biblioteca Central da UFSC a realizar uma pesquisa sobre o nível de ruído em suas dependências. O ruído será medido em todas as seções da Biblioteca, na primeira e na penúltima semanas do semestre, de segunda a sábado, durante todo o horário de funcionamento.

d) No controle de qualidade de uma fábrica de peças, que trabalha 24 horas por dia, sete

dias por semana, um item produzido é retirado de cada máquina, a cada meia hora, para avaliação. O procedimento é feito durante todo o dia, ao longo da semana.

e) O Comando de um Batalhão da Polícia Militar de Santa Catarina quer conhecer a opinião das pessoas que residem em sua área de atuação, no intuito de formular novas escalas de policiamento ostensivo. Para tanto serão feitas entrevistas com as pessoas que se passarem a pé pela frente do Batalhão, de segunda à sexta das 8:30 às 12:00 horas e das 14:00 às 17:30 horas durante duas semanas.

f) Com a finalidade de estudar o perfil dos consumidores de um supermercado, observaram-se os consumidores que compareceram ao supermercado no primeiro sábado do mês.

g) Com a finalidade de estudar o perfil dos consumidores de um supermercado, fez-se a coleta de dados durante um mês, tomando a cada dia um consumidor de cada fila de cada caixa, variando-se sistematicamente o horário de coleta dos dados.

h) Para avaliar a qualidade dos itens que saem de uma linha de produção, observaram-se todos os itens das 14 às 14 horas e trinta minutos.

i) Para avaliar a qualidade dos itens que saem de uma linha de produção, observou-se um item a cada meia hora, durante todo o dia.

j) Para estimar a porcentagem de empresas que investiram em novas tecnologias no último ano, enviou-se um questionário a todas as empresas de um estado. A amostra foi formada pelas empresas que responderam o questionário.

2) Uma determinada faculdade do interior de Santa Catarina possui 6 cursos, estando os alunos matriculados de acordo com a tabela abaixo:

Curso	Direito	Administração	Economia	Agronomia	Veterinária	Computação
Alunos	250	200	150	150	150	100

A diretoria pretende selecionar, por amostragem, alguns alunos para uma atividade extracurricular.

a) Os cursos direito, administração e economia formam um estrato (sócio-econômicos), agronomia e veterinária formam outro (agrários) e computação outro estrato (tecnológicos), extraia uma amostra estratificada proporcional de 20 alunos (use o Microsoft Excel ® ou o Br.Office Calc ®).

b) Através de uma amostragem de conglomerados de 2 estágios extraia uma amostra

aleatória de 21 alunos. Selecione 3 cursos, e depois 7 alunos por curso (use o Microsoft Excel® ou o Br.Office Calc ®).

c) Qual das duas amostras você acredita que tem resultados mais confiáveis? JUSTIFIQUE.

3) Será feito um levantamento por amostragem de uma população de 2000 famílias, para a realização de uma pesquisa.

a) Calcule o tamanho mínimo de uma amostra para que se tenha um erro amostral máximo de 5%.

b) Supondo a população dividida em 2 estratos iguais, qual o tamanho mínimo de amostra para se ter um erro amostral máximo de 5% em cada estrato?

c) Qual seria o erro amostral em cada estrato do item b, se o tamanho da amostra em cada estrato fosse simplesmente o valor definido no item a dividido por 2?

Resumo

O resumo desta Unidade está esquematizado na Figura 36. Veja:

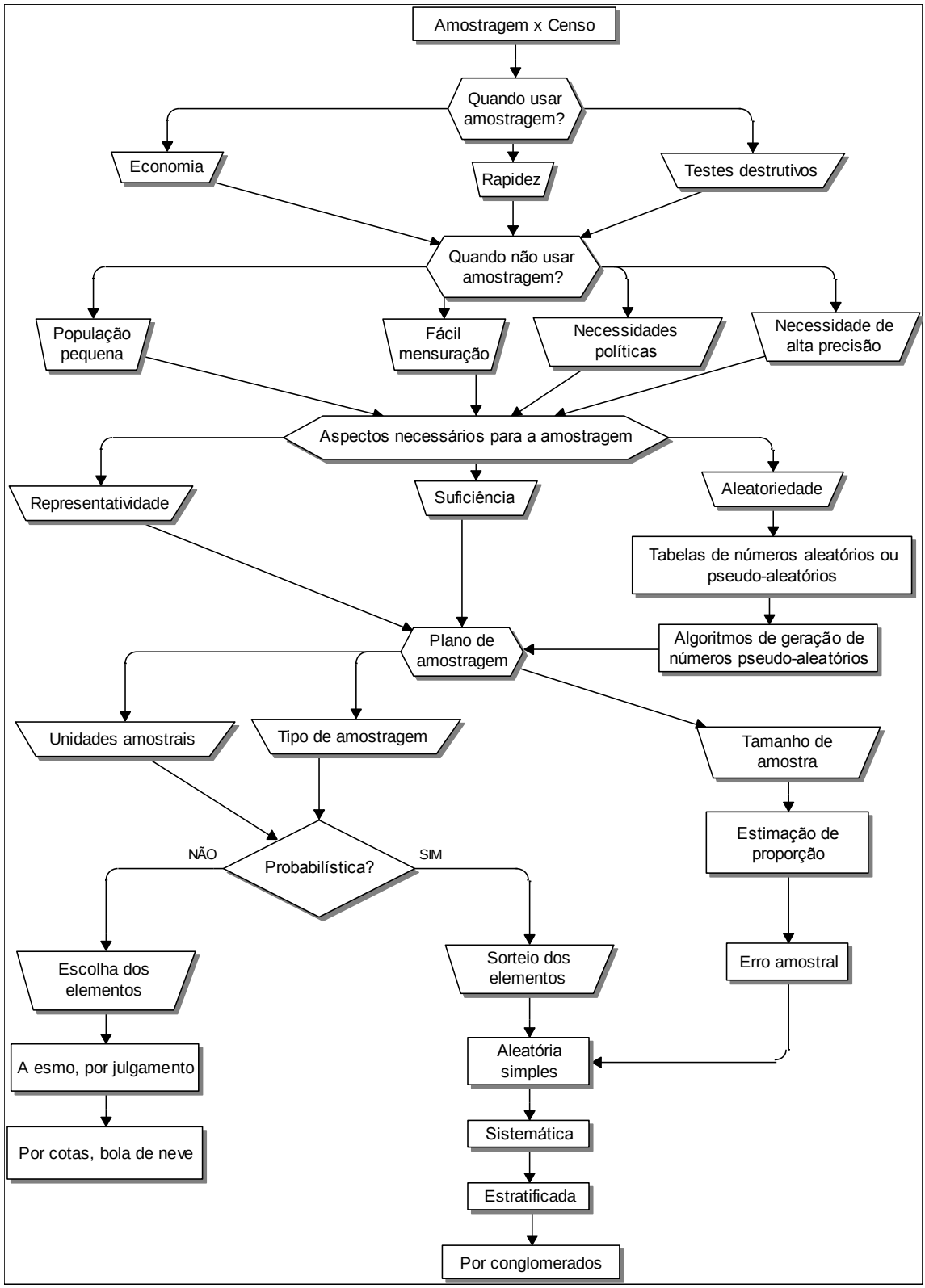


Figura 36 - Resumo da Unidade 3

Fonte: elaborada pelo autor

Caro estudante,

Chegamos ao final da Unidade3. Nela estudamos sobre amostragem e censo e suas formas de utilização, habilidades necessárias para um bom administrador. Essa Unidade foi repleta de Figuras, quadros, representações e exemplos de utilização das técnicas e das diferentes formas de utilização, na íntegra de suas especificidades e deu sustentação para as discussões das próximas unidades. Releia, caso necessário, todos os exemplos, leia as indicações do Saiba mais e discuta com seus colegas. Na realização da atividade de aprendizagem você colocará em prática os ensinamentos repassados. Conte sempre com o acompanhamento da tutoria e das explicações do professor. Lembre-se que não estás sozinho. Conte com a gente!

Unidade 4

Inferência estatística e distribuição amostral

Objetivo

Nesta Unidade você vai aprender os conceitos de inferência estatística e de distribuição amostral, que são a base para o processo de generalização usado pelos administradores em várias tomadas de decisão.

4.1 -Conceito de Inferência Estatística

Caro estudante, vamos relembrar um pouco nossa trajetória ao longo das duasdisciplinas de Estatística Aplicada à Administração.

Na Unidade 1 de Estatística I vimos que através da **Inferência Estatística**, usando os conceitos de Probabilidade (e variáveis aleatórias, Unidade 6 de Estatística I, e Unidade 1 de Estatística II) podemos generalizar os resultados de uma pesquisa por amostragem (Unidade 3 de Estatística II) para a população da qual a amostra foi retirada.

Lembre-se, estamos supondo que a amostra foi retirada por meio de **amostragem probabilística ou aleatória**, temos então um **experimento aleatório**: não sabemos quem fará parte da amostra antes do sorteio (Unidade 3 de Estatística II).

Uma vez retirada a amostra, fazemos análise exploratória dos dados (Unidades 2 e 3 de Estatística I): por exemplo, calculamos média de uma variável quantitativa. Esta média e todas as demais estatísticas serão variáveis aleatórias (pois estão associadas ao **Espaço Amostral** de um experimento aleatório), e poderemos tentar identificar o modelo probabilístico mais apropriado para elas (Unidades 1 e 2 de Estatística II). Mas, neste caso o modelo probabilístico de uma estatística da amostra é chamado de **Distribuição Amostral**.

Conhecer a Distribuição Amostral das principais estatísticas vai nos ser muito útil quando estudarmos os tipos particulares de Inferência Estatística: Estimação de Parâmetros (Unidade 5) e Testes de Hipóteses (Unidade 6) neste livro de Estatística Aplicada à Administração II.

Vamos continuar aprendendo? É muito bom ter você conosco!

Estatística é a ciência que se ocupa de organizar, descrever, analisar e interpretar dados para que seja possível a tomada de decisões e/ou a validação científica de uma conclusão. Os dados são coletados para estudar uma ou mais características de uma População: conjunto das medidas da(s) característica(s) de interesse em todos os elementos que a(s) apresenta(m).

Uma população pode ser representada através de um modelo: este apresenta condições para uso, forma para a distribuição, e parâmetros.

Os dados necessários para a obtenção do modelo podem ser obtidos através de um censo (pesquisa de toda a população), ou através de uma amostra (subconjunto finito) da população.LINK Na Unidade 3 enumeramos as principais razões para usar amostragem.LINK

A amostra deve ser: representativa da população, suficiente (para que o resultado tenha confiabilidade), e aleatória (retirada por sorteio não viciado).

DESTAQUE "A **Inferência Estatística** consiste em fazer *afirmações probabilísticas* sobre as características do modelo probabilístico, que se supõe representar uma população, a partir dos dados de uma amostra aleatória (probabilística) GLOSSÁRIO Amostra aleatória, casual ou probabilística: amostra retirada por meio de um sorteio não viciado, que garante que cada elemento da população terá uma probabilidade maior do que zero de pertencer à amostra. GLOSSÁRIO desta mesma população". DESTAQUE

Fazer uma afirmação probabilística sobre uma característica qualquer é associar à declaração feita uma probabilidade de que tal declaração esteja correta (e, portanto, a probabilidade complementar de que esteja errada). Quando se usa uma amostra da população sempre haverá uma probabilidade de estar cometendo um erro (justamente por ser usada uma amostra): a diferença entre os métodos estatísticos e os outros reside no fato de que os métodos estatísticos permitem calcular essa probabilidade de erro. E para que isso seja possível a amostra da população precisa ser aleatória.

As afirmações probabilísticas sobre o modelo da população podem ser basicamente:

=> estimar quais são os possíveis valores dos parâmetros - GLOSSÁRIO Parâmetros: alguma medida descritiva (média, variância, proporção) dos valores x1, x2, x3, ..., associados à população.Fonte: Barbetta, Reis e Bornia, 2010.Fim GLOSSÁRIO**Estimação de Parâmetros**:GLOSSÁRIO Estimação de Parâmetros: forma de inferência estatística que busca estimar os parâmetros do modelo probabilístico da variável de interesse na população, a partir de dados de uma amostra probabilística desta mesma população. Fonte: Barbetta, Reis e Bornia, 2010.FimGLOSSÁRIO

- qual é o valor da média de uma variável que segue uma distribuição normal?
- qual é o valor da proporção de um dos 2 resultados possíveis de uma variável que segue uma distribuição binomial.

=> testar hipóteses sobre as características do modelo: parâmetros, forma da distribuição de probabilidades, entre outros - **Testes de Hipóteses**.GLOSSÁRIO Testes de hipóteses: forma de inferência estatística que busca testar hipóteses sobre características (parâmetros, forma do modelo) do modelo probabilístico da variável de interesse na população, a partir de dados de uma amostra probabilística desta mesma população. Fonte: Barbetta, Reis e Bornia, 2010.FimGLOSSÁRIO

- o valor da média de uma variável que segue uma distribuição é maior do que um certo valor?
- o modelo probabilístico da população é uma distribuição normal?
- o valor da média de uma variável que segue uma distribuição normal em uma população é diferente da mesma média em outra população?

Estudaremos Estimação de Parâmetros na Unidade 5 e Testes de Hipóteses na Unidade 6.

4.2 – Parâmetros e Estatísticas

Vamos imaginar uma pesquisa como a da Unidade 1 de Estatística Aplicada à Administração 2, opinião dos registrados no CRA-SC sobre os cursos em que se graduaram, desde que tenham se graduado em Santa Catarina. Naquela Unidade, e depois na Unidade 2 de Estatística Aplicada à Administração II, declaramos que era possível realizar uma amostragem probabilística, e vimos um exemplo de como fazer isso.

Independente da pesquisa, uma vez que tenha sido realizada por amostragem probabilística, os dados podem ser estatisticamente generalizados para a população.

Uma vez tendo coletado os dados, é preciso resumi-los e organizá-los de maneira a permitir uma primeira análise, e posterior uso das informações. As técnicas estatísticas que se ocupam desses aspectos constituem a Análise Exploratória de Dados, que estudamos detalhadamente nas Unidades 2 e 3 de Estatística Aplicada à Administração I.

O conjunto de dados pode ser resumido (e apresentado) através das distribuições de frequências, que relacionam os valores que a variável pode assumir com a frequência (contagem) com que foram encontrados naquele conjunto. Esta distribuição pode ser apresentada na forma de uma tabela, ou através de um gráfico (estes dois métodos podem ser usados tanto para variáveis qualitativas quanto para variáveis qualitativas).

Há uma terceira forma de resumir o conjunto de dados, quando a variável sob análise é quantitativa: as medidas de síntese ou estatísticas. GLOSSÁRIO Estatísticas: medidas de síntese da variável calculadas com base nos resultados de uma amostra da população. Se a amostra for probabilística (aleatória) as estatísticas podem ser consideradas variáveis aleatórias. Fonte: Barbetta, Reis e Bornia, 2010.FimGLOSSÁRIO As principais

estatísticas são a média, o desvio padrão, a variância e a proporção. LINK Esta última está relacionada aos percentuais de ocorrência dos valores em uma distribuição de frequências de uma variável qualitativa. LINK

DESTAQUE Atenção, vamos relembrar o que cada uma dessas significa:

- **Média**: média aritmética simples (ver Unidade 3 de Estatística Aplicada à Administração I), trata-se de uma estatística que caracteriza o "centro de massa" do conjunto de dados (Valor Esperado – ver Unidade 1, seção 1.4). Quando é a média populacional recebe o símbolo **μ**, quando é a média amostral recebe o símbolo $\bar{X}$;
- **Variância**: trata-se de uma estatística (ver Unidade 3 de Estatística Aplicada à Administração I) que mede a dispersão em torno da média do conjunto, (em torno do valor esperado – Ver Unidade 1, seção 1.4), possuindo uma unidade que é o quadrado da unidade da média (e dos valores do conjunto). Quando é a variância populacional recebe o símbolo σ^2, quando é a variância amostral recebe o símbolo s^2;
- **Desvio padrão** é a raiz quadrada positiva da variância, tendo, portanto, uma unidade que é igual à unidade da média, sendo muitas vezes preferida para efeito de mensuração da dispersão. Quando é o valor populacional recebe o símbolo **σ**, e quando é o amostral recebe o símbolo **s**.
- **Proporção**: consiste em calcular a razão entre o número de ocorrências do valor de interesse de uma variável qualitativa e o número total de ocorrências registradas no conjunto (de todos os valores que a variável pode assumir); quando é uma proporção populacional recebe o símbolo **π**; quando é uma proporção amostral recebe o símbolo **p**. DESTAQUE

Os valores das medidas de síntese, além de resumirem o conjunto de dados, constituem uma indicação dos prováveis valores dos parâmetros. Assim, em estudos baseados em amostras, é comum utilizar tais medidas de síntese como estatísticas que serão utilizadas para estimar os parâmetros do modelo probabilístico que descreve a população.

O Quadro5 resume os parâmetros e as estatísticas:

Medidas de síntese	Parâmetros (População)	Estatísticas (Amostra)
Média	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\bar{x} = \frac{\sum_{i=1}^{n} X_i}{n}$
Variância	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$
Proporção	$\pi = \frac{f_a}{N}$	$p = \frac{f_a}{n}$

Quadro 5 - Parâmetros e Estatísticas mais comuns

Fonte: elaborado pelo autor

Onde N é o número de elementos da população, n é o número de elementos da amostra, e f_a é a frequência de ocorrência de um dos valores de uma variável qualitativa na população ou na amostra.

As Estatísticas são variáveis aleatórias, pois seus valores podem variar dependendo do resultado da amostra. Se forem variáveis aleatórias, podem ser caracterizadas através de algum modelo probabilístico. Este modelo recebe o nome de distribuição amostral.

4.3 – Distribuição Amostral

Seja uma população qualquer com um parâmetro θ de interesse, correspondendo a uma estatística **T** em uma amostra. Amostras aleatórias são retiradas da população e para cada amostra calcula-se o valor **t** da estatística **T**.

Os valores de **t**LINK NÃO confundir com o t da distribuição t de Student, seção 2.2.4, Unidade 2 LINK formam uma nova população que segue uma distribuição de probabilidades que é chamada de **distribuição amostral** de **T**.

Exemplo 1 - Seja a população abaixo, constituída pelos pesos em kg de oito pessoas adultas:

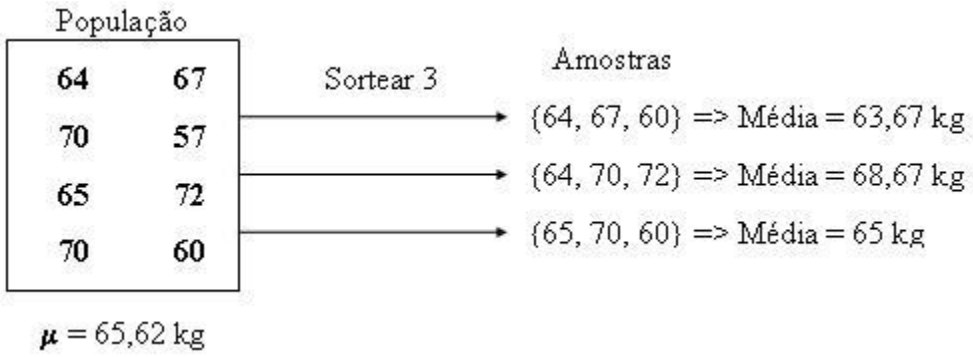


Figura 37 - Distribuição Amostral – Exemplo 1

Fonte: elaborada pelo autor

Observe que foram retiradas três amostras. Para cada amostra foi calculada a média, visando estimar a média populacional, que vale 65,62 kg. Observe que há uma variação na estatística média, pois o processo de amostragem é aleatório: é um experimento aleatório. Esta variação precisa ser considerada quando são realizadas as inferências sobre os parâmetros.

Assim sendo, o conhecimento das distribuições amostrais das principais estatísticas é necessário para fazer inferências sobre os parâmetros do modelo probabilístico da população. Por hora, basta conhecer as distribuições amostrais das estatísticas média de uma variável quantitativa qualquer, e proporção de um dos dois únicos resultados de uma variável qualitativa.

4.3.1 – Distribuição amostral da média

Vamos observar as particularidades da distribuição amostral da média.

