

# **R para Ciência de Dados**

## **Modelos de Análise de Variância (Anova)**

Profa Carolina Paraíba e Prof Gilberto Sassi

Departamento de Estatística  
Instituto de Matemática e Estatística  
Universidade Federal da Bahia

Julho de 2025

# Curso R para Ciência de Dados: Modelos de Análise de Variância (Anova)

## Aula 1: Modelos de Análise de Variância (Anova)

- Estudos Experimentais
- Algumas Definições
- Princípios básicos de Planejamento de Experimentos
- Modelos de Análise de Variância (Anova)
- Modelo Anova um-fator
- Suposições do modelo
- Comparações múltiplas

# Estudos Experimentais

Em um estudo experimental, o ambiente é controlado e algumas variáveis são propositalmente alteradas enquanto outras são mantidas constantes.

Assim, estudos experimentais são conduzidos para verificar relações de causa e efeito entre uma (ou mais) variável explicativa e uma variável resposta.

De modo geral, a relação de causa e efeito pode ser verificada alterando-se os níveis da variável explicativa e observando-se o efeito das mudanças na variável resposta.

# Planejamento de Experimentos

Em seus experimentos agrícolas e biológicos, R. A. Fisher enfrentava uma grande quantidade de variação que muitas vezes confundiam os resultados. Além disso, os experimentos eram demorados e caros de serem realizados. Isso o motivou a encontrar técnicas experimentais capazes de:

- eliminar o máximo possível da variação natural;
- evitar que variações não removidas confundam ou enviesem os efeitos que estão sendo testados;
- detectar causa e efeito com o mínimo de esforço experimental necessário.

## Algumas Definições

- **Fator:** um fator é uma variável explicativa que é estudada no experimento.
- **Níveis de um fator:** para estudar o efeito de um fator na resposta, dois ou mais valores do fator são usados. Esses valores são chamados de níveis.
- **Tratamento (variável independente):** em experimentos com um único fator, os tratamentos são os próprios níveis do fator. Em experimentos com dois ou mais fatores, um tratamento é uma combinação de níveis dos fatores.

## Algumas Definições

- **Unidade experimental:** é a unidade (indivíduo, planta, animal, etc) que é aleatoriamente atribuída a um tratamento.
  - Trata-se da unidade básica do experimento, e define a unidade a ser replicada para aumentar os graus de liberdade.
- **Variável resposta (variável dependente):** denotada por  $Y$ , é a característica da unidade experimental que é medida após (executado) o experimento. A magnitude da resposta depende das configurações das variáveis ou fatores independentes e de variáveis ocultas.

## Algumas Definições

- **Efeito:** é a mudança na resposta causada por uma mudança em um fator ou tratamento.
  - Após a execução de um planejamento de experimento, o efeito pode ser estimado a partir da resposta observada. Esta estimativa é o *efeito calculado*. Antes de realizar o experimento, a pesquisadora pode saber quão grande o efeito deve ser para ter uma importância prática. Isso é chamado de *efeito prático* ou de *tamanho de um efeito prático*.

## Algumas Definições

- **Replicação:** é a repetição da situação experimental por meio da replicação da unidade experimental.
  - Replicações são dois ou mais experimentos conduzidos com as mesmas configurações dos fatores ou tratamentos, mas usando diferentes unidades experimentais. A variável resposta medida pode diferir entre replicações devido à mudanças em variáveis ocultas e à diferenças inerentes nas unidades experimentais.
  - Simplificando: a replicação é o número de vezes que um tratamento aparece em um experimento.

## Algumas Definições

- **Erro experimental:** em estudos experimentais, ocorre que observamos uma variação na variável resposta, em um mesmo tratamento, que não pode ser explicada pelo experimento e pelos fatores em estudo. Essa variação inexplicável na resposta é chamada de erro experimental.

*Observação:* o fato de ser chamado de erro não deve nos levar a supor que seja um *erro* ou *gafe*. Os erros experimentais não são todos iguais a zero porque as variáveis ocultas fazem com que mudem a cada execução do experimento.

## Algumas Definições

De maneira geral, erros experimentais podem ser classificados entre dois tipos:

- **Erro aleatório:** é a variação causada por fatores incontrolláveis (ruído) ou variação aleatória na variável resposta. O erro aleatório muda de um experimento para outro de maneira imprevisível e com média zero. A variação de erros experimentais aleatórios pode ser obtida incluindo replicadas em um delineamento experimental.
- **Erro sistemático:** é uma variação que não está relacionada à variação natural na variável resposta. O exemplo clássico de natureza sistemática é o viés de medição ou erro de medição.

# Princípios básicos de Planejamento de Experimentos

- **Aleatorização (casualização):**

Aleatorização é essencial para a validade dos métodos estatísticos em planejamento de experimentos.

