UNIDADE 6

Testes de Hipóteses

Objetivos específicos de aprendizagem

Ao finalizar esta Unidade, você deverá ser capaz de:

- Escolher o teste de hipótese adequado;
- Formular um teste de hipótese;
- Chegar a uma conclusão sobre uma população a partir dos resultados amostrais; e
- Interpretar os passos e os resultados de um teste de hipótese.

INTRODUÇÃO

Caro estudante,

Vamos conhecer agora os principais testes de hipóteses utilizados na inferência estatística.

Você, como gestor, muitas vezes terá de tomar decisões baseadas na análise de dados a partir de um teste de hipótese. Portanto, esteja atento ao conteúdo que iremos apresentar a você nesta última Unidade, pois ao longo da leitura você certamente perceberá a importância desse assunto quando tratamos de Estatística Aplicada à Administração. Bom estudo!

Na teoria de decisão estatística, os testes de hipóteses assumem uma importância fundamental, já que nos permitem dizer, por exemplo, se duas populações são, de fato, iguais ou diferentes utilizando, para isso, amostras dessas populações. Sendo assim, a tomada de decisão de um gestor público deve estar baseada na análise de dados a partir de um teste de hipótese.

Você pode definir as hipóteses a serem testadas, retirar as amostras das populações a serem estudadas, calcular as estatísticas delas e, por fim, determinar o grau de aceitação de hipóteses baseadas na teoria de decisão, ou seja, se uma determinada hipótese será validada ou não.

Para você decidir se uma hipótese é verdadeira ou falsa, ou seja, se ela deve ser aceita ou rejeitada, considerando-se uma determinada amostra, precisamos seguir uma série de passos que são:

1. Definir a hipótese de igualdade (H_0) e a hipótese alternativa (H_1) para tentar rejeitar H_0 (possíveis erros associados à tomada de decisão).
2. Definir o nível de significância (α).
3. Definir a distribuição amostral a ser utilizada.
4. Definir os limites da região de rejeição e de aceitação.
5. Calcular a estatística da distribuição escolhida a partir dos valores amostrais obtidos e tomar a decisão.

Você deve tomar a decisão baseado na seguinte regra: se o valor da estatística da distribuição calculado estiver na região de rejeição, rejeite a hipótese nula. Caso contrário, se o valor da estatística calculado caiu na região de aceitação, a decisão será que a hipótese nula não poderá ser rejeitada ao nível de significância determinado.

Agora, você terá o detalhamento dos passos na formulação de um teste de hipótese. Esteja bem atento!

Estrutura dos Testes de Hipóteses

Diversos conceitos serão apresentados ao longo do detalhamento dos passos a serem seguidos na formulação de um teste de hipótese.

1) Formular as hipóteses (H_0 e H_1).

Primeiramente, vamos estabelecer as **hipóteses nula e alternativa**. Para exemplificar, você deve considerar um teste de hipótese para uma média. Então, a hipótese de igualdade é chamada de **hipótese de nulidade ou H_0**.

Suponha que você queira testar a hipótese de que o tempo médio de atendimento na retirada de uma guia, em uma prefeitura considerada modelo de atendimento, é igual a 50 segundos. Essa hipótese será simbolizada da seguinte maneira:

H_0: $\mu = 50$ (hipótese de nulidade)

Essa hipótese, na maioria dos casos, será de igualdade.

Se você rejeitar essa hipótese, irá aceitar, nesse caso, outra hipótese, que chamamos de **hipótese alternativa**. Esse tipo de hipótese é simbolizado por **H_1 ou H_a**.

A partir do nosso exemplo, as hipóteses alternativas mais comuns são as apresentadas a seguir:

- H_1: $\mu > 50$ (teste unilateral ou unicaudal à direita).

  O tempo médio de retirada da guia é superior a 50 segundos (>).

- H_1: $\mu < 50$ (teste unilateral ou unicaudal à esquerda).

  O tempo médio de retirada da guia é inferior a 50 segundos (<).

- H_1: m ¹ 50 (teste bilateral ou bicaudal).

  O tempo médio de retirada da guia pode ser superior ou inferior a 50 segundos.

Surge uma dúvida. Qual hipótese alternativa você utilizará? A resposta é bem simples.

A hipótese alternativa será definida por você em razão do tipo de decisão que deseja tomar.

Veja o seguinte exemplo: você inspeciona uma amostra, relativa a uma grande remessa que chega a uma prefeitura, e constata que 8% dela está defeituosa. O fornecedor garante que não haverá mais de 6% de peças defeituosas em cada remessa. O que devemos responder, com auxílio dos testes de significância, é se a afirmação do fornecedor é verdadeira.

As hipóteses que você vai formular são:

H_0: p = 0,06;

H_1: p > 0,06.

A hipótese alternativa somente pode ser maior, pois o fornecedor garante que não haverá mais de 6%.

É importante ressaltar que o sinal de igual para a hipótese H_0 corresponde a um sinal de menor ou igual (nesse exemplo), pois o teste é unilateral à direita (p > 0,06). Portanto, sempre que o teste for unilateral, deve ser feita essa consideração.

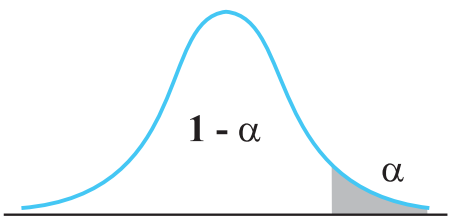
2) Definir o nível de significância.

O nível de significância de um teste é dado pela probabilidade de se cometer erro do tipo I (**ocorre quando você rejeita a hipótese H_0 e essa hipótese é verdadeira**). Com o valor dessa

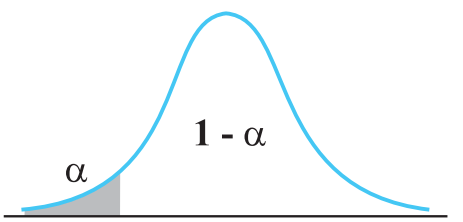
probabilidade fixada, você pode determinar o chamado **valor crítico**, que separa a chamada **região de rejeição** da hipótese H_0, da região de aceitação da hipótese H_0.

No desenho, a seguir, as áreas escuras correspondem à significância do teste, ou seja, a probabilidade de se cometer o chamado erro tipo I (rejeitar H_0 quando ela é verdadeira). Essa probabilidade é chamada de α e geralmente os valores mais utilizados são 0,01 e 0,05. O complementar do nível de significância é chamado de nível de confiança (área clara dos gráficos) e é dado por $1 - \alpha$.

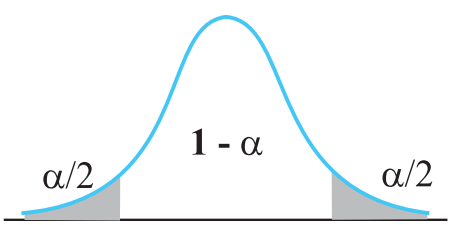
Unilateral à direita:
Ho: $\mu = 50$
H1:: $\mu > 50$



Unilateral à esquerda:
Ho: : $\mu = 50$
H1: : $\mu < 50$



Bilateral:
Ho: : $\mu = 50$
H1: : $\mu \neq 50$



3) Definir a distribuição amostral a ser utilizada.

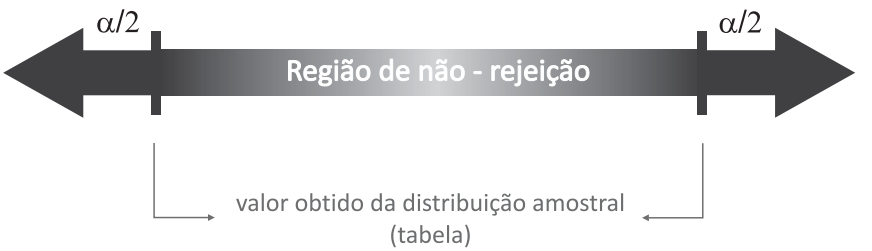
Você definirá a estatística a ser utilizada no teste em razão da distribuição amostral a qual os dados seguem. Se você fizer um teste de hipótese para uma média ou diferença entre médias, utilize a distribuição de Z ou t de *Student*.

Note que o conhecimento das distribuições amostrais vistas na Unidade 5 é muito importante. Caso ainda tenha alguma dúvida, volte e relembre os conceitos das distribuições de t, qui-quadrado e F, e como utilizar as tabelas.

Outro exemplo é se você quiser comparar a variância de duas populações; para tal, deverá trabalhar então com a distribuição F, ou seja, da razão de duas variâncias.

4) Definir os limites da região de rejeição.

Os limites entre as regiões de rejeição e de aceitação da hipótese H_0 você definirá em razão do tipo de hipótese H_1, do valor de α (nível de significância) e da distribuição amostral utilizada. Considerando um teste bilateral, você terá a região de aceitação (não rejeição) com uma probabilidade de $1-\alpha$, e uma região de rejeição com probabilidade α ($\alpha/2 + \alpha/2$).



Por meio da amostra obtida, você deve calcular a estimativa que servirá para aceitar ou para rejeitar a hipótese nula. Neste momento, você deve estar se perguntando: **como irei calcular a estimativa, ou seja, o valor da estatística a partir dos dados amostrais?** A resposta será dada no próximo item.

5) Tomar a decisão

Para tomar a decisão, você deve calcular a estimativa do teste estatístico que será utilizada para rejeitar ou não a hipótese H_0. A estrutura desse cálculo para a média de forma generalista é dada por:

$$Estatistica\ da\ distribuição = \frac{(estimativa\ -\ parâmetro)}{erro\ padrão\ da\ estimativa}$$

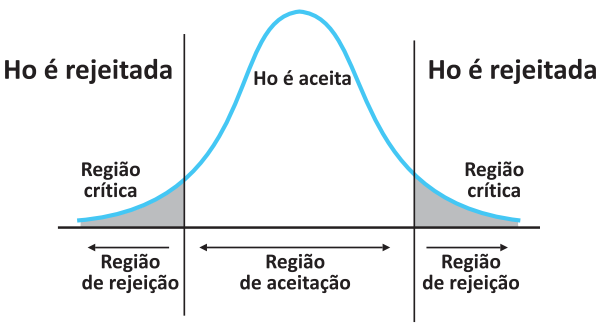
Podemos exemplificar pela distribuição de Z, que será:

Estatística do teste

$$Z_{cal} = \frac{(\bar{x} - \mu)}{(\sigma/\sqrt{n})}$$

Variabilidade das médias

Se o valor da estatística estiver na região crítica (de rejeição), você vai rejeitar H_0, caso contrário, aceite H_0. O esquema a seguir mostra bem a situação de decisão.



Teste de Hipótese para uma Média

Quando você retira uma amostra de uma população e calcula a média dessa amostra, é possível verificar se a afirmação sobre a média populacional é verdadeira. Para tanto, basta verificar se a estatística do teste estará na região de aceitação ou de rejeição da hipótese H_0.

Aqui, você tem duas situações distintas:

Primeira situação: se o desvio padrão da população é conhecido ou a amostra é considerada grande (n >30): a distribuição amostral a ser utilizada será da Normal ou Z e a estatística teste que você utilizará será:

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Onde:

$\bar{x}$: média amostral;

μ : média populacional;

σ : desvio padrão populacional; e

n: tamanho da amostra.

Imagine a seguinte situação: um gestor público sabe que, para montar um determinado negócio em um bairro de Curitiba, é necessário que nele circulem, no mínimo, 1.500 pessoas por dia. Para o tipo de bairro em questão, é possível supor o desvio padrão como sendo igual a 200 pessoas. Uma amostra aleatória formada por 12 observações revelou que passariam pelo local

escolhido 1.400 pessoas por dia, em média. O negócio pode ser montado ou não? Assuma $\alpha = 5\%$ e suponha população normalmente distribuída.

Resolução:

Sempre, em um exercício de tomada de decisão, precisamos da formulação de um teste de hipótese, seguindo os passos apresentados:

1. Formular as hipóteses.
2. Definir o nível de significância.
3. Definir a distribuição amostral a ser utilizada.
4. Definir os limites da região de rejeição (gráfico).
5. Tomar a decisão.

Vamos primeiramente retirar os dados do problema:

$$n = 12; \bar{x} = 1400 \text{ e } \sigma = 200$$

Denominamos esse desvio como populacional, pois, baseados nas características do bairro (conhecimento prévio), podemos supor o valor do desvio.

Vamos estabelecer as hipóteses com base no exercício:

H_o: $\mu = 1500$

H_1: $\mu < 1500$

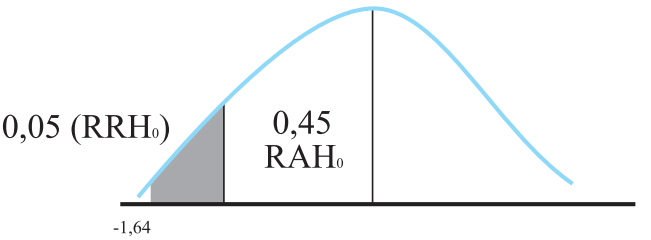
Caso tenhamos uma média igual a 1.500 pessoas, podemos montar o negócio. Mas se aceitarmos a hipótese H_1, não devemos indicar a montagem do negócio.

$\alpha = 0{,}05$

A estatística escolhida é Z. Substituindo os valores da amostra e o da hipótese H_0 na estatística de Z, teremos:

$$Z_c = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{1400 - 1500}{\frac{200}{\sqrt{12}}} = \frac{-100}{57{,}73} = -1{,}73$$

Veja que, mesmo com $n \leq 30$, o desvio padrão populacional foi informado. Quando temos essa situação, devemos sempre usar Z.



O valor $Z_t = -1,64$, que divide a RRH_0 e RAH_0, foi encontrado na tabela Z procurando em seu **interior** o valor 0,4495. Como Z calculado é menor que Z tabelado, ou seja, –1,73 pertence a RRH_0, podemos afirmar com 95% de certeza que transitam menos de 1.500 pessoas por dia no bairro e, assim, verificamos que não é viável montar o negócio no bairro.

Valor mais próximo de 0,45, pois este não existe na tabela.

Agora, antes de prosseguir, você deve resolver a Atividade 1, ao final desta Unidade. Caso tenha alguma dúvida, retorne a situação anterior, aquela que resolvemos juntos.

Segunda situação: se você não conhecer o desvio padrão populacional e a amostra for pequena ($n > 30$), a distribuição amostral a ser utilizada será a t de *Student* e a estatística teste será:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Onde:

$\bar{x}$: média amostral;

μ : média populacional;

s : desvio padrão amostral; e

n: tamanho da amostra.

Uma observação importante: quando trabalhamos com amostras grandes, ou seja, $n \geq 30$, a distribuição de Z e t de *Student* apresentam comportamentos e valores da estatística próximos.

Neste momento, releia os passos anteriores para que não fique nenhuma dúvida em relação à estrutura de um teste de hipótese, pois iremos trabalhar juntos em situações nas quais iremos aplicar os diferentes testes de hipóteses para uma média.

Então, após a releitura do conteúdo apresentado, vamos analisar as situações.

Veja, abaixo, a primeira situação em que utilizaremos o teste de hipótese para uma média usando a estatística de Z (amostras grandes ou variância populacional conhecida). Para resolver essa situação, utilizaremos o teste de hipótese para uma média usando a estatística de t de *Student* (amostra pequena e variância populacional desconhecida).

A Construtora Estrada Forte Ltda. alega ser capaz de produzir concreto com, no máximo, 15 kg de impurezas para cada tonelada fabricada. Mas, segundo a legislação municipal, caso essa quantidade seja maior do que 15 kg, a obra deve ser embargada pela prefeitura. Dezenove amostras de uma tonelada cada uma revelaram possuir impurezas com média amostral igual a 23 kg e desvio padrão igual a 9 kg. Assumindo $\alpha = 5\%$ e população normalmente distribuída, a obra deve ser embargada ou não?

Resolução:

Retirando os dados do problema:

$n = 19$; $\bar{x} = 23$; $s = 9$; $\alpha = 0{,}05$. Vamos estabelecer as hipóteses baseando-nos na afirmação do exercício:

$H_0: \mu = 15$
$H_1: \mu > 15$ Caso a hipótese H_0 seja aceita, a obra não será embargada, pois ela está de acordo com a lei. Caso contrário, a prefeitura embarga a obra.

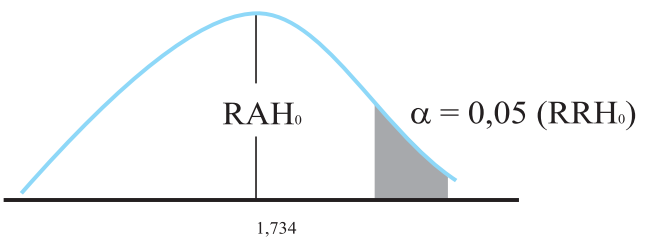
$\alpha = 0,05$

A estatística escolhida é t de *Student*.

Substituindo os valores do problema na expressão, teremos:

Veja que o n foi menor ou igual a 30 ($n \leq 30$), foi informado o desvio padrão amostral e não foi apresentado o desvio padrão populacional. Nessas condições, devemos sempre usar distribuição t de *Student*.

$$t_c = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{23 - 15}{\frac{9}{\sqrt{19}}} = \frac{8}{2,06} = 3,87$$



O valor $t_t = 1,734$ que divide a RRH_0 e RAH_0 foi encontrado na tabela t procurando grau de liberdade 18 (gl = n –1 = 19 = 1) e $\alpha = 0,05$. Como t calculado é maior do que t tabelado, ou seja, 1,734 pertence a RRH_0, podemos afirmar com 95% de certeza que a alegação da construtora não é verdadeira. Eles não são capazes de produzir concreto com, no máximo, 15 kg de impurezas para cada tonelada fabricada. Então, concluímos que a obra deve ser embargada pela prefeitura.

Teste de Hipótese para a Razão de duas Variâncias

Esse teste de hipótese é utilizado para saber se duas variâncias populacionais são estatisticamente iguais ou se uma é maior do que a outra. Utilizando a distribuição F, poderemos formular o teste de hipótese da razão entre duas variâncias e chegar à conclusão baseados apenas nas estimativas calculadas a partir das amostras.

As hipóteses H_0 e H_1 serão:

$H_0 : \sigma_1^2 = \sigma_2^2$ (variâncias das duas populações são iguais)

$H_1 : \sigma_1^2 > \sigma_2^2$ (variância da população 1 é maior do que a da população 2).

> Como estamos utilizando um teste unilateral à direita, por questões didáticas, então, no cálculo da estatística de F, teremos a maior variância dividida pela menor variância.

A maior variância amostral encontrada será chamada de s_1^2 (proveniente de uma amostra de tamanho n_1) e a menor variância amostral será chamada s_2^2 (proveniente de amostra de tamanho n_2).

Vamos considerar duas amostras provenientes de duas populações. Desejamos saber se as variâncias das populações são estatisticamente iguais ou se uma é maior do que a outra. Considere uma significância de 2,5%. Os resultados amostrais são apresentados a seguir:

$s_1^2 = 0{,}5184 \quad \text{com} \quad n_1 = 14$

$s_2^2 = 0{,}2025 \quad \text{com} \quad n_2 = 21$

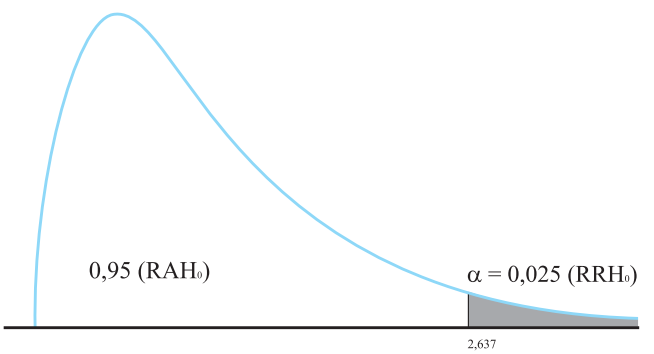
Então, a variável de teste do teste F será:

$$F = \frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2}$$

Como em H_0 estamos considerando que as variâncias populacionais são iguais, então, na expressão acima, as duas variâncias populacionais irão se cancelar. No nosso exemplo, teremos:

$$F = \frac{s_1^2}{s_2^2} = \frac{0{,}5184}{0{,}2025} = 2{,}56$$

O valor tabelado (crítico) da distribuição de F será obtido na tabela da distribuição com uma significância de 2,5%. Considerando como graus de liberdade iguais a 13 ($n_1 - 1$) para o numerador (v_1) e 20 ($n_2 - 1$) para o denominador (v_2), chegaremos ao seguinte resultado: valor tabelado igual a 2,637.



O valor calculado da estatística (2,56) foi menor do que o tabelado (2,637), então, o valor calculado caiu na região de aceitação de H_0. Assim, aceitamos H_0 e consideramos que a variância da população 1 é estatisticamente igual à variância da população 2, ou seja, não ocorre uma diferença entre elas.

Esse teste servirá de base para a escolha do próximo teste (diferença entre médias para amostras independentes), ou seja, escolher o tipo de teste a ser utilizado.

Teste de Hipótese para a Diferença entre Médias

Quando queremos comparar a média de duas populações, retiramos amostras das duas populações, que podem apresentar tamanhos diferentes. Vamos considerar as situações de amostras independentes (as populações não apresentam nenhuma relação entre si) e de amostras dependentes (uma população sofre uma intervenção e avalia-se antes e depois da intervenção para saber se a intervenção resultou algum efeito).

1º caso: amostras independentes e grandes ($n > 30$) ou variâncias populacionais conhecidas.

2º caso: amostras independentes e pequenas ($n \leq 30$), mas que apresentam variâncias populacionais desconhecidas e estatisticamente iguais.

3º caso: amostras independentes e pequenas ($n \leq 30$), mas que apresentam variâncias populacionais desconhecidas e estatisticamente desiguais.

4º caso: amostras dependentes.

Vamos analisar cada uma dessas situações. Lembre-se de que as considerações anteriores em relação aos passos para formulação dos testes de hipóteses permanecem os mesmos.

A grande diferença, como você verá, ocorrerá somente na determinação das hipóteses a serem testadas. A hipótese H_0 será:

$H_0: \mu_1 - \mu_2 = d_0$

Onde:

μ_1: média da população 1;

μ_2: média da população 2; *e*

$\mathbf{d_0}$ corresponde a uma diferença qualquer que você deseja testar.

Geralmente, quando queremos saber se as médias das duas populações são estatisticamente iguais, utilizamos o valor de $\mathbf{d_0}$ igual a zero.

As hipóteses alternativas seguem a mesma linha de raciocínio, como você pode visualizar a seguir.

H_0	H_1
$\mu_1 - \mu_2 = d_0$	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$

É importante ressaltar que, se as hipóteses alternativas forem unilaterais, o sinal da hipótese H_0 será **menor ou igual** ou **maior ou igual** dependendo da hipótese alternativa, apesar de utilizarmos a notação de igual (conforme comentado anteriormente).

Todas as outras considerações em relação aos testes de hipótese permanecem as mesmas. Vamos, então, procurar entender cada situação para os testes de hipóteses para diferença entre médias.

1ª caso: amostras independentes e grandes ($n > 30$) ou variâncias populacionais conhecidas: como estamos trabalhando aqui com amostras grandes ou com desvios padrão populacionais conhecidos, devemos trabalhar com a distribuição amostral de Z (raciocínio semelhante ao utilizado no teste de hipótese para uma média). Portanto, a estatística do teste será dada por:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}}$$

Onde:

$\overline{X}_1$: média da amostra 1; $\overline{X}_2$: média da amostra 2; μ_1: média da população 1; μ_2:média da população 2; σ_1^2: variância da população 1; σ_2^2: variância da população 2; n_1: tamanho da amostra 1 e n_2 tamanho da amostra 2.

Se trabalharmos com amostras grandes, poderemos substituir as variâncias populacionais pelas variâncias amostrais sem nenhum problema.

Vamos, então, ver como podemos aplicar o teste de hipótese para a diferença entre médias nesta situação:

Foram retiradas amostras do valor recebido em milhares de reais de um determinado imposto de duas prefeituras (A e B) de mesmo porte. Os resultados são apresentados no quadro, a seguir. Verifique se as duas prefeituras têm o mesmo recebimento ou se são diferentes, com uma significância de 0,05.

Marcas	A	B
Média	1160	1140
Desvio padrão	90	80
Tamanho amostra	100	100

Como fazer:

Vamos retirar os dados apresentados em nossa situação:

Amostra A: $n = 100; \bar{x} = 1160; s = 90$

Amostra B: $n = 100; \bar{x} = 1140; s = 80$

As hipóteses a serem formuladas são:

H_0: $\mu_a = \mu_b \rightarrow \mu_a - \mu_b = 0$

H_1: $\mu_a \neq \mu_b$

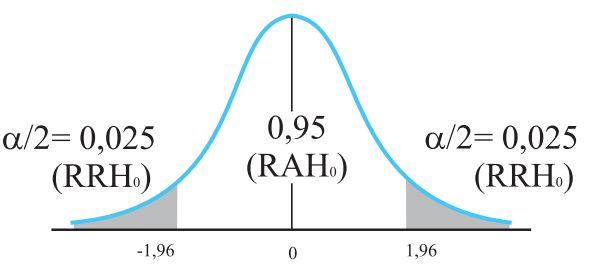
O teste t deve ser bilateral, já que a preocupação está na verificação do fato da média da prefeitura A ser diferente da média da prefeitura B.

$\alpha = 0{,}05$

A estatística usada será Z, pois as amostras são grandes (n > 30), apesar de não termos os desvios padrão populacionais. Sendo assim, nessa situação, ainda utilizamos a estatística de Z.

Substituindo os valores na estatística, teremos:

$$Z_c = \frac{(\bar{X}_a - \bar{X}_b) - (\mu_a - \mu_b)}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}} = \frac{(1160 - 1140) - (0)}{\sqrt{\frac{90^2}{100} + \frac{80^2}{100}}} = 1{,}67$$



Como o valor calculado Zc = 1,67 está entre os valores de –1,96 e 1,96, valores que dividem a RRH_0 da RAH_0, verificamos que o valor calculado Zc = 1,67 pertence a RAH_0 e podemos afirmar, com 95% de certeza, que os valores recebidos pelas duas prefeituras são estatisticamente iguais, ou seja, aquela diferença encontrada entre as amostras foi fruto do acaso.

2ª caso: amostras independentes e pequenas, mas que apresentam variâncias populacionais estatisticamente iguais e desconhecidas: você deve trabalhar com a distribuição t de *Student*, uma vez que as amostras que estamos trabalhando são pequenas, e as variâncias populacionais desconhecidas.

Aqui, estaremos considerando que as variâncias populacionais são estatisticamente iguais, pois essa situação influenciará nos cálculos e, consequentemente, no processo decisório.

Para saber se as variâncias podem ser consideradas iguais, você deve fazer um teste da razão de duas variâncias (teste F), apresentado anteriormente.

A estatística do teste será dada por:

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{1/n_1 + 1/n_2}}$$

Aqui, aparece um termo novo (Sp). Ele corresponde ao desvio padrão ponderado pelos graus de liberdade, ou seja, calculamos um novo desvio padrão cujo fator de ponderação corresponde ao grau de liberdade de cada amostra. Veja a seguir:

$$Sp = \sqrt{\frac{(n_1 - 1)s_1^{\,2} + (n_2 - 1)s_2^{\,2}}{n_1 + n_2 - 2}}$$

Para você encontrar o valor tabelado que limita as regiões de aceitação e de rejeição na tabela t de *Student*, o número de graus de liberdade (v) a ser usado na tabela será dado por:

$$v = n_1 + n_2 - 2$$

Lembre-se de que você pode voltar à tabela t de *Student* quando desejar, ela se encontra na Unidade 5.

Onde:

n_1 e n_2 correspondem aos tamanhos de amostras utilizados.

Vamos resolver, agora, uma situação na qual temos a comparação entre médias de amostras pequenas e variâncias populacionais desconhecidas e estatisticamente iguais.

Situação: em uma comparação de aprovação no vestibular de uma importante universidade, seis estudantes do sexo masculino de colégios da rede pública (amostra A) preencheram o gabarito no tempo médio de 6,4 minutos e desvio padrão de 60 segundos. Outra amostra foi formada por cinco estudantes do sexo feminino selecionados aleatoriamente do mesmo universo (amostra B), com os resultados de tempo de preenchimento do gabarito, de um tempo médio de 5,9 minutos e com desvio padrão de 60 segundos (assuma variâncias populacionais iguais). A Secretaria Municipal de Educação deseja saber se existe diferença ou não entre o sexo dos estudantes para definir se há necessidade de se fazer treinamentos específicos para cada sexo ou um mesmo treinamento para os dois sexos; para, assim, poder reduzir esse tempo e melhorar a *performance* dos estudantes da rede pública no vestibular.

Resolução:

Retirando os dados do nosso exemplo, teremos:

Amostra A: $n = 6; \bar{x} = 6{,}4; s = 1$

Amostra B: $n = 5; \bar{x} = 5{,}9; s = 1$

As hipóteses a serem formuladas são:

$H_0: \mu_a = \mu_b \rightarrow \mu_a - \mu_b = 0$

$H_1: \mu_a \neq \mu_b$

O teste t deve ser bilateral, já que a atenção está voltada para a preocupação em se constatar se, de fato, ocorre diferença entre os estudantes do sexo masculino ou feminino.

Caso isso não seja informado no problema, você deve fazer um teste de hipótese para comparar as variâncias populacionais com base nas variâncias amostrais, como vimos anteriormente.

$$\alpha = 0{,}05$$

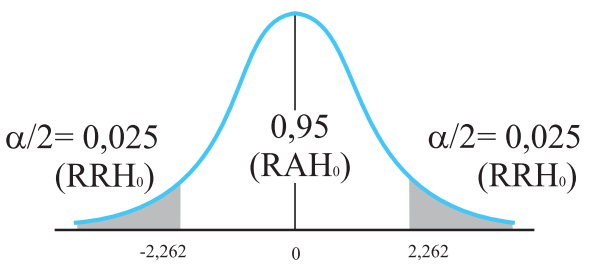
A estatística usada será t, pois as amostras são menores ou iguais a 30 ($n \leq 30$) e a variância populacional é desconhecida. Além disso, consideramos que as variâncias populacionais são estatisticamente iguais, informação que é dada no problema analisado.

Substituindo os valores nas expressões, teremos:

$$Sp = \sqrt{\frac{(n_a - 1).s_a^2 + (n_b - 1).s_b^2}{n_a + n_b - 2}} = \sqrt{\frac{5.1^2 + 4.1^2}{6 + 5 - 2}} = 1$$

$$t_c = \frac{(\bar{x}_a - \bar{x}_b) - (\mu_a - \mu_b)}{Sp\sqrt{\frac{1}{n_a} + \frac{1}{n_b}}} = \frac{(6{,}4 - 5{,}9) - (0)}{1\sqrt{\frac{1}{6} + \frac{1}{5}}} = \frac{0{,}5}{0{,}6055} = 0{,}82$$

$v = n_a + n_b - 2 = 6 + 5 - 2 = 9$ (grau de liberdade)



O valor $t_t = 2{,}262$ que divide a RRH_0 e RAH_0 foi encontrado na tabela t procurando grau de liberdade 9 e $\alpha = 0{,}025$. Como t calculado está entre os valores que dividem a região de aceitação de H_0, ou seja, 0,82 pertence a RAH_0, podemos afirmar com 95% de certeza que o tempo de preenchimento dos estudantes e das estudantes é a mesmo. Então, a prefeitura deve fazer o treinamento independentemente do sexo dos estudantes, ou seja, o mesmo treinamento para todos os estudantes.

Antes de analisar o terceiro caso, realize a Atividade 2, ao final desta Unidade.

3ª caso: amostras independentes e pequenas, mas que apresentam variâncias populacionais estatisticamente desiguais e desconhecidas: a diferença dessa situação para a anterior é que você agora considera que as populações apresentam variâncias estatisticamente desiguais. Para saber se elas são estatisticamente

desiguais ou diferentes, você deve fazer um teste de hipótese para a razão de duas variâncias, visto anteriormente nesta Unidade. Também utilizaremos a estatística do teste a partir da distribuição t de *Student*. A estatística do teste será dada por:

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^{\,2} / n_1 + s_2^{\,2} / n_2}}$$

Outra diferença está no cálculo do número de graus de liberdade, pois, nessa situação, utilizaremos uma aproximação que é dada pela expressão a seguir:

$$v = gl = \frac{(s_1^{\,2} / n_1 + s_2^{\,2} / n_2)^2}{\frac{(s_1^{\,2} / n_1)^2}{n_1 - 1} + \frac{(s_2^{\,2} / n_2)^2}{n_2 - 1}}$$

Se esse valor calculado apresentar valores decimais, você deve fazer o arredondamento para um número inteiro.

Vamos resolver, a seguir, outra situação.

Situação: uma prefeitura deseja reduzir seus custos com combustíveis. Para isso, deseja saber se duas marcas de carro apresentam o mesmo consumo ou se um dos fabricantes apresenta menor consumo. Não confiando nas especificações do fabricante, já que as condições de uso dos veículos pela prefeitura não são ideais. Para tomar a decisão acerca de qual comprar, foi analisada uma amostra de 22 automóveis das duas marcas, obtendo o resultado apresentado, a seguir. Seria possível afirmar que o carro Andaluz é mais econômico, isto é, que apresenta uma média populacional inferior que a do Reluzente? Assuma $\alpha = 5\%$ e população normalmente distribuída.

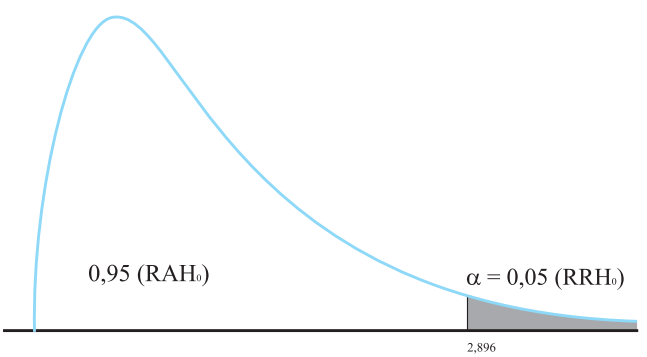
Automóvel	Tamanho da amostra	Média de consumo	Desvio padrão
Andaluz	12 unidades	14 km/l	2 km/l
Reluzente	10 unidades	15 km/l	4 km/l

Resolução:

Nessa situação, faremos um teste de hipótese para diferença entre médias populacionais. Como as amostras são pequenas, precisamos saber se as variâncias são estatisticamente iguais ou não. Para isso, vamos testar se as variâncias populacionais são estatisticamente iguais ou não por meio de teste de F. As hipóteses são:

$$H_0 : \sigma_R^2 = \sigma_A^2$$
$$H_1 : \sigma_R^2 > \sigma_A^2 \qquad \alpha = 0{,}05$$
$$F_c = \frac{s_R^2}{s_A^2} = \frac{16}{4} = 4$$

Como estabelecemos utilizar o teste unilateral no cálculo de F, teremos, então, a maior variância dividida pela menor variância. As variâncias populacionais não estão presentes na fórmula, devida, na hipótese H_0, serem consideradas iguais e, assim, elas se cancelam.



O valor 2,896 foi encontrado na tabela F de 5% com grau de liberdade 9 para o numerador e 11 para o denominador. Como $F_c > 2{,}896$, rejeita-se H_0 e, portanto, as variâncias populacionais são estatisticamente desiguais, ou seja, uma é maior do que a outra.

Agora, vamos testar as médias populacionais:

H_0: $\mu_{andaluz} = \mu_{reluzente} \rightarrow \mu_{andaluz} - \mu_{reluzente} = 0$

H_1: $\mu_{andaluz} < \mu_{reluzente}$

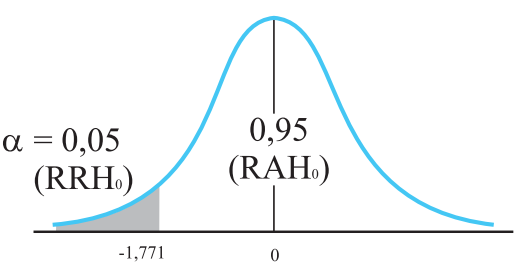
$\alpha = 0{,}05$

Como as amostras são independentes, pequenas e com variâncias populacionais estatisticamente desiguais, usaremos a estatística t.

Vamos encontrar o grau de liberdade:

$$V = \frac{\left(\frac{s_A^2}{n_A} + \frac{s_R^2}{n_R}\right)^2}{\frac{\left(\frac{s_A^2}{n_A}\right)^2}{n_A - 1} + \frac{\left(\frac{s_R^2}{n_R}\right)^2}{n_R - 1}} = \frac{\left(\frac{4}{12} + \frac{16}{10}\right)^2}{\frac{\left(\frac{4}{12}\right)^2}{11} + \frac{\left(\frac{16}{10}\right)^2}{9}} = \frac{3{,}74}{0{,}01 + 0{,}28} = \frac{3{,}74}{0{,}29} = 12{,}89 \cong 13$$

$$t = \frac{(\bar{x}_A - \bar{x}_R) - (\mu_{andaluz} - \mu_{reluzente})}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_R^2}{n_R}}} = \frac{(14 - 15) - (0)}{\sqrt{\frac{4}{12} + \frac{16}{10}}} = \frac{-1}{1{,}39} = -0{,}72$$



O valor $t_t = -1{,}771$, que divide a RRH_0 e RAH_0, foi encontrado na tabela t procurando grau de liberdade 13 e $\alpha = 0{,}05$. Como t calculado ($t = 0{,}72$) pertence a RAH_0, podemos afirmar, com 95% de certeza, que o consumo dos carros Andaluz e Reluzente é o mesmo, ou seja, tanto faz a prefeitura comprar uma marca ou outra que o consumo será o mesmo.

Antes de passarmos ao estudo do quarto caso, resolva a Atividade 3, ao final desta Unidade. Dessa forma, você poderá aplicar os conhecimentos sobre a diferença entre médias.

4ª caso: amostras dependentes: sabemos que amostras dependentes ocorrem quando fazemos uma intervenção e desejamos saber se os resultados antes da intervenção são iguais aos resultados depois da intervenção. Um ponto importante, nessa situação, é que são calculadas, primeiramente, as diferenças de antes e de depois. Essa diferença é chamada de d_i.

Então, você pode ver que:

d_i = valor antes – valor depois

Com base nessas diferenças (d_i), você irá calcular a média ($\overline{D}$) e o desvio padrão dessas diferenças (S_D).

$$\overline{D} = \frac{\sum_{i=1}^{n} d_i}{n} \qquad e \qquad S_D = \frac{\sum_{i=1}^{n} d_i^2 - \frac{\left(\sum_{i=1}^{n} d_i^2\right)}{n}}{n-1}$$

Veja que essas fórmulas são iguais às de cálculo da média e do desvio padrão apresentados anteriormente. Nesse caso, no lugar da variável x são utilizados os valores de di (diferenças).

Com esses valores, a estatística teste será dada por:

$$t = \frac{\overline{D} - d_O}{S_D / \sqrt{n}}.$$

O valor de n corresponde ao número de diferenças calculadas e o grau de liberdade para ser olhado na tabela t de *Student* será dado por n – 1.

Vamos resolver uma situação em que trabalharemos com o caso de amostras dependentes.

Situação: em um estudo procurou-se investigar se a redução em uma gratificação no salário iria diminuir a produtividade dos funcionários de uma prefeitura, considerando uma escala de produtividade de 0 a 12. A tabela a seguir dá os resultados de pessoas selecionadas anteriormente. No nível de 5% de significância, teste a afirmação de que a redução da gratificação reduziu a produtividade, ou seja, que a diferença entre antes e depois deve ser maior do que zero.

Pessoa	A	B	C	D	E	F	G	H
Antes	6,6	6,5	9,0	10,3	11,3	8,1	6,3	11,6
Depois	6,8	2,4	7,4	8,5	8,1	6,1	3,4	2,0

Primeiramente, vamos montar as nossas hipóteses:

H_0: $\mu_D = 0$

H_1: $\mu_D < 0$

Veja que as escolhas dessas hipóteses estão associadas ao que queremos testar. No caso da hipótese H_0: $\mu_D = 0$, estamos testando que as médias das diferenças de antes menos depois são iguais a zero, ou seja, que a redução na gratificação não interferiu na produtividade (a produtividade foi a mesma), já que estamos avaliando os mesmos indivíduos. No caso da hipótese H_1: $\mu_D > 0$, estamos testando que os valores de antes eram maiores do que os valores de depois da redução da gratificação, ou seja, se esta diferença de antes menos de depois for maior do que zero, indica que antes da intervenção os funcionários tinham uma produtividade maior antes do que depois.

Poderíamos testar também, dependendo do caso, as hipóteses H_1: $\mu_D < 0$ ou H_1: $\mu_D \neq 0$.

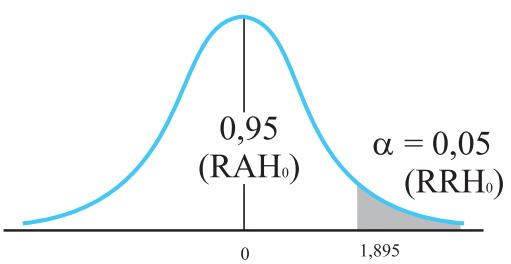
Consideramos um $\alpha = 0{,}05$.

Para calcularmos os valores de $\overline{D}$ e S_D, devemos, primeiramente, calcular as diferenças entre os valores de antes menos de depois de cada indivíduo e com essas diferenças calcular a média das diferenças ($\overline{D}$) e o desvio padrão das diferenças (S_D) para utilizá-las na expressão de t para amostras dependentes. Os resultados das diferenças são apresentados a seguir:

Pessoa	A	B	C	D	E	F	G	H
Antes	6,6	6,5	9	10,3	11,3	8,1	6,3	11,6
Depois	6,8	2,4	7,4	8,5	8,1	6,1	3,4	2
Diferença (antes – depois)	-0,2	4,1	1,6	1,8	3,2	2	2,9	9,6

Como as amostras são dependentes, usaremos a estatística t da seguinte forma:

$$t = \frac{\overline{D} - d_O}{S_D / \sqrt{n}} = \frac{3,125 - 0}{2,9114 / \sqrt{8}} = 3,03$$



O valor $t_t = 1,895$, que divide a RRH_0 e RAH_0, foi encontrado na tabela t quando procurávamos o grau de liberdade, 7 graus de liberdade (n –1, onde n é o número de indivíduos avaliados) e $\alpha = 0,05$. Como t calculado (t = 3,03) pertence a RRH_0, podemos considerar que os valores de produtividade eram maiores antes e, assim, a redução na gratificação influenciou na produtividade dos funcionários da prefeitura.

Teste de Hipótese para a Diferença entre Proporções

Em diversas situações, o que nos interessa é saber se a proporção de sucessos (evento de interesse) em duas populações apresenta a mesma proporção ou não. Nesse caso, os dados seguem uma Distribuição de proporção Bernoulli com média p e variância pq. Portanto, a expressão da estatística teste (no caso utilizaremos a distribuição de Z) será dada por:

Vimos sobre a Distribuição de Bernolli na Unidade 5. Você pode retomar esse conceito.

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}}$$

Onde:

$\hat{p}_1$ e $\hat{p}_2$: correspondem à proporção de sucesso nas amostras 1 e 2, respectivamente; e

p_1 e p_2: correspondem à proporção de sucesso nas populações 1 e 2, respectivamente.

Você deve se lembrar de que a proporção de fracasso (q) é dada por um menos a proporção de sucesso.

Vejamos como aplicar o teste da diferença de proporções.

Situação: uma empresa de pesquisa de opinião pública selecionou, aleatoriamente, 500 eleitores da Bahia e 600 de Pernambuco, e perguntou a cada um se votaria ou não no candidato Honesto Certo nas próximas eleições presidenciais. Responderam afirmativamente 80 eleitores da Bahia e 150 eleitores de

Pernambuco. Existe alguma diferença significativa entre as proporções de eleitores a favor do candidato nos dois Estados? Use nível de significância igual a 6%.

Como fazer:

Bahia: $n = 500;\ \hat{p} = \dfrac{80}{500} = 0{,}16;\ \hat{q} = 0{,}84$

Pernambuco: $n = 600;\ \hat{p} = \dfrac{150}{600} = 0{,}25;\ \hat{q} = 0{,}75$

Vamos estabelecer as hipóteses:

H_0: $p_B = p_p \rightarrow p_B - p_p = 0$

H_1: $p_B \neq p_P \rightarrow p_B - p_p \neq 0$

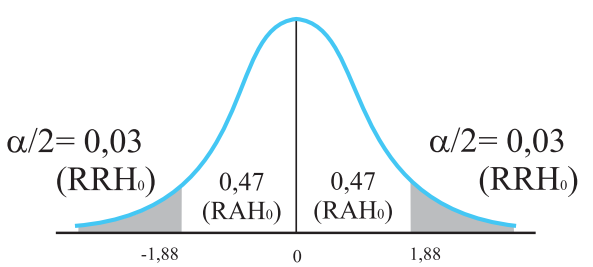
Aqui, seguem as mesmas considerações vistas anteriormente para a formulação das hipóteses.

$\alpha = 0{,}06$

A estatística usada será Z.

$$Z_c = \frac{(\hat{p}_B - \hat{p}_P) - (p_B - p_P)}{\sqrt{\left(\frac{\hat{p}_B.\hat{q}_B}{n_B}\right) + \left(\frac{\hat{p}_P.\hat{q}_P}{n_P}\right)}} = \frac{(0{,}16 - 0{,}25) - (0)}{\sqrt{\left(\frac{0{,}16.0{,}84}{500}\right) + \left(\frac{0{,}25.0{,}75}{600}\right)}} =$$

$$\frac{-0{,}09}{\sqrt{0{,}0002688 + 0{,}0003125}} = \frac{-0{,}09}{0{,}024} = -3{,}73$$



O valor 1,88 foi encontrado no **interior** da tabela Z procurando 0,4699.

Como Z calculado está na região de rejeição de H_0 (menor que –1,88), rejeitamos H_0 e, portanto, podemos afirmar com 94% de certeza que existe diferença significativa entre as proporções de eleitores a favor do candidato nos dois Estados.

Veja que 0,47 não existe na tabela, então, optamos pelo valor mais próximo.

Teste do Qui-Quadrado de Independência

O teste do qui-quadrado de independência está associado a duas variáveis qualitativas, ou seja, a uma análise bidimensional. Muitas vezes, queremos verificar a relação de dependência entre as duas variáveis qualitativas a serem analisadas.

Nesse caso, procuramos calcular a frequência de ocorrência das características dos eventos a serem estudados. Por exemplo, podemos estudar a relação entre o sexo de pessoas (masculino e feminino) e o grau de aceitação do governo estadual (ruim, médio e bom). Então, obteremos, por exemplo, o número de pessoas (frequência) que são do sexo feminino e que acham o governo bom. Todos os cruzamentos das duas variáveis são calculados.

Vamos apresentar a você, como exemplo, os possíveis resultados da situação sugerida anteriormente (dados simulados).

Grau de aceitação				
Sexo	Ruim	Médio	Bom	Total
Masculino	157	27	74	258
Feminino	206	0	10	216
Total	363	27	84	474

Podemos determinar o grau de associação entre essas duas variáveis, ou seja, determinar se o grau de aceitação do governo depende do sexo ou se existe uma relação de dependência.

As hipóteses a serem testadas são:

H_0: variável linha independe da variável coluna (no exemplo anterior, o grau de aceitação independe do sexo das pessoas).

H_1: variável linha está associada à variável coluna (no exemplo anterior, o grau de aceitação depende do sexo das pessoas).

A estatística de qui-quadrado será dada por meio da seguinte expressão:

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(fo_i - fe_i)^2}{fe_i} = \frac{(fo_1 - fe_1)^2}{fe_1} + \frac{(fo_2 - fe_2)^2}{fe_2} + ... + \frac{(fo_k - fe_k)^2}{fe_k}$$

Onde:

k corresponde ao número de classes (frequências encontradas). Você pode verificar que **fo** corresponde à frequência observada, ou seja, ao valor encontrado na tabela de contingência.

Já **fe** corresponde à frequência esperada caso as variáveis não tenham nenhuma relação de dependência, ou seja, caso as duas variáveis sejam independentes. Por causa dessa definição, o cálculo da frequência esperada (**fe**) será obtido por:

$$fe = \frac{(total\ linha).(total\ coluna)}{total\ geral}$$

Nesse caso, os graus de liberdade (v), para que possamos olhar a tabela de qui-quadrado, são dados por:

v = (h–1) (k–1) nas tabelas com h linhas e k colunas

(no exemplo anterior: v = (2–1) x (3–1) = 2 graus de liberdade)

Então, para cada célula da tabela de contingências, você irá calcular a diferença entre **fe** e **fo**. Essa diferença é elevada ao quadrado para evitar que as diferenças positivas e negativas se anulem. A divisão pela frequência esperada é feita para obtermos diferenças em termos relativos.

Vamos entender melhor o teste de qui-quadrado do tipo independência por meio da análise de outra situação.

Situação: o gestor de uma prefeitura deseja saber como seus funcionários atuam no uso do MSN durante o trabalho. Para realizar um programa de conscientização, os gestores públicos precisam saber se o fato de os funcionários usarem pouco ou muito o MSN durante o trabalho depende do sexo das pessoas. Mediante essa informação, a gestão pode definir se fará programas de conscientização para homens e mulheres de forma separada ou em conjunto (um único programa). Para testar essa hipótese, foram selecionados, ao acaso, 96 funcionários de ambos os sexos que usavam pouco ou muito o MSN em razão dessas características na população. Verifique, com uma significância de 5%, a hipótese do gestor público.

Uso do MSN Sexo		
	Pouco	Muito
Homem	8	32
Mulher	16	40

Resolução:

Definindo primeiro as hipóteses H_0 e H_1.

H_0: uso do MSN independe do sexo.

H_1: uso do MSN depende do sexo.

Agora, iremos calcular as frequências esperadas, que são os valores que estão entre parênteses. Confira os cálculos das outras frequências esperadas cujos valores (**fe**) aparecem entre parênteses.

Uso do MSN			
Sexo	Pouco	Muito	
Homem	8 (10)	32 (30)	40
Mulher	16 (14)	40 (42)	56
	24	72	96

$$\frac{56 \cdot 24}{96} = 14$$

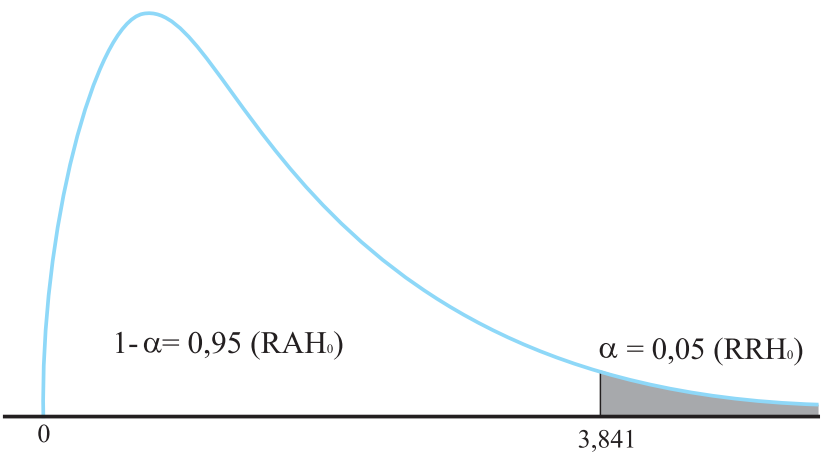
Agora, basta substituir os valores das frequências esperadas e observadas de todas as classes.

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(fo_i - fe_i)^2}{fe_i} = \frac{(8-10)^2}{10} + \ldots\ldots + \frac{(40-42)^2}{42} = 0{,}914$$

O valor do grau de liberdade é apresentado a seguir:

$v = (2 - 1) . (2–1) = 1$ gl

Considerando um $\alpha = 0{,}05$ e olhando na tabela de qui-quadrado para 1 grau de liberdade, teremos:

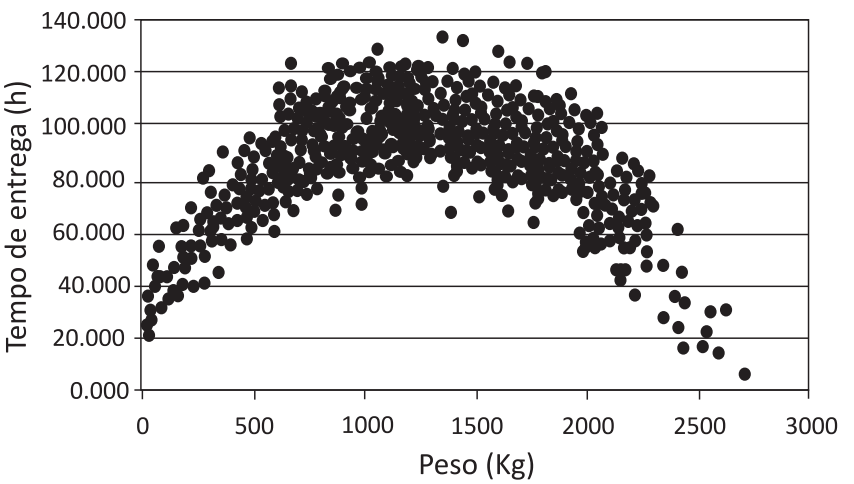


Como o valor calculado (0,914) foi menor do que o tabelado (3,841), então o calculado caiu na região de aceitação de H_0. Portanto, não temos indícios para rejeitar a hipótese H_0, ou seja, o uso do MSN independe do sexo dos funcionários. Dessa forma, o gestor pode fazer um único programa de conscientização tanto para homens quanto para mulheres.

Associação entre variáveis

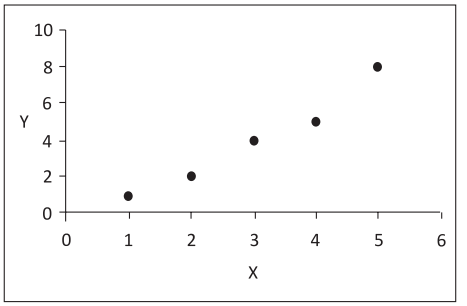
Para verificar o grau de relacionamento entre duas variáveis, ou seja, o grau de associação entre elas, devemos estudar um coeficiente chamado de coeficiente de correlação. Existem vários coeficientes de correlação e, cada um deles, aplicado em casos específicos. Aqui, iremos estudar o coeficiente de correlação de Pearson (r).

Para que possamos ter uma ideia da associação entre as variáveis que estamos estudando, iremos utilizar um gráfico de dispersão como o apresentado, a seguir, pelo qual podemos constatar a relação entre as variáveis: o peso de um pacote e o seu tempo de entrega.

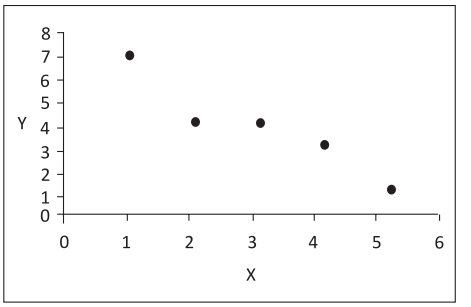


As estimativas correlação podem ser positivas (à medida que a variável x aumenta a variável y também aumenta) ou negativas (à medida que a variável x aumenta a variável y diminui), como você pode ver a partir dos dados e dos gráficos a seguir:

Positiva		Negativa	
1	1	1	1
2	2	2	2
3	3	3	3
4	4	4	4
5	5	5	5



Correlação Positiva



Correlação Negativa

O coeficiente de correlação de Pearson (r) nos dá uma ideia da variação conjunta das variáveis analisadas e pode assumir valores de –1 a +1.

Veja a expressão por meio da qual podemos obter o coeficiente de correlação de Pearson:

No exemplo que iremos trazer mais adiante, você encontrará a explicação dos somatórios dessa expressão. Não se preocupe!

$$r = \frac{\sum x_i y_i - \frac{\sum x_i - \sum y_i}{n}}{\left[\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right] \cdot \left[\sum y_i^2 - \frac{(\sum y_i)^2}{n}\right]}$$

A ocorrência de um valor de r = 0 ou próximo de zero indica apenas que não há correlação **linear** entre as variáveis, porque pode existir uma forte relação não linear entre as variáveis, como no gráfico de dispersão do peso do pacote e o tempo de entrega, na qual temos uma relação não linear.

Vejamos as características que o coeficiente de correlação de Pearson pode apresentar:

- seus valores estão compreendidos entre -1 e 1;
- se o coeficiente for positivo, as duas características estudadas tendem a variar no mesmo sentido.

- se o sinal for negativo, as duas características estudadas tendem a variar em sentido contrário;
- a relação entre duas variáveis é tanto mais estreita quanto mais o coeficiente se aproxima de 1 ou –1; e
- o valor de r é uma estimativa do parâmetro **ρ (rho)**, da mesma forma que a média x é uma estimativa de μ. Para testar se o valor de r é estatisticamente igual ao parâmetro de uma população em que **ρ (rho)** = 0, podemos empregar o teste **t** definido por:

$$t_c = \frac{r - \rho}{\sqrt{1 - r^2}} \cdot \sqrt{n-2}$$

onde:

n : número total de pares;

r^2 : coeficiente de correlação ao quadrado; e

ρ : parâmetro da correlação populacional (considerado igual a zero).

A hipótese H_0 será de que **ρ (rho)** = 0 e a hipótese H_1, que iremos utilizar, será de que **ρ (rho)** ≠ 0.

Vamos analisar a situação, a seguir, para entender melhor esse coeficiente.

Situação:

Vamos determinar o coeficiente de correlação entre a porcentagem de aplicação do total de recursos com Educação em uma prefeitura (x) e o grau de conhecimento médio da população da cidade (y). Para isso, foram avaliadas dez cidades.

Porcentagem de aplicação do total de recursos com educação em uma prefeitura	Grau de conhecimento médio da população da cidade
5	70
10	40
20	27
30	22
40	18
50	16
60	15
70	14
80	13
90	12

Para obtermos a estimativa de correlação, precisamos calcular todos os somatórios presentes na expressão:

$$r = \frac{\sum x_i y_i - \frac{\sum x_i - \sum y_i}{n}}{\sqrt{\left[\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right] \cdot \left[\sum y_i^2 - \frac{(\sum y_i)^2}{n}\right]}}$$

Calculando os somatórios, teremos:

Somatório de todos os valores de x:

$$\sum x_i = x_1 + x_2 + \ldots\ldots + x_{10} = 5 + 10 + \ldots\ldots + 90 = 455$$

Somatório de todos os valores de x elevados ao quadrado:

$$\sum x_i^2 = x_1^2 + x_2^2 + \ldots\ldots + x_{10}^2 = 5^2 + 10^2 + \ldots\ldots + 90^2 = 28525$$

Somatório de todos os valores de y:

$$\sum y_i = y_1 + y_2 + \ldots\ldots + y_{10} = 70 + 40 + \ldots\ldots + 12 = 247$$

Somatório de todos os valores de y elevados ao quadrado:

$$\sum y_i^2 = y_1^2 + y_2^2 + \ldots\ldots + y_{10}^2 = 70^2 + 40^2 + \ldots\ldots + 12^2 = 9027$$

Somatório de todos os valores obtidos por meio do produto dos valores de x e y de cada cidade:

$$\sum x_i y_i = x_1 y_1 + x_2 y_2 ..... + x_{10} y_{10} =$$
$$\sum x_i y_i = 5 \,.\, 70 + 10 \,.\, 40 + ... + 90 \,.\, 12 = 7470$$

Substituindo esses valores na expressão, teremos:

$$r = \frac{7470 - \frac{455.247}{10}}{\sqrt{\left[28525 - \frac{(455)^2}{10}\right] . \left[9027 - \frac{(247)^2}{10}\right]}} = \frac{-3768,5}{4784,28} = -0,7877$$

O valor de r = -0,7877 indica que existe uma associação inversa (negativa) e de média magnitude entre a variação da porcentagem de aplicação do total de recursos com educação em uma prefeitura e o grau de conhecimento médio da população da cidade, ou seja, nesta população de cidades, provavelmente os recursos da educação não estão sendo bem empregados, já que a relação foi negativa quando se esperava uma relação positiva.

Para verificarmos se esse resultado é significativo, vamos fazer o seguinte teste de hipótese:

H_0: **ρ (rho)** = 0

H_1: **ρ (rho)** ≠ 0.

Iremos calcular a estatística por meio da expressão:

$$t_c = \frac{r - \rho}{\sqrt{1 - r^2}} . \sqrt{n - 2}$$

Substituindo os valores na expressão, teremos:

$$t_c = \frac{-0,78770 - 0}{\sqrt{1 - 0,7877^2}} . \sqrt{10 - 2} = -1,25 \cdot 2,82 = 3,525$$

Olhando na tabela de t para 8 graus de liberdade (10-2) e um α=0,025, já que estamos considerando uma significância de 0,05 e o nosso teste é bilateral, teremos um valor tabelado de 2,306. Verificamos que o valor calculado de 3,525 está na região de rejeição da hipótese H_0 e, portanto, iremos aceitar a hipótese

H_1, ou seja, de que **ρ (rho)** ≠ 0. Então, o resultado encontrado na amostra (r) não foi fruto do acaso, considerando uma significância de 5%.

Devemos ter cuidado na interpretação do coeficiente de correlação, pois este não implica necessariamente uma medida de causa e efeito. É mais seguro interpretar o coeficiente de correlação como uma medida de associação. Por exemplo, podemos encontrar uma correlação muito alta entre o aumento dos salários dos professores e o consumo de bebidas alcoólicas através de uma série de anos em uma região. Esse valor de r encontrado foi alto apenas porque pode ser que ambas as variáveis tenham sido afetadas por uma causa comum, ou seja, a elevação do padrão de vida de uma região.

Complementando...

Através do *link* que apresentamos a seguir, você poderá fazer os testes de hipóteses e de estimativas de correlação de Pearson.

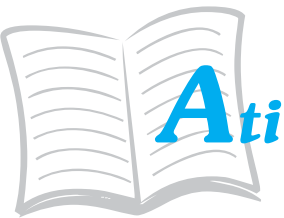
Programa estatístico Bioestat. Disponível em: <http://www.mamiraua.org.br/download/Default.aspx?dirpath=e:\home\mamiraua\Web\download\BioEstat 5 Portugues&tipo=diretorio>. Acesso em: 29 nov. 2010.

Resumindo

Nesta Unidade, conhecemos os principais testes de hipóteses e vimos suas aplicações no dia a dia da gestão de empresas públicas.

Apresentamos a estrutura de um teste de hipótese, de testes de hipóteses para médias, de diferença entre médias e de diferença entre proporções.

Verificamos que o teste de qui-quadrado pode ser utilizado para medir a dependência entre variáveis qualitativas. Dessa forma, você terá plenas condições de aplicar e de interpretar um teste estatístico de maneira correta.



Atividades de aprendizagem

Chegou o momento de analisarmos se você entendeu o que estudamos até aqui! Para saber, procure, resolver as atividades propostas a seguir. Lembre-se: você pode contar com o auxilio de seu tutor.

1. Um fabricante afirma que seus pneus radiais suportam em média uma quilometragem superior a 40.000 km. Uma prefeitura compra os pneus desse fabricante. Existe uma dúvida no setor de compras da prefeitura: “A afirmação do fabricante está correta?”. Para testar essa afirmação, a prefeitura selecionou uma amostra de 49 pneus. Os testes, nessa amostra, forneceram uma média de 43.000 km. Sabe-se que a quilometragem de todos os pneus tem desvio padrão de 6.500 km. Se o comprador (gestor público) testar essa afirmação ao nível de significância de 5%, qual será sua conclusão?
2. Duas técnicas de cobrança de impostos são aplicadas em dois grupos de funcionários do setor de cobrança de uma prefeitura. A técnica A foi aplicada em um grupo de 12 funcionários, resultando em uma efetivação média de pagamento de 76% e uma variância de 50%. Já a técnica B foi aplicada em um grupo de 15 funcionários, resultando em uma efetivação média de 68% e uma variância de 75%. Considerando as variâncias estatisticamente iguais e com uma significância de 0,05, verifique se as efetivações de pagamento são estatisticamente iguais.

3. Um secretário da Educação de uma prefeitura deseja saber se há, no futuro, profissionais promissores em escolas de regiões pobres e de regiões ricas. Uma amostra de 16 estudantes de uma zona pobre resultou, em um teste específico, uma média de 107 pontos e um desvio padrão de 10 pontos. Já 14 estudantes de uma região rica apresentaram uma média de 112 pontos e um desvio padrão de 8 pontos. Você deve verificar se a média dos pontos dos dois grupos é diferente ou igual a fim de que o empresário possa saber se ele deve investir em qualquer uma das áreas ou se uma delas é mais promissora (primeiro verifique se as variâncias são estatisticamente iguais ou diferentes).

Respostas das Atividades de aprendizagem

Unidade 1

1. a) Qualitativa Nominal.

   b) Qualitativa Ordinal.

   c) Quantitativa Discreta.

   d) Quantitativa Contínua.

2. a) Amostragem Sistemática.

   b) Amostragem por Conglomerado.

   c) Amostragem Estratificada.

   d) Amostragem Aleatória Simples.

   e) Amostragem Sistemática.

   f) Amostragem Aleatória Simples.

   g) Amostragem Estratificada.

   h) Amostragem por Conglomerado.

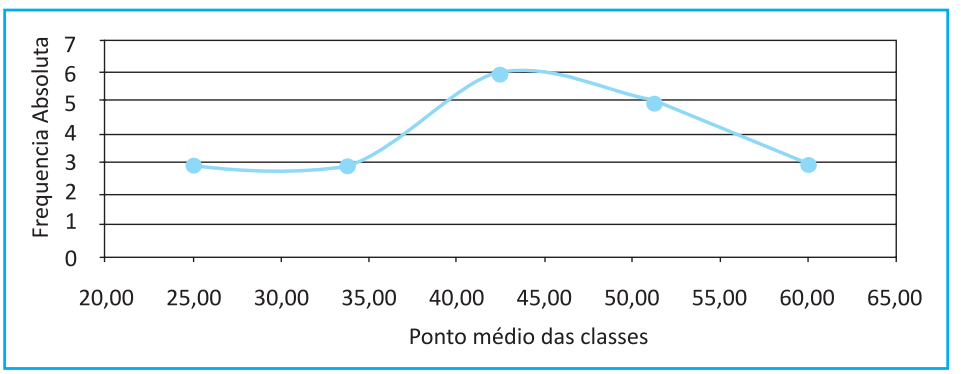
Unidade 2

1. a) n = 20, A = 35, k = 5 (aproximadamente), c = 8,75, Li_1a = 20,925.

Classes	Frequências absolutas
20,625 ├— 29,375	3
29,375 ├— 38,125	3
38,125 ├— 46,875	6
46,875 ├— 55,625	5
55,625 ├— 64,375	3
Total	20

Classes	Frequências acumulada
20,625 ⊢ 29,375	3
29,375 ⊢ 38,125	6
38,125 ⊢ 46,875	12
46,875 ⊢ 55,625	17
55,625 ⊢ 64,375	20

b)



Unidade 3

1. $\bar{x} = \frac{\sum x_i}{n} = \frac{7+42+37+25+38+...+33}{27} = 26,6$

Md = $X_{\left(\frac{n+1}{2}\right)} = X_{\left(\frac{27+1}{2}\right)} = X_{14} = 25$ (elemento de posição 14º)

Mo = 18,23,25 e 28, todos esses valores tem frequência 2 (multimodal)

Variância: $S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{(7-26,6)^2 + ... + (33-26,6)^2}{27-1} = 94,33$

Desvio Padrão: $S = \sqrt{S^2} = \sqrt{94,33} = 9,7$

Coeficiente de Variabilidade:

$$CV = \frac{S}{\bar{x}} \cdot 100 = \frac{9,7}{26,6} \cdot 100 = 36,47\%$$

2. Média	21.0
Mediana	18.0
Moda	10.0
Desvio padrão	12.0
Coeficiente de Variação	57.3

Unidade 4

1. R: 1-(1/3 * 1/5 * 3/10) = 0,98.

2. a) R: 0,125.

   b) R: 0,0694.

   c) R: 0,1388.

3. a) R: 60/100.

   b) R: 40/100.

   c) R: 24/100.

   d) R: 76/100.

Unidade 5

1. R:$P(X = 5) = C_{20}^{5}\, 0{,}1^{5}\, 0{,}9^{15} = 0{,}03192.$

2. Distribuição binomial com n = 4 e p = ½

   a) R: P(x=2) . 2000 = 0,3750 . 2000 = 750 famílias.

   b) R: [P(1) + P(2)] . 2000 = (0,25 + 0,375) . 2000 = 1250 famílias.

   c) R: P(0) . 2000 = 0,0625 . 2000 = 125 famílias.

3. R: 1- [P(0)+P(1)], em que a distribuição de probabilidade é uma Poisson com parâmetro lambda.

   a) $\lambda = 1{,}4$ R= 0,40817

   b) $\lambda = 2{,}8$ R=0,76892

   c) $\lambda = 5{,}6$ R=0,97559

4. Para X = 2200 ➔ $Z = \frac{X - \mu}{\sigma} = \frac{2200 - 2000}{200} = 1{,}00$

Para X = 1700 ➔ $Z = \frac{X - \mu}{\sigma} = \frac{1700 - 2000}{200} = -1{,}50$

5. a) X = 20 ➔ Z = 0

X = 24 ➔ $Z = \frac{24 - 20}{5} = 0{,}8$

P(20 < X < 24) = P(0 < Z < 0,8) = 0,2881 (28,81 %).

b) X = 16 ➔ $Z = \frac{16 - 20}{5} = -0{,}8$

X = 20 ➔ Z = 0

P(16 < X < 20) = P (-0,8 < Z < 0) = P(0 < Z < 0,8) = 0,2881 = 28,81

c) X = 28 ➔ Z = (28 - 20)/ 5 = 1,6

P(X > 28) = P (Z > 1,6) = 0,5 - 0,4452 = 0,0548

6. $1 - \alpha = 0{,}95$ ➔ $\alpha = 0{,}05$ ➔ $\alpha/2 = 0{,}025$

$$e = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 1{,}96 . \frac{3}{\sqrt{100}} = 0{,}588$$

P(26,412 < μ < 27,588) = 0,95

Unidade 6

1. Sugestão: siga os passos para realizar um teste de hipótese:

$$Z = \frac{\overline{X} - \mu_O}{\sigma \ / \sqrt{n}} = \frac{43000 - 40000}{6500 / \sqrt{49}} = 3{,}23 \qquad Z_{\alpha} = Z_{0{,}05} = 1{,}64$$

Conclusão: como o valor calculado foi maior do que o tabelado (1,64), ele caiu na região de rejeição de H_0.

2. $H_0: \mu_A - \mu_B = 0 \qquad H_0: \mu_1 - \mu_2 \neq 0$

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{1/n_1 + 1/n_2}} = \frac{(76 - 68) - 0}{8\sqrt{1/12 + 1/15}} = 2{,}56$$

$t_{0,025}$ = 2,060

Conclusão: como o valor calculado foi maior do que o tabelado (2,060), ele caiu na região de rejeição de H_0.

3. $H_0: \mu_1 - \mu_2 = 0 \qquad H_0: \mu_1 - \mu_2 \neq 0$

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^{\,2}/n_1 + s_2^{\,2}/n_2}} = \frac{(112 - 107) - 0}{\sqrt{8^2/14 + 10^2/16}} = -1{,}52$$

v = 29,7425 = 30 (graus de liberdade obtido pela aproximação).

$t_{0,025}$ = 2,042 (com 30 gl)

Conclusão: como o valor calculado caiu na região de aceitação, as médias são estatisticamente iguais, o que indica que as duas regiões apresentam o mesmo potencial.

Considerações finais

Com os conhecimentos de estatística adquiridos ao longo deste livro, você agora já pode imaginar quantas análises estatísticas podem ser feitas. A análise de dados está presente até em uma simples ligação telefônica que uma empresa de crédito faz para você. A empresa cruza informações como sexo, renda mensal e hábitos de consumo para oferecer um produto na medida certa.

Para fazer tudo isso, é necessário, entretanto, conhecimento básico de estatística para que empresas de Gestão Pública ou não venham a descobrir como transformar quantidades de números e de gráficos em informações que servirão para reduzir os custos e aumentar os lucros. O problema é que falta gente qualificada e com conhecimento de mercado para realizar as análises de dados. Para você trabalhar com conceitos estatísticos em qualquer setor, é necessário desenvolver um raciocínio lógico e, também, administrar informações, além de procurar entender como e por que as coisas acontecem.

Para decidir algo importante, é necessário avaliar os riscos e as oportunidades. Para que isso seja feito com muita precisão, é necessária a estatística!

Assim, você poderá aplicar os conhecimentos de estatística aprendidos em áreas, como a de Recursos Humanos, de Produção, Financeira e muitas outras que você irá identificar à medida que seus conhecimentos na área de Administração forem aumentando.

Espero que você tenha gostado de trabalhar com Estatística e que ela seja uma importante ferramenta a ser utilizada em seu dia a dia.

Um grande abraço e sucesso em sua vida profissional, com bastante estatística, é o que desejamos a você.

Professor Marcelo Tavares

Referências

ARANGO, Hector G. *Bioestatística*: teórica e computacional. Rio de Janeiro: Guanabara Koogan, 2001.

BARBETTA, Pedro Alberto. *Estatística Aplicada às Ciências Sociais*. 4. ed. Florianópolis: Editora da UFSC, 2002.

BEIGUELMAN, Bernardo. *Curso Prático de bioestatística*. Ribeirão Preto: Revista Brasileira de Genética, 1996.

BRAULE, Ricardo. *Estatística Aplicada com Excel*: para cursos de administração e economia. Rio de Janeiro: Campus, 2001.

BUSSAB, Wilton O.; MORETTIN, Pedro. *Estatística Básica*. São Paulo: Atual, 2002.

COSTA NETO, Pedro Luiz de Oliveira. *Estatística*. São Paulo: Edgard Blucher, 2002.

DOWNING, D.; CLARK, J. *Estatística Aplicada*. São Paulo: Saraiva, 2000.

FONSECA, Jairo Simon da; MARTINS, Gilberto de Andrade. *Curso de Estatística*. Rio de Janeiro: LTC, 1982.

FREUD, Jonh E.; SIMON, Gary A. *Estatística aplicada*. Bookman, 2000.

HOUAISS, Instituto Antônio Houaiss. *Dicionário eletrônico Houaiss da Língua Portuguesa*. Versão monousuário, 3.0. Objetiva: junho de 2009. CD-ROM.

LEVINE, David M.; BERENSON, Mark L.; STEPHAN, David F. *Estatística*: teoria e aplicações usando o Microsoft Excel em português. Rio de Janeiro: LTC, 2000.

MORETTIN, Luiz Gonzaga. *Estatística Básica – Probabilidade*. São Paulo: Makron Books, 1999. 1 v.

_____. *Estatística Básica – Inferência*. São Paulo: Makron Books, 1999. 2 v.

SOARES, José F.; FARIAS, Alfredo A.; CESAR, Cibele C. *Introdução à Estatística*. Rio de Janeiro: LTC, 1991.

SPIEGEL, Murray R. *Probabilidade e Estatística*. São Paulo: Mc Graw Hill, 1993.

STEVENSON, William J. *Estatística Aplicada à Administração*. São Paulo: Harper, 1981.

TRIOLA, Mário F. *Introdução à Estatística*. Rio de Janeiro: LTC, 1999.

WONNACOTT, T. H., WONNACOTT, R. J. *Estatística Aplicada à Economia e à Administração*. Rio de Janeiro: LTC, 1981.

MINICURRÍCULO

Marcelo Tavares

Possui Graduação (1989) e Mestrado (1993) pela Universidade Federal de Lavras, e Doutorado pela Escola Superior de Agricultura Luiz de Queiroz/USP (1998). Atualmente, é professor Associado II da Universidade Federal de Uberlândia (UFU). Tem experiência na área de Estatística Aplicada e atua, principalmente, nos seguintes temas: modelagem estatística, estatística, amostragem, controle de qualidade e estatística multivariada. Também foi coordenador do Curso de Especialização em Estatística Empresarial do Núcleo de Estudos Estatísticos e Biométricos da Faculdade de Matemática e, atualmente, é Coordenador da Universidade Federal de Uberlândia na Universidade Aberta do Brasil (UAB) e ministro das disciplinas de Estatística para o Curso de Administração da UFU.