Análise de Variância (ANOVA)

Prof. Carlos Trucíos
ctrucios@gmail.com
ctruciosm.github.io

Instituto de Matemática, Estatística e Computação Científica.
Universidade Estadual de Campinas.

16 de Fevereiro, 2022

1. Introdução
2. Intuição
3. Análise de variância (ANOVA)
4. R e Python

Introdução

- Até agora vimos testes de hipóteses para comparar duas médias populacionais.

- Até agora vimos testes de hipóteses para comparar duas médias populacionais.
- Em algumas circunstâncias, podemos estar interessados em testar se a média de três ou mais populações/grupos são iguais.

- Até agora vimos testes de hipóteses para comparar duas médias populacionais.
- Em algumas circunstâncias, podemos estar interessados em testar se a média de três ou mais populações/grupos são iguais.
- Nestes casos, utilizaremos um procedimento conhecido como **análise de variância** (ou **ANOVA**).

Um teste de habilidades é aplicado em uma amostra de 100 trabalhadores de cada um dos três centros de distribuição (CD) da *Via Varejo*. Roberto Fulcherberguer, o CEO da empresa, gostaria de saber se, em média, os três centros de distribuição possuem funcionários com o mesmo nível de habilidades.

Denotemos por μ_1, μ_2, μ_3 o nível médio de habilidades dos funcionários de cada um dos três CD.

Denotemos por μ_1, μ_2, μ_3 o nível médio de habilidades dos funcionários de cada um dos três CD. Então, o CEO quer testar:

$$H_0 : \mu_1 = \mu_2 = \mu_3 \quad vs. \quad H_1 : H_0 \text{ não é verdade}$$

Denotemos por μ_1, μ_2, μ_3 o nível médio de habilidades dos funcionários de cada um dos três CD. Então, o CEO quer testar:

$$H_0 : \mu_1 = \mu_2 = \mu_3 \quad vs. \quad H_1 : H_0 \text{ não é verdade}$$

- Nenhum dos testes vistos até agora são úteis.

Denotemos por μ_1, μ_2, μ_3 o nível médio de habilidades dos funcionários de cada um dos três CD. Então, o CEO quer testar:

$$H_0 : \mu_1 = \mu_2 = \mu_3 \quad vs. \quad H_1 : H_0 \text{ não é verdade}$$

- Nenhum dos testes vistos até agora são úteis.
- Utilizaremos ANOVA para testar esta hipótese

Suponha que as seguintes condições acontecem:

- **Independência:** entre as observações (a pontuação que cada trabalhador obteve é independente da obtida por outro funcionário).

Suponha que as seguintes condições acontecem:

- **Independência:** entre as observações (a pontuação que cada trabalhador obteve é independente da obtida por outro funcionário).
- **Normalidade:** a variável de interesse de cada grupo é normalmente distribuida.

Suponha que as seguintes condições acontecem:

- **Independência:** entre as observações (a pontuação que cada trabalhador obteve é independente da obtida por outro funcionário).
- **Normalidade:** a variável de interesse de cada grupo é normalmente distribuida.
- **Igualdade de variâncias:** os grupos tem a mesma variância (variâncias desconhecidas mas iguais).

Intuição

- Se as três médias populacionais fossem iguais, esperamos que as três médias amostrais estejam próximas entre si.

- Se as três médias populacionais fossem iguais, esperamos que as três médias amostrais estejam próximas entre si.
- Quanto mais próximas entre si estejam as médias amostrais, teremos maior evidência a favor de H_0.

- Se as três médias populacionais fossem iguais, esperamos que as três médias amostrais estejam próximas entre si.
- Quanto mais próximas entre si estejam as médias amostrais, teremos maior evidência a favor de H_0.
- Por outro lado, quando mais diferirem entre si as médias amostrais, teremos maior evidência para dizer que as médias populacionais não são iguais.

- Se as três médias populacionais fossem iguais, esperamos que as três médias amostrais estejam próximas entre si.
- Quanto mais próximas entre si estejam as médias amostrais, teremos maior evidência a favor de H_0.
- Por outro lado, quando mais diferirem entre si as médias amostrais, teremos maior evidência para dizer que as médias populacionais não são iguais.
- Em outras palavras, se a variabilidade entre as médias amostrais for pequena, teremos evidência favoravel a $H_0 : \mu_1 = \mu_2 = \mu_3$ (não rejeitar H_0), enquanto que, se a variabilidade entre as médias amostrais for grande, teremos evidência contraria a H_0 (rejeitar H_0).

Intuição

- Se as três médias populacionais fossem iguais, esperamos que as três médias amostrais estejam próximas entre si.
- Quanto mais próximas entre si estejam as médias amostrais, teremos maior evidência a favor de H_0.
- Por outro lado, quando mais diferirem entre si as médias amostrais, teremos maior evidência para dizer que as médias populacionais não são iguais.
- Em outras palavras, se a variabilidade entre as médias amostrais for pequena, teremos evidência favoravel a $H_0 : \mu_1 = \mu_2 = \mu_3$ (não rejeitar H_0), enquanto que, se a variabilidade entre as médias amostrais for grande, teremos evidência contraria a H_0 (rejeitar H_0).
- Se H_0 for verdade, isto implica que todas as amostras vem de uma mesma $N(\mu, \sigma^2)$

- Assim, sob H_0, podemos pensar em $\bar{X}_1$, $\bar{X}_2$ e $\bar{X}_3$ como uma a.a extraida de uma $N(\mu, \sigma^2_{\bar{X}})$.