Exemplo 2 - Suponha uma variável quantitativa cujos valores constituem uma população com os seguintes valores: (**2, 3, 4, 5**)

Para esta população, que tem uma distribuição uniforme, podemos observar que os parâmetros são: μ **= 3,5** σ^2 **= 1,25** (usou-se **n** no denominador por ser uma população)

Se retirarmos todas as amostras aleatórias de 2 elementos (com reposição) possíveis desta população (**n = 2**), teremos os seguintes resultados:LINK Há 16 amostras possíveis. LINK

(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)**

O cálculo das médias de todas as amostras acima resultará na matriz abaixo:

$$\bar{X}\begin{Bmatrix} (2,0) & (2,5) & (3,0) & (3,5) \\ (2,5) & (3,0) & (3,5) & (4,0) \\ (3,0) & (3,5) & (4,0) & (4,5) \\ (3,5) & (4,0) & (4,5) & (5,0) \end{Bmatrix}$$

Se estas médias forem plotadas em um histograma (Figura 38):

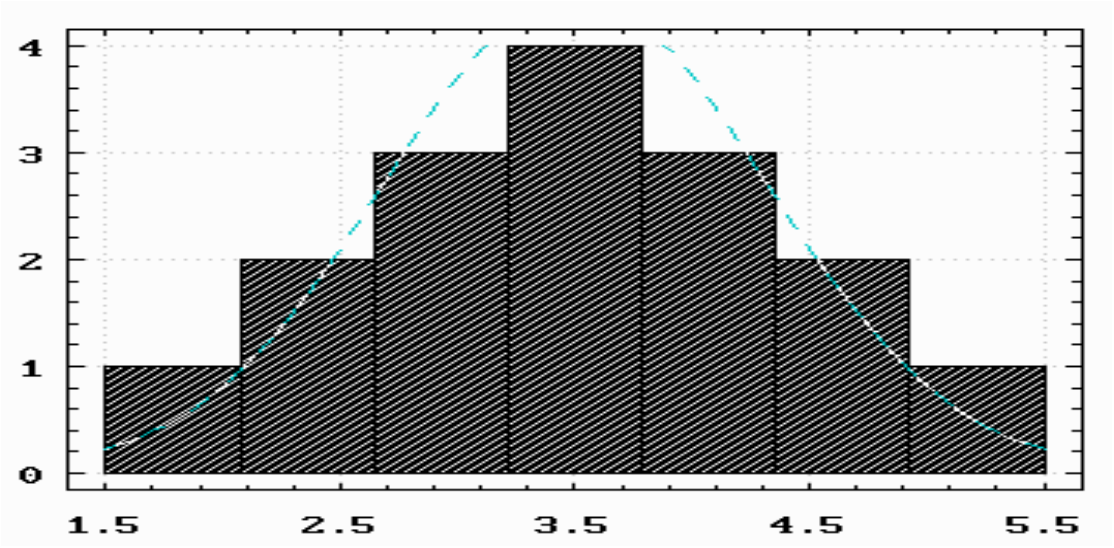


Figura 38 - Histograma de médias amostrais

Fonte: adaptada pelo autor de Statsoft ®

Se forem calculados a média e a variância das médias de todas as amostras o resultado será:

$$\bar{X} = 56/16 = 3{,}5 = \mu \qquad V(\bar{x}) = 0{,}625 = \frac{1{,}25}{2} = \frac{\sigma^2}{n}$$

Observe como a distribuição das médias amostrais da variável pode ser aproximada por um modelo normal (não obstante a distribuição da variável na população não ser normal), e que o valor esperado das médias amostrais (média das médias) é igual ao valor da média populacional da variável e a variância das médias amostrais é igual ao valor da variância populacional da variável dividida pelo tamanho da amostra. Quanto maior o tamanho da amostra (quanto maior **n**) mais o histograma acima vai se aproximar de um modelo normal, independentemente do formato da distribuição da variável na população.

Podemos então enunciar os teoremas:

Teorema das Combinações Lineares

Se a variável de interesse segue uma distribuição normal na população a distribuição amostral das médias de amostras aleatórias retiradas desta população também será normal, independentemente do tamanho destas amostras.

Teorema Central do Limite

Se a variável de interesse não segue uma distribuição normal na população (ou não se sabe qual é a sua distribuição) a distribuição amostral das médias de amostras aleatórias retiradas desta população será normal se o tamanho destas amostras for suficientemente grande, LINK Este "suficientemente grande" varia de distribuição para distribuição, como foi visto uma distribuição uniforme precisa de uma amostra pequena (n =2 no caso) para que a aproximação seja possível, outras distribuições precisam de amostras maiores. Alguns autores costumam chamar de "grandes amostras" aquelas que possuem mais de 30 elementos, a partir deste tamanho a aproximação poderia ser feita sem maiores

preocupações. LINK com uma média igual à média populacional e uma variância igual à variância populacional dividida pelo tamanho da amostra.

Para o caso da Proporção podemos chegar a uma conclusão semelhante.

4.3.2 – Distribuição amostral da proporção

Vamos estudar as particularidades da distribuição amostral da proporção através de um exemplo.

Exemplo 3.Seja uma variável qualitativa que pode assumir apenas dois valores, e que constitui a seguinte população:(□, □, □, □,■)

Vamos supor que há interesse no valor ■ (este valor seria o nosso "sucesso"). A proporção deste valor na população (o valor do parâmetro) será $\pi = 1/5$.

Se retirarmos todas as amostras aleatórias de 2 elementos (com reposição) possíveis desta população (**n = 2**) teremos os seguintes resultados:LINK Há 25 amostras possíveis. LINK

(□,□)	(□,□)	(□,□)	(□,□)	(□, ■)
(□,□)	(□,□)	(□,□)	(□,□)	(□, ■)
(□,□)	(□,□)	(□,□)	(□,□)	(□, ■)
(□,□)	(□,□)	(□,□)	(□,□)	(□, ■)
(□, ■)	(□, ■)	(□, ■)	(□, ■)	(■, ■)

Figura 39 - Amostras de tamanho 2 para proporção

Fonte: elaborada pelo autor

Observe que se definirmos a variável como o número de "sucessos" (número de ■) esta seguirá um modelo binomial: há apenas dois resultados possíveis para cada realização,

há um número limitado de realizações (**n** = 2 no caso), e cada realização independe da outra (porque a amostra é aleatória com reposição).

Calculando a proporção de ■ em cada uma das amostras, e chamando esta proporção amostral de **p**, teremos os seguintes resultados:

$$p = \begin{matrix} (0) & (0) & (0) & (0) & (1/2) \\ (0) & (0) & (0) & (0) & (1/2) \\ (0) & (0) & (0) & (0) & (1/2) \\ (0) & (0) & (0) & (0) & (1/2) \\ (1/2) & (1/2) & (1/2) & (1/2) & (1) \end{matrix}$$

Calculando a média (valor esperado) e a variância das proporções acima teremos:

$$\bar{X} = E(p) = \frac{1}{5} = \pi \qquad s^2 = 0{,}08 = \frac{\left(\frac{1}{5}\right) \times \left(1 - \frac{1}{5}\right)}{2} = \frac{\pi \times (1-\pi)}{n}$$

Observe que o valor esperado (média) das proporções amostrais é igual ao valor da proporção populacional de ■, e que a variância das proporções amostrais é igual ao produto da proporção populacional de ■ por seu complementar, dividido pelo tamanho da amostra.LINK Voltaremos a analisar o significado deste resultado quando estudarmos Estimação por Ponto. LINK

Lembre-se de que um modelo binomial pode ser aproximado por um modelo normal se algumas condições forem satisfeitas: se o produto do número de realizações pela probabilidade de "sucesso" (**n** × **p**) E o produto do número de realizações pela probabilidade de "fracasso" (**n** × [**1 - p**]) forem ambos maiores ou iguais a 5. LINK Isto também é decorrência do Teorema Central do Limite. LINK E esta distribuição normal teria média igual a **n** × **p** e variância igual a **n** × **p** × (**1 - p**). Se estivermos interessados apenas na proporção (probabilidade de "sucesso") e não no número de "sucessos" as expressões anteriores podem ser divididas por **n** (o tamanho da amostra): média = **p** e variância = [**p** × (**1**- **p**) / **n**].

Por causa do Teorema Central do Limite é que o modelo normal é tão importante. É claro que ele representa muito bem uma grande variedade de fenômenos, mas é devido à sua utilização em Inferência Estatística que o seu estudo é imprescindível. Ressalte-se, porém, que a sua aplicação costuma resumir-se ao que se chama de Inferência Paramétrica, inferências sobre os parâmetros dos modelos probabilísticos que descrevem as variáveis na população. Para fazer inferências sobre outros aspectos que não os parâmetros, ou quando as amostras utilizadas não forem suficientemente grandes para se assumir a validade do Teorema Central do Limite, é preciso usar técnicas de Inferência Não Paramétrica (que nós não veremos nesta disciplina).

Tô afim de saber:

- Sobre distribuição amostral - BARBETTA, P.A., REIS, M.M., BORNIA, A.C. **Estatística para Cursos de Engenharia e Informática**. 2ª ed. - São Paulo: Atlas, 2008, capítulo 7;

STEVENSON, Willian J. **Estatística Aplicada à Administração**. São Paulo: Ed. Harbra, 2001, capítulo 7;

ANDERSON, D.R., SWEENEY, D.J., WILLIAMS, T.A., **Estatística Aplicada à Administração e Economia.** 2ª ed. – São Paulo: Thomson Learning, 2007, capítulo 7.

- Sobre a utilização do Microsoft Excel ® para estudar distribuições amostrais veja LEVINE, D. M., STEPHAN, D., KREHBIEL, T. C., BERENSON, M. L. **Estatística:** Teoria e Aplicações - Usando Microsoft Excel em Português. 5ª ed. – Rio de Janeiro: LTC, 200, capítulo 5.

Resumo

O resumo desta Unidade está mostrado na Figura40:

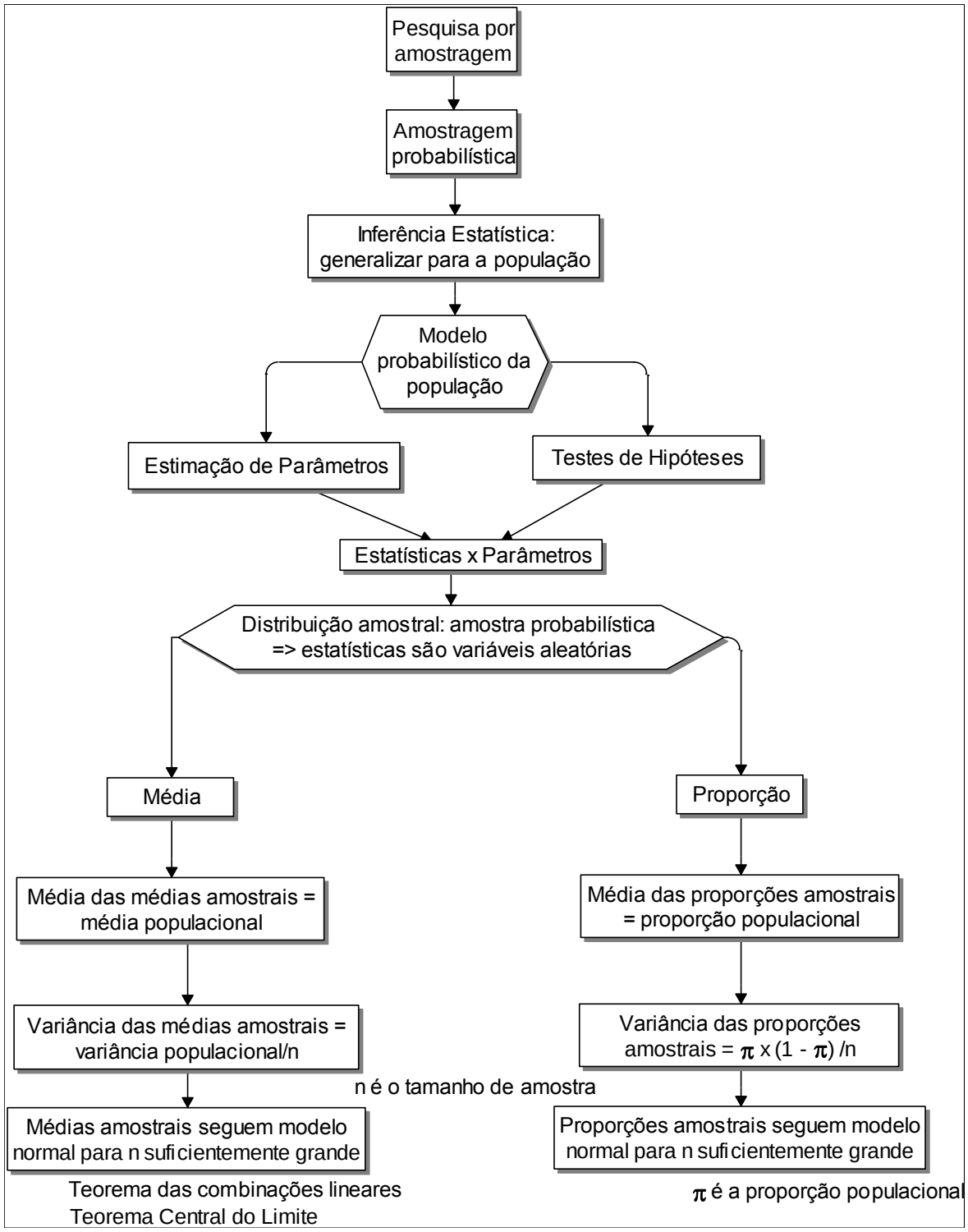


Figura 40 - Resumo da Unidade 4

Fonte: elaborado pelo autor

Atividades de aprendizagem

1) Uma variável tem média 200 e desvio padrão 12 na população, com uma distribuição bastante assimétrica.

a) Imagine que serão retiradas amostras aleatórias de 2 elementos desta população.

a.1 – Encontre a média das médias amostrais. (R. 200)

a.2 – Encontre o desvio padrão das médias amostrais (R. 8,485)

a.3 – A distribuição das médias amostrais será aproximadamente normal? JUSTIFIQUE.

b) Imagine que serão retiradas amostras aleatórias de 225 elementos desta população.

b.1 – Encontre a média das médias amostrais. (R. 200)

b.2 – Encontre o desvio padrão das médias amostrais (R. 0,8)

b.3 – A distribuição das médias amostrais será aproximadamente normal? JUSTIFIQUE.

2) O censo indicou que 60% dos homens de um município são casados. Se fossem retiradas amostras aleatórias de 200 elementos da população de homens.

a) Qual é a média da proporção amostral de casados? (R. 0,60).

b) Qual é o desvio padrão da proporção amostral de casados? (R. 0,0346).

c) A distribuição das proporções amostrais será aproximadamente normal? JUSTIFIQUE.

d) Supondo que a distribuição das proporções amostrais possa ser considerada normal, qual é a probabilidade de uma proporção de uma das amostras aleatórias diferir em mais de 5% (para mais ou para menos) da proporção populacional? (R. aproximadamente 0,1484)

3) Sabe-se que 50% dos edifícios construídos em uma grande cidade apresentam problemas estéticos relevantes em menos de 5 anos após a entrega da obra. Considerando a seleção de uma amostra aleatória simples com 200 edifícios com 5 anos, qual é a probabilidade de menos de 90 deles apresentarem problemas estéticos relevantes (considerar que não tenha havido obras de reparo nos edifícios selecionados)? (R.: aproximadamente 0,0783).

Caro estudante,

Esta Unidade foi muito importante para o seu aprendizado, pois lhe dará base para chegar à Inferência Estatística propriamente dita, assunto que será tema de discussão nas Unidades 5 e 6. Vimos até agora sobre a inferência estatística e distribuição amostral, seu modelo probabilístico e testes de hipóteses. Chegamos ao final desta Unidade e a continuidade da aprendizagem proposta desde o início deste material. Interaja com seus colegas, responda a atividade de aprendizagem e visite o Ambiente Virtual de Ensino-Aprendizagem, espaço este que contemplará suas possíveis dúvidas.Procure seu tutor e solicite todas as informações necessárias para o seu aprendizado. Bons estudos!!!

Unidade 5

Estimação de parâmetros

Objetivo

Nesta Unidade você vai conhecer e aplicar os conceitos de Estimação de Parâmetros por Ponto e por Intervalo de Média e Proporção, e aprenderá como calcular o tamanho mínimo de amostra necessário para a Estimação por Intervalo.

Prezado estudante!

Na Unidade 4 você viu o conceito de Distribuição Amostral, e observou a importância do modelo normal. Nesta Unidade você vai aprender como aplicar estes conceitos no primeiro tipo particular de Inferência Estatística, a **Estimação de Parâmetros**: por ponto e por intervalo.

Parâmetros são medidas de síntese de variáveis quantitativas na População que estamos pesquisando. Por ser inviável ou inconveniente pesquisar toda a População coletamos uma amostra para estudá-la. Os resultados da amostra podem ser então usados para fazer afirmações probabilísticas sobre o parâmetro de interesse: definir um intervalo possível para os valores do parâmetro e calcular a probabilidade de que o valor real do parâmetro esteja dentro dele (esta é a Estimação por Intervalo).

Vamos aprender como estimar os parâmetros média de uma variável quantitativa e proporção de um dos valores de uma variável qualitativa. Além disso, você vai ver como é possível definir de forma mais acurada o tamanho mínimo de uma amostra aleatória para estimar média e proporção (para esta última apresentamos uma primeira expressão de cálculo Unidade 3).

5.1 – Estimação por Ponto

Uma vez tendo decidido que modelo probabilístico é mais adequado para representar a variável de interesse na População resta obter os seus parâmetros. Nos estudos feitos com base em amostras é preciso escolher qual das estatísticas da amostra será o melhor estimador para cada parâmetro do modelo.

A Estimação por PontoGLOSSÁRIO Estimação por ponto: tipo de estimação de parâmetros que procura identificar qual é o melhor estimador para um parâmetro populacional a partir das várias estatística amostrais disponíveis, seguindo alguns

critérios.Fonte: Barbetta, Reis e Bornia, 2010. Fim GLOSSÁRIOconsiste em determinar qual será o melhor estimador para o parâmetro de interesse.

Como os parâmetros serão estimados através das estatísticas, estimadores,de uma amostra aleatória, e como para cada amostra aleatória as estatísticas apresentarão diferentes valores, os estimadores também terão valores aleatórios. Em outras palavras um Estimador é uma variável aleatória que pode ter um modelo probabilístico para descrevê-la.

Naturalmente haverá várias estatísticas **T** que poderão ser usadas como estimadores de um parâmetro θqualquer. Como escolher qual das estatísticas será o melhor estimador para o parâmetro?

Há basicamente três critérios para a escolha de um estimador: o estimador precisa ser justo, consistente e eficiente.

1) Um Estimador **T** é um estimador justo (não tendencioso) de um parâmetro θ quando o valor esperado de **T** é igual ao valor do parâmetro θ a ser estimado: **E(T)** $=\theta$

2) Um Estimador T é um estimador consistente de um parâmetro θ quando além ser um estimador justo a sua variância tende a zero à medida que o tamanho da amostra aleatória aumenta: $\lim_{n\to\infty} V(T) = 0$.

3) Se há dois Estimadores justos de um parâmetro o mais eficiente é aquele que apresentar a menor variância.

Conforme foi dito na introdução desta Unidade, estamos interessados em estimar dois parâmetros: média e proporção populacional. Vamos então buscar os estimadores mais apropriados para ambos.

5.1.1 – Estimação por ponto dos principais parâmetros

Os principais parâmetros que vamos avaliar aqui são: média de uma variável que segue um modelo normal (ou qualquer modelo se a amostra for suficientemente grande) em uma população (média populacional - **μ**) e proporção de ocorrência de um dos valores de uma variável que segue um modelo Binomial em uma população (proporção populacional - **π**). Em suma escolher quais estatísticas amostrais são mais adequadas para estimar estes parâmetros, usando os critérios definidos acima.

Lembrando-se dos Exemplos 2, e 3 da Unidade 4, algumas constatações que lá foram feitas passarão a fazer sentido agora.

Vamos supor que houvesse a intenção de estimar a média populacional da variável do Exemplo 2. Qual das estatísticas disponíveis seria o melhor estimador?

Lembrem-se de que após retirar todas as amostras aleatórias possíveis daquela população, calculamos a média de cada amostra, e posteriormente a média dessas médias. Constatou-se que o valor esperado das médias amostrais (média das médias) é igual ao valor da média populacional da variável e a variância das médias amostrais é igual ao valor da variância populacional da variável dividida pelo tamanho da amostra:

$$E(\bar{x}) = \mu \qquad V(\bar{x}) = \frac{\sigma^2}{n}$$

O melhor estimador da média populacional **μ** é a média amostral $\bar{x}$, pois se trata de um estimador justo e consistente:

- Justo porque o valor esperado da média amostral será a média populacional;
- Consistente porque se o tamanho da amostra **n** tender ao infinito a variância da média amostral (do Estimador) tenderá a zero.

Agora vamos supor que houvesse a intenção de estimar a proporção populacional do valor ■ da variável do Exemplo 3. Qual das estatísticas disponíveis seria o melhor estimador?

Lembrem-se de que após retirar todas as amostras aleatórias possíveis daquela população, calculamos a proporção de ■ em cada amostra, e posteriormente a média dessas proporções. Constatou-se que o valor esperado das proporções amostrais (média das proporções) é igual ao valor da proporção populacional do valor ■ da variável e a variância das proporções amostrais é igual ao valor do produto da proporção populacional do valor ■ da variável pela sua complementar dividida pelo tamanho da amostra:

$$E(p) = \pi \qquad V(p) = \frac{\pi \times (1 - \pi)}{n}$$

O melhor estimador da proporção populacional **π** é a proporção amostral **p**, pois se trata de um estimador justo e consistente:

- Justo porque o valor esperado da proporção amostral será a proporção populacional;
- Consistente porque se o tamanho da amostra **n** tender ao infinito a variância da proporção amostral (do Estimador) tenderá a zero.

Poderíamos fazer um procedimento semelhante para estimar outros parâmetros, como, por exemplo, a variância populacional de uma variável. Este procedimento não será demonstrado, mas o melhor estimador da variância populacional será a variância amostral se for usado **n - 1**no denominador da expressão de cálculo. Somente assim a variância amostral será um estimador justo (não viciado) da variância populacional.

Como o desvio padrão é a raiz quadrada da variância é comum estimar o desvio padrão populacional extraindo a raiz quadrada da variância amostral.

O problema da Estimação por Ponto é que geralmente só dispomos de uma amostra aleatória. Intuitivamente, qual será a probabilidade de que a média ou proporção amostral, de uma amostra aleatória, coincida exatamente com o valor do parâmetro? É como pescar usando uma lança de bambu... É preciso muita habilidade para pegar o peixe... Mas, se você puder usar uma rede, ficará bem mais fácil. Esta "rede" é a Estimação por Intervalo.

5.2 – **Estimação por Intervalo de Parâmetros**

Geralmente uma inferência estatística é feita com base em uma única amostra: na maior parte dos casos é totalmente inviável retirar todas as amostras possíveis de uma determinada população.

Intuitivamente percebemos que as estatísticas calculadas nessa única amostra, mesmo sendo os melhores estimadores para os parâmetros de interesse, terão uma probabilidade infinitesimal de coincidir exatamente com os valores reais dos parâmetros. Então a Estimação por Ponto dos parâmetros é insuficiente, e as estimativas assim obtidas servirão apenas como referência para a Estimação por Intervalo.

A Estimação por Intervalo consiste em colocar um Intervalo de Confiança (I.C.) em torno da estimativa obtida através da Estimação por Ponto.