Aleatorização é o processo de atribuir os tratamentos às unidades experimentais de forma que todos tenham a mesma chance de ocupar cada unidade experimental: cada unidade experimental terá a mesma chance de receber qualquer tratamento.

A aleatorização garante a eliminação do erro sistemático. Além disso, garante que qualquer componente de erro que ainda persista nas observações seja de natureza puramente aleatória. A aleatorização também garante (geralmente) a suposição de que os erros sejam independentes.

# Princípios básicos de Planejamento de Experimentos

- **Replicação:**

A replicação é o ato de realizar um experimento novamente, desde o início ao fim, não simplesmente tomando uma nova medição (duplicata) da variável resposta na mesma unidade experimental.

Cada repetição de um experimento é chamada de replicação.

O número de replicações será o número de vezes que o tratamento aparece no experimento.

O processo de replicação permite estimar o erro experimental e aumenta a precisão do experimento (e o poder dos testes).

As replicações permitem confirmar a resposta das unidades experimentais aos tratamentos.

# Princípios básicos de Planejamento de Experimentos

- **Blocos (controle local, blocking):**

Necessário na presença de fatores, controláveis ou incontroláveis e sem interesse prático, que podem afetar os resultados do experimento.

O controle local é útil na presença de fatores que fazem com que as unidades experimentais sejam heterogêneas.

Basicamente, a estratégia é agrupar as unidades homogêneas em blocos, dispondo as unidades experimentais heterogêneas em sub-unidades homogêneas, com a finalidade de diminuir o erro experimental.

# Princípios básicos de Planejamento de Experimentos

Aleatorização, replicações e blocos são três ferramentas que podem ser usadas para reduzir ou eliminar o erro aleatório associado a fatores incontroláveis.

## **Simplificando:**

- Aleatorização: os tratamentos são atribuídos ao acaso às unidades experimentais. (*validade dos teste estatísticos*)
- Replicações: unidades sujeitas às mesmas condições experimentais. (*possibilita testes estatísticos*)
- Bloco: técnica que aumenta a precisão do experimento quando uma porção do total avaliado é mais homogênea do que todo o grupo experimental. (*eficiência*)

# Modelos de Análise de Variância (Anova)

É comum termos interesse em comparar mais de um par de médias simultaneamente. Podemos, também, querer comparar médias de grupos quando um ou mais fatores são controlados.

A Anova consiste em particionar a variação total dos dados em partes atribuíveis à causas controladas (tratamentos) e à causas não controladas (variáveis ocultas, fatores desconhecidos), o erro.

A hipótese nula em Anova é que as médias de todos os tratamentos são iguais. A hipótese alternativa é que pelo menos uma das médias é diferente.

# Modelos de Análise de Variância (Anova)

**Exemplo 1.** Suponha que plantas sejam cultivadas em vasos e três tipos diferentes de fertilizantes sejam aplicados nos vasos. Após um período de tempo, o peso seco das plantas é registrado. Os dados estão na Tabela 1.

**Tabela 1:** Peso seco, em gramas, do Gerânio “Dilys”, submetido a três fertilizantes.

Fertilizante		
A	B	C
1.02	1.00	0.99
0.79	1.21	1.36
1.00	1.22	1.17
0.59	0.96	1.22
0.97	0.79	1.12

# Modelos de Análise de Variância (Anova)

```
library(readxl)
library(tidyverse)

dados <- read_xlsx("../dados/dados_aula1.xlsx",
                  sheet = "geranio")

glimpse(dados)

## Rows: 15
## Columns: 2
## $ peso          <dbl> 1.02, 0.79, 1.00, 0.59, 0.9
## $ fertilizante <chr> "a", "a", "a", "a", "a", "b"
```

```
dados <- dados |>  
  mutate(fertilizante = as.factor(fertilizante))
```

```
glimpse(dados)
```

```
## Rows: 15
```

```
## Columns: 2
```

```
## $ peso          <dbl> 1.02, 0.79, 1.00, 0.59, 0.9
```

```
## $ fertilizante <fct> a, a, a, a, a, b, b, b, b, b,
```