- Assim, sob H_0, podemos pensar em $\bar{X}_1$, $\bar{X}_2$ e $\bar{X}_3$ como uma a.a extraida de uma $N(\mu, \sigma^2_{\bar{X}})$.
- Então podemos estimar μ por

$$\bar{X} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3}{3}$$

e podemos estimar $\sigma^2_{\bar{X}}$ por

$$\hat{\sigma}^2_{\bar{X}} = \frac{(\bar{X}_1 - \bar{X})^2 + (\bar{X}_2 - \bar{X})^2 + (\bar{X}_3 - \bar{X})^2}{3-1}.$$

- Assim, sob H_0, podemos pensar em $\bar{X}_1$, $\bar{X}_2$ e $\bar{X}_3$ como uma a.a extraida de uma $N(\mu, \sigma^2_{\bar{X}})$.
- Então podemos estimar μ por

$$\bar{X} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3}{3}$$

e podemos estimar $\sigma^2_{\bar{X}}$ por

$$\hat{\sigma}^2_{\bar{X}} = \frac{(\bar{X}_1 - \bar{X})^2 + (\bar{X}_2 - \bar{X})^2 + (\bar{X}_3 - \bar{X})^2}{3-1}.$$

- Assim, sob H_0, podemos pensar em $\bar{X}_1$, $\bar{X}_2$ e $\bar{X}_3$ como uma a.a extraida de uma $N(\mu, \sigma^2_{\bar{X}})$.
- Então podemos estimar μ por

$$\bar{X} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3}{3}$$

e podemos estimar $\sigma^2_{\bar{X}}$ por

$$\hat{\sigma}^2_{\bar{X}} = \frac{(\bar{X}_1 - \bar{X})^2 + (\bar{X}_2 - \bar{X})^2 + (\bar{X}_3 - \bar{X})^2}{3-1}.$$

Como $\sigma^2_{\bar{X}} = \sigma^2/n$, temos que $\hat{\sigma}^2 = n\hat{\sigma}^2_{\bar{X}}$.

- Então, $\hat{\sigma}^2$ é um estimador não viesado de σ^2

- Então, $\hat{\sigma}^2$ é um estimador não viesado de σ^2
- Por outro lado, $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$ e $\hat{\sigma}_3^2$ são também estimadores não viesados de σ^2

- Então, $\hat{\sigma}^2$ é um estimador não viesado de σ^2
- Por outro lado, $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$ e $\hat{\sigma}_3^2$ são também estimadores não viesados de σ^2
- Note que $\dfrac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2 + \hat{\sigma}_3^2}{3}$ também é um estimador não viesado de σ^2

- Então, $\hat{\sigma}^2$ é um estimador não viesado de σ^2
- Por outro lado, $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$ e $\hat{\sigma}_3^2$ são também estimadores não viesados de σ^2
- Note que $\dfrac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2 + \hat{\sigma}_3^2}{3}$ também é um estimador não viesado de σ^2
- Se H_0 for verdadeira tanto $\hat{\sigma}^2$ quanto $(\hat{\sigma}_1^2 + \hat{\sigma}_2^2 + \hat{\sigma}_3^2)/3$ serão ambas próximas entre si.

- Então, $\hat{\sigma}^2$ é um estimador não viesado de σ^2
- Por outro lado, $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$ e $\hat{\sigma}_3^2$ são também estimadores não viesados de σ^2
- Note que $\dfrac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2 + \hat{\sigma}_3^2}{3}$ também é um estimador não viesado de σ^2
- Se H_0 for verdadeira tanto $\hat{\sigma}^2$ quanto $(\hat{\sigma}_1^2 + \hat{\sigma}_2^2 + \hat{\sigma}_3^2)/3$ serão ambas próximas entre si.
- Isso significa que o quociente entre eles deve ser próximo de um.

ANOVA baseia-se em analisar essas duas quantidades e através dessa análise seremos capazes de testar as hipóteses.

ANOVA baseia-se em analisar essas duas quantidades e através dessa análise seremos capazes de testar as hipóteses.

- $\hat{\sigma}^2$ baseia-se na variabilidade existente entre as próprias médias amostrais (chamada **variância entre tratamentos**).

ANOVA baseia-se em analisar essas duas quantidades e através dessa análise seremos capazes de testar as hipóteses.

- $\hat{\sigma}^2$ baseia-se na variabilidade existente entre as próprias médias amostrais (chamada **variância entre tratamentos**).
- A outra, baseia-se na variabilidade dos dados existente dentro de cada grupo (chamada **variância dentro dos tratamentos**).