O Intervalo de Confiança GLOSSÁRIO Intervalo de confiança: faixa de valores da estatística usada como estimador, dentro da qual há uma probabilidade conhecida de que o verdadeiro valor do parâmetro esteja. Sinônimo de estimação por intervalo. Fonte: Barbetta, Reis e Bornia, 2010. FimGLOSSÁRIOterá uma certa probabilidade chamada de Nível de confiança (que costuma ser simbolizado como **1 - α**) de conter o valor real do parâmetro LINK fazer uma Estimação por Intervalo de um parâmetro é efetuar uma afirmação probabilística sobre este parâmetro, indicando uma faixa de possíveis valores. LINK e a probabilidade de que esta faixa realmente contenha o valor real do parâmetro. A probabilidade de que o Intervalo de Confiança não contenha o valor real do parâmetro é chamada de Nível de Significância (**α**), e o valor desta probabilidade será o complementar do Nível de Confiança.GLOSSÁRIO Nível de confiança: probabilidade de que o intervalo de confiança contenha o valor real do parâmetro a estimar, espera-se que seja um valor alto, de no mínimo 90%. Fonte: Moore, McCabe, Duckworth e Sclove, 2006. Fim GLOSSÁRIO É comum definir o Nível de Significância como uma probabilidade máxima de erro, um risco máximo admissível.

A determinação do Intervalo de Confiança para um determinado parâmetro resume-se basicamente a definir o Limite Inferior e o Limite Superior do intervalo, supondo um determinado Nível de Confiança ou Significância.GLOSSÁRIO Nível de Significância: complementar do nível de confiança, a probabilidade de que o intervalo de confiança não contenha o valor real do parâmetro. Probabilidade de erro espera-se que seja um valor baixo, de no máximo 10%. Fonte: Barbetta, Reis e Bornia, 2010. Fim GLOSSÁRIO

A definição dos limites dependerá também da distribuição amostral da estatística usada como referência para o intervalo e do tamanho da amostra utilizada.

Para os dois parâmetros em que temos maior interesse (média populacional **μ** e proporção populacional $\boldsymbol{\pi}$) a distribuição amostral dos estimadores (média amostral $\bar{\mathbf{x}}$ e proporção amostral **p**, respectivamente) pode ser aproximada por uma distribuição normal: o Intervalo de Confiança será então simétrico em relação ao valor calculado da estimativa (média ou proporção amostral), com base na amostra aleatória coletada (Figura 66):

Figura 41 - Intervalo de Confiança para um modelo normal
Fonte: elaborada pelo autor

Onde: $\mathbf{L_i}$ é o limite inferior e $\mathbf{L_s}$ é o limite superior do Intervalo de Confiança; **1 - α** é o Nível de Confiança estabelecido, observando que o valor do Nível de Significância **α** é dividido igualmente entre os valores abaixo de $\mathbf{L_i}$ e acima de $\mathbf{L_s}$.

Para obter os limites em função do Nível de Confiança devemos utilizar a distribuição normal padrão (variável **Z** com média zero e variância um): fixar um certo valor de probabilidade, obter o valor de Z correspondente, e substituir o valor em **Z = (x -**

“média**”) “**desvio padrão**”**, LINK Foram colocados entre aspas porque os valores dependerão dos parâmetros sob análise e de outros fatores. LINK para obter o valor **x** (valor correspondente ao valor de **Z** para a probabilidade fixada). Observe a Figura 42:

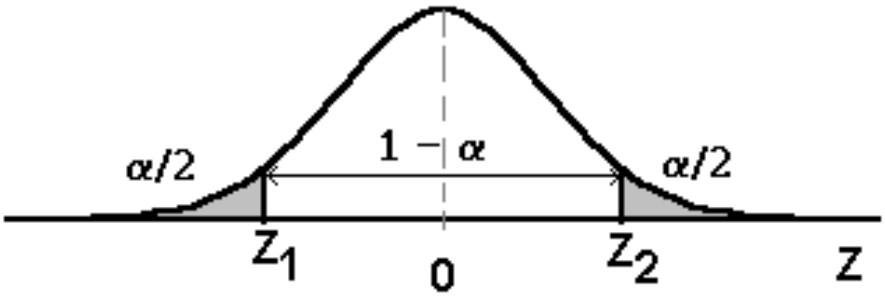


Figura 42 - Intervalo de Confiança para a distribuição normal padrão

Fonte: elaborada pelo autor

O limite $\mathbf{L_i}$ (inferior) corresponde a $\mathbf{Z_1}$ e o limite $\mathbf{L_s}$ (superior) corresponde a $\mathbf{Z_2}$. O ponto central 0 (zero) corresponde ao valor calculado da Estimativa. Como a variável **Z** tem distribuição normal com média igual a zero (lembrando que a distribuição normal é simétrica em relação à média) os valores de $\mathbf{Z_1}$ e $\mathbf{Z_2}$ serão iguais em módulo ($\mathbf{Z_1}$ será negativo e $\mathbf{Z_2}$ positivo):

$\mathbf{Z_1}$ será um valor de **Z** tal que $P(Z \le Z_1) = \frac{\alpha}{2}$, e $\mathbf{Z_2}$ será um valor tal que $P(Z \le Z_2) = 1 - \frac{\alpha}{2}$

Então obteremos os valores dos limites através das expressões:

Z_1 = (L_i - “média”)/ “desvio padrão” => L_i = “média” + Z_1× “desvio padrão”

Z_2 = (L_s - “média”)/ “desvio padrão” => L_s = “média” + Z_2× “desvio padrão”

Como $\mathbf{Z_1}$ = - $\mathbf{Z_2}$, podemos substituir:

L_i = “média” - Z_2× “desvio padrão” L_s = “média” + Z_2× “desvio padrão”

E este valor $\mathbf{Z_2}$ costuma ser chamado de $\mathbf{Z_{crítico}}$, porque corresponde aos limites do intervalo:

L_i = “média” - $Z_{crítico}$× “desvio padrão” L_s = “média” + $Z_{crítico}$× “desvio padrão”

Reparem que o mesmo valor é somado, e subtraído da "média". Este valor é chamado de semi-intervalo ou precisão do intervalo, ou margem de erro, $\mathbf{e_0}$:

$$e_0 = Z_{crítico} \times \text{"desvio padrão"}$$

Resta agora definir corretamente o valor da "média" e do "desvio padrão" para cada um dos parâmetros em que estamos interessados (média e proporção populacional). Com base nas conclusões obtidas na Estimação por Ponto isso será simples. Contudo, há alguns outros aspectos que precisarão ser esmiuçados.

5.2.1 – Estimação por Intervalo da Média Populacional

Lembrando das expressões anteriores:

$$L_i = \text{"média"} - Z_{crítico} \times \text{"desvio padrão"} = \text{"média"} - e_0$$

$$L_s = \text{"média"} + Z_{crítico} \times \text{"desvio padrão"} = \text{"média"} + e_0$$

Neste caso a "média" será a média amostral $\bar{x}$ (ou mais precisamente o seu valor):

$$\mathbf{P(\bar{x} - e_0 \leq \mu \leq \bar{x} + e_0) = 1 - \alpha}$$

O valor de $\mathbf{e_0}$ dependerá de outros aspectos.

a) Se a variância populacional σ^2 da variável (cuja média populacional queremos estimar) for conhecida.

Neste caso a variância amostral da média poderá ser calculada através da expressão:

$V(\bar{x}) = \frac{\sigma^2}{n}$, e, por conseguinte, o "desvio padrão" será $\text{desvio padrao} = \frac{\sigma}{\sqrt{n}}$

E $\mathbf{e_0}$ será: $\quad e_0 = Z_{critico} \times \frac{\sigma}{\sqrt{n}}$

Bastará então fixar o Nível de Confiança (ou de Significância) para obter $\mathbf{Z_{crítico}}$ através da Tabela disponível no Ambiente Virtual e calcular $\mathbf{e_0}$.

b) Se a variância populacional σ^2 da variável for desconhecida.

Naturalmente este é o caso mais encontrado na prática. Como se deve proceder?

Dependerá do tamanho da amostra.

b.1 - Grandes amostras (mais de 30 elementos)

Nestes casos procede-se como no item anterior, apenas fazendo com que σ = **s**, ou seja, considerando que o desvio padrão da variável na população é igual ao desvio padrão da variável na amostra (suposição razoável para grandes amostras).

b.2 - Pequenas amostras (até 30 elementos)

Nestes casos a aproximação do item b.1 não será viável. Terá que ser feita uma correção na distribuição normal padrão (**Z**) através da distribuição **t de Student** que estudamos na Unidade 2.

Quando a variância populacional da variável é desconhecida e a amostra tem até 30 elementos substitui-se σ por **s** e **Z** por $\mathbf{t_{n-1}}$ em todas as expressões para determinação dos limites do intervalo de confiança, obtendo:

Li = "média" - $t_{n-1,crítico}$× "desvio padrão" = "média" - e_0

Ls = "média" + $t_{n-1,crítico}$× "desvio padrão" = "média" + e_0

E $\mathbf{e_0}$ será: $$e_0 = t_{n-1,critico} \times \frac{s}{\sqrt{n}}$$

Os valores de $\mathbf{t_{n-1,crítico}}$ podem ser obtidos de forma semelhante aos de $\mathbf{Z_{crítico}}$, definindo o Nível de Confiança (ou de Significância), mas precisam também da definição do número de graus de liberdade (**n - 1**): tendo estes valores basta procurar o valor da Tabela 2 do Ambiente Virtual ou em um programa computacional.

Se o tamanho da amostra (n) for superior a 5% do tamanho da população (N) os valores de e_0 precisam ser corrigidos. Caso contrário os limites dos intervalos não serão acusados. A correção é mostrada na equação a seguir:

$$e_{0_{corrigido}} = e_0 \times \sqrt{\frac{N-n}{N-1}}$$

Exemplo 1 - Retirou-se uma amostra aleatória de 4 elementos de uma produção de cortes bovinos no intuito de estimar a média do peso do corte. Obteve-se média de 8,2 kg e desvio padrão de 0,4 kg. Supondo população normal.Determinar um intervalo de confiança para a média populacional com 1% de significância.

O parâmetro de interesse é a média populacional **μ** do peso do corte.

Adotou-se um nível de significância de 1%, então **α** = 0,01e **1 - α** = 0,99.LINK Este valor pode ser arbitrado pelo usuário ou pode ser uma exigência do problema sob análise, ou até mesmo uma exigência legal. Os níveis de significância mais comuns são de 1%, 5% ou mesmo 10%. LINK

As estatísticas disponíveis são: **média amostral** = 8,2 kg **s** = 0,4 kg **n** = 4 elementos.

Definição da variável de teste: como a variância populacional é DESCONHECIDA, e o tamanho da amostra é menor do que 30 elementos, não obstante a população ter distribuição normal, a distribuição amostral da média será **t** de Student, e a variável de teste será $\mathbf{t_{n-1}}$.

Encontrar o valor de $\mathbf{t_{n-1,crítico}}$: como o Intervalo de Confiança para a média é bilateral, teremos uma situação semelhante à da Figura 43:

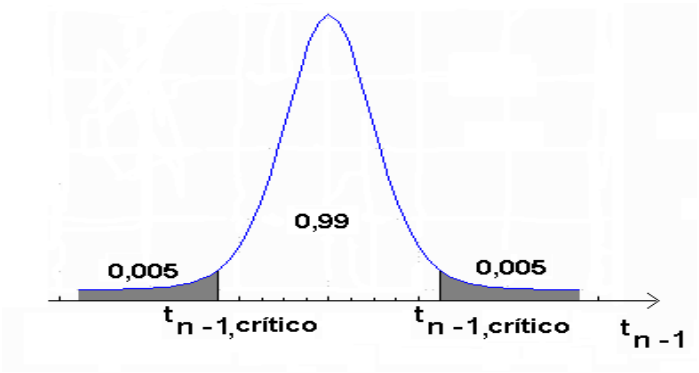


Figura 43- Distribuição t de Student para 99% de confiança

Fonte: elaborada pelo autor a partir de Statgraphics ®

Para encontrar o valor crítico devemos procurar na tabela da distribuição de Student, na linha correspondente a **n-1** graus de liberdade, ou seja, em 4 - 1 = 3 graus de liberdade. O valor da probabilidade pode ser visto na Figura acima: os valores críticos serão $\mathbf{t_{3;0,005}}$e $\mathbf{t_{3;0,995}}$os quais serão iguais em módulo. E o valor de $\mathbf{t_{n-1,crítico}}$será igual a **5,84** (em módulo)

Determinam-se os limites do intervalo, através da expressão abaixo (cujo resultado será somado e subtraído da média amostral) para determinar os limites do intervalo:

$$e_0 = \frac{t_{n-1,crítico} * s}{\sqrt{n}} = \frac{5,84 * 0,4}{\sqrt{4}} = 1,168kg$$

$$L_I = \bar{x} - e_0 = 8,2 - 1,168 = 7,032kg \quad L_S = \bar{x} + e_0 = 8,2 + 1,168 = 9,368kg$$

Então o intervalo de 99% de confiança para a média populacional da dimensão é [7,032;9,368] kg. **Interpretação**: há 99% de probabilidade de que a verdadeira média populacional do peso de corte esteja entre 7,032 e 9,368 kg.

5.2.2 – Estimação por Intervalo da Proporção Populacional

Anteriormente declaramos que o melhor estimador para a proporção populacional $\boldsymbol{\pi}$ é a proporção amostral **p**. E que esta proporção amostral teria média igual a $\boldsymbol{\pi}$ e variância igual a [$\boldsymbol{\pi}$ x (**1 - $\boldsymbol{\pi}$**)]/**n** onde **n** é o tamanho da amostra aleatória. A distribuição da

proporção amostral **p** é binomial, e sabe-se que a distribuição binomial pode ser aproximada por uma normal se algumas condições forem satisfeitas:
Se $\mathbf{n}\times\boldsymbol{\pi}\geq 5$E $\mathbf{n}\times(\mathbf{1}-\boldsymbol{\pi})\geq 5$.

Ora, se $\boldsymbol{\pi}$ fosse conhecido não estaríamos aqui nos preocupando com a sua Estimação por Intervalo, assim vamos verificar se é possível aproximar a distribuição binomial de **p** por uma normal se:

$\mathbf{n}\times\mathbf{p}\geq 5$E $\mathbf{n}\times(\mathbf{1}-\mathbf{p})\geq 5$, ou seja, usando o próprio valor da proporção amostral observada (trata-se de uma aproximação razoável).

Se e somente se estas duas condições forem satisfeitas poderemos usar as expressões abaixo (lembrando das expressões anteriores):

L_i = "média" - $Z_{crítico}\times$ "desvio padrão" = "média" - e_0

L_s = "média" + $Z_{crítico}\times$ "desvio padrão" = "média" + e_0

Neste caso a "média" será a proporção amostral (ou mais precisamente o seu valor):

$$P(p-e_0\leq\mu\leq p+e_0)=1-\alpha$$

E o valor do "desvio padrão" será igual a $\sqrt{\frac{\pi\times(1-\pi)}{n}}$. Novamente, como $\boldsymbol{\pi}$ é desconhecido, usaremos a proporção amostral **p** como aproximação.

Então $\mathbf{e_0}$ será: $$e_0=Z_{critico}\times\sqrt{\frac{p\times(1-p)}{n}}$$

Bastará então fixar o Nível de Confiança (ou de Significância), $\mathbf{Z_{crítico}}$ e calcular $\mathbf{e_0}$.

Novamente, precisamos corrigir o valor de e_0 para o caso de população finita:

$$e_{0_{corrigido}}=e_0\times\sqrt{\frac{N-n}{N-1}}$$

Em suma a Estimação por Intervalo da média e da proporção populacional consiste basicamente em calcular a amplitude do semi-intervalo (o $\mathbf{e_0}$), de acordo com as condições do problema sob análise.

- Para a média, observar se é viável considerar que a distribuição da variável na população é normal, ou que a amostra seja suficientemente grande para que a distribuição das médias amostrais possa ser considerada normal;
- Se isso for verificado, identificar se a variância populacional da variável é conhecida: caso seja deverá ser usada a variável **Z** da distribuição normal padrão, para qualquer tamanho de amostra;
- Se variância populacional da variável é desconhecida há duas possibilidades: para amostras com mais de 30 elementos usar a variável **Z**, e fazer a variância populacional igual à variância amostral da variável; se a amostra tem até 30 elementos usar a variável $\mathbf{t_{n-1}}$ da distribuição de Student;
- Para a proporção, observar se é possível fazer a aproximação pela distribuição normal.

Exemplo 2. Retirou-se uma amostra aleatória de 1000 peças de um lote. Verificou-se que 35 eram defeituosas.Determinar um intervalo de confiança de 95% para a proporção peças defeituosas no lote.

O parâmetro de interesse é a proporção populacional $\boldsymbol{\pi}$ de peças defeituosas.

Adotou-se um nível de significância de 5%, então $\boldsymbol{\alpha} = 0{,}05$e $\mathbf{1 - \alpha} = 0{,}95$

As estatísticas são: proporção amostral de peças defeituosas $\mathbf{p} = 35/1000$**n** = 1000 elementos.

Definição da variável de teste: precisamos verificar se é possível fazer a aproximação pela normal, então**n** x **p** = 1000 x 0,035 = 35 > 5 e **n** x (**1- p**) = 1000 x 0,965 = 965 > 5. Como ambos os produtos satisfazem as condições para a aproximação podemos usar a variável **Z** da distribuição normal padrão

Encontrar o valor de $\mathbf{Z_{crítico}}$: como o Intervalo de Confiança para a média é bilateral, teremos uma situação semelhante à da Figura :

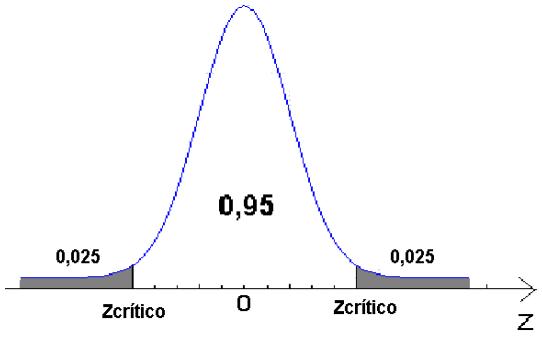


Figura 44 - Distribuição normal padrão para 95% de confiança

Fonte: elaborada pelo autor

Para encontrar o valor crítico devemos procurar na tabela da distribuição normal padrão pela probabilidade 0,975 (0,95+0,025). O valor da probabilidade pode ser visto na Figura 84acima: os valores críticos serão $\mathbf{Z_{0,025}}$e $\mathbf{Z_{0,975}}$os quais serão iguais em módulo. E o valor de $\mathbf{Z_{crítico}}$será igual a 1,96 (em módulo).

Passa-se agora a determinação dos limites do intervalo, através da expressão abaixo, cujo resultado será somado e subtraído da proporção amostral de peças defeituosas, para determinar os limites do intervalo:

$$e_0 = Z_{critico} \times \sqrt{\frac{p \times (1-p)}{n}} = 1{,}96 \times \sqrt{\frac{0{,}035 \times 0{,}965}{1000}} = 0{,}0114$$

$$L_I = p - e_0 = 0{,}035 - 0{,}0114 = 0{,}0236 \quad L_S = p + e_0 = 0{,}035 + 0{,}0114 = 0{,}0464$$

Então, o intervalo de 95% de confiança para a proporção populacional de peças defeituosas é [2,36%;4,64%]. Interpretação: há 95% de probabilidade de que a verdadeira proporção populacional de plantas atacadas pelo fungo esteja entre 2,36% e 4,64%.

5.3 – Tamanho mínimo de amostra para Estimação por Intervalo

Como foi observado nos itens anteriores a determinação dos limites de um Intervalo de Confiança(determinação do $\mathbf{e_0}$) depende do tamanho da amostra aleatória coletada, além do Nível de Confiança e da distribuição amostral do estimador utilizado. Nada podemos fazer quanto à distribuição amostral do estimador, o Nível de Confiança nós podemos controlar, seria interessante definir então uma precisão (um valor para $\mathbf{e_0}$) para o Intervalo de Confiança: é muito comum querermos estabelecer previamente qual será a faixa de variação de um determinado parâmetro, com uma certa confiabilidade.

Contudo, para um mesmo tamanho de amostra:

- se aumentarmos o Nível de Confiança (reduzirmos o Nível de Significância) teremos um valor crítico maior, o que aumentará o valor de $\mathbf{e_0}$, resultando em um Intervalo de Confiança mais "largo", com menor precisão.
- se resolvermos aumentar a precisão (menor valor de $\mathbf{e_0}$), obter um Intervalo de Confiança mais "estreito", teremos uma queda no Nível de Confiança.

A solução para o dilema acima é obter um **tamanho mínimo de amostra** capaz de atender simultaneamente ao Nível de Confiança (ou de Significância) e à precisão ($\mathbf{e_0}$) especificados. Como as expressões de $\mathbf{e_0}$ são em função do tamanho de amostra (**n**), seria razoável pensar em reordená-las de forma a fazer com que o tamanho de amostra seja função do Nível de Confiança e da precisão ($\mathbf{e_0}$).

5.3.1 – Tamanho mínimo de amostra para Estimação por Intervalo da Média Populacional

a) Variância populacional conhecida

$$e_0 = Z_{critico} \times \frac{\sigma}{\sqrt{n}} \qquad \text{isolando } \mathbf{n}: \qquad n = \left(\frac{Z_{critico} \times \sigma}{e_0} \right)^2$$

Neste caso basta especificar o valor de $\mathbf{e_0}$ (na mesma unidade do desvio padrão populacional σ), e o Nível de Confiança (que será usado para encontrar o $\mathbf{Z_{crítico}}$) e calcular o tamanho mínimo de amostra.

b) Variância populacional desconhecida

$$e_0 = t_{n-1,critico} \times \frac{s}{\sqrt{n}} \qquad \text{isolando } \mathbf{n}: \qquad n = \left(\frac{t_{n-1,critico} \times s}{e_0} \right)^2$$

O procedimento neste caso seria semelhante exceto por um pequeno problema: "se estamos calculando o tamanho da amostra como podemos conhecer **n - 1** e o desvio padrão amostral **s**?"

