Probabilidade

Testes de Hipótese

Para facilitar, vamos utilizar:
As Hipóteses:
H0: os pacotes tem 500g de peso médio e 12g de desvio padrão.
H1: os pacotes tem 490g de peso médio e 12g de desvio padrão.

A Região Crítica ou de Rejeição

$$H_0 : RC = \{se\ \bar{x} < 494g\}$$

A Regra de Decisão:
Se na amostra aleatória de 25 pacotes, o peso médio amostral for menor que 494g, rejeita-se H0.

Repare que mesmo tomando esses cuidados, podemos estar cometendo erros:
Rejeitar uma hipótese verdadeira!
Aceitar uma hipótese falsa!

As probabilidades dos erros são:

$$P(erro\ tipo\ 1) = P(\bar{x}\ está\ na\ RC \mid H_0\ é\ verdadeiro) = \alpha$$

$$P(erro\ tipo\ 2) = P(\bar{x}\ não\ está\ na\ RC \mid H_1\ é\ verdadeiro) = \beta$$

No exemplo dos pacotes de café,
$\alpha = 0{,}091 = 9{,}1\%$ e $\beta = 0{,}023 = 2{,}3\%$.

A quantidade α também é conhecida como nível de significância.

TESTES DE HIPÓTESE

Uma maneira equivalente de tomarmos a decisão é através do cálculo da **probabilidade de significância, nível descritivo ou valor p (p-value).**

O que se faz é indicar a probabilidade de ocorrer valores da estatística mais extremos do que o observado, sob a suposição da hipótese H0 ser verdadeira.

TESTES DE HIPÓTESE

No exemplo dos pacotes de café, sendo a média da amostra igual a 495g,

$$p - value = P(\bar{x} < 495 \mid \mu_0 = 500) = 4{,}7\%$$

Adotando o nível de significância de 5%, rejeitamos a hipótese H0.

TESTES DE HIPÓTESE

- Quando a variância populacional é desconhecida, o teste de hipóteses é baseado na estatística t-Student.

- Também existem testes de hipóteses para variâncias e proporções.

TESTE PARA DUAS POPULAÇÕES

Em geral, nosso interesse é comparar duas ou mais populações.

Por exemplo, uma nova estratégia de marketing foi proposta e queremos saber se existe diferença entre ela e a atual em termos de comportamento em determinado grupo de consumidores.

TESTE PARA DUAS POPULAÇÕES

Usamos o teste t-Student ou teste t para duas amostras independentes.

As suposições para utilização do teste t são:

- as amostras são aleatórias e independentes;
- as observações são normalmente distribuídas e
- as variâncias dos dois grupos é a mesma.

TESTE PARA DUAS POPULAÇÕES

Caso as variâncias sejam diferentes, também é possível fazer a comparação com uma pequena modificação.

A comparação entre as populações também pode ser feita através das variâncias com a estatística F.

Os testes t e F são os baseados na suposição de normalidade da população.

Quando as amostras aleatórias são de tamanho grande e a distribuição dos dados não é normal, o Teorema Central do Limite possibilita o uso desses testes.

ANOVA COM UM FATOR

Quando o interesse é comparar mais de duas populações adotamos os chamados modelos de médias ou **modelo de análise de variância com um fator (One-Way ANOVA)**.

ANOVA → **An**alysis **o**f **Va**riance

ANOVA COM UM FATOR

Por exemplo, queremos comparar:

- ✓ Marcas de café para **verificar a aceitabilidade de uma nova** antes de seu lançamento no mercado;
- ✓ Estratégias de Marketing, em 6 empresas do mesmo segmento;
- ✓ Instituições Financeiras, entre as 4 privadas maiores.

ANOVA COM UM FATOR

No modelo ANOVA estuda-se a relação entre **uma variável de interesse ou resposta contínua e uma variável explicativa discreta**.

Antes da coleta de informações para testar hipóteses é importante saber que situações ou populações devem ser avaliadas.

Os **Delineamento de Experimentos e de Amostras** é uma importante estratégia de Estatística que tem suas bases em Metodologia Científica.