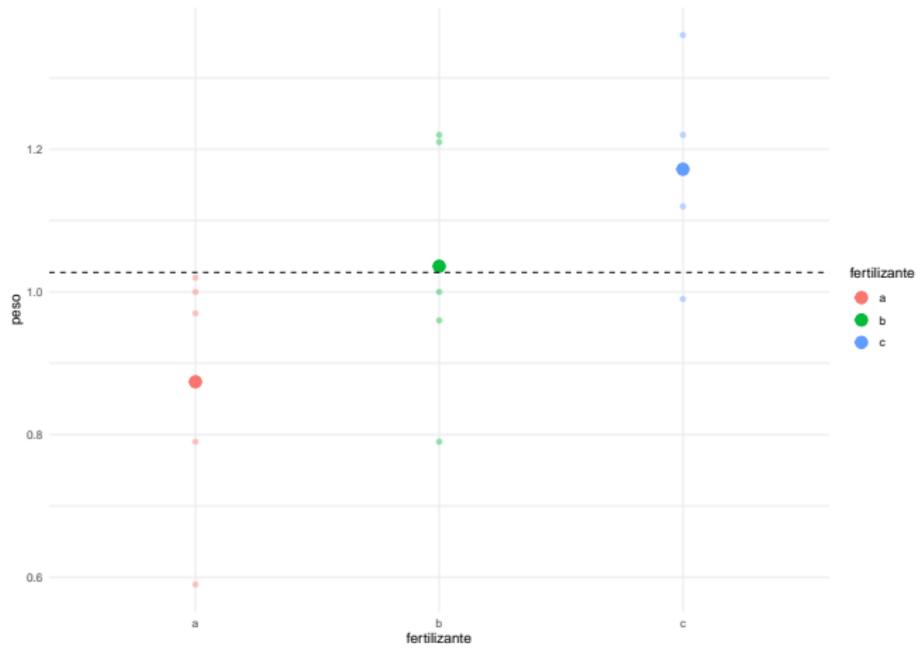
```
medias <- datos |>
  group_by(fertilizante) |>
  summarise(peso = mean(peso))
medias
```

```
## # A tibble: 3 x 2
##   fertilizante  peso
##   <fct>         <dbl>
## 1 a             0.874
## 2 b             1.04
## 3 c             1.17
```

```
mean(dados$peso)
```

```
## [1] 1.027333
```

```
ggplot(dados, aes(x = fertilizante, y = peso,  
                  color = fertilizante)) +  
  geom_point(alpha = .4) +  
  geom_point(data = medias, size = 4) +  
  geom_hline(yintercept = mean(dados$peso),  
            linetype = "dashed",  
            color = "black") +  
  theme_minimal()
```



```
sum_dados <- dados |>
  group_by(fertilizante) |>
  summarise(media = mean(peso),
            n = n(), variancia = var(peso))
sum_dados
```

```
## # A tibble: 3 x 4
##   fertilizante media      n variancia
##   <fct>         <dbl> <int>     <dbl>
## 1 a             0.874     5     0.0335
## 2 b             1.04     5     0.0329
## 3 c             1.17     5     0.0184
```

```
mean(dados$peso)
```

```
## [1] 1.027333
```

```
var(dados$peso)
```

```
## [1] 0.04013524
```

## Modelo Anova um-fator

**Exemplo 2.** Uma fabricante de cereais desejava testar quatro modelos de embalagens diferentes para um novo cereal matinal. Vinte lojas, com volumes de vendas aproximadamente iguais, foram selecionadas como unidades experimentais. Cada loja recebeu aleatoriamente um dos modelos de embalagem, sendo que cada modelo foi atribuído a cinco lojas. Ocorreu um incêndio em uma loja durante o período do estudo, então essa loja teve que ser retirada do estudo. Assim, um dos modelos foi testado em apenas quatro lojas. Para este experimento, foram escolhidas lojas semelhantes em localização e volume de vendas. Condições relevantes que podem afetar as vendas, como preço, quantidade e localização das prateleiras, e ações promocionais especiais, foram mantidas iguais para todas as lojas do experimento. As vendas, em número de casos, foram observadas durante o período de estudo e os resultados estão registrados na Tabela 2. ■

## Modelo Anova um-fator

**Tabela 2:** Vendas de caixas de cereais usando quatro modelos de embalagem.

Modelo	Loja				
	1	2	3	4	5
1	11	17	16	14	15
2	12	10	15	19	11
3	23	20	18	17	
4	27	33	22	26	28

# Modelo Anova um-fator

## **Estudo experimental com um fator:**

- Os tratamentos correspondem aos níveis do fator.
- Aleatorização é usada para designar os tratamentos às unidades experimentais.

Este estudo é um *delineamento completamente aleatorizado com um fator* (delineamento inteiramente casualizado com um fator).

## Modelo Anova um-fator

Os elementos básicos do modelo Anova para um delineamento completamente aleatorizado com um fator (*modelo Anova um-fator*) são simples:

- Para cada nível do fator, há uma distribuição de probabilidade das respostas.
  - No caso do exemplo anterior, há uma distribuição de probabilidade para cada tipo de modelo de embalagem.

# Modelo Anova um-fator

O modelo Anova um-fator assume que:

- 1 cada distribuição de probabilidade é normal.
- 2 todas as distribuições têm a mesma variância.
- 3 As respostas de cada nível do fator são realizações aleatórias da distribuição de probabilidade correspondente ao fator e são independentes das respostas para cada outro nível do fator.