Quando a variância populacional da variável é desconhecida o usual é retirar uma **amostra piloto**GLOSSÁRIO Amostra piloto: amostra teste, de tamanho arbitrado pelo pesquisador, a partir da qual são calculadas estatísticas necessárias para a determinação do tamanho mínimo de amostra. Fonte: Costa Neto, 2002. Fim GLOSSÁRIOcom um tamanho **n*** arbitrário. A partir dos resultados desta amostra são calculadas as estatísticas (entre elas o desvio padrão amostral **s**) que são substituídas na expressão acima.

Se $\mathbf{n \leq n^*}$ então a amostra piloto é suficiente para o Nível de Confiança e a precisão exigidos.

Se $\mathbf{n > n^*}$ então a amostra piloto é insuficiente para o Nível de Confiança e a precisão exigidas, sendo então necessário retornar à população e retirar os elementos necessários para completar o tamanho mínimo de amostra. O processo continua até que a amostra seja considerada suficiente.

Conforme visto na Unidade 3, se o tamanho da população for conhecido é recomendável corrigir o tamanho da amostra obtida, seja para o intervalo de confiança de média ou proporção, através da seguinte fórmula:

$$n_{corrigido} = \frac{N \times n}{N + n} \qquad \text{onde } N \text{ é o tamanho da população}$$

Assim procedendo, evitamos o inconveniente de obter um tamanho de amostra superior ao tamanho da população, o que pode ocorrer se N não for muito grande.

Exemplo 3 – Considere os dados do Exemplo 1. Para estimar a média, com 1% de significância e precisão de 0,2 kg, esta amostra é suficiente?

Como a variância populacional é desconhecida, e o tamanho da amostra é menor do que 30 elementos, não obstante a população ter distribuição normal, a distribuição amostral da média será **t** de Student, e a variável de teste será $\mathbf{t_{n-1}}$. Assim será usada a seguinte expressão para calcular o tamanho mínimo de amostra para a estimação por intervalo da média populacional.

$$n = \left(\frac{t_{n-1,critico} \times s}{e_0} \right)^2$$

O nível de significância é o mesmo do item a. Sendo assim, o valor crítico continuará sendo o mesmo: $\mathbf{t_{n-1,crítico}} = 5{,}84$. O desvio padrão amostral vale 0,4 kg, e o valor de $\mathbf{e_0}$, a precisão foi fixado em 0,2 kg. Basta então substituir os valores na expressão:

$$n = \left(\frac{t_{n-1,critico} \times s}{e_0} \right)^2 = \left(\frac{5{,}84 \times 0{,}4}{0{,}2} \right)^2 = 136{,}42 \cong 137 \text{ elementos}$$

Conclui-se que a amostra retirada é insuficiente, pois é menor do que o valor calculado acima.

5.3.2 – Tamanho mínimo de amostra para Estimação por Intervalo da Proporção Populacional

Para a proporção populacional teremos:

$$e_0 = Z_{critico} \times \sqrt{\frac{p \times (1-p)}{n}} \qquad \text{isolando n:} \qquad n = \left(\frac{Z_{critico}}{e_0} \right)^2 \times p \times (1-p)$$

É necessário especificar o Nível de Confiança (ou de Significância) que será usado para encontrar o $\mathbf{Z_{crítico}}$, e o valor de $\mathbf{e_0}$ (tomando o cuidado de que tanto $\mathbf{e_0}$ quanto **p** e **1- p** estejam **todos** como proporções adimensionais ou como percentuais) para que seja possível calcular o valor do tamanho mínimo de amostra.

Da mesma forma que no caso da Estimação da média quando a variância populacional é desconhecida teremos que recorrer à uma amostra piloto. No cálculo do tamanho mínimo de amostra para a Estimação por Intervalo da proporção populacional há porém uma solução alternativa: utiliza-se uma estimativa exageradaLINK Esta solução somente é usada quando a natureza da pesquisa é tal que não é possível retirar uma amostra piloto: a retirada de uma amostra piloto e a eventual retirada de novos elementos da população poderiam prejudicar muito o resultado da pesquisa. Paga-se então o preço de ter uma amostra substancialmente maior do que talvez fosse necessário. LINK da amostra, supondo o máximo valor possível para o produto **p**× (**1 - p**), que ocorrerá quando ambas as proporções forem iguais a 0,5 (50%).

Conforme visto na Unidade 3, se o tamanho da população for conhecido é recomendável corrigir o tamanho da amostra obtida, seja para o intervalo de confiança de média ou proporção, através da seguinte fórmula:

$$n_{corrigido} = \frac{N \times n}{N + n}$$ onde N é o tamanho da população

Assim procedendo, evitamos o inconveniente de obter um tamanho de amostra superior ao tamanho da população, o que pode ocorrer se N não for muito grande.

Exemplo 4 - Considere o caso do Exemplo 2. Supondo 99% de confiança e precisão de 1%, esta amostra é suficiente para estimar a proporção populacional?

De acordo com o Exemplo 2 é possível utilizar a aproximação pela distribuição normal. A expressão para o cálculo do tamanho mínimo de amostra para a proporção populacional será:

$$n = \left(\frac{Z_{critico}}{e_0} \right)^2 \times p \times (1-p)$$

Os valores de **p** e **1 - p** já são conhecidos: **p** = 0,035 **1 - p** = 0,965

O nível de confiança exigido é de 99%: para encontrar o valor crítico devemos procurar na tabela da distribuição normal padrão pela probabilidade 0,995 (0,99+0,005); os valores críticos serão $\mathbf{Z_{0,005}}$ e $\mathbf{Z_{0,995}}$ os quais serão iguais em módulo. E o valor de $\mathbf{Z_{crítico}}$ será igual a 2,58 (em módulo).

A precisão foi fixada em 1% (0,01). Substituindo os valores na expressão acima:

$$n = \left(\frac{Z_{critico}}{e_0} \right)^2 \times p \times (1-p) = \left(\frac{2,58}{0,01} \right)^2 \times 0,035 \times 0,965 = 2248,14 \cong 2249$$

Observe que o tamanho mínimo de amostra necessário para atender a 99% de confiança e precisão de 1% deveria ser de 2249 elementos. Como a amostra coletada possui apenas 1000 elementos ela é insuficiente para a confiança e precisão exigidas. Recomenda-se o retorno à população para a retirada aleatória de mais 1249 peças.

Visto tudo o que estudamos, agora você já pode acompanhar atentamente os resultados das pesquisas de opinião veiculadas na mídia. Apenas mais um pequeno adendo.

5.4 - "Empate técnico"

Estamos acostumados a ouvir declarações do tipo "os candidatos A e B estão tecnicamente empatados na preferência eleitoral". O que significa isso? Geralmente as pesquisas de opinião eleitoral consistem em obter as proporções de entrevistados que declara votar neste ou naquele candidato, naquele momento. Posteriormente as proporções são generalizadas estatisticamente para a população, através do cálculo de intervalos de confiança para as proporções de cada candidato. Se os intervalos de confiança das proporções de dois ou mais candidatos apresentam grandes superposições declara-se que há

um "empate técnico": as diferenças entre eles devem-se provavelmente ao acaso, e para todos os fins estão em condições virtualmente iguais, naquele momento.

Exemplo 3 - Imagine que uma pesquisa de opinião eleitoral apresentasse os seguintes resultados (intervalos de confiança para a proporção que declara votar no candidato) sobre a prefeitura do município de Tapioca. Quais candidatos estão tecnicamente empatados (Quadro 5)?

Opinião	Limite inferior %	Limite superior %
Godofredo Astrogildo	31%	37%
Filismino Arquibaldo	14%	20%
Urraca Hermengarda	13%	19%
Salustiano Quintanilha	22%	28%
Indecisos	11%	17%

Quadro 5 - Resultados de uma pesquisa eleitoral municipal

Fonte: fictícia, elaborado pelo autor.

Filismino e Urraca estão tecnicamente empatados, pois seus intervalos de confiança apresentam grande sobreposição. Godofredo está muito na frente, pois o limite inferior de seu intervalo é maior do que o limite superior de Salustiano, que está em segundo lugar. É importante ressaltar que o número de indecisos é razoável, variando de 11 a 17%, quando eles se decidirem poderão mudar completamente o quadro da eleição, ou garantir a vitória folgada de Godofredo.

Tô afim de saber:

- Sobre propriedades e características desejáveis de um estimador,

BARBETTA, P.A., REIS, M.M., BORNIA, A.C. **Estatística para Cursos de Engenharia e Informática**. 3ª ed. - São Paulo: Atlas, 2010, capítulo 7.

- Sobre estimadores e intervalos de confiança para variância,

TRIOLA, M**. Introdução à Estatística**, Rio de Janeiro: LTC, 1999, capítulo 6.

- Para entender melhor o conceito de distribuição amostral, e sua relação com estimação de parâmetros, veja o arquivo Estima.xls, e suas instruções, no ambiente virtual.

- Sobre a utilização do Microsoft Excel ® para realizar estimação por intervalo,

LEVINE, D. M., STEPHAN, D., KREHBIEL, T. C., BERENSON, M. L. **Estatística**: Teoria e Aplicações - Usando Microsoft Excel em Português. 5ª ed. – Rio de Janeiro: LTC, 200, capítulo 6.

Atividades de aprendizagem

1) O tempo médio de atendimento em uma agência lotérica está sendo analisado por técnicos. Uma amostra de 40 clientes foi sistematicamente monitorada em relação ao tempo que levavam para serem atendidos, obtendo-se as seguintes estatísticas: tempo médio de atendimento de 195 segundos e desvio padrão de 15 segundos.

Considerando que o tempo de utilização segue uma distribuição normal:

a) Faça uma estimação por intervalo para o tempo médio de utilização para toda a população de clientes da agência lotérica, utilizando um nível de confiança de 95%. R. 190,35 a 199,65 segundos

b) Se a legislação estabelecesse que em média o tempo seja de 180 segundos para atendimento, a lotérica está atendendo ao padrão? JUSTIFIQUE.

b) A amostra utilizada seria suficiente para uma precisão de 1 minuto? R. Sim, n = 1.

2) O tempo de montagem de determinados conectores utiliza um processo já há algum tempo, que dura em média 3,5 segundos. Está sendo analisada a possibilidade de troca deste processo para um outro que se afirma possuir um tempo de montagem menor. Para esta análise foram observados os tempos de montagem de conectores por um operário padrão utilizando o novo processo e foram anotados os seguintes valores (em segundos):2,5 2,5 2,6 3,0 3,2 3,5 3,7 3,7 2,1 2,4 2,7 2,8 3,1 3,1 3,6 3,6 2,5 2,9 2,8 3,8

Considerando a situação exposta acima e utilizando um nível de confiança de 95% :

a)Estime o tempo médio de montagem dos conectores utilizando o novo processo. R. 2,767 a 3,243 s.

b) Considerando que o tempo médio aceitável seja de 3 minutos, o novo processo atende ao padrão? JUSTIFIQUE.

c) Calcule o tamanho mínimo da amostra que seria necessária para estimar a média com 95% de confiança e precisão de 0,5 segundos. R. n = 5

3) Buscando melhorar a qualidade do serviço, uma empresa estuda o tempo de atraso na entrega dos pedidos recebidos. Supondo que o tempo de atraso se encontra normalmente distribuído, e conhecendo o tempo de atraso dos últimos 20 pedidos, descritos abaixo (em dias), determine:

5 1 0 3 6 10 2 3 4 1 5 3 1 6 6 9 0 0 1 0

a) Estime o atraso médio na entrega dos pedidos com confiança de 90%. **R.: 2,136 a 4,464 dias.**

b) Um dos clientes da empresa propôs romper o contrato, pois reclama que os atrasos são muito grandes, ele aceitaria em média 2 dias. Com base nosresultados do item a, a empresa deve se preocupar com a possibilidade de rompimento do contrato? JUSTIFIQUE.

c) Para a situação do item ao tamanho da amostra é suficiente, se é necessária uma precisão de 0,5 dias, para o mesmo nível de confiança? **R. Não, n = 109**

4) A satisfação da população de um estado em relação a determinado governo foi pesquisada através de uma amostra com a opinião de 1000 habitantes do estado. Destes, 585 se declararam insatisfeitas com a administração estadual. Admitindo-se um nível de significância de 5%, solucione os itens abaixo.

a)Estime o percentual da população que está insatisfeita com a administração estadual. **R. 55,45% a 61,55%**

b) Com base no resultado do item a você considera a população do estado satisfeita com o governo? JUSTIFIQUE.

c)Qual o tamanho da amostra necessária para a estimação se a empresa responsável pela pesquisa estipulou uma folga máxima de 2,5% ?**R. n = 1493**

5) Uma fábrica está convertendo as máquinas que aluga para uma versão mais moderna. Até agora foram convertidas 40 máquinas. O tempo médio de conversão foi de 24 horas, com desvio padrão de 3 horas.

a) Determine um intervalo de 98% de confiança para o tempo médio de conversão. R. 22,895 h a 25,105 h

b) A direção da fábrica esperava uma média de 20 h para a conversão. A equipe responsável atingiu este padrão? JUSITIFIQUE.

Adaptado de STEVENSON, W.J. Estatística Aplicada à Administração, São Paulo: Harper do Brasil, 2001.

6) Um banco possui 800 terminais de auto-atendimento instalados no estado de SC. Avaliando 48 terminais, 6 apresentaram defeitos.

a) Estime a proporção de terminais com defeitos. **R. 3,144% a 21,86%**

b) Você considera o intervalo de confiança obtido na letra a preciso para estimar a proporção de terminais com defeitos? JUSTIFIQUE.

Adaptado de STEVENSON, W.J. Estatística Aplicada à Administração, São Paulo: Harper do Brasil, 2001.

7) Em uma pesquisa de mercado, acerca da preferência pelo produto X, 300 consumidores foram entrevistados, sendo que 100 declararam consumir o produto.

a) O fabricante quer que você determine um intervalo de 95% para a proporção populacional de pessoas que consomem o produto. **R. 28% a 38,67%**

b) Um dos diretores do fabricante exige que o intervalo de confiança para a proporção populacional tenha 99% de confiança, com um erro máximo de 2,5%. Qual seria o tamanho de amostra necessário para atingir tais critérios?R. 2358

Adaptado de BUSSAB, W.O., MORETTIN, P. A. Estatística Básica, 8ª ed. São Paulo: Saraiva,2013.

8) A Polícia Rodoviária Estadual fez recentemente uma pesquisa secreta sobre as velocidades desenvolvidas na SC 401 das 23h às 2h. No período de observação, 100 carros passaram por um aparelho de radar a uma velocidade média de 112 km/h, com desvio padrão de 22 km/h.

a) Construa um intervalo de 95% de confiança para a média da população. **R. 107,69 km/h a 116,31 km/h**

b) O comando da Polícia divulgaria os resultados do item a apenas se a margem de erro fosse inferior a 10 km/h. Na sua opinião os resultados podem ser divulgados? JUSTIFIQUE.

Adaptado de STEVENSON, W.J. Estatística Aplicada à Administração, São Paulo: Harper do Brasil, 2001.

9) Uma máquina produz peças classificadas como boas ou defeituosas. Retirou-se uma amostra de 1000 peças da produção, verificando-se que 35 eram defeituosas. O controle de qualidade para a linha de produção para rearranjo dos equipamentos envolvidos quando o percentual de defeituosos é superior a 3%.

a) Determinar um intervalo de 95% de confiança para a proporção de peças defeituosas. **R. 2,361% a 4,639%**

b) Com base nos resultados do item a o controle de qualidade deve parar a produção? JUSTIFIQUE.

C) Se há interesse em obter um intervalo de 95% de confiança, com precisão de 1,5%, para a proporção de peças defeituosas, qual deve ser o tamanho mínimo de amostra?**R. 577**

10) Os índices dos alunos dos cursos de Economia e de Administração estão sendo avaliados, no sentido de definirem se há diferença entre os cursos. Para tanto foram analisados os índices de 10 alunos de cada curso, escolhidos aleatoriamente dentre os regularmente matriculados e anotados seus valores, onde se obteve:

Economia	média 7,3	desvio padrão 2,6
Administração	média7,1	desvio padrão 3,1

a) Estime os valores médios dos índices de cada curso com 95% de confiança. **R. Economia: 5,44 a 9,16; Administração: 4,88 a 9,32.**

b) Com base nos resultados do item a, há diferenças significativas entre as médias dos índices dos dois cursos? JUSTIFIQUE.

c) Para o mesmo nível de confiança de a. Será que 10 alunos é uma amostra suficiente, em ambos os cursos, para estimar seus índices médios, com uma precisão igual a 1? Quais seriam os tamanhos de amostra necessários? **R. Economia: insuficiente, n = 35. Administração: insuficiente, n = 50.**

11) O CRA de SC está conduzindo uma pesquisa sobre a opinião dos acadêmicos de administração sobre seus cursos. Suspeita-se que haja diferença entre as proporções de satisfeitos de instituições públicas e privadas: os acadêmicos das públicas seriam mais satisfeitos. Para avaliar esta suposição foi conduzida uma pesquisa por amostragem, entrevistando alunos de duas instituições públicas, SHUFSC e GASE, e de três privadas, PATÁPIO de SÁ, UNIMALI e UNILUS. Os resultados estão na tabela a seguir:

Medidas	UNIVERSIDADES				
	SHUFSC	GASE	PATÁPIO	UNIMALI	UNILUS
n	120	165	185	194	189
p	0,55	0,48	0,32	0,49	0,25
N (população)	890	900	1500	1200	1800

Usando 1% de significância responda os itens a seguir:

a) Estime a proporção populacional de satisfeitos com o seu curso, em cada universidade.

R. SHUFSC 44,11% a 65,89%, GASE 38,94% a 57,06%, PATÁPIO 23,73% a 40,27%, UNIMALI 40,53% a 57,47%, UNILUS 17,32% a 32,68%.

b) De acordo com os resultados do item a, a suposição do CRA é confirmada? JUSTIFIQUE.

c) Para uma margem de erro de 2% qual deveria ser o tamanho de amostra para estimar a proporção de satisfeitos em cada universidade?* **R. SHUFSC 732, GASE 740, PATÁPIO 1060, UNIMALI 931, UNILUS 1141.**

Resumo

O resumo desta Unidade está mostrado na Figura45:

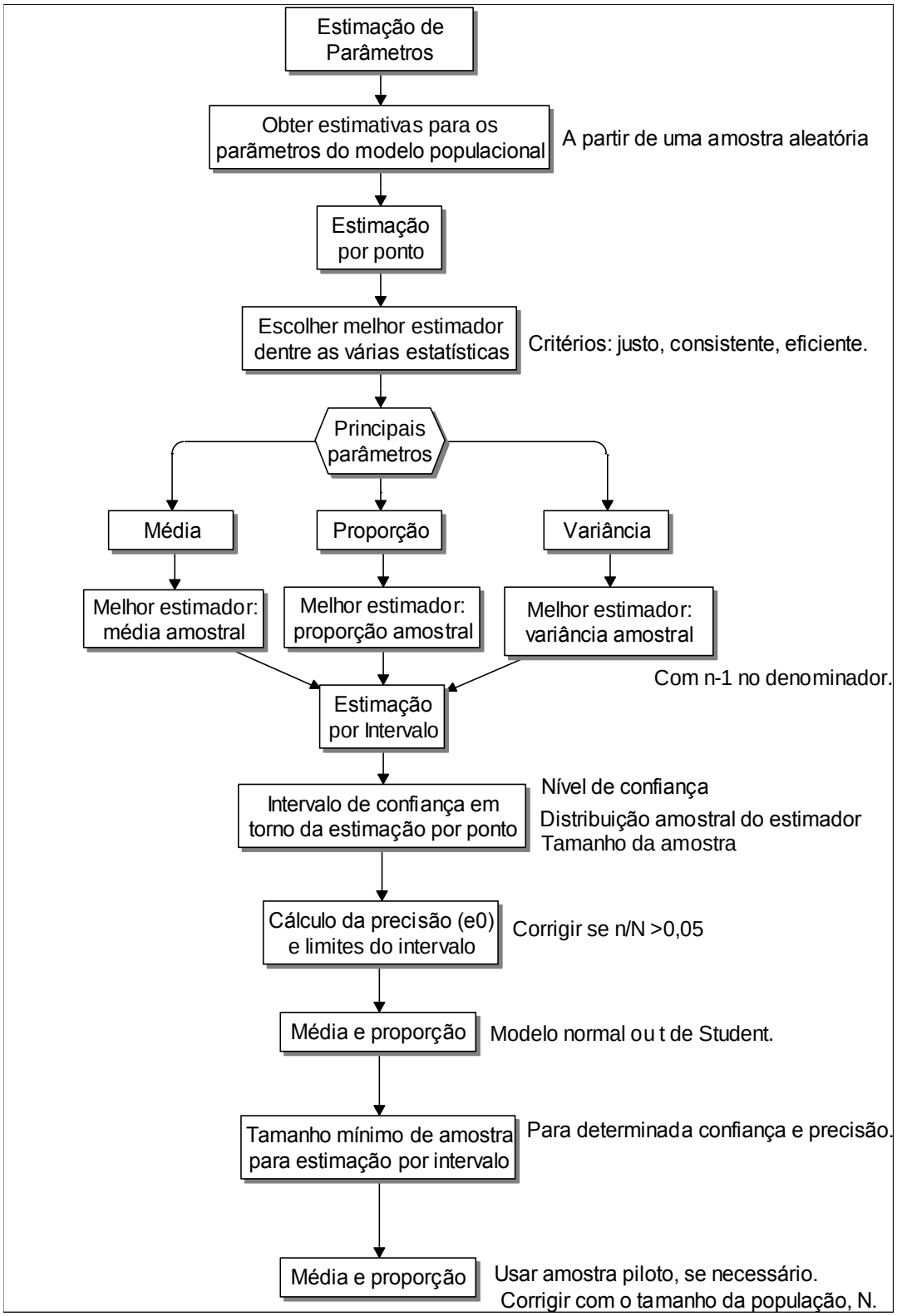


Figura 45 - Resumo da Unidade 5

Fonte: elaborado pelo autor

Vimos nessa Unidade sobre os conceitos de Estimação de Parâmetros. Aprendemos a estimar os parâmetros média de uma variável quantitativa e proporção de um dos valores de uma variável qualitativa. Além de definir o tamanho mínimo de uma amostra aleatória para estimar média e proporção. Veremos mais sobre este assunto na última Unidade deste livro. Estamos próximos do final do nosso material e é de suma importância a continuidade da interação comseus colegas e professor. Não deixe de ver as tabelas indicadas no livro e disponíveis no Ambiente Virtual de Ensino-Aprendizagem e de realizar a atividade de aprendizagem.