Basicamente temos que **verificar como as comparações entre populações podem ser feitas** de modo a **avaliar corretamente a influência de cada tratamento na resposta de interesse**.

ANOVA COM UM FATOR

Por exemplo, se uma nova estratégia de Marketing for adotada, é importante saber qual o retorno esperado em relação às já utilizadas.

Vamos supor que o serviço será oferecido através de telemarketing e checkout.

Com base nesses pressupostos, temos 4 situações para avaliar.

ANOVA COM UM FATOR

- Telemarketing x atual estratégia;
- Telemarketing x nova estratégia;
- **Checkout x atual estratégia;**
- **Checkout x nova estratégia.**

Para cada grupo identificado será selecionada uma amostra de clientes que serão submetidos a uma combinação de estratégias e mídia.

O modelo de análise de variância com um fator de classificação corresponde a escrever **a resposta de cada indivíduo das amostras como uma combinação das médias populacionais de cada população de mesma variância**.

Uma representação da média é na forma de efeitos, ou seja, a média de cada população é dada por:

$$\mu + \tau_i$$

ANOVA COM UM FATOR

Sendo *a* populações e *n* o tamanho da amostra em cada população, escrevemos a observação como:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad \begin{cases} i=1,2,\ldots,a \\ j=1,2,\ldots,n \end{cases}$$

Y_{ij}

- o valor da variável no j-ésimo indivíduo da amostra da i-ésima população;

μ

- constante para todas as observações (média geral);

τ_i

- efeito da i-ésima população (diferença entre a média da i-ésima população e a média geral, μ);

ε_{ij}

- erro aleatório do j-ésimo indivíduo da amostra da i-ésima população. Eles tem média zero e variância $\boldsymbol{\sigma^2}$.

Suposições do modelo de análise de variância:

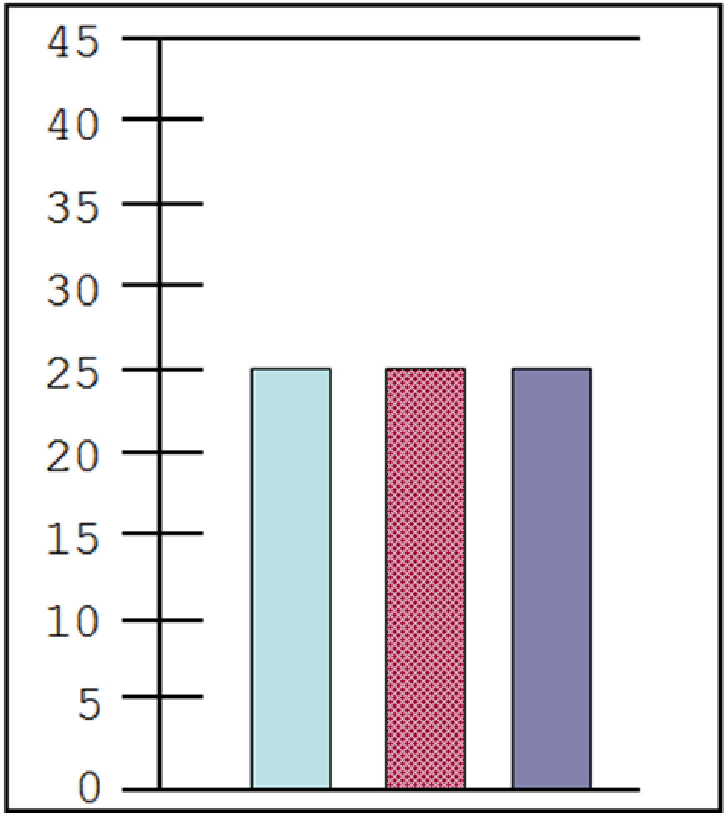
- ✓ as observações são resultados de amostras aleatórias ou os indivíduos são aleatorizados antes da aplicação dos tratamentos;

- ✓ os erros aleatórios tem média 0 (zero) e variância σ^2 constante para todos os níveis do fator, e são independentes.