Observe que as distribuições de probabilidade diferem apenas com respeito a média. Assim, diferenças nas médias, refletem os efeitos dos níveis do fator. Por esse motivo, a análise se concentra nas respostas médias para os diferentes níveis de fator.

# Modelo Anova um-fator

## Notações:

- $a$ : número de níveis do fator (número de tratamentos).  
Usamos o índice  $i$  para denotar um tratamento ( $i = 1, \dots, a$ ).
- $n_i$ : número de replicatas para o  $i$ -ésimo nível do fator. Usamos o índice  $j$  para denotar uma replicata para nível particular do fator ( $j = 1, \dots, n_i$ ). Assim, temos que  $y_{ij}$  denota o valor da variável resposta na  $j$ -ésima replicata do  $i$ -ésimo nível do fator.
- $n = \sum_{i=1}^a n_i$ : número total de observações no estudo.

## Modelo Anova um-fator

### Notações:

Vamos definir a notação  $y_{i.}$  por

$$y_{i.} = \sum_{j=1}^{n_i} y_{ij},$$

de tal forma que o ponto (.) indica que o somatório é no índice.

Assim, a média do  $i$ -ésimo tratamento é definida por

$$\bar{y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}. \quad (1)$$

A média geral é definida por

$$\bar{y}_{..} = \frac{1}{n} \sum_{i=1}^{n_i} \sum_{j=1}^{n_i} y_{ij}. \quad (2)$$

## Modelo Anova um-fator

O modelo Anova um-fator (de médias) é definido por

$$Y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1, \dots, a; \quad j = 1, \dots, n_i, \quad (3)$$

onde  $y_{ij}$  é o valor observado da variável resposta na  $j$ -ésima replicata do  $i$ -ésimo tratamento,  $\mu_i$  é a média do  $i$ -ésimo tratamento (os  $\mu$ 's são parâmetros) e  $\epsilon_{ij}$  é um erro aleatório que incorpora todas as outras fontes de variação no experimento. Os  $\epsilon_{ij}$  são independentes com distribuição  $N(0, \sigma^2)$ .

# Modelo Anova um-fator

- 1 No modelo (3), o valor observado de  $Y$  na  $j$ -ésima replicata para o  $i$ -ésimo tratamento é a soma de duas componentes:
  - i um termo constante  $\mu_i$ ;
  - ii um erro aleatório  $\epsilon_{ij}$ .

# Modelo Anova um-fator

## **Interpretação das médias dos níveis dos fatores:**

Em estudos experimentais, a média do nível do fator  $i$ ,  $\mu_i$ , representa a resposta média que seria obtida se o tratamento  $i$  fosse aplicado a todas as unidades na população de unidades experimentais sobre as quais serão feitas inferências.

A variância  $\sigma^2$  representa a variabilidade das respostas se qualquer tratamento fosse aplicado à toda a população de unidades experimentais.

## Modelo Anova um-fator

```
library(readxl)
library(tidyverse)

dados <- read_xlsx("../dados/dados_aula1.xlsx",
                  sheet = "embalagem")

dados <- dados |>
  mutate(embalagem = as.factor(embalagem))

sum_dados <- dados |>
  group_by(embalagem) |>
  summarise(n = n(),
            media = mean(vendas),
            variancia = var(vendas))
```

```
sum_dados
```

```
## # A tibble: 4 x 4
##   embalagem      n media  variancia
##   <fct>      <int> <dbl>      <dbl>
## 1 1          5  14.6         5.3
## 2 2          5  13.4        13.3
## 3 3          4  19.5         7
## 4 4          5  27.2        15.7
```

```
mean(dados$ vendas)
```

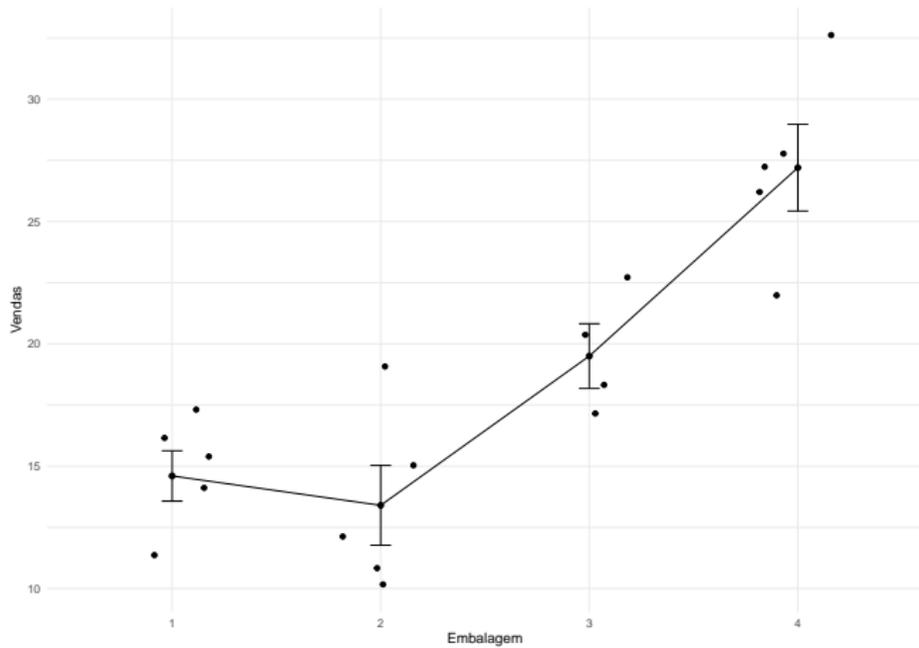
```
## [1] 18.63158
```

```
var(dados$ vendas)
```

```
## [1] 41.46784
```

```
library(ggpubr)
```

```
ggline(dados, x = "embalagem", y = "vendas",  
       add = c("mean_se", "jitter"),  
       ylab = "Vendas", xlab = "Embalagem") +  
theme_minimal()
```