Unidade 6

Testes de Hipóteses

Objetivo

Nesta Unidade você vai conhecer e aplicar os conceitos de Testes de Hipóteses, especialmente para média de uma variável quantitativa, proporção de um dos valores de uma variável quantitativa, e associação entre duas variáveis qualitativas. Você aprenderá também qual é a importância de tais conceitos para o dia a dia do administrador.

Caro estudante, você viu anteriormente que uma determinada população pode ser descrita através de um modelo, que apresenta características e parâmetros. Muitas vezes estes parâmetros são desconhecidos e há interesse em estimá-los para obter um melhor conhecimento sobre a população: retira-se então uma amostra aleatória da população e através das técnicas de Estimação de Parâmetros (Unidade 5) procura-se obter uma estimativa de algum parâmetro de interesse, e associamos uma probabilidade de que a estimativa esteja correta. Nesta última e importantíssima Unidade veremos que a Estimação de Parâmetros é uma subdivisão da Inferência Estatística (que consiste em fazer afirmações probabilísticas sobre o modelo da população a partir de uma amostra aleatória desta população), a outra grande subdivisão constitui os **Testes de Hipóteses**. Vamos saber mais!

Contrariamente à Estimação de Parâmetros os Testes de Hipóteses permitem fazer inferências sobre outras características do modelo da população além dos parâmetros, como, por exemplo, a forma do modelo da população. Quando os Testes são feitos sobre os parâmetros da população são chamados de **Testes Paramétricos**GLOSSÁRIO. Testes paramétricos – testes de hipóteses sobre parâmetros do modelo da variável sob análise Fonte: elaborado pelo autor. Fim GLOSSÁRIO, e quando são feitos sobre outras características são chamados de **Testes Não Paramétricos**GLOSSÁRIO. Testes paramétricos – testes de hipóteses sobre outros aspectos do modelo da variável sob análise ou alternativas aos testes paramétricos quando as condições para uso destes não forem satisfeitas. Fonte: elaborado pelo autor. Fim GLOSSÁRIO. TO AFIM DE SABER Na realidade a denominação correta deveria ser "Testes dependentes de distribuição de referência" (porque para fazer inferências sobre os parâmetros devemos supor que o modelo probabilístico populacional é normal, por exemplo, ou que a distribuição amostral do parâmetro pode ser aproximada por uma normal), e "Testes livres de distribuição" (porque os Testes Não Paramétricos não exigem que os dados tenham uma aderência a certo modelo). FIM Não obstante vamos nos restringir aos Testes Paramétricos: de uma média de uma variável quantitativa e de uma proporção de um dos valores de uma variável qualitativa.

Vimos que uma determinada população pode ser descrita através de um modelo probabilístico, que apresenta características e parâmetros. Muitas vezes estes parâmetros são desconhecidos e há interesse em estimá-los para obter um melhor conhecimento sobre a população: retira-se então uma amostra aleatória da população e através das técnicas de **Estimação de Parâmetros** (Unidade 5) procura-se obter uma estimativa de algum parâmetro de interesse, e associamos uma probabilidade de que a estimativa esteja correta. A Estimação de Parâmetros é uma subdivisão da Inferência Estatística (que consiste em fazer afirmações probabilísticas sobre o modelo probabilístico da população a partir de uma amostra aleatória desta população), a outra grande subdivisão constitui os **Testes de Hipóteses**.

Contrariamente à Estimação de Parâmetros os Testes de Hipóteses permitem fazer inferências sobre outras características do modelo probabilístico da população além dos parâmetros (como por exemplo a forma do modelo probabilístico da população). Quando os Testes são feitos sobre os parâmetros da população são chamados de Testes Paramétricos, e quando são feitos sobre outras características são chamados de Testes Não Paramétricos Não obstante vamos nos restringir aos Testes Paramétricos LINK A você estudante interessado em Testes Não Paramétricos recomendo a referência: SIEGEL, S. **Estatística Não Paramétrica (para as Ciências do Comportamento)** ; McGraw-Hill, São Paulo, 1975. É uma boa referência no assunto, em português. LINK.

Imagine-se que um determinado pesquisador está interessado em alguma característica de uma população. Devido a estudos prévios, ou simplesmente por bom senso (melhor ponto de partida para o estudo) ele estabelece que a característica terá um determinado comportamento. Formula então uma hipótese estatística sobre a característica da população, e esta hipótese é aceita como válida até prova estatística em contrário.

Para testar a hipótese é coletada uma amostra aleatória representativa da população, sendo calculadas as estatísticas necessárias para o teste. Naturalmente, devido ao fato de ser utilizada uma amostra aleatória, haverá diferenças entre o que se esperava, sob a condição da hipótese verdadeira, e o que realmente foi obtido na amostra. A questão a ser respondida

é: as diferenças são significativas o bastante para que a hipótese estatística estabelecida seja rejeitada? Esta não é uma pergunta simples de responder: dependerá do que está sob teste (que parâmetro, por exemplo), da confiabilidade desejada para o resultado, entre outros. Basicamente, porém, será necessário comparar as diferenças com uma referência, a distribuição amostral de um parâmetro, por exemplo, que supõe que a hipótese sob teste é verdadeira: a comparação costuma ser feita através de uma estatística de teste que envolve os valores da amostra e os valores sob teste.

A tomada de decisão é feita da seguinte forma:

- se a diferença entre o que foi observado na amostra e o que era esperado (sob a condição da hipótese verdadeira) não for **significativa** a hipótese será **aceita**.
- se a diferença entre o que foi observado na amostra e o que era esperado (sob a condição da hipótese verdadeira) for significativa a hipótese será **rejeitada**.

O valor a partir do qual a diferença será considerada significativa será determinado pelo **Nível de Significância**GLOSSÁRIO Nível de Significância: probabilidade arbitrada pelo pesquisador, valor máximo de erro admissível para rejeitar a hipótese nula sendo ela verdadeira, espera-se que seja um valor baixo, de no máximo 10%. Fonte:Barbetta, Reis e Bornia, 2010; Moore, McCabe, Duckworth e Sclove, 2006. Fim GLOSSÁRIOdo teste. O Nível de Significância geralmente é fixado pelo pesquisador, muitas vezes de forma arbitrária, e também será a probabilidade de erro do Teste de Hipóteses: a probabilidade de cometer um erro no teste, rejeitando uma hipótese válida. Como a decisão do teste é tomada a partir dos dados de uma amostra aleatória da população há sempre a probabilidade de estar cometendo um erro, mas com a utilização de métodos estatísticos é possível calcular o valor desta probabilidade LINK Usando outros métodos (empíricos) não há como ter idéia da chance de erro (pode ser um erro de 0% ou de 5000%...). LINK.

O Nível de Significância é uma probabilidade, portanto é, um número real que varia de 0 a 1 (0 a 100%), e como é a probabilidade de se cometer um erro no teste é interessante que seja o mais próximo possível de zero: valores típicos são 5%, 10%, 1% e até menores dependendo do problema sob análise. Contudo, não é possível usar um Nível de

Significância igual a zero porque devido ao uso de uma amostra aleatória sempre haverá chance de erro, a não ser que a amostra fosse do tamanho da população. O complementar do Nível de Significância é chamado de **Nível de Confiança**, pois ele indica a confiabilidade do resultado obtido, a probabilidade de que a decisão tomada esteja correta

Você deve estar lembrado destes dois conceitos de Estimação de Parâmetros: Nível de Confiança era a probabilidade de que o Intervalo de Confiança contivesse o valor real do parâmetro, e Nível de Significância, complementar daquele, era a probabilidade de que o Intervalo não contivesse o parâmetro, em suma a probabilidade da Estimação estar correta ou não, respectivamente.

6.1 – Tipos de Hipóteses

Para realizar um Teste de Hipóteses é necessário definir (enunciar) duas Hipóteses Estatísticas complementares (que abrangem todos os resultados possíveis): a chamada **Hipótese Nula** (denotada por $\mathbf{H_0}$) e a **Hipótese Alternativa** (denotada por $\mathbf{H_1}$ ou $\mathbf{H_a}$). Enunciar as hipóteses é o primeiro e possivelmente mais importante passo de um Teste de Hipóteses, pois todo o procedimento dependerá dele.

A Hipótese Nula ($\mathbf{H_0}$) é a hipótese estatística aceita como verdadeira até prova estatística em contrário: pode ser o ponto de partida mais adequado para o estudo, ou exatamente o contrário do que o pesquisador quer provar (ou o contrário daquilo que o preocupa).

A Hipótese Alternativa ($\mathbf{H_1}$), que será uma hipótese complementar de $\mathbf{H_0}$, fornecerá uma alternativa à hipótese nula: muitas vezes é justamente o que o pesquisador quer provar (ou o que o preocupa).

Quando as hipóteses são formuladas sobre os parâmetros do modelo probabilístico da população o Teste de Hipóteses é chamado de Paramétrico. Quando as hipóteses são formuladas sobre outras características do modelo o Teste é chamado de Não Paramétrico.

A decisão do teste consiste em aceitar ou rejeitar a Hipótese Nula (**H_0**): vai-se aceitar ou não a hipótese até então considerada verdadeira.

É importante ter a noção exata do que significa aceitar ou rejeitar a Hipótese Nula (**H_0**). A decisão é tomada sobre esta hipótese e não sobre a Hipótese Alternativa porque é a Hipótese Nula é que considerada verdadeira (até prova em contrário). Quando se aceita a Hipótese Nula significa que não há provas suficientes para rejeitá-la. Já quando a decisão é por rejeitar a Hipótese Nula há evidências suficientes de que as diferenças obtidas (entre o que era esperado e o que foi observado na amostra) não ocorreram por acaso. Usando uma analogia com o direito dos EUA, aceitar **H_0** seria comparável a um veredito de não culpado "*not guilty*", ou seja, não há provas suficientes para condenar o réu. Por outro lado rejeitar **H_0** seria comparável a um veredito de culpado "*guilty*", ou seja, as provas reunidas são suficientes para condenar o réu. O Nível de Significância será a probabilidade assumida de **Rejeitar H_0 sendo H_0 verdadeira.**

6.2 – Tipos de Testes Paramétricos

A formulação das hipóteses é o ponto inicial do problema, e deve depender única e exclusivamente das conclusões que se pretende obter com o teste. A formulação da hipótese alternativa determinará o tipo de teste: se **Unilateral**GLOSSÁRIO. Teste unilateral – teste no qual a região de rejeição da hipótese nula está concentrada em apenas um dos lados da distribuição amostral da variável de teste. Fonte: Barbetta, Reis e Bornia, 2010. Fim GLOSSÁRIOou **Bilateral**GLOSSÁRIO. Teste bilateral – teste no qual a região de rejeição da hipótese nula está dividida em duas partes, em cada um dos lados da distribuição amostral da variável de teste. Fonte: Barbetta, Reis e Bornia, 2010. Fim GLOSSÁRIO.

Se a formulação da hipótese alternativa indicar que o parâmetro é maior ou menor do que o valor de teste (valor considerado verdadeiro até prova em contrário) o teste será **Unilateral**: somente há interesse se as diferenças entre os dados da amostra e o valor de teste forem em uma determinada direção. Se a formulação da hipótese alternativa indicar

que o parâmetro é diferente do valor de teste o teste será **Bilateral**: há interesse nas diferenças em qualquer direção. As hipóteses então seriam:

Testes Unilaterais

H_0 **: parâmetro = valor de teste**

H_1 **: parâmetro < valor de teste**

H_0 **: parâmetro = valor de teste**

H_1 **: parâmetro > valor de teste**

Testes Bilaterais

H_0 **: parâmetro = valor de teste**

H_1 **: parâmetro ≠ valor de teste**

A escolha do tipo de teste dependerá das condições do problema sob estudo, sejam as três situações abaixo:

a) Um novo protocolo de atendimento foi implementado no Banco RMG, visando reduzir o tempo que as pessoas passam na fila do caixa. O protocolo será considerado satisfatório se a média do tempo de fila for menor do que 30 minutos. Um teste **Unilateral** seria o adequado.

b) Cerca de 2000 formulários de pedidos de compra estão sendo analisados. Os clientes podem ficar insatisfeitos se houver erros nos formulários. Neste caso admite-se que a proporção máxima de formulários com erros seja de 5%. Ou seja, um valor maior do que 5% causaria problemas. Um teste **Unilateral** seria o adequado.

c) Uma peça automotiva precisa ter 100 mm de diâmetro, exatamente. Neste caso, a dimensão não pode ser maior ou menor do que 100 mm (em outras palavras não pode ser diferente de 100 mm), pois isso indicará que a peça não está de acordo com as especificações. Um teste **Bilateral** seria o adequado.

Após definir as hipóteses é coletada uma amostra aleatória da população para seu teste.

DESTAQUE É importante ressaltar que a montagem das hipóteses deve depender apenas das conclusões que se pretende obter e jamais de uma eventual evidência amostral disponível. DESTAQUE

A decisão de aceitar ou rejeitar $\mathbf{H_0}$ dependerá das **regiões de aceitação e rejeição de H0**, que por sua vez dependem dos seguintes fatores:

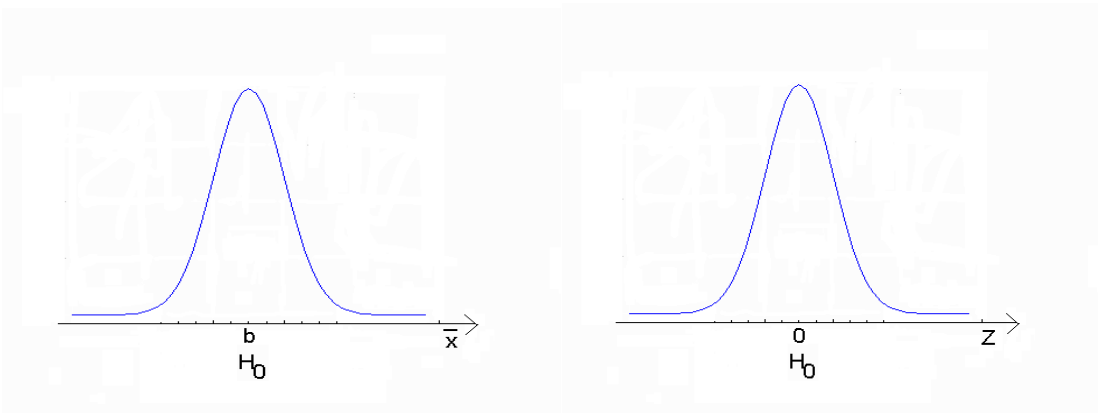
- do parâmetro sob teste (e da estatística ou variável de teste usada para testá-lo).
- do tipo de teste, se Unilateral (à esquerda ou à direita) ou Bilateral.
- do valor de teste (valor do parâmetro considerado verdadeiro até prova em contrário).
- do Nível de Significância ($\boldsymbol{\alpha}$) ou Nível de Confiança ($\mathbf{1 - \alpha}$) adotado.
- de um valor crítico da estatística ou variável de teste a partir do qual a hipótese será rejeitada, e este valor dependerá por sua vez do Nível de Significância, do tipo de teste e da **Distribuição Amostral** do parâmetro.

A **Região de Aceitação de $\mathbf{H_0}$** será a faixa de valores da estatística (ou da variável de teste) associada ao parâmetro em que as diferenças entre o que foi observado na amostra e o que era esperado não são significativas.A **Região de Rejeição de $\mathbf{H_0}$** será a faixa de valores da estatística (ou da variável de teste) associada ao parâmetro em que as diferenças entre o que foi observado na amostra e o que era esperado **são significativas**.

Esta abordagem é chamada de abordagem **clássica** dos testes de hipóteses. Há também a do **valor-p**, muito usada por programas computacionais: "a probabilidade de significância, ou valor p, é definida como a probabilidade da estatística do teste acusar um resultado tão ou mais distante do esperado como o resultado ocorrido na particular amostra observada, supondo H_0 como a hipótese verdadeira" (Barbetta, Reis, Bornia, 2010). O valor-p obtido é comparado com o nível de significância: se for MENOR do que o nível de significância, rejeita-se H_0, se for maior ou igual aceita-se H_0.

Para entender melhor os conceitos acima, observe a situação a seguir:

Há interesse em realizar um teste de hipóteses sobre o comprimento médio de uma das dimensões de uma peça mecânica. O valor nominal da média (aceito como verdadeiro até prova em contrário) é igual a "**b**" (valor genérico), $\mathbf{H_0}$: μ = **b.** Supondo que a distribuição amostral do estimador do parâmetro (distribuição de $\overline{\mathbf{X}}$) seja normal, e será centrada em **b**: é possível fazer a conversão para a distribuição normal padrão (média zero e desvio padrão 1, variável **Z**) (Figuras 46 e 47).



$\mathbf{H_0}$: $\boldsymbol{\mu}$**= b** $\mathbf{H_0}$: μ **= 0**

Figura 46 - Hipótese nula: média populacional = b Figura 47 - Hipótese nula média = 0

Fonte: elaboradas pelo autor.

O valor de **b** (média da dimensão e média de $\overline{\mathbf{X}}$) corresponde a zero, média da variável **Z**. Dependendo da formulação da Hipótese Alternativa haveria diferentes Regiões de Rejeição de $\mathbf{H_0}$.

Se a Hipótese Alternativa fosse $\mathbf{H_1}$: $\boldsymbol{\mu}$**< b** ($\mathbf{H_1}$: $\boldsymbol{\mu}$**< 0**), ou seja, se o teste fosse Unilateral à esquerda a Região de Rejeição de $\mathbf{H_0}$ seria (Figura 48):

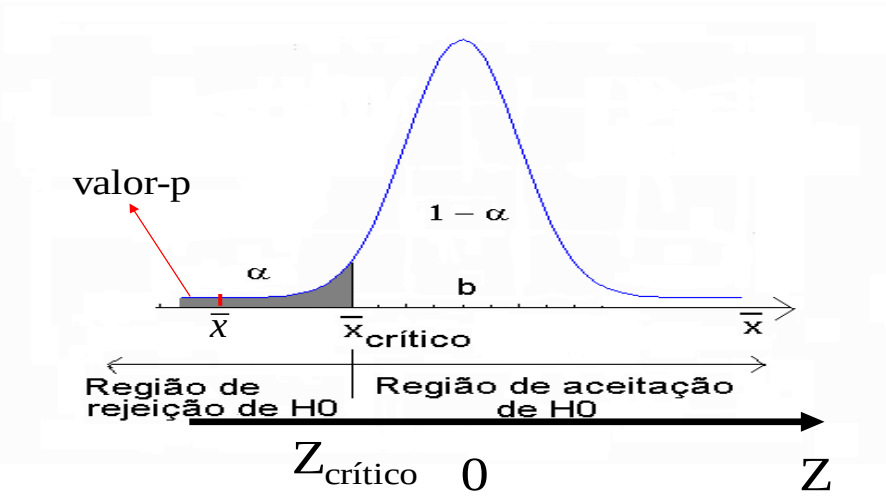


Figura 48 – H_1: μ< b μ< 0

Fonte: elaboradas pelo autor.

Observe que há um valor crítico de $\bar{x}$: se a média amostral estiver abaixo dele a Hipótese Nula será rejeitada, acima será aceita. A determinação do valor é feita com base no Nível de Significância, a área abaixo da curva normal até o valor crítico de $\bar{x}$. Geralmente obtém-se o valor crítico da variável de teste (**Z** neste caso) através de uma tabela ou pacote computacional, que corresponde ao valor crítico de $\bar{x}$, faz-se a transformação de variáveis $Z = \frac{(\bar{x}-\mu_0)}{\sigma}$ e obtém-se o valor crítico de $\bar{x}$. **μ_0** é o valor sob teste (**b** no exemplo) e **σ** é um desvio padrão (cujo valorserá explicitado posteriormente).

Pela abordagem **clássica** a decisão será tomada comparando valor da média amostral $\bar{x}$ com o valor crítico desta mesma média: se for menor do que o valor crítico$\bar{x}_{critico}$, ou seja, está na região de **RejeiçãodeH₀**), então se rejeita a Hipótese Nula. É muito comum também tomar a decisão comparando o valor da variável de teste (**Z** neste caso), obtido com base nos dados da amostra, com o valor crítico $Z_{critic\,o}$ desta mesma variável (obtido de uma tabela ou programa computacional): se for menor do que o valor crítico rejeita-se a Hipótese Nula. Observe que o valor do Nível de Significância **α** é colocado na curva referente à Hipótese Nula, porque é esta que é aceita como válida até prova em contrário. Observe também que a faixa de valores da região de Rejeição pertence à curva da Hipótese Nula, assim o valor **α** é a probabilidade de Rejeitar **H_0** sendo

H_0verdadeira. LINK Probabilidade de tomar uma decisão errada fixada pelo pesquisador.LINK

Neste ponto é importante ressaltar um ponto que costuma passar despercebido. Se a decisão for tomada com base na variável de teste (**Z**, por exemplo) é interessante notar que, como o teste é Unilateral à esquerda o valor $Z_{criti\,co}$ será NEGATIVO, uma vez que a região de Rejeição de **H_0** está à ESQUERDA de 0 (menor do que zero). No teste Unilateral à direita, que veremos a seguir, o valor de $Z_{critico}$ será positivo, pois a região de Rejeição de **H_0** estará à DIREITA de 0 (maior do que zero). Se por exemplo o Nível de Significância fosse de 5% (0,05) o valor de $Z_{critico}$ para o teste Unilateral à esquerda seria -1,645. Se houvesse interesse em obter o valor de $\bar{x}_{critico}$ correspondente bastaria usar a expressão $Z = (\bar{x} - \mu_0)/\sigma$ substituindo **Z** por -1,645. LINK O sinal correto é importante para que o valor de coerente com a posição da região de Rejeição de **H_0**. LINK

Pela abordagem do **valor-p**calcula-se a probabilidade de que $\bar{x}$ assuma valores *menores* que aquele obtido na amostra. Veja na figura 48 uma posição hipotética de $\bar{x}$: o valor-p é a probabilidade de que ela seja menor (porque o teste é unilateral à esquerda) do que aquele valor. Se o valor-p for MENOR do que 0,05 (nível de significância α), rejeita-se H_0.