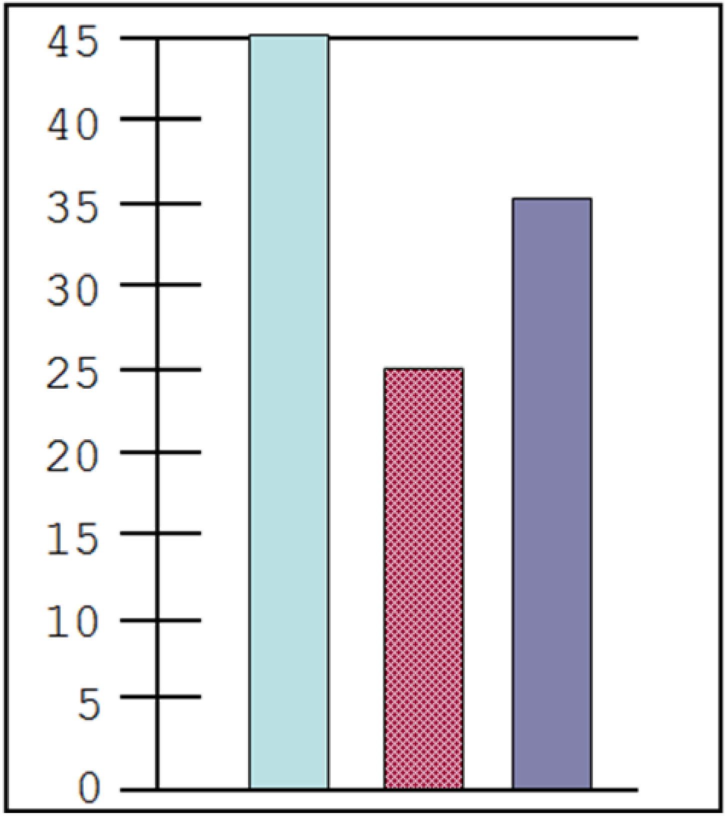
ANOVA COM UM FATOR

Hipótese nula da ANOVA

H_0: Todas as médias são iguais



H_1: Pelo menos uma média é diferente



As informações a respeito da **variação em torno da média geral ou variação total** é particionada em variação devido ao modelo e **variação devido ao erro aleatório** que definem as **causas ou fontes de variação (Source of Variation)**.

A quantificação das informações relacionadas às **fontes de variação – Source** são apresentadas na tabela chamada **Quadro de Análise de Variância** ou **ANOVA table**.

As informações do **Quadro de Análise de Variância** são expressas por:

- ✓ **Graus de Liberdade – DF - Degrees of Freedom;**
- ✓ **Soma de Quadrados – SS - Sum of Squares;**
- ✓ **Quadrado Médio – MS - Mean Square;**
- ✓ **Estatística de Teste – F Value e**
- ✓ **Valor p – p-value como Pr > F.**

Fonte de Variação Total corrigido ou Corrected total:

- **SST : Total Sum of Squares**.

- Soma de quadrados das diferenças entre o valor observado da variável resposta e a média geral. Muitas vezes chamado por Total Corrigido (Corrected Total).

- DF: an-1.

Fonte de Variação Modelo ou Model:

- ✓ **SSM: Model Sum of Squares** Soma de quadrados das diferenças entre a média de cada população e a média geral;

- ✓ **DF: a-1**;

- ✓ **MSM: Mean Square of Model** = SSM/(a-1);

Resíduo:

- **SSE: Error Sum of Squares** Soma de quadrados das diferenças entre cada valor observado e a média de cada população;
- **DF: a_n-a**;
- **MSE: Mean Square of Error** Quadrado médio do resíduo: SSE/(a_n-a).

ANOVA COM UM FATOR

Quando a hipótese de igualdade de médias é rejeitada, deseja-se saber que níveis do fator são diferentes.

Para tanto são adotados métodos específicos de **comparações múltiplas** que têm a característica de **manter o nível de significância sob controle**, o que não acontece se utilizamos o teste t diversas vezes.

Métodos de Comparações Múltiplas:

- ✓ Duncan,
- ✓ Tukey,
- ✓ Bonferroni.

O Duncan e Tukey são mais indicados para se fazer comparações duas a duas e Bonferroni as demais comparações.