## Modelo Anova um-fator

### Modelo de efeitos:

Considere o modelo definido por

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad i = 1, \dots, a; \quad j = 1, \dots, n_i, \quad (4)$$

onde  $Y_{ij}$  é a resposta,  $\mu$  é uma constante geral,  $\tau_i$  é o efeito do  $i$ -ésimo tratamento e  $\epsilon_{ij}$  é um erro aleatório. Usualmente,  $\epsilon_{ij} \sim N(0, \sigma^2)$ , para  $i = 1, \dots, a$  e  $j = 1, \dots, n_i$ .

O modelo da equação (4) é conhecido como *modelo (de efeitos) Anova um-fator*.

Note que no modelo (4), as médias dos tratamentos são expressas por

$$\mu_i = \mu + \tau_i, \quad i = 1, \dots, a. \quad (5)$$

## Modelo Anova um-fator

### Definição de $\mu$ :

A decomposição da média do fator,  $\mu_i$ , em duas componentes, uma constante geral  $\mu$  e um efeito do tratamento  $\tau_i$ , depende da especificação de  $\mu$ , que pode ser definida de várias maneiras. A mais comum é:

- Média geral (não ponderada):  $\mu$  é definida como a média das médias dos tratamentos:

$$\mu = \frac{1}{a} \sum_{i=1}^a \mu_i. \quad (6)$$

Essa definição implica a seguinte restrição aos  $\tau_i$ 's:

$$\sum_{i=1}^a \tau_i = 0. \quad (7)$$

# Análise de variância e teste-F para médias de tratamentos

## Hipóteses:

Nosso interesse é testar se as médias dos  $a$  tratamentos são iguais:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \mu_3 = \dots = \mu_a \\ H_1 : \mu_i \neq \mu_j \text{ para pelo menos um par } (i, j). \end{cases} \quad (8)$$

No modelo de efeitos (4), os  $\tau_i$ 's são “desvios” das médias  $\mu_i$ 's com relação à média geral  $\mu$ . Portanto, podemos escrever as hipóteses

$$\begin{cases} H_0 : \tau_1 = \tau_2 = \tau_3 = \dots = \tau_a = 0 \\ H_1 : \tau_i \neq 0 \text{ para pelo menos um } i. \end{cases} \quad (9)$$

## Análise de variância e teste-F para médias de tratamentos

Partição da variância:

$$\sum_{i=1}^a \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^a \sum_{j=1}^{n_i} (\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^a \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2, \quad (10)$$

onde

- $SQT = \sum_{i=1}^a \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$ : soma de quadrados total; com  $n - 1$  graus de liberdade.
- $SQTR = \sum_{i=1}^a \sum_{j=1}^{n_i} (\bar{y}_{i.} - \bar{y}_{..})^2$ : soma de quadrados de tratamentos; com  $a - 1$  graus de liberdade.
- $SQR = \sum_{i=1}^a \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$ : soma de quadrados residual;  $n - a$  graus de liberdade.

## Análise de variância e teste-F para médias de tratamentos

### Quadrados médios:

Os quadrados médios são obtidos dividindo-se cada soma de quadrados por seus respectivos graus de liberdade. Portanto:

- Quadrado médio dos tratamentos:

$$QMTR = \frac{SQTR}{a - 1} = \frac{\sum_{i=1}^a \sum_{j=1}^{n_i} (\bar{y}_{i.} - \bar{y}_{..})^2}{a - 1}. \quad (11)$$

- Quadrado médio residual:

$$QMR = \frac{SQR}{n - a} = \frac{\sum_{i=1}^a \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2}{n - a}. \quad (12)$$

# Análise de variância e teste-F para médias de tratamentos

## Quadrados médios:

- QMR é um estimador não viciado de  $\sigma^2$ , a variância dos termos de erro  $\epsilon_{ij}$ , independentemente das médias  $\mu_i$ 's serem iguais ou diferentes.
- Quando os efeitos dos tratamentos,  $\tau_i$ , são todos iguais a zero, então todas as médias dos tratamentos  $\mu_i$  são iguais, e portanto iguais à média geral  $\mu$ .