Se a Hipótese Alternativa fosse **H_1: μ> b** (**H_1: μ> 0**), ou seja, se o teste fosse Unilateral à direita a Região de Rejeição de **H_0** seria (Figura 49)

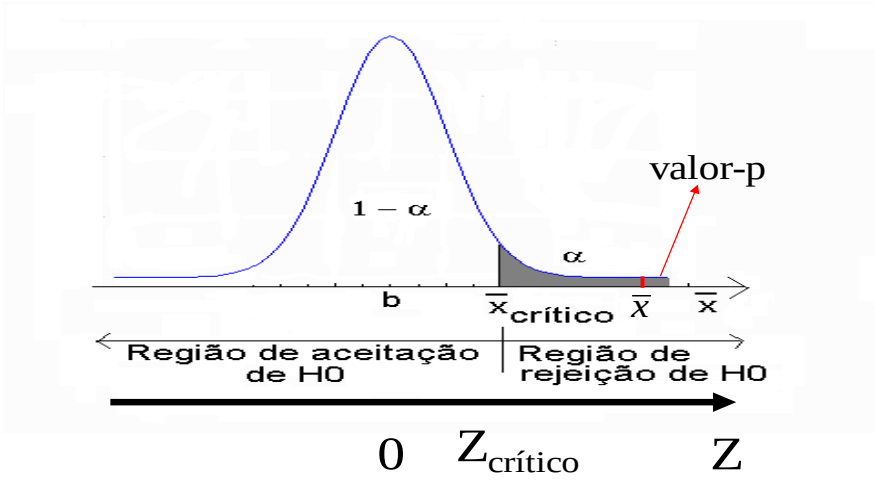


Figura 49 – H_1: μ> b μ > 0

Fonte: elaboradas pelo autor

Neste caso o valor crítico está à direita: se a média amostral $\bar{x}$ ou a variável de teste **Z** tiverem valores superiores aos respectivos valores críticos a Hipótese Nula será rejeitada, pois os valores "caíram" na região de Rejeição de $\mathbf{H_0}$. Como foi notado anteriormente o valor de $Z_{critico}$ será positivo, pois é maior do que zero: usando o mesmo Nível de Significância de 5% o valor de $Z_{critico}$ seria 1,645, igual em módulo ao anterior, uma vez que a distribuição normal padrão é simétrica em relação à sua média que é igual a zero.

Pela abordagem do **valor-p** é preciso calcular a probabilidade de que $\bar{x}$ assuma valores *maiores* que aquele obtido na amostra. Veja na figura 49 uma posição hipotética de $\bar{x}$: o valor-p é a probabilidade de que ela seja maior (porque o teste é unilateral à direita) do que aquele valor. Se o valor-p for MENOR do que 0,05 (nível de significância α), rejeita-se H_0.

Se a Hipótese Alternativa fosse $\mathbf{H_1}$: **μ≠ b** ($\mathbf{H_1}$: **μ≠ 0**), ou seja, o teste fosse Bilaterala Região de Rejeição de $\mathbf{H_0}$ seria (Figura 50)

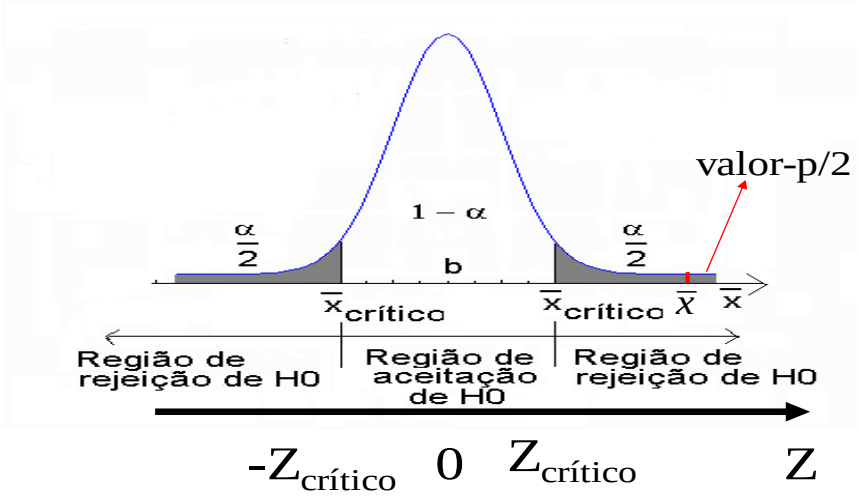


Figura 50 – H_1: μ≠ b μ ≠ 0

Fonte: elaboradas pelo autor.

Neste caso a região de Rejeição se divide em duas iguais (probabilidades iguais à metade do Nível de Significância **α**), semelhante ao que acontece na Estimação por Intervalo. Existirão dois valores críticos, um abaixo do valor de teste e outro acima: se a média amostral $\bar{x}$ ou a variável de teste **Z** tiverem valores acima do valor crítico "superior" ou abaixo do valor crítico "inferior" a Hipótese Nula será rejeitada, pois os valores "caíram" em uma das 2 regiões de Rejeição. Se for usada a variável de teste **Z** os valores críticos serão iguais em módulo, pois estão à mesma distância do valor sob teste (zero).

Pela abordagem do **valor-p** é preciso calcular a probabilidade de que $\bar{x}$ assuma valores *maiores* que aquele obtido na amostra e depois multiplicar esta probabilidade por 2. Veja na figura 50 uma posição hipotética de $\bar{x}$: o valor-p é a probabilidade de que ela seja maior do que aquele valor*multiplicada por 2* (porque o teste é bilateral). Se o valor-p for MENOR do que 0,05 (nível de significância α), rejeita-se H_0.

Recordando as três situações que foram abordadas anteriormente, seria interessante definir completamente as Hipóteses Estatísticas. Nos dois primeiros casos optou-se por um Teste Unilateral e no terceiro por um Teste Bilateral.

a) Um novo protocolo de atendimento foi implementado no Banco RMG, visando reduzir o tempo que as pessoas passam na fila do caixa. O protocolo será considerado satisfatório se a média do tempo de fila for menor do que 30 minutos. Um teste **Unilateral** seria o adequado. Mas Unilateral à Esquerda ou à Direita? Como está grifado na frase anterior haverá problema se a média do tempo fosse menor do que 30, resultando:

Teste **unilateral à esquerda**

$\mathbf{H_0}$: $\boldsymbol{\mu} = 30$ onde $\boldsymbol{\mu_0} = 30$ (valor de teste)

$\mathbf{H_1}$: $\boldsymbol{\mu} < 30$ Teste Unilateral à Esquerda.

b) Cerca de 2000 formulários de pedidos de compra estão sendo analisados. Os clientes podem ficar insatisfeitos se houver erros nos formulários. Neste caso admite-se que a proporção máxima de formulários com erros seja de 5%. Ou seja, um valor maior do que 5% causaria problemas. Um teste **Unilateral** seria o adequado. Neste caso, um teste de proporção, o problema será um valor maior do que 5%, resultando:

Teste unilateral à direita

$\mathbf{H_0}$: $\boldsymbol{\pi}$ = 5% onde $\boldsymbol{\pi_0}$ = 5% (valor de teste)

$\mathbf{H_1}$: $\boldsymbol{\pi}$> 5%

c) Uma peça automotiva precisa ter 100 mm de diâmetro, exatamente. Neste caso, a dimensão não pode ser maior ou menor do que 100 mm (em outras palavras não pode ser diferente de 100 mm) pois isso indicará que a peça não está de acordo com as especificações. Um teste **Bilateral** seria o adequado, resultando:

Teste Bilateral

$\mathbf{H_0}$: $\boldsymbol{\mu}$ = 100 mm onde μ_0 = 100 mm (valor de teste)

$\mathbf{H_1}$: $\boldsymbol{\mu}\neq$ 100 mm

Para a definição apropriada das hipóteses é imprescindível a correta identificação do valor de teste, pois se trata de um dos aspectos mais importantes: o resultado da amostra será comparado ao valor de teste.

Lembrando novamente que a tomada de decisão depende da correta determinação da região de Rejeição (e, por conseguinte, de Aceitação) da Hipótese Nula (ou do cálculo do valor-p), e isso, por sua vez, depende diretamente da formulação das Hipóteses Estatísticas.

6.3 - Testes de Hipóteses sobre a Média de uma Variável em uma População

Neste caso há interesse em testar a hipótese de que o parâmetro média populacional (**μ**) de uma certa variável quantitativa seja maior, menor ou diferente de um certo valor. Para a realização deste teste é necessário que uma das duas condições seja satisfeita:

- sabe-se, ou é razoável supor, que a variável de interesse segue um modelo normal na população: isso significa que a distribuição amostral da média também será normal, permitindo realizar a inferência estatística paramétrica.
- a distribuição da variável na população é desconhecida, mas a amostra retirada desta população é considerada "suficientemente grande" LINK Há muita controvérsia a

respeito do que seria uma amostra "suficientemente grande", mas geralmente uma amostra com pelo menos 30 elementos costuma ser considerada grande o bastante para que a distribuição amostral da média possa ser aproximada por uma normal. LINKo que, de acordo com o Teorema Central do Limite, permite concluir que a distribuição amostral da média é normal.

- supõe-se também que a amostra é representativa da população e foi retirada de forma aleatória.

Tal como na Estimação de Parâmetros por Intervalo existirão diferenças nos testes dependendo do conhecimento ou não da variância populacional da variável.

a) Se a variância populacional ($\boldsymbol{\sigma^2}$) da variável (cuja média populacional queremos testar) for conhecida.

Neste caso a variância amostral da média poderá ser calculada através da expressão: $V(\bar{x}) = \frac{\sigma^2}{n}$, e, por conseguinte, o "desvio padrão" LINK O desvio padrão é a raiz quadrada positiva da variância. LINK será desvio padrão $= \frac{\sigma}{\sqrt{n}}$

A variável de teste será a variável **Z** da distribuição normal padrão, lembrando que:

$$Z = \frac{valor - "média"}{"desviopadrão"}$$

A "**média**" será o valor de teste (suposto verdadeiro até prova em contrário), denotado por $\boldsymbol{\mu_0}$. O **valor** (obtido pela amostra) será a média amostral (que é o melhor estimador da média populacional) denotada por $\bar{x}$, e o "desvio padrão" será o valor obtido anteriormente. Sendo assim a expressão da variável de teste **Z**assumirá um determinado valor que chamaremos de $Z_{calculado}$:

$$Z_{calculado} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Pela abordagem **clássica** compara-se o valor da variável de teste com o valor crítico ($\mathbf{Z_{crítico}}$ que depende do Nível de Significância adotado) de acordo com o tipo de teste (as expressões abaixo também estão no apêndice):

Se $H_1\mu:>\mu_0$ **Rejeitar H_0 se** $\mathbf{Z_{calculado}>Z_{crítico}}$ $\mathbf{(\bar{x}>\bar{x}_{crítico})}$

Se $H_1\mu:<\mu_0$ **Rejeitar H_0 se** $\mathbf{Z_{calculado}<Z_{crítico}}$

LINK Neste caso $Z_{crítico}$ será negativo, já que a região de Rejeição de H_0 está à esquerda de zero. LINK$\mathbf{(\bar{x}<\bar{x}_{crítico})}$

Se H_1 $\mu:\neq\mu_0$ **Rejeitar H_0 se** $\mathbf{|Z_{calculado}|\neq |Z_{crítico}|}$

Pela abordagem do **valor-p** calcula-se a probabilidade associada ao valor da variável de teste:

Se H_1 $\mu:> \mu_0$ **Rejeitar H_0 se** $\mathbf{P(Z >Z_{calculado}) < \alpha}$

Se H_1 $\mu:< \mu_0$ **Rejeitar H_0 se** $\mathbf{P(Z <Z_{calculado}) < \alpha}$

LINK Neste caso calcula-se a probabilidade de Z ser MENOR do que o $Z_{calculado}$ pois o teste é unilateral à esquerda. LINK

Se H_1 $\mu:\neq \mu_0$ **Rejeitar H_0 se** $\mathbf{2\times P(Z > |Z_{calculado}|) < \alpha}$

LINK Neste caso multiplica-se por 2 a probabilidade de Z ser MAIOR do que o valor em módulo de $Z_{calculado}$ pois o teste é bilateral. LINK

b) Se a variância populacional σ^2 da variável for desconhecida.

Naturalmente este é o caso mais encontrado na prática. Como se deve proceder? Dependerá do tamanho da amostra.

b.1 - Grandes amostras (mais de 30 elementos)

Nestes casos procede-se como no item anterior, apenas fazendo com que $\sigma = s$, ou seja, considerando que o desvio padrão da variável na população é igual ao desvio padrão da variável na amostra (suposição razoável para grandes amostras).

b.2 - Pequenas amostras (até 30 elementos)

Nestes casos a aproximação do item b.1 não será viável. Terá que ser feita uma correção na distribuição normal padrão (**Z**) através da distribuição **t de Student**. Esta distribuição já é conhecida (ver Unidades 2 e 5). Trata-se de uma distribuição de

probabilidades que possui média zero (como a distribuição normal padrão, variável **Z**), mas sua variância é igual a **n/(n-2)**, ou seja a variância depende do tamanho da amostra. Quanto maior for o tamanho da amostra mais o quociente acima se aproxima de 1 (a variância da distribuição normal padrão), e mais a distribuição t de Student aproxima-se da distribuição normal padrão. A partir de **n = 30**, já é possível considerar a variância da distribuição **t** de Student aproximadamente igual a 1. LINK E talvez este seja o motivo de se considerar mais de 30 elementos como sendo uma amostra suficientemente grande. LINK

A variável de teste será então $\mathbf{t_{n-1}}$ (**t** com **n - 1** graus de liberdade), e assumirá um valor que chamaremos de $t_{n-1,calculado}$:

$$t_{n-1,calculado} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

onde **s** é o desvio padrão amostral e os outros valores têm o mesmo significado da expressão anterior.

Pela abordagem **clássica** compara-se o valor da variável de teste com o valor crítico ($\mathbf{t_{n-1,crítico}}$ que depende do Nível de Significância adotado) de acordo com o tipo de teste (as expressões abaixo também estão no apêndice):

Se H_1 μ:> μ_0 **Rejeitar H_0 se** $\mathbf{t_{n-1,calculado} > t_{n-1,crítico}}$ **($\bar{x} > \bar{x}_{crítico}$)**

Se H_1 μ:< μ_0 **Rejeitar H_0 se** $\mathbf{t_{n-1,calculado} < t_{n-1,crítico}}$

LINK Neste caso $t_{n-1,crítico}$ será negativo, já que a região de Rejeição de H_0 está à esquerda de zero. LINK**($\bar{x} < \bar{x}_{crítico}$)**

Se H_1 μ:≠ μ_0 **Rejeitar H_0 se** $\mathbf{|t_{n-1,calculado}| \neq |t_{n-1,crítico}|}$

Pela abordagem do **valor-p** calcula-se a probabilidade associada ao valor da variável de teste:

Se H_1 μ:> μ_0 **Rejeitar H_0 se** $\mathbf{P(t_{n-1} > t_{n-1,calculado}) < \alpha}$

Se H_1 μ:< μ_0 **Rejeitar H_0 se** $\mathbf{P(t_{n-1} < t_{n-1,calculado}) < \alpha}$

LINK Neste caso calcula-se a probabilidade de Z ser MENOR do que o $t_{n-1,calculado}$ pois o teste é unilateral à esquerda. LINK

Se H_1 μ:≠ μ_0 **Rejeitar H_0 se** $\mathbf{2 \times P(|t_{n-1}| > |t_{n-1,calculado}|) < \alpha}$

LINK Neste caso multiplica-se por 2 a probabilidade de t_{n-1} ser MAIOR do que o valor em módulo de $t_{n-1,calculado}$ pois o teste é bilateral. LINK

Exemplo 1 - Uma peça automotiva precisa ter 100 mm de diâmetro, exatamente. Foram medidas 15 peças, aleatoriamente escolhidas. Obteve-se média de 100,7 mm e variância de 0,01 mm^2. Supõe-se que a dimensão segue distribuição normal na população. A peça está dentro das especificações? Usar 1% de significância.

Enunciar as hipóteses.Conforme visto na seção 6.2 o teste mais adequado para este caso é um Teste Bilateral:

$\mathbf{H_0}$: $\boldsymbol{\mu}$ = 100 mm onde $\boldsymbol{\mu_0}$ = 100 mm (valor de teste)

$\mathbf{H_1}$:$\boldsymbol{\mu} \neq$ 100 mm

Nível de significância.O problema declara que é necessário usar 1% de significância (se não fosse especificado, outro valor poderia ser usado).

Variável de teste.Uma vez que a variância populacional da variável é DESCONHECIDA (o valor fornecido é a variância amostral), e a amostra retirada apresenta apenas 15 elementos (portanto menos de 30) a variável de teste será $\mathbf{t_{n-1}}$ da distribuição **t** de Student.

Através dos valores da amostra avaliar o valor da variável. Neste ponto é preciso encontrar o valor da variável de teste:

$$t_{n-1,calculado} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

O valor de teste $\boldsymbol{\mu_0}$ é igual a 100 mm, a média amostral $\overline{\mathbf{X}}$ vale 100,7 mm, o tamanho de amostra **n** é igual a 15 e o desvio padrão amostral **s** é a raiz quadrada de 0,01 mm^2. Substituindo na equação acima:

$$t_{n-1,calculado} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = t_{15-1,calculado} = t_{14,calculado} = \frac{100{,}7-100}{\sqrt{0{,}01}/\sqrt{15}} = 27{,}11$$

então $|\mathbf{t_{14,calculado}}| = 27{,}11$

Pela abordagem clássica é preciso definir a região de aceitação de $\mathbf{H_0}$(Figura 51).

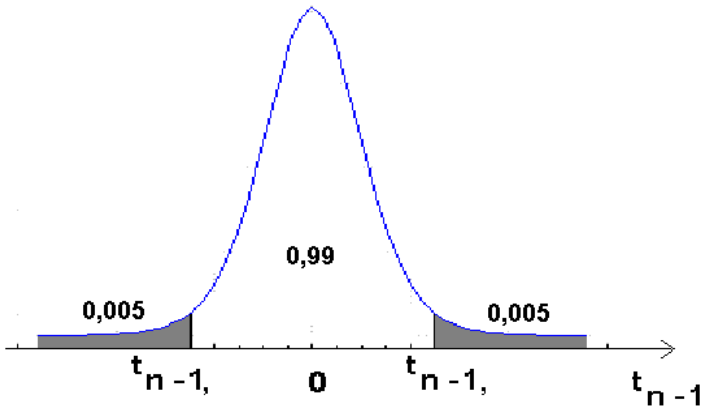


Figura 51 - Regiões de rejeição e aceitação da hipótese nula - Teste bilateral de média
Fonte: elaborada pelo autor

Observe que por ser um teste Bilateral o Nível de Significância α foi dividido em dois, metade para cada região de rejeição de $\mathbf{H_0}$. Para encontrar o valor crítico devemos procurar na tabela da distribuição de Student ou em um pacote computacional, na linha correspondente a **n-1** graus de liberdade, ou seja, em 15 - 1 = 14 graus de liberdade. O valor da probabilidade pode ser visto na figura ao lado: os valores críticos serão $\mathbf{t_{14;0,005}}$e $\mathbf{t_{14;0,995}}$os quais serão iguais em módulo. E o valor de $\mathbf{t_{n-1,crítico}}$será igual a 2,977 (em módulo).

Pela abordagem do valor-p é preciso calcular a probabilidade de que $|t_{n-1}|$ seja maior do que $|t_{n-1,calculado}|$, em outras palavras, $P(t_{14} > 27,11)$. Se procurarmos na tabela t de Student para 14 graus de liberdade disponível no ambiente virtual veremos que o maior valor encontrado é 4,140, correspondente a uma probabilidade 0,0005. Como 27,11 é bem maior do que 4,140 (mais de 6 vezes), a probabilidade associada deve ser praticamente igual a zero, mesmo multiplicando-a por 2 por ser o teste bilateral o valor-p poderá ser considerado praticamente igual a zero (usando o Microsoft Excel ® chegamos a $1,68 \times 10^{-13}$, um número muito pequeno).

Decidir pela aceitação ou rejeição de $\mathbf{H_0}$.Como se trata de um teste bilateral:

Pela abordagem clássica

Rejeitar H_0 se $|t_{n-1,\ calculado}| > |t_{n-1,crítico}|$

Como $|t_{14}| = 27{,}11 > |t_{n-1,crítico}| = |t_{14,0,995}| = 2{,}977$

Rejeitar H_0 a 1% de Significância (há 1% de chance de erro)

Pela abordagem do valor-p

Rejeitar H_0 se $2\times P(|t_{n-1}| > |t_{n-1,calculado}|) < \alpha$

Como valor-p $\cong 0 < \alpha = 0{,}01$

Rejeitar H_0 a 1% de Significância (há 1% de chance de erro)

Interpretar a decisão no contexto do problema. Há provas estatísticas suficientes de que a dimensão da peça não está dentro das especificações. LINK Cuidado com os casos de FRONTEIRA, em que o valor da variável de teste é muito próximo do valor crítico (abordagem clássica) ou o valor-p muito próximo de α (abordagem do valor-p). Nesses casos a rejeição ou aceitação de **H_0** pode ocorrer por acaso. Sempre que apresentar o resultado recomende que uma nova amostra seja retirada para avaliar novamente o problema. Mas neste caso rejeita-se H_0 com folga. LINK

Exemplo 2 - Um novo protocolo de atendimento foi implementado no Banco RMG, visando reduzir o tempo que as pessoas passam na fila do caixa. O protocolo será considerado satisfatório se a média do tempo de fila for **menor** do que 30 minutos. Suponha que o tempo que 35 clientes (selecionados aleatoriamente) passaram na fila foi monitorado, resultando em uma média de 29 minutos e desvio padrão de 5 minutos. O protocolo pode ser considerado satisfatório a 5% de significância?