Após o ajuste do modelo, é importante verificar se suas suposições estão satisfeitas:

- ✓ os erros tem distribuição normal e
- ✓ os erros tem variância constante.

Desenho esquemático dos resíduos ou resíduos padronizados são indicadores da duas suposições.

ANOVA com dois fatores (two-way)

Algumas vezes é importante verificar **a influência simultânea de dois fatores** na variável resposta.

Exemplos:

- ✓ estudar o efeito de **duas estratégias** de Marketing e três **tipos de mídia** no retorno esperado;
- ✓ verificar a influência da **semana do mês** (1ª, 2ª, 3ª, 4ª) e o **número de ligações** (1, 2 ou 3) na semana sobre a taxa de pagamentos atrasados, em certa instituição financeira.

ANOVA COM DOIS FATORES

Neste caso temos o modelo ANOVA com dois fatores.

A incorporação dos efeitos dos dois fatores, por exemplo A e B, na variável resposta pode ser feita através de:

- ✓ uma média geral,
- ✓ mais o efeito do Fator A,
- ✓ mais o efeito do Fator B.

ANOVA COM DOIS FATORES

Considerando que:

- o fator A tem a níveis,
- o fator B tem b níveis e
- o número de repetições em cada combinação de níveis é n,

A média da variável resposta, yijk pode ser expressa através do modelo :

$$\mu+\tau i+\beta j$$

Que representa a média populacional do i-ésimo nível do fator A e j-ésimo nível do fator B.

ANOVA COM DOIS FATORES

$$y_{ijn} = \mu + \tau_i + \beta_j + \varepsilon_{ijn}$$

para $i = 1,2,...,a$

$j = 1,2,...,b$

- μ - é a média geral,
- τ_i - é o efeito do *i-ésimo* nível do Fator A,
- β_j - é o efeito do *j-ésimo* nível do Fator B,
- ε_{ijn} - é o erro aleatório, independentes, com média 0 e variância σ^2 e
- n – é o número de repetições.

ANOVA COM DOIS FATORES

O **ajuste** do modelo com dois fatores é feito pelo **método de mínimos quadrados**, obtendo-se as estimativas dos parâmetros e os valores das observações preditos pelo modelo.

As **fontes de variação** que compõem o Quadro de Análise de Variância continuam divididas em: **variação total, variação devido ao modelo e variação devido ao erro aleatório**.

As modificações acontecem nas somas de quadrados e graus de liberdade do modelo e do resíduo.

Fonte de Variação do Modelo:

- ✓ *MSS - Model Sum of Squares*
- ✓ *DF*: *a+b-2*
- ✓ *MQM* (*Mean Square of Model*) : *MSS*/(*a+*b-2)

Dividida entre os dois fatores:

- ✓ Fator A *DF: a-1* *MSA = SSA/(a-1)*
- ✓ Fator B *DF: b-1* *MSB = SSB/(b-1)*

Resíduo:

- ✓ *SSE: Error Sum of Squares*.
- ✓ *DF: abn-(a+b-2).*
- ✓ *MSE: Mean Square of Error* Quadrado Médio do Resíduo = *SSE/[abn-(a+b-2)].*

Estatística F usada para verificar se o modelo é significante:

$$F = MSM/MSE$$

Dado que o modelo é significante, é interessante saber se os fatores também são:

- O teste para verificação da igualdade das médias dos níveis do fator A é baseada na estatística.

F= MSA/MSE

- O teste para verificação da igualdade das médias dos níveis do fator B é baseada na estatística.

F= MSB/MSE

ANOVA COM DOIS FATORES

As suposições do modelo ANOVA relacionadas com **variância constante e normalidade do erro** podem ser verificadas através de gráficos dos resíduos padronizados.

Usamos então diagramas de dispersão do resíduo versus valor predito, e resíduo versus níveis de fator.

Para **normalidade** dos resíduos padronizados.

Para **variância constante**, espera-se que os resíduos padronizados tenham o **mesmo espalhamento** para cada nível de cada fator.

ANOVA COM DOIS FATORES E INTERAÇÕES

Algumas vezes percebemos que o efeito de diferentes de níveis de um fator não é o mesmo para os níveis do segundo fator.