Assim, *quando não há efeitos de tratamento*, tanto o QMR quanto o QMTR são estimadores não viciados de  $\sigma^2$ . Por outro lado, *quando há efeitos de tratamento*, o QMTR tende a ser maior do que o QMR.

# Análise de variância e teste-F para médias de tratamentos

## Quadrados médios:

Podemos utilizar essa “propriedade” ou característica para testar a hipótese de igualdade das médias dos tratamentos comparando o QMTR e o QMR.

- Se o QMTR e o QMR são iguais (em magnitude), então há um indicativo de que as médias são iguais.
- Se QMTR é substancialmente maior do que o QMR, então há um indicativo de que as médias não são iguais.

## Análise de variância e teste-F para médias de tratamentos

### Estatística teste:

A estatística teste para escolher entre as hipóteses em (9) é definida por

$$F_0 = \frac{QMTR}{QMR}. \quad (13)$$

Sob  $H_0$ , tanto o QMR quanto o QMTR são estimadores não viciados de  $\sigma^2$ . Portanto, um valor de  $F_0$  próximo a 1 fornece evidência a favor de  $H_0$ .

Se  $H_0$  não é verdadeira, o QMTR tende a ser maior do que o QMR. Portanto, um valor grande de  $F_0$  fornece evidência contra  $H_0$ .

## Análise de variância e teste-F para médias de tratamentos

Sob  $H_0$ , segue que  $F_0 \sim F(a - 1, n - a)$ .

A um nível de significância  $\alpha \in (0, 1)$ , rejeitamos  $H_0$  se

$$F_0 > F_{\alpha; a-1; n-a}. \quad (14)$$

# Análise de variância e teste-F para médias de tratamentos

## Tabela Anova:

**Tabela 3:** Tabela Anova: modelo com um fator (efeitos fixos).

Fonte de Variação	Soma de Quadrados	Graus de Liberdade	Quadrado Médio	$F_0$
Tratamento	$SQTR$	$a - 1$	$QMTR$	$\frac{QMTR}{QMR}$
Resíduo	$SQR$	$n - a$	$QMR$	
Total	$SQT$	$n - 1$		

# Análise de variância e teste-F para médias de tratamentos

## Usando $p$ -valor para tomar decisão em testes de hipóteses:

- O  $p$ -valor é o menor nível de significância para o qual rejeita-se  $H_0$  com os dados observados.
- Como decidir entre  $H_0$  e  $H_1$  usando o  $p$ -valor do teste?
  - Usamos e interpretamos o  $p$ -valor do seguinte modo:
    - Se o  $p$ -valor for grande, a amostra não fornece evidências suficientes para rejeitar  $H_0$ , então podemos decidir por  $H_0$ ;
    - Se o  $p$ -valor for pequeno, a amostra fornece evidências suficientes para rejeitar  $H_0$ , então podemos decidir por  $H_1$ .
  - Assim, para para um teste de hipóteses com nível de significância  $\alpha \in (0, 1)$  pré-fixado, rejeitamos  $H_0$  se o  $p$ -valor do teste for menor do que  $\alpha$ .

## Suposições do modelo

**Resíduos:**

$$e_{ij} = y_{ij} - \hat{y}_{ij}; \quad i = 1, \dots, a, j = 1, \dots, n_i. \quad (15)$$

Temos que,

$$\hat{y}_{ij} = \hat{\mu} + \hat{\tau}_i = \bar{y}_i.; \quad i = 1, \dots, a, j = 1, \dots, n_i. \quad (16)$$

Portanto,

$$e_{ij} = y_{ij} - \bar{y}_i.; \quad i = 1, \dots, a, j = 1, \dots, n_i. \quad (17)$$

**Resíduos padronizados:**

$$d_{ij} = \frac{e_{ij}}{\sqrt{QMR}}; \quad i = 1, \dots, a, j = 1, \dots, n_i. \quad (18)$$

# Suposições do modelo

## Normalidade:

- Testes de normalidade (Shapiro-Wilks, Kolmogorov-Smirnov, Anderson-Darling).
- Gráfico normal probabilístico dos resíduos.

## Independência:

- Gráficos dos resíduos contra a ordem de observação.
- Teste de correlação serial (Durbin-Watson).

## Homoscedasticidade:

- Testes homogeneidade de variâncias (Bartlett, Levene).
- Gráfico dos resíduos contra os valores ajustados.