Análise de variância (ANOVA)

Sejam as hipóteses:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k. \quad vs. \quad H_1 : H_0 \text{ não é verdade},$$

em que μ_i $(i = 1, \ldots, k)$ é a média da i-ésima população. $(k \geq 3)$

Sejam as hipóteses:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k. \quad vs. \quad H_1 : H_0 \text{ não é verdade},$$

em que μ_i $(i = 1, \ldots, k)$ é a média da i-ésima população. $(k \geq 3)$

Sejam:

- n_i: tamanho da a.a. extraida da i-ésima população;
- X_{ij}: j-ésimo elemento da a.a extraida da i-ésima população;
- $\bar{X}_{i.}$: média amostral da i-ésima população;
- $\hat{\sigma}_i^2$: variância amostral da i-ésima população.

Sejam as hipóteses:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k. \quad vs. \quad H_1 : H_0 \text{ não é verdade},$$

em que μ_i $(i = 1, \ldots, k)$ é a média da i-ésima população. $(k \geq 3)$

Sejam:

- n_i: tamanho da a.a. extraida da i-ésima população;
- X_{ij}: j-ésimo elemento da a.a extraida da i-ésima população;
- $\bar{X}_{i\cdot}$: média amostral da i-ésima população;
- $\hat{\sigma}_i^2$: variância amostral da i-ésima população.

Por outro lado, denotemos por $\bar{\bar{X}}$ a média global de todas as observações,

$$\bar{\bar{X}} = \frac{\sum_{i=1}^{k} \sum_{i=1}^{n_i} X_{ij}}{n_T},$$

em que $n_T = n_1 + n_2 + \cdots + n_k$

Vamos supor que as seguintes condições acontecam:

1. Independendência entre observações
2. Normalidade
3. Igualdade de variâncias.

Vamos supor que as seguintes condições acontecam:

1. Independência entre observações
2. Normalidade
3. Igualdade de variâncias.

Sob essas suposições, vamos a obter a estatística de teste.

A soma de quadrados total (SQT) é dada por:

$$SQT = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{\bar{X}})^2$$

A soma de quadrados total (SQT) é dada por:

$$SQT = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{\bar{X}})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.} + \bar{X}_{i.} - \bar{\bar{X}})^2$$

A soma de quadrados total (SQT) é dada por:

$$SQT = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{\bar{X}})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\cdot} + \bar{X}_{i\cdot} - \bar{\bar{X}})^2$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[(X_{ij} - \bar{X}_{i\cdot})^2 + (\bar{X}_{i\cdot} - \bar{\bar{X}})^2 + 2(X_{ij} - \bar{X}_{i\cdot})(\bar{X}_{i\cdot} - \bar{\bar{X}})\right]$$

A soma de quadrados total (SQT) é dada por:

$$SQT = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{\bar{X}})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i.} + \bar{X}_{i.} - \bar{\bar{X}})^2$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[(X_{ij} - \bar{X}_{i.})^2 + (\bar{X}_{i.} - \bar{\bar{X}})^2 + 2(X_{ij} - \bar{X}_{i.})(\bar{X}_{i.} - \bar{\bar{X}})\right]$$

$$= \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i.})^2}_{I} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{X}_{i.} - \bar{\bar{X}})^2}_{II} + 0$$

Vocês encontrarão em alguns livros texto que **I** e **II** são chamados de SQE (Soma de Quadrados dos Erros) e SQTr (Soma de Quadrados dos Tratamentos), respectivamente. Assim,

$$SQT = SQE + SQTr$$

Note que:

$$\frac{\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i.})^2}{\sigma^2}$$

Note que:

$$\frac{\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i.})^2}{\sigma^2} = \frac{(n_i-1)\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i.})^2}{\sigma^2(n_i-1)}$$

Note que:

$$\frac{\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i.})^2}{\sigma^2} = \frac{(n_i-1)\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i.})^2}{\sigma^2(n_i-1)} = \frac{(n_i-1)\hat{\sigma}_i^2}{\sigma^2}$$

Note que:

$$\frac{\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i.})^2}{\sigma^2} = \frac{(n_i-1)\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i.})^2}{\sigma^2(n_i-1)} = \frac{(n_i-1)\hat{\sigma}_i^2}{\sigma^2} \sim \chi^2_{n_i-1}$$

Note que:

$$\frac{\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i.})^2}{\sigma^2} = \frac{(n_i-1)\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i.})^2}{\sigma^2(n_i-1)} = \frac{(n_i-1)\hat{\sigma}_i^2}{\sigma^2} \sim \chi^2_{n_i-1}$$

Então,

$$\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i.})^2}{\sigma^2}$$

Note que:

$$\frac{\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i.})^2}{\sigma^2} = \frac{(n_i-1)\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i.})^2}{\sigma^2(n_i-1)} = \frac{(n_i-1)\hat{\sigma}_i^2}{\sigma^2} \sim \chi^2_{n_i-1}$$

Então,

$$\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i.})^2}{\sigma^2} = \frac{SQE}{\sigma^2}$$