Isto é, no exemplo das estratégias de Marketing, o efeito das estratégias não é o mesmo em cada tipo de mídia.

Nesse caso é necessário incluir no modelo a interação entre os fatores estratégias e mídia.

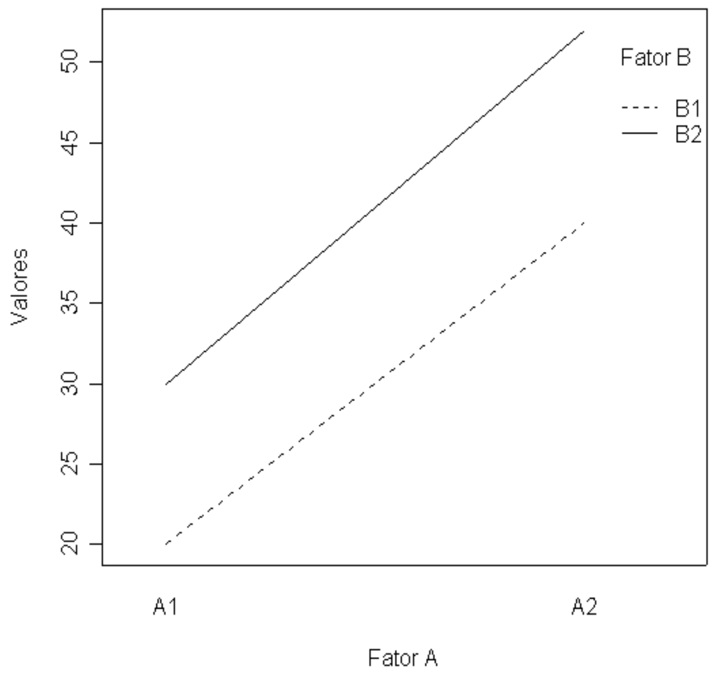
Esse comportamento pode ser observado graficamente.

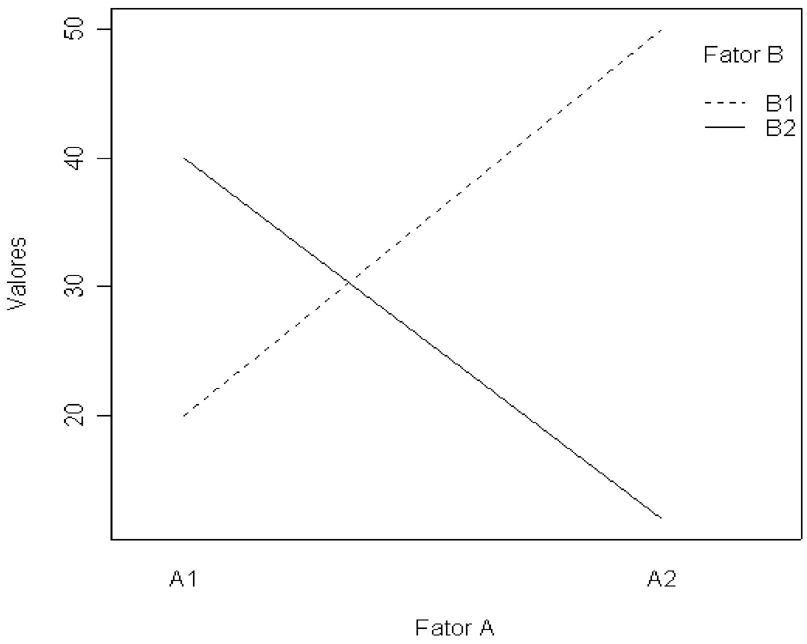
ANOVA COM DOIS FATORES E INTERAÇÕES

Sendo dois níveis em cada um dos dois fatores, calculamos os valores das 4 médias que correspondem a combinação dos níveis do fatores A e B:

- ✓ A1 e B1,
- ✓ A1 e B2,
- ✓ A2 e B1 e
- ✓ A2 e B2.