## Comparações Múltiplas

Quando o teste-F do modelo Anova um-fator é rejeitado, temos um indicativo de que existem diferenças significativas entre pelo menos duas das médias dos tratamentos (o que não significa necessariamente que todas as médias são significativamente diferentes entre si).

Para testar a diferença entre médias de tratamentos, tomadas duas a duas, deve ser realizada utilizando-se procedimentos de *comparações múltiplas*.

Testes de comparações múltiplas são semelhantes ao teste-t usual de comparação de médias, com a diferença de que controlam o nível de significância ao levar em consideração o número de comparações a realizadas.

# Comparações Múltiplas

Principais procedimentos de comparações múltiplas:

- Método de Tukey
- Método de Scheffé
- Método de Fisher
- Correção de Bonferroni
- Método de Dunnett

## Exemplos

**Exemplo 3.** Uma pesquisadora conduziu um experimento farmacêutico para testar o efeito de um novo medicamento para colesterol. A pesquisadora selecionou 15 indivíduos aleatoriamente de uma população e cada pessoa foi atribuída aleatoriamente a um de três grupos de tratamento. Indivíduos no grupo 1 receberam dose do medicamento de 0 mg/dia; indivíduos no grupo 2 receberam dose de 50 mg/dia; indivíduos no grupo 3 receberam dose de 100 mg/dia. Após 30 dias, a pesquisadora mediu o nível de colesterol de cada paciente. Os dados são apresentados na planilha *colesterol* do arquivo *dados\_aula1.xlsx*. ■

## Exemplos

### Exemplo 4.

Uma psicóloga clínica realizou um experimento para comparar dois tratamentos para depressão, terapia cognitivo-comportamental (TCC) e terapia centrada no cliente (CCT), contra uma condição de controle. Os participantes foram alocados aleatoriamente aos tratamentos. Após 12 semanas, os escores de depressão dos pacientes foram medidos usando a escala de depressão CESD. Os dados são apresentados na planilha *depressao* do arquivo *dados\_aula1.xlsx*. ■

## Exemplos

**Exemplo 5.** Quatro marcas de baterias de lanternas foram testadas em cinco lanternas cada: vinte lanternas foram selecionadas ao acaso, divididas em quatro grupos de cinco lanternas e, então, cada grupo de lanternas foi testada com uma das quatro marcas de bateria. O tempo de vida das baterias, em horas, está na planilha *bateria* do arquivo *dados\_aula1.xlsx*. ■

## Exemplos

**Exemplo 6.** Três grupos de cinco ratos foram selecionados ao acaso serem colocados em um labirinto padrão, mas com portas de cores diferentes. A resposta é o tempo necessário para completar o labirinto e os dados estão na planilha *ratos* do arquivo *dados\_aula1.xlsx*. ■

## Exemplos

**Exemplo 7.** Uma pesquisadora realizou um experimento para investigar a aptidão física de engenheiros contratados pelo Corpo de Bombeiros de uma cidade. Para medir a condição física, a fisioterapeuta escolheu cinco engenheiros, cada um com 5, 10, 15 e 20 anos de experiência. Ela, então, registrou o número de flexões que cada pessoa conseguiu realizar em 60 segundos. Os dados estão na planilha *fitness* do arquivo *dados\_aula1.xlsx*. ■

## Exemplos

**Exemplo 8.** Um desenvolvedora videogames testou um novo jogo em três grupos diferentes. Cada grupo representa um mercado-alvo diferente para o jogo. A desenvolvedora coletou pontuações de uma amostra aleatória de cada grupo. Os dados estão na planilha *jogo* do arquivo *dados\_aula1.xlsx*. ■