Note que:

$$\frac{\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i.})^2}{\sigma^2} = \frac{(n_i - 1)\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i.})^2}{\sigma^2 (n_i - 1)} = \frac{(n_i - 1)\hat{\sigma}_i^2}{\sigma^2} \sim \chi^2_{n_i - 1}$$

Então,

$$\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i.})^2}{\sigma^2} = \frac{SQE}{\sigma^2} \sim \chi^2_{\underbrace{n_1 + n_2 + \cdots + n_k}_{n_T} - \underbrace{(1 + 1 + \cdots + 1)}_{k}}$$

Por outro lado, sob H_0:

$$\bar{X}_{i\cdot} \sim N\left(\mu, \frac{\sigma^2}{n_i}\right) \quad e \quad \bar{\bar{X}} \sim N\left(\mu, \frac{\sigma^2}{n_T}\right),$$

Por outro lado, sob H_0:

$$\bar{X}_{i\cdot} \sim N\Big(\mu, \frac{\sigma^2}{n_i}\Big) \quad e \quad \bar{\bar{X}} \sim N\Big(\mu, \frac{\sigma^2}{n_T}\Big),$$

$$\frac{\sqrt{n_i}(\bar{X}_{i\cdot} - \mu)}{\sigma} \sim N(0,1) \quad e \quad \frac{\sqrt{n_T}(\bar{\bar{X}} - \mu)}{\sigma} \sim N(0,1),$$

Por outro lado, sob H_0:

$$\bar{X}_{i\cdot} \sim N\Big(\mu, \frac{\sigma^2}{n_i}\Big) \quad e \quad \bar{\bar{X}} \sim N\Big(\mu, \frac{\sigma^2}{n_T}\Big),$$

$$\frac{\sqrt{n_i}(\bar{X}_{i\cdot} - \mu)}{\sigma} \sim N(0,1) \quad e \quad \frac{\sqrt{n_T}(\bar{\bar{X}} - \mu)}{\sigma} \sim N(0,1),$$

$$\frac{n_i(\bar{X}_{i\cdot} - \mu)^2}{\sigma^2} \sim \chi^2_1 \quad , \quad \frac{n_T(\bar{\bar{X}} - \mu)^2}{\sigma^2} \sim \chi^2_1$$

Por outro lado, sob H_0:

$$\bar{X}_{i\cdot} \sim N\Big(\mu, \frac{\sigma^2}{n_i}\Big) \quad e \quad \bar{\bar{X}} \sim N\Big(\mu, \frac{\sigma^2}{n_T}\Big),$$

$$\frac{\sqrt{n_i}(\bar{X}_{i\cdot} - \mu)}{\sigma} \sim N(0,1) \quad e \quad \frac{\sqrt{n_T}(\bar{\bar{X}} - \mu)}{\sigma} \sim N(0,1),$$

$$\frac{n_i(\bar{X}_{i\cdot} - \mu)^2}{\sigma^2} \sim \chi_1^2 \quad , \quad \frac{n_T(\bar{\bar{X}} - \mu)^2}{\sigma^2} \sim \chi_1^2 \quad e \quad \sum_{i=1}^{k} \frac{n_i(\bar{X}_{i\cdot} - \mu)^2}{\sigma^2} \sim \chi_k^2$$

Assim,

$$\sum_{i=1}^{k} \frac{n_i(\bar{X}_{i\cdot} - \mu)^2}{\sigma^2} =$$

Assim,

$$\sum_{i=1}^{k} \frac{n_i(\bar{X}_{i\cdot} - \mu)^2}{\sigma^2} = \sum_{i=1}^{k} \frac{n_i(\bar{X}_{i\cdot} - \bar{\bar{X}} + \bar{\bar{X}} - \mu)^2}{\sigma^2} =$$

Assim,

$$\sum_{i=1}^{k} \frac{n_i(\bar{X}_{i\cdot} - \mu)^2}{\sigma^2} = \sum_{i=1}^{k} \frac{n_i(\bar{X}_{i\cdot} - \bar{\bar{X}} + \bar{\bar{X}} - \mu)^2}{\sigma^2} = \sum_{i=1}^{k} \frac{n_i(\bar{X}_{i\cdot} - \bar{\bar{X}})^2}{\sigma^2} + \frac{n_T(\bar{\bar{X}} - \mu)^2}{\sigma^2}$$

Assim,

$$\sum_{i=1}^{k}\frac{n_i(\bar{X}_{i\cdot}-\mu)^2}{\sigma^2}=\sum_{i=1}^{k}\frac{n_i(\bar{X}_{i\cdot}-\bar{\bar{X}}+\bar{\bar{X}}-\mu)^2}{\sigma^2}=\sum_{i=1}^{k}\frac{n_i(\bar{X}_{i\cdot}-\bar{\bar{X}})^2}{\sigma^2}+\frac{n_T(\bar{\bar{X}}-\mu)^2}{\sigma^2}$$

$$\underbrace{\sum_{i=1}^{k}\frac{n_i(\bar{X}_{i\cdot}-\mu)^2}{\sigma^2}}_{\chi^2_k}=\sum_{i=1}^{k}\frac{n_i(\bar{X}_{i\cdot}-\bar{\bar{X}})^2}{\sigma^2}+\underbrace{\frac{n_T(\bar{\bar{X}}-\mu)^2}{\sigma^2}}_{\chi^2_1}$$

Assim,

$$\sum_{i=1}^{k} \frac{n_i(\bar{X}_{i\cdot} - \mu)^2}{\sigma^2} = \sum_{i=1}^{k} \frac{n_i(\bar{X}_{i\cdot} - \bar{\bar{X}} + \bar{\bar{X}} - \mu)^2}{\sigma^2} = \sum_{i=1}^{k} \frac{n_i(\bar{X}_{i\cdot} - \bar{\bar{X}})^2}{\sigma^2} + \frac{n_T(\bar{\bar{X}} - \mu)^2}{\sigma^2}$$

$$\underbrace{\sum_{i=1}^{k} \frac{n_i(\bar{X}_{i\cdot} - \mu)^2}{\sigma^2}}_{\chi^2_k} = \sum_{i=1}^{k} \frac{n_i(\bar{X}_{i\cdot} - \bar{\bar{X}})^2}{\sigma^2} + \underbrace{\frac{n_T(\bar{\bar{X}} - \mu)^2}{\sigma^2}}_{\chi^2_1}$$

Logo, $\displaystyle\sum_{i=1}^{k} \frac{n_i(\bar{X}_{i\cdot} - \bar{\bar{X}})^2}{\sigma^2} =$

Assim,

$$\sum_{i=1}^{k} \frac{n_i(\bar{X}_{i\cdot} - \mu)^2}{\sigma^2} = \sum_{i=1}^{k} \frac{n_i(\bar{X}_{i\cdot} - \bar{\bar{X}} + \bar{\bar{X}} - \mu)^2}{\sigma^2} = \sum_{i=1}^{k} \frac{n_i(\bar{X}_{i\cdot} - \bar{\bar{X}})^2}{\sigma^2} + \frac{n_T(\bar{\bar{X}} - \mu)^2}{\sigma^2}$$

$$\underbrace{\sum_{i=1}^{k} \frac{n_i(\bar{X}_{i\cdot} - \mu)^2}{\sigma^2}}_{\chi^2_k} = \sum_{i=1}^{k} \frac{n_i(\bar{X}_{i\cdot} - \bar{\bar{X}})^2}{\sigma^2} + \underbrace{\frac{n_T(\bar{\bar{X}} - \mu)^2}{\sigma^2}}_{\chi^2_1}$$

Logo, $\sum_{i=1}^{k} \frac{n_i(\bar{X}_{i\cdot} - \bar{\bar{X}})^2}{\sigma^2} = \frac{SQTr}{\sigma^2} \sim \chi^2_{k-1}$

Lembre-se

Sejam X e Y duas v.as independentes t.q $X \sim \chi^2_m$ e $Y \sim \chi^2_n$. Então

$$\frac{X/m}{Y/n} \sim F_{m,n}$$

Nós temos que $\dfrac{SQTr}{\sigma^2} \sim \chi^2_{k-1}$ e $\dfrac{SQE}{\sigma^2} \sim \chi^2_{n_T-k}$. Assim, se elas forem independentes teriamos que

$$\frac{\dfrac{SQTr}{\sigma^2(k-1)}}{\dfrac{SQE}{\sigma^2(n_T-k)}} \sim F_{k-1,n_T-k}$$

Lembre-se

Sejam X e Y duas v.as independentes t.q $X \sim \chi^2_m$ e $Y \sim \chi^2_n$. Então

$$\frac{X/m}{Y/n} \sim F_{m,n}$$

Nós temos que $\frac{SQTr}{\sigma^2} \sim \chi^2_{k-1}$ e $\frac{SQE}{\sigma^2} \sim \chi^2_{n_T-k}$. Assim, se elas forem independentes teriamos que

$$\frac{\frac{SQTr}{\sigma^2(k-1)}}{\frac{SQE}{\sigma^2(n_T-k)}} \sim F_{k-1,n_T-k}$$

(Invocando o teorema de Cochran, pode-se mostrar que são independentes).

Teorema de Cochran

Sejam $Z_1, \cdots, Z_v$ v.as iid $\sim N(0,1)$ e seja

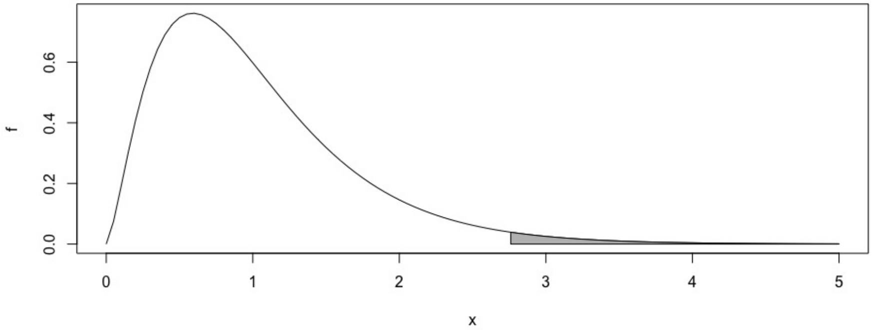
$$\sum_{i=1}^{v} Z_i^2 = Q_1 + Q_2 + \cdots + Q_s,$$

com $s \leq v$ e Q_i com v_i graus de liberdade. Então $Q_1, \cdots, Q_s$ são v.as χ^2 independentes com $v_1, \cdots, v_s$ graus de liberdades, respectivamente, se e somente se, $v = v_1 + \cdots + v_s$

Fonte de Variação	g.l	Soma dos Q.	Q. Médios	F
Tratamento	$k-1$	SQTr	$QMTr = \frac{SQTr}{k-1}$	$\frac{QMTr}{QME}$
Erro	$n_T - k$	SQE	$QME = \frac{SQE}{n_T - k}$	
Total	$n_T - 1$	SQT		

Fonte de Variação	g.l	Soma dos Q.	Q. Médios	F
Tratamento	$k-1$	SQTr	$QMTr = \frac{SQTr}{k-1}$	$\frac{QMTr}{QME}$
Erro	$n_T - k$	SQE	$QME = \frac{SQE}{n_T - k}$	
Total	$n_T - 1$	SQT		

Rejeitamos H_0 se $F = \frac{QMTr}{QME} > F_{1-\alpha,k-1,n_T-k}$


0.0
0.2
0.4
0.6
f
0
1
2
3
4
5
x

Note que se $n_i = n \quad \forall i$:

- $SQTr = \sum_{i=1}^{k} n_i(\bar{X}_{i\cdot} - \bar{\bar{X}})^2 = n\sum_{i=1}^{k}(\bar{X}_{i\cdot} - \bar{\bar{X}})^2$, e então

  $QMTr = \dfrac{SQTr}{k-1} = \dfrac{n\sum_{i=1}^{k}(\bar{X}_{i\cdot} - \bar{\bar{X}})^2}{k-1}$ é o $\hat{\sigma}^2$ obtido no *caso de estudo* (ver slide 10).

- De forma semelhante, $QME = \dfrac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\cdot})^2}{n_T - k}$, mas como $n_i = n$ e $n_T = n_1 + \cdots + n_k = kn$ temos que,

$$QME = \frac{\sum_{i=1}^{k}\overbrace{\sum_{j=1}^{n}(X_{ij} - \bar{X}_{i\cdot})^2}^{(n-1)\hat{\sigma}_i^2}}{k(n-1)} = \frac{\sum_{i}^{k}\hat{\sigma}_i^2}{k}.$$

Voltando ao *caso de estudo*. O CEO da *Via Varejo* quer testar:

$$H_0 : \mu_1 = \mu_2 = \mu_3 \quad vs. \quad H_1 : H_0 \text{ não é verdade}$$

Voltando ao *caso de estudo*. O CEO da *Via Varejo* quer testar:

$$H_0 : \mu_1 = \mu_2 = \mu_3 \quad vs. \quad H_1 : H_0 \text{ não é verdade}$$

Sejam os valores:

- **Amostra do CD1**: $\bar{x}_1 = 79$, $\hat{\sigma}_1 = 5.83$, $n_1 = 100$
- **Amostra do CD2**: $\bar{x}_2 = 74$, $\hat{\sigma}_2 = 4.47$, $n_2 = 100$
- **AMostra do CD3**: $\bar{x}_3 = 66$, $\hat{\sigma}_3 = 5.66$, $n_3 = 100$

Voltando ao *caso de estudo*. O CEO da *Via Varejo* quer testar:

$$H_0 : \mu_1 = \mu_2 = \mu_3 \quad vs. \quad H_1 : H_0 \text{ não é verdade}$$

Sejam os valores:

- **Amostra do CD1**: $\bar{x}_1 = 79$, $\hat{\sigma}_1 = 5.83$, $n_1 = 100$
- **Amostra do CD2**: $\bar{x}_2 = 74$, $\hat{\sigma}_2 = 4.47$, $n_2 = 100$
- **AMostra do CD3**: $\bar{x}_3 = 66$, $\hat{\sigma}_3 = 5.66$, $n_3 = 100$

Fonte de Variação	g.l	Soma dos Q.	Q. Médios	F
Tratamento	$k-1$	SQTr	$QMTr = \frac{SQTr}{k-1}$	$\frac{QMTr}{QME}$
Erro	$n_T - k$	SQE	$QME = \frac{SQE}{n_T - k}$	
Total	$n_T - 1$	SQT		

- **Amostra do CD1**: $\bar{x}_1 = 79$, $\hat{\sigma}_1 = 5.83$, $n_1 = 100$
- **Amostra do CD2**: $\bar{x}_2 = 74$, $\hat{\sigma}_2 = 4.47$, $n_2 = 100$
- **AMostra do CD3**: $\bar{x}_3 = 66$, $\hat{\sigma}_3 = 5.66$, $n_3 = 100$

- **Amostra do CD1**: $\bar{x}_1 = 79$, $\hat{\sigma}_1 = 5.83$, $n_1 = 100$
- **Amostra do CD2**: $\bar{x}_2 = 74$, $\hat{\sigma}_2 = 4.47$, $n_2 = 100$
- **AMostra do CD3**: $\bar{x}_3 = 66$, $\hat{\sigma}_3 = 5.66$, $n_3 = 100$

Como $k = 3$ grupos e $n_i = n \quad i = 1, \ldots, 3$:

- $\bar{\bar{x}} = \dfrac{79 + 74 + 66}{3} = 73$

- **Amostra do CD1**: $\bar{x}_1 = 79$, $\hat{\sigma}_1 = 5.83$, $n_1 = 100$
- **Amostra do CD2**: $\bar{x}_2 = 74$, $\hat{\sigma}_2 = 4.47$, $n_2 = 100$
- **AMostra do CD3**: $\bar{x}_3 = 66$, $\hat{\sigma}_3 = 5.66$, $n_3 = 100$

Como $k = 3$ grupos e $n_i = n \quad i = 1, \dots, 3$:

- $\bar{\bar{x}} = \dfrac{79 + 74 + 66}{3} = 73$
- $QMTr = n\dfrac{\sum_{i=1}^{3}(\bar{x}_{i\cdot} - \bar{\bar{x}})^2}{3 - 1} =$
  $100 \times \dfrac{(79 - 73)^2 + (74 - 73)^2 + (66 - 73)^2}{2} = 4300$

- **Amostra do CD1**: $\bar{x}_1 = 79$, $\hat{\sigma}_1 = 5.83$, $n_1 = 100$
- **Amostra do CD2**: $\bar{x}_2 = 74$, $\hat{\sigma}_2 = 4.47$, $n_2 = 100$
- **AMostra do CD3**: $\bar{x}_3 = 66$, $\hat{\sigma}_3 = 5.66$, $n_3 = 100$

Como $k = 3$ grupos e $n_i = n \quad i = 1, \ldots, 3$:

- $\bar{\bar{x}} = \dfrac{79 + 74 + 66}{3} = 73$
- $QMTr = n\dfrac{\sum_{i=1}^{3}(\bar{x}_{i.} - \bar{\bar{x}})^2}{3-1} =$
  $100 \times \dfrac{(79-73)^2 + (74-73)^2 + (66-73)^2}{2} = 4300$
- $QME = \dfrac{\sum_{i}^{3} \hat{\sigma}_j^2}{3} = \dfrac{5.83^2 + 4.47^2 + 5.66^2}{3} = 28.66847$

- $F = QMTr/QME = 4300/28.66847 = 149.9906$

- $F = QMTr/QME = 4300/28.66847 = 149.9906$
- F $= 149.9906$

- $F = QMTr/QME = 4300/28.66847 = 149.9906$
- F $= 149.9906$

- $F = QMTr/QME = 4300/28.66847 = 149.9906$
- F $= 149.9906$

```
k = 3; n = 100; alpha = 0.05
qf(1-alpha, k-1, 3*n-k)
```

```
## [1] 3.026153
```

- $F = QMTr/QME = 4300/28.66847 = 149.9906$
- F $= 149.9906$

```
k = 3; n = 100; alpha = 0.05
qf(1-alpha, k-1, 3*n-k)
```

```
## [1] 3.026153
```

$149.9906 > 3.026153$**?**

- $F = QMTr/QME = 4300/28.66847 = 149.9906$
- F = 149.9906

```
k = 3; n = 100; alpha = 0.05
qf(1-alpha, k-1, 3*n-k)
```

```
## [1] 3.026153
```

149.9906 $>$ 3.026153**?** Sim, então rejeitamos H_0.

Exemplo: Quatro observações foram selecionadas aleatoriamente de três populações diferentes. Os dados obtidos são os seguintes:

Observação	Grupo 1	Grupo 2	Grupo 3
1	165	174	169
2	149	164	154
3	156	180	161
4	142	158	148

Teste

$$H_0 : \mu_1 = \mu_2 = \mu_3 \quad vs. \quad H_1 : H_0 \text{ não é verdadeira}$$

Suponha que todas as suposições necessárias são verdade.

```r
x1 = c(165, 149, 156, 142)  # amostra G1
x2 = c(174, 164, 180, 158)  # amostra G2
x3 = c(169, 154, 161, 148)  # amostra G3
x = c(x1,x2,x3)             # todos os elementos
# Calculamos as médias:
m_g = mean(x)               # média global
m_1 = mean(x1)              # média da amostra G1
m_2 = mean(x2)              # média da amostra G2
m_3 = mean(x3)              # média da amostra G3
# Tamanhos de amostra em cada grupo
n1 = length(x1)             # Obs na amostra G1
n2 = length(x2)             # Obs na amostra G2
n3 = length(x3)             # Obs na amostra G3
```

$$SQT = SQTr + SQE$$

$$SQT = \sum_{j=1}^{k} \sum_{i}^{n_j} (x_{ij} - \bar{\bar{x}})^2 \quad SQTr = \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{\bar{x}})^2$$

$$SQT = SQTr + SQE$$

$$SQT = \sum_{j=1}^{k}\sum_{i}^{n_j}(x_{ij} - \bar{\bar{x}})^2 \quad SQTr = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{\bar{x}})^2$$

```
# Soma de Quadrados Totais
SQT = sum((x-m_g)^2)     # Cuidado! sum((x-m_g)^2) != sum(x-mg)^2
# Soma de Quadrados dos Tratamentos
SQTr = n1*(m_1-m_g)^2 + n2*(m_2-m_g)^2 + n3*(m_3-m_g)^2
# Soma de Quadrados dos Erros
SQE = SQT -  SQTr
# Imprimindo resultados
c(SQT, SQTr, SQE)
```

```
## [1] 1364  536  828
```

Fonte de Variação	g.l	Soma dos Q.	Q. Médios	F
Tratamento	$k-1$	SQTr	$QMTr = \frac{SQTr}{k-1}$	$\frac{QMTr}{QME}$
Erro	$n_T - k$	SQE	$QME = \frac{SQE}{n_T - k}$	
Total	$n_T - 1$	SQT		

Fonte de Variação	g.l	Soma dos Q.	Q. Médios	F
Tratamento	$k-1$	SQTr	$QMTr = \frac{SQTr}{k-1}$	$\frac{QMTr}{QME}$
Erro	$n_T - k$	SQE	$QME = \frac{SQE}{n_T - k}$	
Total	$n_T - 1$	SQT		

Fonte de Variação	g.l	Soma dos Q.	Q. Médios	F
Tratamento	2	536	$QMTr = 268$	2.9130435
Erro	9	828	$QME = 92$	
Total	11	1364		

Então rejeitamos $H_0 : \mu_1 = \mu_2 = \mu_3$ se $F = 2.9130435 > F_{1-\alpha,k-1,n_T-k}$

Então rejeitamos $H_0 : \mu_1 = \mu_2 = \mu_3$ se $F = 2.9130435 > F_{1-\alpha,k-1,n_T-k}$

```
alpha = 0.05; k = 3; nT = n1 + n2 + n3
qf(1-alpha, k-1, nT-k)
```

```
## [1] 4.256495
```

Então rejeitamos $H_0 : \mu_1 = \mu_2 = \mu_3$ se $F = 2.9130435 > F_{1-\alpha,k-1,n_T-k}$

```
alpha = 0.05; k = 3; nT = n1 + n2 + n3
qf(1-alpha, k-1, nT-k)
```

```
## [1] 4.256495
```

$2.9130435 > 4.256495$ **?** Não, então não rejeitamos H_0.

R e Python

Os dados do exemplo anterior estão disponíveis aqui.

Os dados do exemplo anterior estão disponíveis aqui.

```
dados <- read.csv("anova_dados.csv", sep = ";")
oneway.test(V1 ~ grupo, data = dados, var.equal = TRUE)
```

```
##
##  One-way analysis of means
##
## data:  V1 and grupo
## F = 2.913, num df = 2, denom df = 9, p-value = 0.1058
```

```
import pandas as pd
import scipy.stats as stats

df = pd.read_csv("anova_dados.csv", sep = ";")

stats.f_oneway(df['V1'][df['grupo'] == 'Grupo 1'],
               df['V1'][df['grupo'] == 'Grupo 2'],
               df['V1'][df['grupo'] == 'Grupo 3'])
```

```
## F_onewayResult(statistic=2.9130434782608696, pvalue=0.1057966
```

- Utilizamos ANOVA para testar ($k \geq 3$):

$$H_0 : \mu_1 = \cdots = \mu_k \quad vs. \quad H_1 : H_0 \text{ não é verdade}$$

- Utilizamos ANOVA para testar ($k \geq 3$):

$$H_0 : \mu_1 = \cdots = \mu_k \quad vs. \quad H_1 : H_0 \text{ não é verdade}$$

- As suposições do ANOVA são: normalidade, independencia das observações e variância constante.

- Utilizamos ANOVA para testar ($k \geq 3$):

$$H_0 : \mu_1 = \cdots = \mu_k \quad vs. \quad H_1 : H_0 \text{ não é verdade}$$

- As suposições do ANOVA são: normalidade, independencia das observações e variância constante.
- Tanto **R** quanto **Python** possuem funções para realizar ANOVA.

- Anderson, D. R; Sweeney, D. J.; e Williams, T. A. (2008). *Estatística Aplicada à Administração e Economia*. 2ed. Cengage Learning. **Cap 10**
- Devore, J. L. (2018). *Probabilidade e Estatística para Engenharia e Ciências*. 9ed, Cengage. **Cap 10**
- Morettin, P. A; e Bussab, W. de O. (2004). *Estatística Básica*. 5ed, Saraiva. **Cap 13**