ANOVA COM DOIS FATORES E INTERAÇÕES





ANOVA COM DOIS FATORES E INTERAÇÕES

Modelo de médias com dois fatores e interação pode ser representado por:

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk}$$

para i = 1,2,...,a ; j = 1,2,...,b e k = 1,2,...,n

Onde:

a – número de níveis do fator A;

b – número de níveis do fator B e

n – número de indivíduos observados em cada combinação dos níveis de A e B.

ANOVA COM DOIS FATORES E INTERAÇÕES

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk}$$

y_{ijk} – valor da variável no *i-ésimo* nível de A, *j-ésimo* nível de B e na *k-ésima* repetição,

μ - média geral,

τ_i - efeito do *i-ésimo* nível de A,

β_j - efeito no *j-ésimo* nível de B,

$(\tau\beta)_{ij}$ - efeito da interação no *i-ésimo* nível de A e *j-ésimo* nível de B,

ε_{ijk} - erro aleatório, independente, com média 0 e variância σ^2.

ANOVA COM DOIS FATORES E INTERAÇÕES

O ajuste do modelo é feito pelo **método de mínimos quadrados**, obtendo-se as **estimativas** dos parâmetros e os **valores preditos** das observações.

As fontes de variação que compõem o Quadro de Análise de Variância continuam divididas em: **variação total, variação devido ao modelo e variação devido ao erro aleatório**.

As modificações acontecem nas somas de quadrados e graus de liberdade do modelo e do resíduo.

ANOVA COM DOIS FATORES E INTERAÇÕES

Fonte de Variação do Modelo:

- *MSS: Model Sum of Squares .*
- *DF*: $a+b+(a-1)(b-1)-2$
- MSM: *Mean Square of Model* = MSS/[$a+b+(a-1)(b-1)-2$]

A fonte de variação do modelo é dividida em 3:

Fator A	DF: $a-1$	MSA = SSA/($a-1$)
Fator B	DF: b-1	MSB = SSB/($b-1$)
Interação	DF: (a-1)(b-1)	MSAB = SSAB/[(a-1)(b-1)]

Resíduo:

- *SSE: Error Sum of Squares;*
- DF: *n-(a+b+(a-1)(b-1)-2);*
- *MSE: Mean Square of Error*

MSE = SSE/[*n-(a+b+(a-1)(b-1)-2*)].

- Estatística F para verificar se o modelo é significante:

F= MSM/MSE

ANOVA COM DOIS FATORES E INTERAÇÕES

Dado que o modelo é significante, é interessante saber se os fatores e a interação também o são.

- O teste para verificação da **igualdade das médias dos níveis do fator A** é baseado na estatística

$$F = MSA/MSE$$

- O teste para verificação da **igualdade das médias dos níveis do fator B** é baseado na estatística

$$F = MSB/MSE$$

- Dado que os fatores são significantes, o teste verificar a **significância da interação** é baseado na estatística

$$F = MSAB/MSE$$

Nosso interesse está em:

- ✓ **estudar a frequência** de ocorrência de variáveis categorizadas, **representá-las graficamente** e **verificar se existe um certo padrão.**

- ✓ **relacionar** duas variáveis, ou seja **conforme uma varia, como se dá o comportamento da segunda,** isto é, queremos conhecer o **grau de dependência ou associação** entre elas.

TABELAS E TESTES QUI-QUADRADO

Em Pesquisa de Marketing, nosso interesse pode ser:

- ✓ Verificar se **a preferência é igual para produtos de determinada linha.**

- ✓ Relacionar a **preferência por um produto** com a **condição socioeconômica** ajuda nos processos de ***segmentação do mercado*** **e de** ***planejamento de marketing*****.**

- ✓ Relacionar a **quantidade de gastos promocionais** com os **resultados das vendas ou de participação no mercado** ajuda no *planejamento futuro* desse tipo de gasto, tendo em vista **os *objetivos de vendas ou de participação no mercado.***

TABELAS E TESTES QUI-QUADRADO

Analisando apenas uma variável, além de obter a tabela de frequências e gráficos, nosso interesse é saber **se as categorias podem ter a mesma probabilidade de ocorrência ou não.**

Se no total houver ***C*** categorias da variável e a probabilidade da ocorrência de cada uma é p_i, $i=1, \ldots, C$, a hipótese a ser testada é

$$H_0: p_i = p_j \,,\ i \neq j,\ i,j=1, \ldots, C.$$
$$H_1: p_i \neq p_j \,,\ \text{para algum } i \neq j.$$

TABELAS E TESTES QUI-QUADRADO

O **teste quiquadrado** para avaliar esta situação, é baseado em:

- ***uma distribuição teórica***, que no contexto tem igual probabilidade para cada categoria,
- a **distribuição de frequências observadas** para a particular amostra e
- uma ***medida de afastamento dos valores observados dos valores esperados*** da distribuição teórica.