## Usando o R

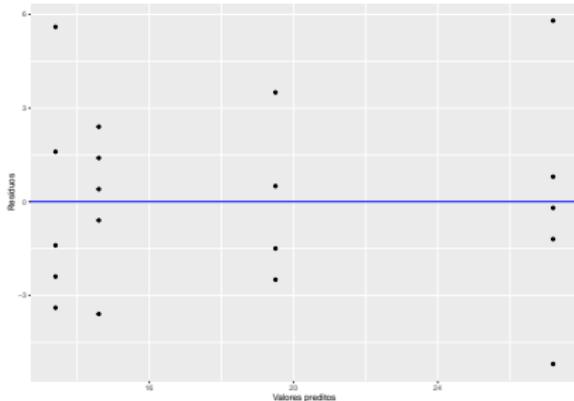
```
library(readxl)  
library(ggthemes)  
library(ggpubr)  
library(MASS)  
library(car)  
library(DescTools)  
library(agricolae)  
library(tidymodels)  
library(tidyverse)
```

```
fit <- aov(vendas ~ embalagem, data = dados)

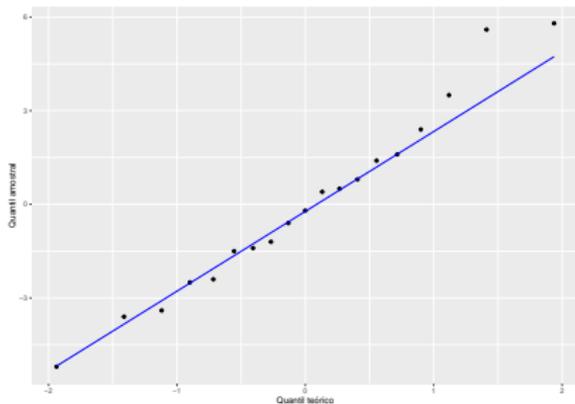
fit_lm <- lm(vendas ~ embalagem , data = dados)

tab <- augment (fit_lm)
```

```
ggplot(tab, aes(x = .fitted, y = .resid)) +  
  geom_point() +  
  geom_abline(intercept = 0, slope = 0,  
             color = 'blue') +  
  labs(x = "Valores preditos", y = "Resíduos")
```



```
ggplot(tab, aes(sample = .resid)) +  
  geom_qq() +  
  geom_qq_line(color = "blue") +  
  labs(x = "Quantil teóric", y = "Quantil amostral")
```



```
shapiro.test(fit$residuals)
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: fit$residuals  
## W = 0.97231, p-value = 0.8216
```

```
bartlett.test(vendas ~ embalagem, data = dados)
```

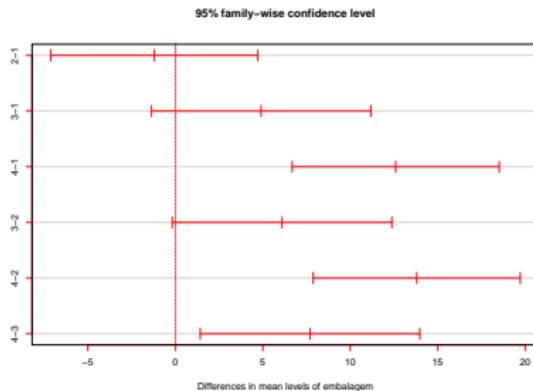
```
##  
## Bartlett test of homogeneity of variances  
##  
## data: vendas by embalagem  
## Bartlett's K-squared = 1.3144, df = 3, p-value = 0.7257
```

```
leveneTest(vendas ~ embalagem, data = dados)
```

```
## Levene's Test for Homogeneity of Variance (center = median)  
##      Df F value Pr(>F)  
## group 3  0.2417 0.8659  
##      15
```

```
tukey <- TukeyHSD(fit, conf.level = 0.95)
```

```
plot(tukey, col = "red")
```



```
ScheffeTest (fit, which = "embalagem", conf.level = 0.9)
```

```
##
```

```
## Posthoc multiple comparisons of means: Scheffe Test
```

```
## 90% family-wise confidence level
```

```
##
```

```
## $embalagem
```

```
##      diff      lwr.ci      upr.ci      pval
```

```
## 2-1 -1.2 -6.8134440  4.413444 0.95067
```

```
## 3-1  4.9 -1.0539565 10.853956 0.21253
```

```
## 4-1 12.6  6.9865560 18.213444 0.00023 ***
```

```
## 3-2  6.1  0.1460435 12.053956 0.08949 .
```

```
## 4-2 13.8  8.1865560 19.413444 8.6e-05 ***
```

```
## 4-3  7.7  1.7460435 13.653956 0.02478 *
```

```
##
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
```

```
comp_bon <- LSD.test(fit, "embalagem", alpha = 0.05, p.adj = "bonferroni",
                    console = FALSE)
comp_bon
```

```
## $statistics
##      MSerror Df      Mean      CV
## 10.54667 15 18.63158 17.43042
##
## $parameters
##      test p.adjusted name.t ntr alpha
## Fisher-LSD bonferroni embalagem 4 0.05
##
## $means
##      vendas      std r      se      LCL      UCL Min Max  Q25 Q50  Q75
## 1 14.6 2.302173 5 1.452354 11.50438 17.69562 11 17 14.00 15 16.00
## 2 13.4 3.646917 5 1.452354 10.30438 16.49562 10 19 11.00 12 15.00
## 3 19.5 2.645751 4 1.623782 16.03899 22.96101 17 23 17.75 19 20.75
## 4 27.2 3.962323 5 1.452354 24.10438 30.29562 22 33 26.00 27 28.00
##
## $comparison
## NULL
##
## $groups
##      vendas groups
## 4 27.2 a
## 3 19.5 b
## 1 14.6 b
## 2 13.4 b
##
## attr(,"class")
## [1] "group"
```