Enunciar as hipóteses. Conforme visto na seção 6.2 o teste mais adequado para este caso é um Teste Unilateral à Esquerda:

H_0 : **μ** = 30 onde **μ_0** = 30 (valor de teste)

H_1 : **μ**< 30

Nível de significância. O problema declara que é necessário usar 5% .

Variável de teste. Uma vez que a variância populacional da variável é DESCONHECIDA (o valor fornecido é o desvio padrão AMOSTRAL), mas a amostra retirada apresenta 35 elementos (portanto mais de 30) a variável de teste será **Z** da distribuição normal.

Através dos valores da amostra avaliar o valor da variável. Neste ponto é preciso encontrar o valor da variável de teste:

$$Z_{calculado} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

O valor de teste $\boldsymbol{\mu_0}$ é igual a 30, a média amostral $\overline{\mathbf{X}}$ vale 29, o tamanho de amostra **n** é igual a 35 e o desvio padrão amostral **s** é 5. Substituindo na equação acima:

$$Z_{calculado} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{29 - 30}{5/\sqrt{35}} = -1{,}183$$

Pela abordagem clássica é preciso definir a região de aceitação de $\mathbf{H_0}$ (Figura 52).

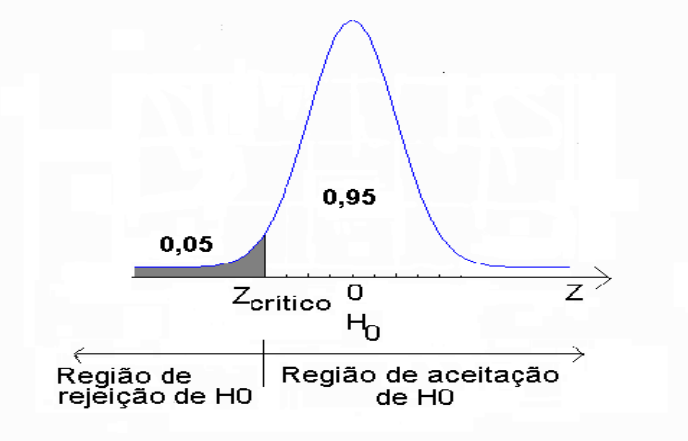


Figura 52 - Regiões de aceitação e de rejeição - Teste unilateral à esquerda

Fonte: elaborada pelo autor

Observe que por ser um teste Unilateral à Esquerda o Nível de Significância **α** está todo concentrado em um dos lados da distribuição, definindo a região de rejeição de $\mathbf{H_0}$. Para encontrar o valor crítico devemos procurar na tabela da distribuição normal ou em um pacote computacional, pela probabilidade acumulada 0,95. Ou procurar a probabilidade

complementar 0,05 e mudar o sinal do valor encontrado, pois o $\mathbf{Z_{crítico}}$ aqui é menor do que zero. O valor crítico será igual a -1,645.

Pela abordagem do valor-p é preciso calcular a probabilidade de que Z seja menor do que $Z_{,calculado}$, em outras palavras, P(Z < -1,183). Lembrando da simetria da distribuição normal padrão (que tem média zero), sabemos que P(Z < -1,183) é igual a P(Z > 1,183). Se procurarmos na tabela da normal padrão disponível no ambiente virtual,veremos que a probabilidade vale 0,1190 (usando o Microsoft Excel ® chegamos a 0,1184).

Decidir pela aceitação ou rejeição de $\mathbf{H_0}$. Como se trata de um teste Unilateral à esquerda:

Pela abordagem clássica

Rejeitar H_0 se$Z_{calculado}$<Zcrítico Como $\mathbf{Z_{calculado}}$ = -1,185 >**Zcrítico =** -1,645

Aceitar $\mathbf{H_0}$ a 5% de Significância (há 5% de chance de erro)

Pela abordagem do valor-p

Rejeitar H_0 se$P(Z < Z_{calculado}) < \alpha$

Como valor-p ≅ 0,1190 > α = 0,05

Aceitar H_0 a 5% de Significância (há 5% de chance de erro)

Interpretar a decisão no contexto do problema. Não há provas estatísticas suficientes para concluir que o protocolo tem um desempenho satisfatório.

6.4 - Testes de Hipóteses sobre a Proporção de uma Variável em uma População

Neste caso há interesse em testar a hipótese de que o parâmetro proporção populacional ($\boldsymbol{\pi}$) de um dos valores de uma certa variável seja maior, menor ou diferente de um certo valor. Para a realização deste teste, tal como será descrito é necessário que duas condições sejam satisfeitas:

- que o produto $\mathbf{n \times \pi_0}$ seja maior ou igual a 5;
- que o produto $\mathbf{n \times (1 - \pi_0)}$ seja maior ou igual a 5.

Onde **n** é o tamanho da amostra e $\boldsymbol{\pi_0}$ é a proporção sob teste (de um dos valores da variável). Se ambas as condições forem satisfeitas a distribuição amostral da proporção que é binomial (uma Bernoulli repetida **n** vezes) pode ser aproximada por uma normal. Obviamente supõe-se que a amostra é representativa da população e foi retirada de forma aleatória, e que a variável pode assumir apenas dois valores, aquele no qual há interesse e o seu complementar.

Se as condições acima forem satisfeitas a distribuição amostral da proporção poderá ser aproximada por uma normal com:

Média =μ_0 **Desvio Padrão =** $\sqrt{\frac{\pi_0 \times (1-\pi_0)}{n}}$

Lembrando-se da expressão da variável **Z**:

$$Z = \frac{\text{valor} - \text{"média"}}{\text{"desvio padrão"}}$$

O **valor** será a proporção amostral (que é o melhor estimador da proporção populacional) do valor da variável denotada por **p**. A "**média"** e o "**desvio padrão"** são os valores definidos acima, então a expressão de **Z**assumirá um valor que chamaremos de $Z_{calculado}$:

$$Z_{calculado} = \frac{p - \pi_0}{\sqrt{\frac{\pi_0 \times (1-\pi_0)}{n}}}$$

Pela abordagem **clássica** compara-se o valor da variável de teste com o valor crítico ($\mathbf{Z_{crítico}}$ que depende do Nível de Significância adotado) de acordo com o tipo de teste (as expressões abaixo também estão no apêndice):

Se H_1 π:>π_0 **Rejeitar H_0 se** $\mathbf{Z_{calculado} > Z_{crítico}}$ **(**$\boldsymbol{p > p_{crítico}}$**)**

Se H_1 π:<π_0 **Rejeitar H_0 se** $\mathbf{Z_{calculado} < Z_{crítico}}$

LINK Neste caso $Z_{crítico}$ será negativo, já que a região de Rejeição de H_0 está à esquerda de zero. LINK**($p<p_{crítico}$)**

Se $H_1 \pi: \neq \pi_0$ Rejeitar H_0 se $|Z_{calculado}| \neq |Z_{crítico}|$

Pela abordagem do **valor-p** calcula-se a probabilidade associada ao valor da variável de teste:

Se $H_1 \pi: > \pi_0$ Rejeitar H_0 se $P(Z > Z_{calculado}) < \alpha$

Se $H_1 \pi: < \pi_0$ Rejeitar H_0 se $P(Z < Z_{calculado}) < \alpha$

LINK Neste caso calcula-se a probabilidade de Z ser MENOR do que o $Z_{calculado}$ pois o teste é unilateral à esquerda. LINK

Se $H_1 \pi: \neq \pi_0$ Rejeitar H_0 se $2 \times P(Z > |Z_{calculado}|) < \alpha$

LINK Neste caso multiplica-se por 2 a probabilidade de Z ser MAIOR do que o valor em módulo de $Z_{calculado}$ pois o teste é bilateral. LINK

Exemplo 3 - Cerca de 2000 formulários de pedidos de compra estão sendo analisados. Os clientes podem ficar insatisfeitos se houver erros nos formulários. Neste caso admite-se que a proporção máxima de formulários com erros seja de 5%. Suponha que dentre os 2000 formulários 7% apresentavam erros. A proporção máxima foi ultrapassada a 1% de significância?

Enunciar as hipóteses. Conforme visto na seção 6.2, o teste mais adequado para este caso é um Teste Unilateral à Direita:

H_0 :π = 5% onde **π_0** = 5% (valor de teste)

H_1 :π> 5%

Nível de significância. O problema declara que é necessário usar 1% de significância (se não fosse especificado, outro valor poderia ser usado).

Variável de teste. Como se trata de um teste de proporção é necessário verificar o valor dos produtos:

n x π_0 = 2000 x 0,05 = 100 e **n x (1 - π_0) = 2000 x 0,95 = 1900**. Como ambos são maiores do que 5 é possível aproximar pela normal, e a variável de teste será **Z**.

Através dos valores da amostra avaliar o valor da variável. Neste ponto é preciso encontrar o valor da variável de teste:

$$Z_{calculado} = \frac{p - \pi_0}{\sqrt{\frac{\pi_0 \times (1-\pi_0)}{n}}}$$

O valor de teste **π_0** é igual a 0,05 (5%), a proporção amostral **p** vale 0,07 (7%), e o tamanho de amostra **n** é igual a 2000. Substituindo na equação acima:

$$Z_{calculado} = \frac{p - \pi_0}{\sqrt{\frac{\pi_0 \times (1-\pi_0)}{n}}} = \frac{0,07 - 0,05}{\sqrt{\frac{0,05 \times (0,95)}{2000}}} = 4,104$$

Pela abordagem clássica é preciso definir a região de aceitação de **H_0** (Figura 53).

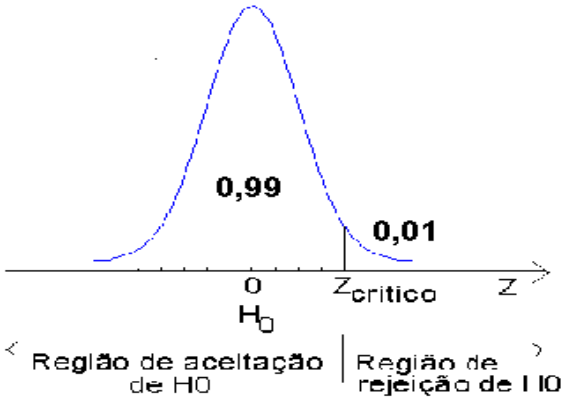


Figura 53 - Regiões de aceitação e de rejeição - Teste unilateral à direita
Fonte: elaborada pelo autor

Observe que por ser um teste Unilateral à Direita o Nível de Significância **α** está todo concentrado em um dos lados da distribuição, definindo a região de rejeição de **H_0**. Para encontrar o valor crítico devemos procurar na tabela da distribuição normal, pela probabilidade acumulada 0,01 (o **Zcrítico** aqui é maior do que zero). O valor crítico será aproximadamente igual a 2,33.

Pela abordagem do valor-p é preciso calcular a probabilidade de que Z seja maior do que $Z_{,calculado}$, em outras palavras, P(Z > 4,104). Se procurarmos na tabela da normal padrão disponível no ambiente virtual, veremos o valor de Z mais próximo é 4,0, e que a probabilidade de Z ser maior do que 4,0 vale 0,0000317 (usando o Microsoft Excel ® chegamos a 0,000020303).

Decidir pela aceitação ou rejeição de $\mathbf{H_0}$. Como se trata de um teste Unilateral à direita:

Pela abordagem clássica

Rejeitar H_0 se$Z_{calculado}$>Zcrítico Como $\mathbf{Z_{calculado}}$ = 4,104 >**Zcrítico =** 2,33

Rejeitar $\mathbf{H_0}$ a 1% de Significância (há 1% de chance de erro)

Pela abordagem do valor-p

Rejeitar H_0 se$P(Z > Z_{calculado}) < \alpha$

Como valor-p ≅0,0000317< α = 0,01

Rejeitar H_0 a 1% de Significância (há 1% de chance de erro)

Interpretar a decisão no contexto do problema. Há provas estatísticas suficientes de que a proporção está acima do máximo admitido LINK Este não é um caso de fronteira LINK. Provavelmente os vendedores/compradores precisarão passar por novo treinamento.

Agora vamos ver um tipo de teste estatístico muito utilizado pelos administradores, para avaliar o relacionamento entre duas variáveis qualitativas: o teste de associação (independência de quiquadrado).

6.5 – Teste de associação de quiquadrado

O teste do quiquadradoGLOSSÁRIO Teste de associação (independência) de quiquadrado – teste que permite avaliar se duas variáveis qualitativas, cujas frequências estão dispostas em uma tabela de contingências, apresentam associação significativa ou

não. Fonte: Barbetta, Reis e Bornia, 2010. Fim GLOSSÁRIO, também chamado de teste de independência de quiquadrado, está vinculado à análise de duas variáveis qualitativas. Vamos ver alguns conceitos antes de apresentar o teste de associação de quiquadrado.

6.5.1 – Variáveis qualitativas e tabelas de contingência

É comum haver interesse em saber se duas variáveis quaisquer estão relacionadas, e o quanto estão relacionadas, seja na vida prática, seja em trabalhos de pesquisa, por exemplo:

- se a satisfação com um produto está relacionada à faixa etária do consumidor;
- se a função exercida por uma pessoa em uma organização está associada a seu gênero.

Na Unidade 2 de Estatística Aplicada à Administração I apresentamos técnicas para tentar responder as perguntas do parágrafo anterior.

Variáveis Qualitativas são as variáveis cujas realizações são atributos, categorias (Unidades 1 e 2 de Estatística Aplicada à Administração I). Como exemplo de variáveis qualitativas tem-se: sexo de uma pessoa (duas categorias, masculino e feminino), grau de instrução (analfabeto, primeiro grau incompleto, etc.), opinião sobre um assunto (favorável, desfavorável, indiferente).

Em estudos sobre variáveis qualitativas é extremamente comum registrar as frequências de ocorrência de cada valor que as variáveis podem assumir, e quando há duas variáveis envolvidas é comum registrar-se a frequência de ocorrência dos cruzamentos entre valores: por exemplo, quantas pessoas do sexo masculino são favoráveis a uma certa proposta de lei, quantas são desfavoráveis, quantas pessoas do sexo feminino são favoráveis. E, para facilitar a análise dos resultados estes resultados costumam ser dispostos em uma Tabela de Contingências. A Tabela de Contingências relaciona os possíveis valores de uma variável qualitativa com os possíveis valores da outra, registrando quantas ocorrências foram verificadas de cada cruzamento.

Exemplo 4 – O Quadro 6 mostra uma tabela de contingências relacionando as funções exercidas e o sexo de 474 funcionários de uma organização.

	Função			
Sexo	Escritório	Serviços gerais	Gerência	Total
Masculino	157	27	74	258
Feminino	206	0	10	216
Total	363	27	84	474

Quadro 6 - Tabela de contingências de Função por Sexo

Fonte: elaborado pelo autor

Podemos apresentar os percentuais calculados em relação aos totais das colunas no Quadro 7:

	Função			
Sexo	Escritório	Serviços gerais	Gerência	Total
Masculino	43,25%	100%	88,10%	54%
Feminino	56,75%	0%	11,90%	46%
Total	100%	100%	100%	100%

Quadro 7 - Tabela de contingências de Função por Sexo: percentuais por colunas

Fonte: elaborado pelo autor

Seria interessante saber se as duas variáveis são estatisticamente dependentes, e o quão forte é esta associação. Repare que os percentuais de homens e mulheres em cada função são diferentes dos percentuais marginais (de homens e mulheres no total de funcionários), sendo que em duas funções (Serviços gerais e Gerência) as diferenças são bem grandes.

O teste de associação de quiquadrado é uma das ferramentas estatísticas mais utilizadas quando se deseja estudar o relacionamento entre duas variáveis qualitativas. Permite verificar se duas variáveis qualitativas são independentes, se as proporções de ocorrência dos valores das variáveis observadas estão de acordo com o que era esperado, etc. Neste texto haverá interesse em usar o teste para avaliar se duas variáveis qualitativas são independentes.

Como todo teste de hipóteses o teste de associação de quiquadrado consiste em comparar os valores observados em uma amostra com os valores de uma referência (referência esta que supõe que a hipótese nula seja válida).

As frequências observadas da variável são representadas em uma tabela de contingências, e a Hipótese Nula (**H_0**) do teste será que as duas variáveis não diferem em relação às frequências com que ocorre uma característica particular, ou seja, as variáveis são independentes, que será testada contra a Hipótese Alternativa (**H_1**) de que as variáveis não são independentes.

O teste pode ser realizado porque o grau de dependência pode ser quantificado descritivamente através de uma estatística, que se chama justamente quiquadrado (χ^2) na população, mas na amostra é chamada de q^2 cuja expressão é:

$$q^2 = \sum_{i=1}^{L} \sum_{j=1}^{C} \left[\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \right]$$

Sendo $E_{ij} = \frac{\text{total da linha i} \times \text{total da coluna j}}{\text{total geral}}$

Onde:

- E_{ij} é a frequência esperada, sob a condição de independência entre as variáveis, em uma célula qualquer da tabela de contingências. Todas as frequências esperadas precisam ser maiores ou iguais a 5 para que o resultado do teste seja válido. LINK Se isso não ocorrer

recomenda-se agrupar as categorias (de uma ou outra variável, ou de ambas) até obter todas as frequências pelo menos iguais a 5.LINK

- O_{ij} é a frequência observada em uma célula qualquer da tabela de contingências;
- L é o número total de linhas da tabela de contingências (número de valores que uma das variáveis pode assumir);
- C é o número total de colunas da tabela (número de valores que a outra variável pode assumir).

Então, para cada célula da tabela de contingências calcula-se a diferença entre a frequência observada e a esperada. Para evitar que as diferenças positivas anulem as negativas elas são elevadas ao quadrado. E para evitar que uma diferença grande em termos absolutos, mas pequena em termos relativos, "inflacione" a estatística, ou que uma diferença pequena em termos absolutos, mas grande em termos relativos, tenha sua influência reduzida, divide-se o quadrado da diferença pela frequência esperada. Somam-se os valores de todas as células da tabela e obtêm-se o valor da estatística total, que chamaremos de $\mathbf{q^2_{calculado}}$: quanto maior $\mathbf{q^2_{calculado}}$, mais o Observado se afasta do Esperado, portanto maior a dependência.

Sob a hipótese de independência entre as duas variáveis (H_0) a estatística $\mathbf{q^2}$ seguirá o modelo quiquadrado com (L-1) × (C – 1) graus de liberdade, que estudamos na Unidade 2, prometendo usá-la aqui na Unidade 6. O número de graus de liberdade assume este valor porque para calcular as frequências esperadas não é necessário calcular os valores de todas as células, as últimas podem ser calculadas por diferença já que os totais são fixos. Por exemplo, para duas variáveis que somente podem assumir 2 valores cada, o número de graus de liberdade seria igual a 1 [(2-1)×(2-1)]: bastaria calcular a frequência esperada de uma célula e obter as outras por diferença em relação ao total.

Da mesma forma que nos testes de hipóteses anteriores podemos usar a abordagem clássica ou a do valor-p para tomar a decisão de rejeitar ou aceitar H_0 com base na evidência amostral, mas teremos um processo mais simples: o teste de associação do quiquadrado para avaliar se duas variáveis são independentes será **sempre**unilateral.

Pela abordagem **clássica**, definido o nível de significância α é possível encontrar o $q^2_{crítico}$para determinado grau de liberdade. Por exemplo, para o caso em que há 3 graus de liberdade, e o Nível de Significância fosse de 5% (a região de Rejeição de H_0 ficará**sempre** À DIREITA), o valor crítico seria (lembre-se da Unidade 2) (Figura 54):

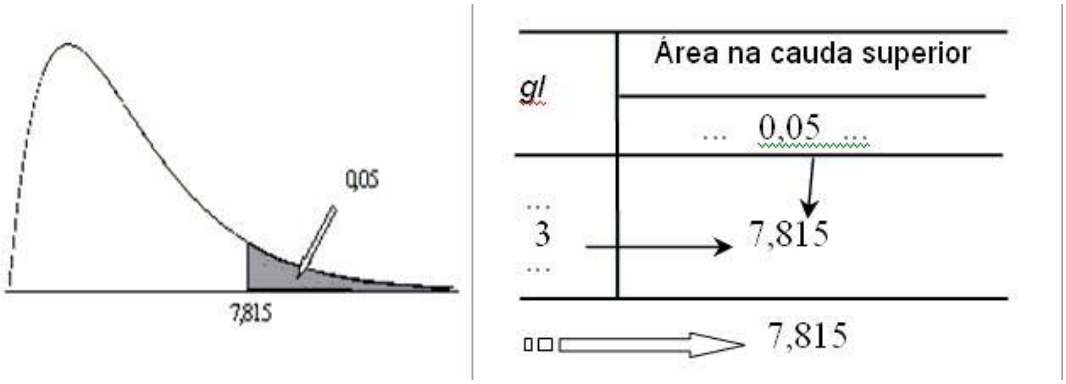


Figura 54- Uso da tabela da distribuição quiquadrado. Ilustração com gl = 3 e área na cauda superior de 5%

Fonte: adaptado pelo autor de Barbetta, Reis e Bornia (2010)

A Hipótese Nula será rejeitada se $q^2_{calculado} > q^2_{crítico}$, para **(L - 1) × (C - 1)** graus de liberdade.

Pela abordagem do valor-p é preciso encontrar a probabilidade do valor associado à variável de teste com **(L - 1) × (C - 1)** graus de liberdade:

rejeita-se se H_0 se $P(q^2 > q^2_{calculado}) < \alpha$ (nível se significância).

Exemplo 4 - Para os dados mostrados no Quadro 6, supondo que os resultados são provenientes de uma amostra aleatória, aplique um teste estatístico apropriado para avaliar se as variáveis são independentes a 1% de significância.

Terá que ser usado o teste de associação de quiquadrado, pois os dados estão em uma tabela de contingências, relacionando duas variáveis qualitativas.

Enunciar as Hipóteses:

H_0: as variáveis sexo e função são independentes

H_1: as variáveis sexo e função não são independentes

Nível de significância: determinado pelo problema, igual a 1% (0,01).