Para a situação de 3 categorias,

Categorias	Valores observados	Valores Esperados	Medida
1	O_1	E_1	$\frac{(O_1 - E_1)^2}{E_1}$
2	O_2	E_2	$\frac{(O_2 - E_2)^2}{E_2}$
3	O_3	E_3	$\frac{(O_3 - E_3)^2}{E_3}$

A estatística quiquadrado é definida como:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

TABELAS E TESTES QUI-QUADRADO

O **teste de quiquadrado** também pode ser utilizado na análise de **tabelas de dupla entrada**.

Em uma **primeira situação**, o objetivo é saber se as duas variáveis categorizadas **são relacionadas ou não**.

Então, a partir de ***uma amostra aleatória*** dos indivíduos da população, as variáveis categorizadas são anotadas.

Se elas **não são relacionadas**, dizemos que **as variáveis são independentes.**

Para verificar e testar a hipótese de que as variáveis são relacionadas, associadas ou independentes assumimos que:

- ✓ X é uma variável categorizada em R categorias $A_1, A_2, \ldots, A_R$ e a probabilidade de ocorrência do resultado $X{=}A_r$ é $P[X{=}A_r] = p_r$;

- ✓ Y é uma variável categorizada em C categorias $B_1, B_2, \ldots, B_C$ e a probabilidade de ocorrência do resultado $Y{=}B_c$ é $P[Y{=}B_c] = q_c$;

TABELAS E TESTES QUI-QUADRADO

A probabilidade de ocorrência do resultado $X = A_r$ e $Y = B_c$ é indicada por:

$$P[X = A_r\,,\ Y = B_c] = p_{rc}$$

Quando as variáveis **são independentes,**

$$p_{rc} = P[X = A_r\,,\ Y = B_c] = P[X = A_r] \times P[Y = B_c] = p_r q_c$$

O teste de hipóteses de que as variáveis são independentes tem por base as hipóteses:

- ✓ H_0: $p_{rc} = p_r q_c$ para qualquer r e c.
- ✓ H_1: $p_{rc} \neq p_r q_c$ para algum r e c.

Em uma **segunda situação**, o interesse é saber se o **comportamento da variável categorizada é o mesmo em diversas subpopulações.**

Nesse caso, o delineamento por amostras estratificadas é o mais indicado, ou seja, **de cada subpopulação** se retira **uma amostra aleatória**, e as frequências observadas de cada categoria são comparadas.

Sendo R populações e C categorias, a probabilidade da categoria c na subpopulação r é dada por:

$$p_{rc} = \mathrm{P}[X = A_c \mid r]$$

O **teste de hipótese de populações homogêneas** é formulado por:

- ✓ H_0: $p_{1c} = p_{2c} = \ldots = p_{Rc}$ para todo c.
- ✓ H_1: pelo menos uma igualdade não é verificada.

A estatística quiquadrado para avaliar as hipóteses nessas duas situações tem a mesma forma.

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Uma medida de quantificação da associação entre duas variáveis categorizadas é o **coeficiente de contingência de Cramer**, que tem por base a estatística quiquadrado e é calculado por:

$$0 < C < 1$$

TABELAS E TESTES QUI-QUADRADO

O coeficiente de contingência varia de 0 a 1.

Quando mais próximo de 1 mais forte é a associação entre as variáveis.

Quanto mais próximo de zero, maior a indicação que a associação é fraca ou nula.

O teste de Mantel-Hanszel e o coeficiente de correlação de Spearman são específicos para variáveis categorizadas ordinais.