Retirar as amostras aleatórias e montar a tabela de contingências (isso já foi feito) (repetida a tabela de contingências no Quadro 6):

	Função			
Sexo	Escritório	Serviços gerais	Gerência	Total
Masculino	157	27	74	258
Feminino	206	0	10	216
Total	363	27	84	474

Quadro 8 - Tabela de contingências de Função por Sexo

Fonte: elaborado pelo autor

Na tabela acima se encontram os totais marginais e o total geral:

L1 = total Masculino = 258 L2 = total Feminino = 216 C1 = total Escritório = 157

C2 = total S.Gerais = 27 C3 = total gerência = 84 N = total geral =474

Repare que somando os totais das linhas o resultado é o total geral, e que somando os totais das colunas o resultado é o total geral também.

Calcular as frequências esperadas. Calculando as frequências esperadas de acordo com a fórmula vista anteriormente:

Masculino - Escritório $E = (258 \times 363)/ 474 = 197,58$

Masculino - Serviços Gerais $E = (258 \times 27)/ 474 = 14,70$

Masculino - Gerência $E = (258 \times 84)/ 474 = 45,72$

Feminino - Escritório $E = (216 \times 363)/ 474 = 165,42$

Feminino - Serviços Gerais $E = (216 \times 27)/ 474 = 12,30$

Feminino - Gerência $E = (216 \times 84)/ 474 = 38,28$

Calculando a estatística $\mathbf{q^2}$para cada célula. Agora podemos calcular as diferenças entre as frequências e as demais operações, que serão mostradas nos Quadros 9, 10 e 11.

O - E	Função		
Sexo	Escritório	Serviços gerais	Gerência
Masculino	157 - 197,58	27 - 14,70	74 - 45,72
Feminino	206 - 165,42	0 - 12,30	10 - 38,28

Quadro 9– Diferença entre frequências observadas e esperadas de Função por Sexo

Fonte: elaborado pelo autor.

$(O-E)^2$	Função		
Sexo	Escritório	Serviços gerais	Gerência
Masculino	1646,921	151,383	799,672
Feminino	1646,921	151,383	799,672

Quadro 10 – Diferença entre frequências observadas e esperadas de Função por Sexo elevadas ao quadrado

Fonte: elaborado pelo autor

Finalmente:

$(O-E)^2/E$	Função		
Sexo	Escritório	Serviços gerais	Gerência
Masculino	8,336	10,301	17,490
Feminino	9,956	12,304	20,891

Quadro 11 – Estatísticas q^2 de Função por Sexo

Fonte: elaborado pelo autor

Agora podemos somar os valores:

$$q^2_{calculado} = 8,336 + 10,301 + 17,490 + 9,956 + 12,304 + 20,891 = 79,227$$

Os graus de liberdade: (**número de linhas -1**) × (**número de colunas - 1**) = (**2 -1**)×(**3-1**)= **2**

Então $\mathbf{q^2}$calculado = 79,227 para 2 graus de liberdade.

Abordagem clássica. O $\mathbf{q^2}$ crítico será: procurando na Tabela 3 do ambiente, ou em um pacote computacional, para 2 graus de liberdade e 99% de confiança (1% de significância): $\mathbf{q^2_{crítico}}$ = 9,21, ver Figura 55.

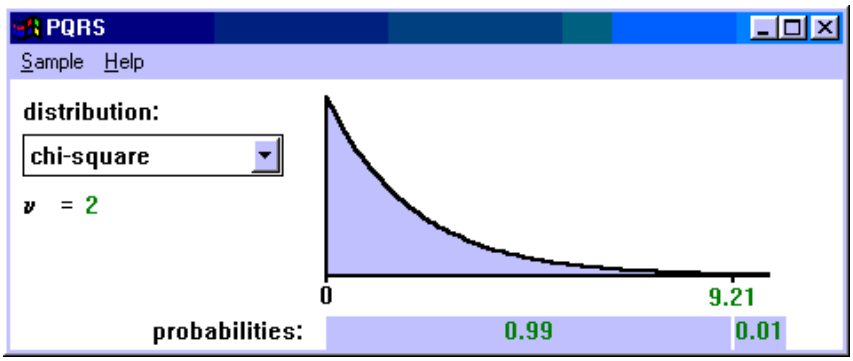


Figura 55 - Valor crítico de q^2 para 2 graus de liberdade e 1% de significância

Fonte: adaptada pelo autor de PQRS®

Como $\mathbf{q^2_{calculado}}$ é maior do que $\mathbf{q^2_{crítico}}$para 2 graus de liberdade rejeita-se$\mathbf{H_0}$ a 1% de significância.

Abordagem do valor-p. O $q^2_{calculado}$ vale 79,227. Procurando na tabela da distribuição quiquadrado que está no ambiente virtual para 2 graus o maior valor encontrado é 13,82, que corresponde a uma probabilidade igual a 0,001. Como 79,227 é muito maior do que 13,82 a probabilidade de q^2 ser maior do que 79,227 deve ser bem menor do que 0,001 (através do Microsoft Excel ® a probabilidade é praticamente igual a zero). Então o valor-p deve ser praticamente igual a zero. Como o valor – $p \cong 0 < \alpha = 0,01$, rejeita-se H_0 a 1% de significância.

HÁ evidência estatística suficiente que indicam que as variáveis função e sexo não são independentes. Isso confirma nossas suspeitas iniciais, devido às grandes diferenças nas frequências da tabela.

No tópico Tô afim de saber... você terá indicações de vários outros tipos de hipóteses que não foram mencionados nesta Unidade. As referências lá citadas serão extremamente valiosas se você tiver que:

- aplicar testes para avaliar se há diferenças entre médias de duas ou mais populações.
- aplicar testes para avaliar se há diferenças entre proporções de duas populações.
- aplicar testes não paramétricos, por exemplo testes de aderência dos dados a um determinado modelo probabilístico.

Com este tópico terminamos nossa jornada... Agora é com vocês. Boa sorte!

Tô afim de saber:

- Sobre tipos de erro, poder, em testes de hipóteses –

BARBETTA, P.A., REIS, M.M., BORNIA, A.C. **Estatística para Cursos de Engenharia e Informática**. 3ª ed. - São Paulo: Atlas, 2010, capítulo 8;

STEVENSON, Willian J. **Estatística Aplicada à Administração**. São Paulo: Ed. Harbra, 2001, capítulo10.

- Sobre testes de uma variância -

BARBETTA, P.A., REIS, M.M., BORNIA, A.C. **Estatística para Cursos de Engenharia e Informática**. 3ª ed. - São Paulo: Atlas, 2010, capítulo 8;

TRIOLA, M. **Introdução à Estatística**, Rio de Janeiro: LTC, 1999, capítulo 7.

- Sobre testes de comparação de duas médias -

BARBETTA, P.A., REIS, M.M., BORNIA, A.C. **Estatística para Cursos de Engenharia e Informática**. 3ª ed. - São Paulo: Atlas, 2010, capítulo 9.

- Sobre testes de comparação de duas proporções,

MOORE, D.S., McCABE, G.P., DUCKWORTH, W.M., SCLOVE, S. L., **A prática da estatística empresarial**: como usar dados para tomar decisões. Rio de Janeiro: LTC, 2006, capítulo 8.

- Sobre Análise de Variância, comparação de várias médias,

BARBETTA, P.A., REIS, M.M., BORNIA, A.C. **Estatística para Cursos de Engenharia e Informática**. 3ª ed. - São Paulo: Atlas, 2010, capítulo 9.

STEVENSON, Willian J. **Estatística Aplicada à Administração**. São Paulo: Ed. Harbra, 2001, capítulo 11.

MOORE, D.S., McCABE, G.P., DUCKWORTH, W.M., SCLOVE, S. L., **A prática da estatística empresaria**l: como usar dados para tomar decisões. Rio de Janeiro: LTC, 2006, capítulos 14 e 15.

TRIOLA, M. **Introdução à Estatística**, Rio de Janeiro: LTC, 1999, capítulo 11.

- Sobre testes não paramétricos,

BARBETTA, P.A., REIS, M.M., BORNIA, A.C. **Estatística para Cursos de Engenharia e Informática**. 3ª ed. - São Paulo: Atlas, 2010, capítulo 10,

STEVENSON, Willian J. **Estatística Aplicada à Administração**. São Paulo: Ed. Harbra, 2001, capítulo 13;

SIEGEL, S. **Estatística Não Paramétrica** (para as Ciências do Comportamento). São Paulo: McGraw-Hill, 1975.

- Sobre a utilização do Microsoft Excel ® para realizar testes de hipóteses,

LEVINE, D. M., STEPHAN, D., KREHBIEL, T. C., BERENSON, M. L. **Estatística**: Teoria e Aplicações - Usando Microsoft Excel em Português. 5ª ed. – Rio de Janeiro: LTC, 200, capítulo 6.

Atividades de aprendizagem

1) O tempo médio de atendimento em uma agência lotérica está sendo analisado por técnicos. Uma amostra de 40 clientes foi sistematicamente monitorada em relação ao tempo que levavam para serem atendidos, obtendo-se as seguintes estatísticas: tempo médio de atendimento de 195 segundos e desvio padrão de 15 segundos.Considerando que o tempo de utilização segue uma distribuição normal. O dono da agência garante que o tempo médio de atendimento é de 3 minutos (se for maior ele se compromete a contratar mais um atendente). Aplicando o teste estatístico apropriado, com base nos dados da amostra, a afirmação do dono é verdadeira, ou ele deve contratar um novo atendente? Use um nível de significância de 1%? **R.: Sim, Z = 6,32**

2) O tempo de montagem de determinados conectores utiliza um processo já há algum tempo, que dura em média 3,5 segundos. Está sendo analisada a possibilidade de troca deste processo para um outro que se afirma possuir um tempo de montagem menor. Para esta análise foram observados os tempos de montagem de conectores por um operário padrão utilizando o novo processo e foram anotados os seguintes valores (em segundos):2,5 2,5 2,6 3,0 3,2 3,5 3,7 3,7 2,1 2,4 2,7 2,8 3,1 3,1 3,6 3,6 2,5 2,9 2,8 3,8

Aplicando o teste estatístico apropriado, considerando a situação exposta acima, com um nível de confiança de 95%, a empresa deve mudar para o novo processo ou manter o atual? **R. Deve mudar. t = -4.36**

3) Buscando melhorar a qualidade do serviço, uma empresa estuda o tempo de atraso na entrega dos pedidos recebidos. Supondo que o tempo de atraso se encontra normalmente distribuído, e conhecendo o tempo de atraso dos últimos 20 pedidos, descritos abaixo (em dias), determine:

5 1 0 3 6 10 2 3 4 1 5 3 1 6 6 9 0 0 1 0

Um cliente enfurecido quer testar estatisticamente a hipótese (declarada pela empresa) de que o atraso médio será de no máximo 1 dia. Ele argumenta que deve ser maior, e quer uma confiança de 99% para um teste estatístico apropriado. Com base nos dados da amostra, o cliente tem razão na sua reclamação? **R. Sim, t = 3,42.**

4) A satisfação da população em relação a determinado governo foi pesquisada através de uma amostra com a opinião de 1000 habitantes do estado. Destes, 585 se declararam insatisfeitas com a administração estadual. Admitindo-se um nível de significância de 5%, solucione os itens abaixo.

A atual administração decidiu que se o percentual de descontentamento fosse superior a 50% deveria ser redirecionado o plano governamental. Aplicando o teste estatístico apropriado, utilizando a informação amostral, o que você conclui? **R. Redirecionar o plano. Z = 5,375.**

5. Uma firma está convertendo as máquinas que aluga para uma versão mais moderna. Até agora foram convertidas 40 máquinas. O tempo médio de conversão foi de 24 horas, com desvio padrão de 3 horas.O fabricante das novas máquinas afirma que a conversão em média dura no máximo 25 horas. Aplicando o teste estatístico apropriado, com base nas conversões feitas até o momento, a 1% de significância, a afirmação do fabricante é verdadeira? **R. Sim. Z = -2,1082**
Adaptado de STEVENSON, W.J. Estatística Aplicada à Administração, São Paulo: Harper do Brasil, 2001.

6) Em uma pesquisa de mercado, acerca da preferência pelo produto X, 300 consumidores foram entrevistados, sendo que 100 declararam consumir o produto.
O fabricante do produto X afirma que é a marca líder no mercado, que mais de 40% dos consumidores a preferem. Aplique o teste estatístico apropriado e com base nos dados verifique se a afirmação é válida. Usar 1% de significância.**R. Não. Z = -2,35**
Adaptado de BUSSAB, W.O., MORETTIN, P. A. Estatística Básica, 8ª ed. São Paulo: Saraiva,2013.

7) Uma máquina produz peças classificadas como boas ou defeituosas. Retirou-se uma amostra de 1000 peças da produção, verificando-se que 35 eram defeituosas. O controle de qualidade pára a linha de produção para rearranjo dos equipamentos envolvidos quando o percentual de defeituosos é superior a 3%. Aplique o teste estatístico apropriado e baseado nos dados amostrais verifique se a linha de produção deve ser parada? **R. Não. Z = 0,9268**.

8) Em 600 lançamentos de um dado obteve-se a face 6 em 123 lançamentos.
a) Aplique o teste estatístico apropriado e verifique se a 5% de significância há razão para desconfiar que o dado é viciado quanto a face 6 ? **R. Sim. Z= 2,519**
b) E a 1% de significância? **R. Não.**

9) Uma amostra aleatória entre homens e mulheres foi analisada com o objetivo de pesquisar-se o comportamento de "fumar cigarros". Verificou-se que de 27 homens, 15 eram fumantes, e que de 33 mulheres, 12 tinham o hábito de fumar.

Teste a hipótese de que o sexo influencia o comportamento de fumar, a um nível de 5% de significância. **R. Não associada. $q^2 = 2,210$**

10) Dentre os alunos de uma sala alguns não frequentavam as aulas, apenas comparecendo às provas. Na tabela abaixo estão apresentados seus resultados:

	Aprovados	Reprovados	Total
"frequentadores"	22	8	30
"ausentes"	10	18	28
Total	32	26	58

Utilizando 1- α = 99 %. Aplique o teste apropriado para verificar se a presença nas aulas está associada aos resultados finais dos alunos?

R. Associada. $\chi 2 = 8,287$)

11) Queremos saber se há associação entre três meios de comunicação, em termos de lembrança do consumidor da propaganda veiculada. O resultado de um estudo sobre propaganda mostrou:

Capacidade de lembrança	Meio de comunicação			
	Revista	TV	Rádio	Total
Lembram da propaganda	25	93	7	125
Não lembram da propaganda	73	10	108	191
Total	98	103	115	316

a) Usando 1% de significância e o teste estatístico apropriado é possível concluir que há associação entre a capacidade de lembrança e o meio de comunicação usado? (**R.: Sim. $\chi 2 = 172,8536$)**

b) Observando os resultados acima, qual meio de comunicação você recomendaria para veicular uma propaganda, para maximizar a capacidade de lembrança do público alvo? Por quê?

Adaptado de LEVINE, D.M., BERENSON, M.L., STEPHAN, D., **Estatística**: Teoria e Aplicações usando Microsoft ® Excel em Português. Rio de Janeiro: LTC, 2000.

Resumo

O resumo desta Unidade está demonstrado na Figura 59:

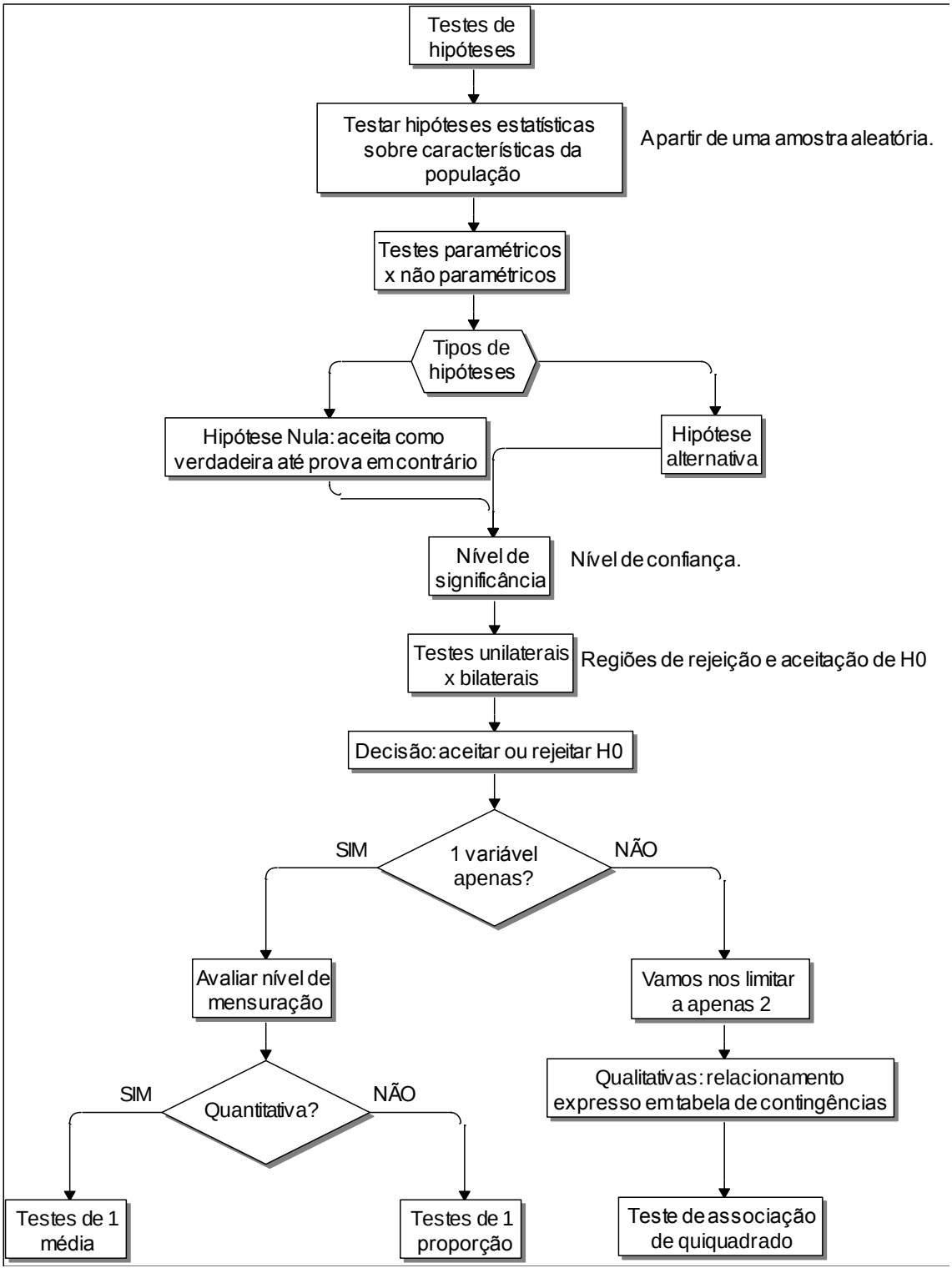


Figura 59 - Resumo da Unidade 6

Fonte: elaborado pelo autor

Chegamos ao final da disciplina de Estatística Aplicada a Administração II.Estudamos nessa última Unidade os testes de hipóteses, tipos de hipóteses e suas variáveis. A Unidade foi explorada com Figuras, exemplos e Quadros para melhor representar o conteúdo oferecido. Além do material produzido pelo professor você tem em mãos uma riquíssima fonte de referencias para saber mais sobre o assunto. Explore os conhecimentos propostos. Não tenha esta Unidade como fim, mas sim o começo de uma nova trajetória em sua vida acadêmica. Bons estudos e boa sorte!

Referências

ANDERSON, D.R., SWEENEY, D.J., WILLIAMS, T.A., **Estatística Aplicada à Administração e Economia**. 2ª ed. – São Paulo: Thomson Learning, 2007

BARBETTA, P.A., REIS, M.M., BORNIA, A.C. **Estatística para Cursos de Engenharia e Informática**. 2ª ed. - São Paulo: Atlas, 2008.

BARBETTA,P. A. **Estatística Aplicada às Ciências Sociais**. 7ª. ed. – Florianópolis: Ed. da UFSC, 2007.

COSTA NETO, P.L. da O. Estatística. 2ª ed, São Paulo: Edgard Blücher, 2002.

LOPES, P. A. **Probabilidades e Estatística**. Rio de Janeiro: Reichmann e Affonso Editores, 1999.

MARCONI, Marina de Andrade, LAKATOS, Eva Maria. Técnicas de Pesquisa - 5a ed. São Paulo: Atlas, 2003.

MONTGOMERY, D. C. Introdução ao Controle Estatístico da Qualidade. 4.ed. Rio de Janeiro: LTC, 2004.

MOORE, D.S., McCABE, G.P., DUCKWORTH, W.M., SCLOVE, S. L., **A prática da estatística empresarial**: como usar dados para tomar decisões. Rio de Janeiro: LTC, 2006.

STEVENSON, Willian J. **Estatística Aplicada à Administração**. São Paulo: Ed. Harbra, 2001.

TRIOLA, M. Introdução à Estatística, Rio de Janeiro: LTC, 1999.

VIRGILITTO, S. B. **Estatística Aplicada** – Técnicas básicas e avançadas para todas as áreas do conhecimento. São Paulo: Alfa-Omega, 2003.

Minicurrículo e foto do autor

Minicurrículo:

MARCELO MENEZES REIS é formado em Engenharia Elétrica pela Universidade Federal de Santa Catarina - UFSC, bacharel em Administração de Empresas pela Universidade para o Desenvolvimento de Santa Catarina – UDESC, registro no CRA-SC 4049, Especialização em Seis Sigma (Beyond Six Sigma Certification Program) na University of South Florida- USF (EUA), mestre em Engenharia Elétrica pela Universidade Federal de Santa Catarina, e doutor em Engenharia de Produção pela Universidade Federal de Santa Catarina. Professor Adjunto, lotado no Departamento de Informática e Estatística da Universidade Federal de Santa Catarina, desde 1995. Tem ministrado disciplinas de estatística em vários cursos de graduação e pós-graduação da Universidade, incluindo os de Administração.

Foto: