MATERIAL DIDÁTICO

# Parte I

# Estatística Experimental – Medicina Veterinária

Faculadade de Ciências Agrárias e Veterinárias

Campus de Jaboticabal – SP

Gener Tadeu Pereira

2º SEMESTRE DE 2015

ÍNDICE

INTRODUÇÃO AO R.....2
AULA 1 – ESTATÍSTICA ESCRITIVA.....10
1º EXERCÍCIO PRÁTICO ESTATÍSTICA EXPERIMENTAL..... 33
AULA 2 – TESTES DE SIGNIFICÂNCIA.....35
2º EXERCÍCIO PRÁTICO DE ESTATÍSTICA EXPERIMENTAL..... 46
AULA 3- DELINEAMENTO INTEIRAMENTE CASUALIZADO (DIC).....48
3º EXERCÍCIO PRÁTICO DE ESTATÍSTICA EXPERIMENTAL..... 70
AULA 4 TESTE DE COMPARAÇÕES MÚLTIPLAS.....73
4º EXERCÍCIO PRÁTICO DE ESTATÍSTICA EXPERIMENTAL..... 92
AULA 5 TESTES F PLANEJADOS.....93
5º EXERCÍCIO PRÁTICO DE ESTATÍSTICA EXPERIMENTAL..... 104
AULA 6 DELINEAMENTO EM BLOCOS CASUALIZADOS (DBC).....107
6º EXERCÍCIO PRÁTICO DE ESTATÍSTICA EXPERIMENTAL..... 124

# 1. Introdução ao R

## 1.1 O que é o R?

R é uma linguagem e ambiente para calcular estatísticas e gráficos. Ele é um projeto "GNU" o qual é similar à linguagem ***S*** *e ambiente* a qual foi desenvolvida na *"Bell Laboratories"* , formalmente (AT&T) por John Chambers e colaboradores. R pode ser considerado uma implantação diferente do S. Existem algumas diferenças importantes, mas muitos dos códigos escritos para o S rodam sem modificações no R.

O R fornece uma grande variedade de técnicas estatísticas (modelagem linear e não linear, testes estatísticos clássicos, análise de séries-temporais, ...) e gráficas, e é altamente extensível.

Um dos pontos fortes de R é a facilidade com que bem projetados gráficos para publicações de qualidade pode ser produzidos, incluindo símbolos matemáticos e fórmulas.

O programa R esta disponível como um programa livre ("*Free Software"*) sob os termos da "*Free Software Foundation's GNU General Public License"*.

## 1.2 Instalando o R

Geralmente, o sistema R consite de duas partes. Uma é denominada de *Sistema básico do R* para o núcleo da linguagem R e bibliotecas fundamentais associadas. A outra consiste de contribições de usuários que desenvolvem *pacotes* que são aplicações mais especializadas. Ambas as partes podem obtidas do " *Comprehensive R Archive Network* " (CRAN) do site:

http://CRAN.r-project.org

A instalação do sistema R é descrito a seguir

### Instalando o Sistema básico do R

Usuários do *Windows* podem baixar a última versão do R no endereço

http://www.vps.fmvz.usp.br/CRAN/

Em *"Dowload and Install R",* acione o "*link"* que corresponde ao sistema operacional do seu computador ( no caso do Windows – *"Dowload R for Windows"* e depois no *link base.* Depois de baixar (salvar) o arquivo executável, basta executá-lo e seguir a rotina de instalação. Neste mesmo endereço são disponibilizadas versões do R nas plataformas do *"Linux" ,*e *"MacOS X"*.

O endereço acima é o local disponível mais próximo de Jaboticabal, no caso a USP/Pirassununga, SP.

## 2. Manipulando dados

O R usa vários tipos de dados que frequentemente são usados na maioria de suas funções ou cálculos. A seguir uma listagem dos mais importantes tipos de dados e objetos:

Um ***vetor*** pode ser numérico, complexo, um vetor de caracteres, ou um vetor lógico. Estes vetores são sequências de números, números complexos, caracteres, ou valores lógicos, respectivamente;

Uma “array” é um vetor com um atributo de dimensão, em que atributo de dimensão é um vetor de inteiros não negativos. Se a amplitude de um vetor é *k*, então o vetor é *k*-dimensional. A “array” mais comumente usada é uma *matriz,*a qual é uma “array” bi-dimensional de números.

Um *fator* objeto é usado para definir variáveis categóricas (nominais ou ordenadas). Estes podem ser vistos como vetores inteiros em que cada valor inteiro tem um label correspondente.

Uma *lista* é uma coleção de objetos ordenados. Um vetor pode conter somente elementos de um tipo, mas uma lista pode ser usada para criar uma coleção de vetores ou objetos misturados.

Um *data frame* é uma lista de vetores ou fatores com uma mesma amplitude tal que cada linha corresponde a uma observação.

Os seguintes códigos usam as funções c( ), factor( ), data.frame( ), e list( ) para exemplificar os diferentes tipos de dados

```
> vet1 <- 4                          # Vetor numérico de dimensão 1
> vet1
[1] 4
> vet2 <- c(1,2,3.4,5.6,7)        # Vetor numérico de dimensão 5
> vet2
[1] 1.0 2.0 3.4 5.6 7.0
> # Criando um vetor de caracteres
> vet3 <- c("João", "Antônio", "Costa", "Silva", "Oliveira")
> # Criando um vetor de valores lógicos
> vet4<- c(TRUE, TRUE,TRUE, FALSE, TRUE)
> f<- factor(vet3)                  # Definindo um vetor baseado em vet3
> f
[1] João     Antônio  Costa    Silva    Oliveira
Levels: Antônio Costa João Oliveira Silva
> x<- 1+2i                           # Entrando com um número complexo
> x
[1] 1+2i

> sqrt(-1)                           #  O R não reconhece número complexo
[1] NaN
Mensagens de aviso perdidas:
In sqrt(-1) : NaNs produzidos
```

```
> sqrt(-1+0i)                    #  Mas reconhece este
[1] 0+1i
> m<- matrix(1:6,,ncol=2)
> m
     [,1] [,2]
[1,]    1    4
[2,]    2    5
[3,]    3    6
> list(vet2, comp=x, m)          # Combina 3 dferentes objetos
[[1]]
[1] 1.0 2.0 3.4 5.6 7.0

$comp
[1] 1+2i

[[3]]
     [,1] [,2]
[1,]    1    4
[2,]    2    5
[3,]    3    6

> data.frame(vet2, f, vet4)       # Cria um data.frame
  vet2        f        vet4
1  1.0      João   TRUE
2  2.0  Antônio   TRUE
3  3.4     Costa    TRUE
4  5.6      Silva    FALSE
5  7.0   Oliveira   TRUE
```

Tabela 2.1 Operadores matemáticos e funções

| Símbolo/ funções | Descrição |
|---|---|
| **+** | adição |
| **-** | subtração |
| ***** | multiplicação |
| **/** | divisão |
| **^** ou ** | exponenciação |
| **%%** | modulus |
| **% / %** | divisão inteira |
| **abs (x)** | valor absoluto |
| **sqrt (x)** | raiz quadrada |
| **ceiling (x)** | menor valor não menor que x |
| **floor (x)** | maior inteiro não maoir que x |
| **trunc (x)** | trunca x descartando as decimais |
| **round (x, digits=0)** | arrendonda x para nº de dígitos dec. |
| **signif (x, digits=6)** | arredonda x para nº de dígitos signif. |
| **cos (x), sin (x)** e **tan (x)** | coseno, seno e tangente |
| **log (x)** | logaritmo natural (base = e) |
| **log (x, base=2)** | logarítmo com base 2 |
| **log10 (x)** | logarítmo decimal (base = 10) |
| **exp (x)** | função exponencial |
| **% * %** | multiplicação de matrizes |

## 2.1 Usando funções matemáticas e operações

**Problema:** Voce deseja aplicar funções matemáticas numéricas básicas ou usar opradores aritméticos em seus cálculos.

**Solução:** O R contém um grande número de operadores aritméticos e funções matemáticas e algumas delas são listadas na Tabela 2.1. Os códigos abaixo são exemplos do uso destas funções.

```
> 5/2 + 2*(5.1 - 2.3)          # Soma, subtração, multiplicação e divisão
[1] 8.1
> 2**8                         # 2 elevado a potência 8
[1] 256
> 1.61^5                       # 1.61 elevado a potência 5
[1] 10.81756
> 10 %% 3                      # 10 modulus 3 tem resto 1
[1] 1
> 10 %/% 3                     # divisão inteira
[1] 3
> abs( -3.1)                   # valor absoluto de -3.1
[1] 3.1
> ceiling(4.3)                 # menor inteiro maior que 4.3
[1] 5
> floor(4.3)                   # maior inteiro menor que 4.3
[1] 4
```

```
> trunc(4.3)                  # remove as decimais
[1] 4
> round(4.5)                  # arredonda para 0 casas decimais
[1] 4
> round(4.51)
[1] 5
> round(4.51,digits=1)
[1] 4.5
> # Angulos para funções trigonometricas use radianos - não graus
> cos(pi/2)                   # coseno de pi/2
[1] 6.123032e-17
> sin(pi/4)                   # seno de pi/4
[1] 0.7071068
> tan(pi/6)                   # tangente de pi/6
[1] 0.5773503
> log(5)                      # logarítmo natural de 5
[1] 1.609438
> log(5,base=2)               # logarítmo de 5 na base 2
[1] 2.321928
> log10(5)                    # logarítmo de 5 na base 10
[1] 0.69897
> exp(log(5) + log(3))        # função exponencial
[1] 15
> # Criando duas matrizes
> x <- matrix(1:6,ncol=3)
> y <- matrix(c(1, 1, 0, 0, 0, 1), ncol=2)
> x
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
> y
     [,1] [,2]
[1,]    1    0
[2,]    1    0
[3,]    0    1
> x%*%y                       # Multiplicação de matrizes
     [,1] [,2]
[1,]    4    5
[2,]    6    6
> y%*%x                       # Multiplicação de matrizes
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    1    3    5
[3,]    2    4    6
```

## 2.2 Trabalhando com funções comuns

**Problema:** Voce precisa identificar e/ou usar algumas funções comuns.

**Solução:** O grande número de funções no R pode trazer alguma dificuldade para os usuários na identificação da função a qual fornece certa facilidade. A Tabela 2.2 lista algumas funções mais frequentemente usadas encontradas no R.

| Símbolo/função | Descrição |
|---|---|
| **mean (x)** | média |
| **median (x)** | mediana |
| **sd (x)** | desvio padrão |
| **var (x)** | variância |
| **summary (x)** | sumário estatístico |
| **IQR (x)** | intervalo inter-quartílico |
| **cor (x,y)** | correlação entre x e y |
| **length (x)** | amplitude de x |
| **sum (x)** | soma |
| **cumsum (x)** | soma acumulativa |
| **sort (x)** | ordenação de x |
| **order (x)** | postos de x |
| **min (x)** e **max (x)** | mínimo e maximo de x |
| **quantile (x)** | quantis de x |
| **is.na (x)** | teste para observações perdidas |
| nrow (df) e **ncol (df)** | nº linhas e colunas no data frame df |

```
> x<- c(6, 8, 1:4)
> x
[1] 6 8 1 2 3 4
> mean(x)                  # média
[1] 4
> median(x)                # mediana
[1] 3.5
> sd (x)                   # desvio padrão
[1] 2.607681
> IQR (x)                  # amplitude inter-quartile
[1] 3.25
> summary (x)              # sumário estatístico
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    2.25    3.50    4.00    5.50    8.00
> y<- (1:6)**2             # novo vetor
> cor(x, y)                # correlação de Pearson
[1] -0.3783846
> cor (x, y, method="spearman")   # correlação de Spearman
[1] -0.3714286
> cor(x, y,method= "kendall")     # corerelação e Kendall
[1] -0.06666667
> length (x)                      # amplitude de x
[1] 6
```

```
> sum (x)                          # soma dos elementos em x
[1] 24
> cumsum (x)                       # soma acumulativa de x
[1]  6 14 15 17 20 24
> sort (x)                         # ordenação do vetor x
[1] 1 2 3 4 6 8
> order (x)                        # postos dos elementos
[1] 3 4 5 6 1 2
> min (x)                          # valor mínimo de x
[1] 1
> max (x)                          # valor máximo de x
[1] 8
> quantile (x)                     # fornece os quartis de x
  0%  25%  50%  75% 100%
1.00  2.25  3.50  5.50  8.00
> quantile (x, probs= c(0.15, 0.25, 0.99)) # quantis especificados
 15%  25%  99%
1.75  2.25  7.90
> x[c(2,4)] <- NA                  # redefine x na 2ª e 4ª posição com NA
> is.na(x)                         # identifica os valores perdidos
[1] FALSE  TRUE FALSE  TRUE FALSE FALSE
> df <-data.frame(x=c(1, 2, 3, 4), y=c(2, 1, 3, 2))
> df
   x  y
1 1  2
2 2  1
3 3  3
4 4  2
> nrow(df)                         # número de linhas no data frame df
[1] 4
> ncol(df)                         # número de colunas no data frame df
[1] 2
```

Várias funções definidas acima usam argumentos opcionais. Em particular a opção na.rm pode ser colocada como TRUE para assegurar que observações perdidas são desconsideradas nos cálculos como é mostrado a seguir. Do mesmo modo, a opção decreasing = TRUE para as funções sort e options revertem a ordem.

```
> x                                # vetor com valores perdidos
[1]  6 NA  1 NA  3  4
> sum(x)                           # sum não é avaliável
[1] NA
> sum(x, na.rm=TRUE)               # a menos que os NA's sejam removidos
[1] 14
> mean(x)                          # mean tmambém não é executado
[1] NA
> mean(x, na.rm = TRUE)            # a menos que os NA's sejam removidos
[1] 3.5
> order(x)                         # order coloca os NA's por último
[1] 3 5 6 1 2 4
```

```
> sort(x)                          # sort remove todos os NA's
[1] 1 3 4 6
> sort(x, na.last = TRUE, decreasing=TRUE)  # a menos que queremos
mantê-los
```

## 2.3 Importando dados

**Problema:** Voce deseja importar um conjunto de dados (arquivo de dados) armazenado em arquivo texto ASCII (arquivos com extensão .txt)

**Solução:** Dados armazenados em simples arquivos textos podem ser lidos no R pela função **read.table( )**. Por default, as observações devem ser listadas em colunas e os campos individuais são separados por um ou mais caracteres de espaço em branco, e cada linha no arquivo corresponde a uma linha no arquivo "data frame". Considere o arquivo, **dados.txt**, com o seguinte conteúdo

| animal | sexo | acido | digestao |
|---|---|---|---|
| NA | m | 30.3 | 70.6 |
| pituco | m | 29.8 | 67.5 |
| rufus | f | NA | 87 |
| pretinha | f | 4.1 | 89.9 |
| princesa | f | 4.4 | . |
| rose | f | 2.8 | 93.1 |
| "" | f | 3.8 | 96.7 |

O código default para observações perdidas é a sequência de caracteres NA, o qual aparece na 1ª observação da 1ª coluna e na 3ª linha da 3ª coluna. Os outros símbolos, " " e . , tem de ser especificados na opção **na.strings( )**. O arquivo acima é lido pelo R com o seguinte comando:

```
> dadosin<-read.table("dados.txt", header=TRUE,
+                  na.strings=c("NA","","."))
> dadosin
   animal   sexo acido digestao
1   rufus     m  30.3     70.6
2  pituco     m  29.8     67.5
3    <NA>     f    NA     87.0
4 pretinha    f   4.1     89.9
5 princesa    f   4.4      NA
6    rose     f   2.8     93.1
7    <NA>     f   3.8     96.7
```

O primeiro argumento é o nome do arquivo, e o segundo (**header=TRUE**) é opcional e deve ser usado somente se a primeira linha do arquivo texto é composta por nomes das variáveis (cabeçalho) e o terceiro argumento especifica os símbolos **NA**, "", e . como valores perdidos.

O R procura o arquivo **dados.txt** no diretório corrente de trabalho estabelecido inicialmente pelos comandos: na janela da console acione as abas **Arquivo/Mudar diretório/....** . Se o diretório de trabalho é o drive d, então o caminho é estabelecido por **Arquivo/Mudar diretório/d:/**. O caminho pode ser estabelecido no próprio comando de leitura, ou seja,

```
dadosin<-read.table("d:/dados.txt", header=TRUE,
                    na.strings=c("NA","","."))
```

Outras funções usadas para obter e estabelecer o diretório de trabalho são **getwd( )** e **setwd( )**

```
> getwd()                  # obtém o diretório de trabalho corrente
[1] "C:/ "
 setwd("d:/")              # estabelece o diretório de trabalho par d:/
```

## 3. Estatística Descritiva

**3.1 Símbolos:** conjunto de dados e da somatória

**Conjunto de dados**:

Considere uma variável aleatória de interesse representada pela letra maiúscula **Y** e os valores específicos assumidos por esta variável aleatória pelas letras minúsculas ***y***. Para distinguir um valor do outro, utilizamos um subscrito ***i***. Por exemplo, $y_1, y_2, ..., y_n$. Em geral, um valor típico da variável aleatória será designado por $y_i$ e o valor final desta amostra por $y_n$, sendo que n representa o tamanho da amostra.

Uma notação compacta para representar a soma de todos os valores de uma variável aleatória de interesse, por exemplo, ***Y***, é

$$\sum_{i=1}^{n} y_i = y_1 + y_2 + \ldots + y_n$$

A letra grega ***Σ*** **(**sigma) é usada como símbolo da soma para a soma e $y_i$ para o valor da observação *i*, denominado de sinal de soma, será usado extensivamente neste curso.

Alguns exemplos e propriedades da somatória:

A soma de n números $y_1, y_2, ..., y_n$ , como vimos, pode ser expressa por

$$\sum_{i=1}^{n} y_i = y_1 + y_2 + \ldots + y_n$$

A soma dos quadrados de n números $y_1, y_2, ..., y_n$ é:

$$\sum_{i=1}^{n} y_i^2 = y_i^2 + y_2^2 + \ldots + y_n^2$$

A soma dos produtos de dois conjuntos de n números $x_1, x_2, ..., x_n$ e $y_1, y_2, ..., y_n$:

$$\sum_{i=1}^{n} x_i y_i = x_1 y_1 + x_2 y_2 + \ldots x_n y_n$$

Exemplo: Considere um conjunto de 3 números: 1, 3 e 6. Os números são simbolizados por: $Y = \{y_1, y_2, y_3\} = \{1,3,6\}$

A soma e a soma dos quadrados destes números são:

$$\sum_{i=1}^{n} y_i = 1 + 3 + 6 = 10, \qquad \sum_{i=1}^{n} y_i^2 = 1^2 + 3^2 + 6^2 = 46$$

Considere outro conjunto de números $x_1 = 2,\ x_2 = 4\ e\ x_3 = 5.$

A soma dos produtos de *x* e *y* é:

$$\sum_{i=1}^{3} x_i y_i = (2)(1) + (4)(3) + (5)(6) = 44$$

As três principais regras da adição são:

1 A soma da adição de dois conjuntos de números é igual à adição das somas

$$\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$$

2 A soma dos produtos de uma constante *k* e uma variável *Y* é igual ao produto da constante pela soma dos valores da variável ($y_i$)

$$\sum_{i=1}^{n} k\, y_i = k \sum_{i=1}^{n} y_i$$

3 A soma de n constantes com valor *k* é igual ao produto $nk$

$$\sum_{i=1}^{n} k = k + k + \ldots + k = n\,k$$

**Atenção:** notem que o cálculo da expressão $\sum_{i=1}^{n} y_i^2 = y_1^2 + y_2^2 + \ldots + y_n^2$, denominada de "***soma de quadrados"*** é diferente do cálculo da expressão $\left(\sum_{i=1}^{n} y_i\right)^2 = (y_1 + y_2 + \ldots + y_n)^2$, denominada de "***quadrado da soma"***.

Outras notações:

$y_+ = \sum_{i=1}^{n} y_i = y_1 + y_2 + \ldots + y_n$ , e $\qquad \bar{y} = \frac{y_+}{n} = \frac{\sum_{i=1}^{n} y_i}{n}$

Notação com dois subescritos. Considere dois grupos de dados

1. grupo controle: { 5, 7, 5, 4 } , o qual é representado por $\{ y_{11} = 5, y_{12} = 7, y_{13} = 5, y_{14} = 4 \}$ ,

2. grupo tratado: { 7 , 9 , 6 , 9 , 8 } , o qual é representado por $\{ y_{21} = 7, y_{22} = 9, y_{23} = 6, y_{24} = 9, y_{25} = 8 \}$ ,
   sendo, i =1, 2, representando os grupos e j = 1, 2,..., $r_i$ representando as repetições dentro de cada grupo.

   Calcular o valor da expressão $\frac{\sum_{i=1}^{2} \left(\sum_{j=1}^{r_i} y_{ij}\right)^2}{r_i}$

Exemplo de Tabela de dupla entrada. Qualquer observação é representada por $y_{ij}$ , sendo que, o índice *i* refere-se às linhas (*i=1, 2,..., k*) e o índice *j* refere-se às colunas (*j=1, 2, ..., r*).

| | Colunas | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Linhas | 1 | 2 | 3 | ... | j | ... | r | Total | Média |
| 1 | $y_{11}$ | $y_{12}$ | $y_{13}$ | ... | ... | ... | $y_{1r}$ | $y_{1+}$ | $\bar{y}_{1+}$ |
| 2 | $y_{21}$ | $y_{22}$ | $y_{23}$ | ... | ... | ... | $y_{2r}$ | $y_{2+}$ | $\bar{y}_{2+}$ |
| 3 | $y_{31}$ | $y_{32}$ | $y_{33}$ | ... | ... | ... | $y_{3r}$ | $y_{3+}$ | $\bar{y}_{3+}$ |
| . | . | . | . | . | ... | . | . | . | . |
| . | . | . | . | . | ... | . | . | . | . |
| . | . | . | . | . | ... | . | . | . | . |
| i | . | . | . | . | $y_{ij}$ | . | . | $y_{j+}$ | $\bar{y}_{j+}$ |
| . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . |
| k | $y_{k1}$ | $y_{k2}$ | $y_{k3}$ | ... | ... | ... | $y_{kr}$ | $y_{k+}$ | $\bar{y}_{k+}$ |
| Total | $y_{+1}$ | $Y_{+2}$ | $y_{+3}$ | ... | $y_{+j}$ | ... | $y_{+r}$ | $y_{++}$ | |
| Média | $\bar{y}_{+1}$ | $\bar{y}_{+2}$ | $\bar{y}_{+3}$ | ... | $\bar{y}_{+j}$ | ... | $\bar{y}_{+r}$ | | $\bar{y}_{++}$ |

$y_{+j}$ *é o total da* $j-$*ésima coluna*; $\bar{y}_{+j}$ *= é a média da* $j-$*ésima coluna*;

$y_{i+}$ *é o total da* $i-$*ésima linha*; $\bar{y}_{i+}$ *= é a média da* $i-$*ésima linha*;

$y_{++}$ *=é o total geral (soma de todas as observações )* ; $\bar{y}_{++}$ *=é a média geral*

### 3.2 Medidas de tendência central

Um dos aspectos mais importantes do estudo de um conjunto de dados é a posição do valor central. Qualquer valor numérico que representa o centro de um conjunto de dados é denominado de medida de locação ou medida de tendência central. As duas medidas mais comumente utilizadas é média aritmética, ou simplesmente a média, e a mediana.

---

#### 3.2.1 Média aritmética.

A mais familiar medida de tendência central é a média aritmética. Ela é a medida descritiva que a maioria das pessoas tem em mente quando elas falam de média.

A média pode ser expressa como

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{y_1 + y_2 + \ldots + y_n}{n} = \frac{y_+}{n}$$

Vamos supor que a variável aleatória Y assume os seguintes valores, { 10, 54, 21, 33, 53 }, então a média destes 5 valores é dada por:

$$\bar{y} = \frac{\sum_{i=1}^{5} y_i}{5} = \frac{10+54+21+33+53}{5} = \frac{171}{5} = 34{,}2$$

Script no R para o cálculo da média

```
> y <-c(10,54,21,33,53)
> media.1<-sum(y)/length(y)      # calculo da média pela definição
> media.1
[1] 34.2
> media.2<-mean(y)               # pela função mean( )
```

```
> media.2
[1] 34.2
```

Propriedades da média;

a) Única. Para um conjunto de dados existe uma e somente uma média aritmética.

b) Simplicidade. A média aritmética é fácil de ser entendida e fácil de ser calculada.

c) Dado que toda observação do conjunto de dados entra no seu cálculo, ela é afetada por cada valor. Valores extremos têm influência na média e, em algumas situações podem ocorrer distorções, o que pode torná-la uma medida indesejável como medida de tendência central.

---

### 3.2.2 Mediana.

Uma alternativa à média aritmética como medida de tendência central é a mediana. A mediana de um conjunto de valores finitos é o valor que ocupa a posição central dos dados ordenados, ou seja, aquele valor o qual divide o conjunto de dados em duas partes iguais tal que o número de valores iguais ou maiores que a mediana é igual ao número de valores menores ou iguais que a mediana. Temos que considerar duas situações:

$$\tilde{y} = \begin{cases} y_{(k+1)} & se\ n = 2k+1 \quad (n\ é\ impar) \\ \frac{1}{2}(y_{(k)} + y_{(k+1)}) & se\ n = 2k \quad (n\ é\ par) \end{cases}$$

Exemplos:

1. Considere os dados 10, 54, 21, 33, 53 , com n=5 observações, e a seqüência ordenada fica 10, 21, 33, 53, 54. A mediana é calculada como sendo a observação que ocupa a 3ª posição da seqüência ordenada, ou seja,

$$n = 2k+1 \Rightarrow k = (n-1)/2,\ ou\ seja,\ k = 2 \Rightarrow \tilde{y} = y_{(2+1)} = y_{(3)} = 33$$

2. Considere os dados 10, 54, 21, 33, 53, 55, e a seqüência ordenada fica 10, 21, 33, 53, 54, 55. Como o número de observações é par e a mediana é calculada como sendo a média das observações que ocupam a posição central, ou seja,

$$n = 2k \Rightarrow k = (n)/2,\ ou\ seja,\ k = 3 \Rightarrow \tilde{y} = \frac{1}{2}(y_{(3)} + y_{(3+1)}) = \frac{1}{2}(y_{(3)} + y_{(4)})$$

$$= \frac{1}{2}(33 + 53) = 43$$

Script no R para o cálculo da mediana

```
> mediana<-median(y)            # calculo da mediana pela função
median( )
> mediana
[1] 33
```

Propriedades da mediana;

a) Única. Assim como a média, para um conjunto de dados existe uma e somente uma mediana.

b) Simplicidade. A mediana é fácil de ser calculada.

c) Ela não é drasticamente afetada por valores extremos, como a média.

### 3.2.3 Moda.

A moda é comumente definida como a observação mais freqüente do conjunto de dados. Se todas as observações são diferentes não existe moda; por outro lado um conjunto de dados pode ter mais de uma moda. Exemplo: considere o conjunto de dados

**{98, 102, 100, 100, 99, 97, 96, 95, 99, 100}**, então a moda é ***mo*** = 100, e no conjunto de dados, abaixo,

**{ 20, 21, 20, 20, 34, 22, 24, 27, 27, 27}** existe duas modas 20 e 27 (bimodal).

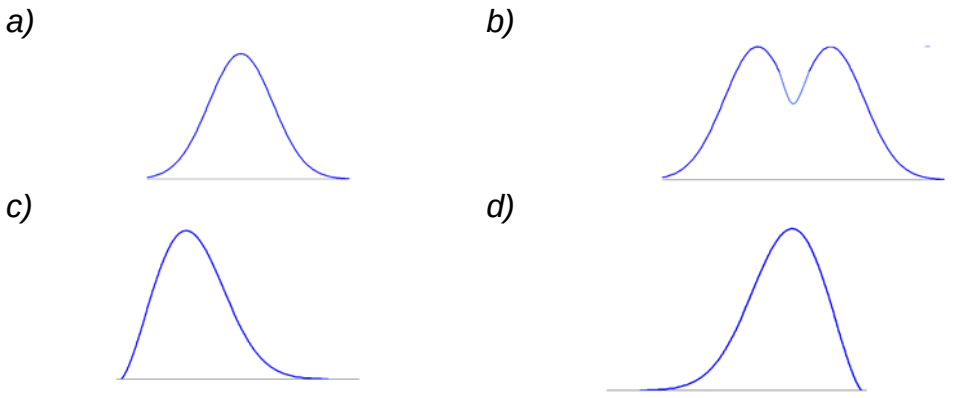



Figura 1.1 Distribuições de freqüência mostrando as medidas de tendência central. Distribuições em a) e b) são simétricas, c) é positivamente assimétrica, e d) é negativamente assimétrica. As distribuições a), c), e d) são unimodal, e a distribuição b) é bimodal.

## 3.3 Medidas de dispersão

Apesar das medidas de tendência central fornecerem uma idéia do comportamento de um conjunto de dados, elas podem esconder valiosas informações. Essas medidas não são suficientes para descrever ou discriminar diferentes conjunto de dados. Por exemplo, a Figura 3.1 mostra os polígonos de freqüência duas variáveis que possuem a mesma média, mas diferentes valores de dispersão. A variável B, a qual tem maior variabilidade que a variável A, é mais espalhada. A dispersão de um conjunto de dados se refere à variedade que eles exibem. Uma medida de dispersão fornece informação a respeito da quantidade de variabilidade presente no conjunto de dados.

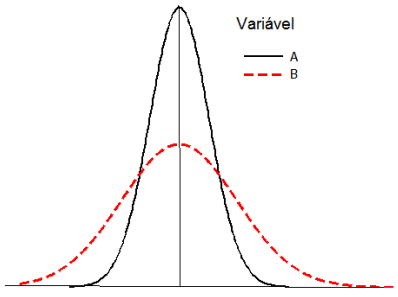


Figura 3.1 Dois polígonos de freqüência com a mesma média, mas com diferentes quantidades de dispersão.

Se todos os valores do conjunto de dados são iguais, não existe dispersão; se eles são diferentes, a dispersão está presente nos dados. A quantidade de dispersão pode ser pequena, quando os dados, embora diferentes, são muito próximos.

### 3.3.1 Amplitude

A amplitude é definida como a diferença entre o maior e o menor valor do conjunto de dados. O problema desta mediada é que ela só leva em conta dois valores do conjunto de dados e, assim, seria mais conveniente considerarmos uma mediada que utilizasse todas as observações do conjunto de dados. A primeira idéia que ocorre é considerar o desvio de cada observação em relação a um ponto de referência e então calcular a sua média. Se tomarmos a média aritmética como este ponto de referência, temos a seguinte situação:

Seja o conjunto de dados $y_1, y_2, ..., y_n$ *e* $\bar{y}$, a média destes dados. Definiremos por $d_i = y_i - \bar{y}$, os desvios destas observações em relação à sua média. Por exemplo, considere os dados $y_1 = 4,\ y_2 = 5,\ y_3 = 6$ *e* $y_4 = 9$. Assim temos:

$$\bar{y} = \frac{4+5+6+9}{4} = 6,$$

$$d_1 = (4-6) = -2,\ d_2 = (5-6) = -1,\ d_3 = (6-6) = 0,\ d_4 = (9-6) = 3$$

Reparem que a soma dos desvios é igual a zero, ou seja, $\sum_{i=1}^{4} d_i = 0$. Isto pode ser provado algebricamente, da seguinte forma,

$$\sum_{i=1}^{n} d_i = \sum_{i=1}^{n} (y_i - \bar{y}) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \bar{y} = \sum_{i=1}^{n} y_i - n\bar{y} = \sum_{i=1}^{n} y_i - n\frac{\sum_{i=1}^{n} y_i}{n} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} y_i = 0$$

Portanto a soma destes desvios não seria nada informativa sobre a dispersão dos dados. Definiremos então, uma medida que utiliza o quadrado dos desvios em relação à média.

### 3.3.2 Variância e desvio-padrão

A variância de um conjunto de dados, é definida como média dos desvios das observações em relação à média ao quadrado, ou seja,

$$s^2 = \frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \ldots + (y_n - \bar{y})^2}{n-1}$$

Para manter a mesma unidade dos dados originais, é conveniente definirmos o *desvio-padrão* como sendo a raiz quadrada positiva da variância $s^2$,

$$s = \sqrt{\frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \ldots (y_n - \bar{y})^2}{n-1}}$$

A variância amostral é frequentemente calculada usando-se a fórmula mais rápida e prática

$$s^2 = \frac{1}{n-1}\left\{ y_1^2 + y_2^2 + \ldots + y_n^2 - \frac{(y_1 + y_2 + \ldots + y_n)^2}{n} \right\} =$$

$$= \frac{1}{n-1}\left\{ \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \right\}$$

Exemplo: Os pesos (em pounds) de uma amostra aleatória de trutas em um lago são:

1,19; 0,93; 2,40; 1,71; 0,89; 1,74; 1,06; 1,16; 1,47; 1,15

A média aritmética destes dados é

$$\bar{y} = \frac{1}{10}(1{,}19 + 0{,}93 + \ldots + 1{,}15) = \frac{13{,}7}{10} = 1{,}37 \ pounds \ .$$

E a variância é

$$s^2 = \frac{1}{10-1}\left\{ (1{,}19 - 1{,}37)^2 + (0{,}93 - 1{,}37)^2 + \ldots + (1{,}15 - 1{,}37)^2 \right\}$$

$$= 0{,}2187 \ \ (pounds)^2$$

Alternativamente, temos

$$s^2 = \frac{1}{10-1}\left\{ 1{,}19^2 + 0{,}93^2 + \ldots + 1{,}15^2 - \frac{(1{,}19 + 0{,}93 + \ldots + 1{,}15)^2}{10} \right\} =$$

$$= \frac{1}{9}\left\{ 20{,}74 - \frac{13{,}70}{10} \right\} = 0{,}2187 \ \ (pounds)^2, e$$

$$s = \sqrt{0{,}2187} = 0{,}47 \ pounds.$$

Script no R para os cálculos acima

```
> # entrando com os dados pelo comando concaternar c( )
> peso <- c(1.19, 0.93, 2.40, 1.71, 0.89, 1.74, 1.06, 1.16, 1.47, 1.15)

> # cálculo da média pela definição com os comandos sum() e length()
> m.peso1 <- sum(peso)/length(peso)
> m.peso1
[1] 1.37

> m.peso2 <- mean(peso)        > # cálculo da média pela função mean()
> m.peso2
[1] 1.37

 # mais detalhes da função mean() execute o comando ??mean()

# 3 formas de se calcular a variância pelas fórmulas do item 3.3.2
> v1.peso <- sum((peso-mean(peso))^2)/(length(peso)-1)
> v1.peso
[1] 0.2187111

> v2.peso <- (sum(peso^2)-sum(peso)^2/length(peso))/(length(peso)-1)
> v2.peso
[1] 0.2187111
```

```
> # cálculo pela função var( )
> v3.peso <- var(peso)
> v3.peso
[1] 0.2187111

> # cálculo do desvio padrão pela definição
> sd1.peso <- sqrt(v3.peso)
> sd1.peso
[1] 0.4676656
> # cálculo do desvio padrão pela função sd( )
> sd2.peso <- sd(peso)
> sd2.peso
[1] 0.4676656
```

### 3.3.3 Quartis

Alguns quartis são definidos de modo análogo à mediana. Assim como a mediana divide o conjunto de dados em duas partes, os quartis dividem os dados em quatro partes. O segundo quartil, representado por $Q_2$ é igual à mediana, então $Q_2 = \tilde{y}$. O primeiro quartil, $Q_1$ é definido como aquele valor do conjunto de dados tal que não mais que 25% dos dados têm valores menores que $Q_1$ e não mais que 75% dos dados têm valor maior que $Q_1$. O terceiro quartil, $Q_3$, pode ser definido de maneira similar. Assim como a mediana, mais de uma observação pode satisfazer a definição dos quartis. As seguintes fórmulas podem ser utilizadas para calcular o primeiro e o terceiro quartis de um conjunto de dados

$$Q_1 = \frac{n+1}{4} \quad \textit{ésima observação ordenada}$$

$$Q_3 = \frac{3(n+1)}{4} \textit{ ésima observação ordenada}$$

```
> q25<-quantile(peso,0.25)        # 1º quartil
> q25
   25%
1.0825
> q50<-quantile(peso,0.50)        # 2º quartil
> q50
  50%
1.175
> q75<-quantile(peso,0.75)        # 3º quartil
> q75
 75%
1.65
```

### 3.3.4 Gráfico "*BOX-PLOT*"

O gráfico tipo *Box-plot* é um recurso visual útil de comunicação da informação contida em conjunto de dados. O objetivo de um gráfico tipo *Box-Plot* é mostrar as principais características de um conjunto de dados. Para interpretar um gráfico *Box-Plot* adequadamente, os valores devem ser visualizados como pontos de linha horizontal/vertical localizada no centro do

gráfico. Valores grandes correspondem a grandes pontos na horizontal/vertical. Existem três componentes importantes no gráfico *Box-plot*:

- A caixa, a qual contém 50% dos valores, começa no primeiro quartil $Q_1$ e termina no terceiro quartil, $Q_3$.
- As duas pontas (*whiskers*), se extendem acima e abaixo da caixa até a localização da maior e da menor observação que estão dentro da distância de 1.5 vezes o intervalo interquartil.
- Os valores atípicos "*outliers*", são os valores fora das pontas.

Exemplo: Considere os dados a seguir, os quais se referem a peso (g) de tumores cancerígenos extraídos do abdome de 57 cães

68 63 42 27 30 36 28 32 79 27 22 23 24 25 44 65 43 25 74
51 36 42 28 31 28 25 45 12 57 51 12 32 49 38 42 27 31 50
38 21 16 24 69 47 23 22 43 27 49 28 23 19 46 30 43 49 12

O conjunto ordenado fica:

12 12 12 16 19 21 22 22 23 23 23 24 24 25 25 25 27 27 27
27 28 28 28 28 30 30 31 31 32 32 36 36 38 38 42 42 42 43
43 43 44 45 46 47 49 49 49 50 51 51 57 63 65 68 69 74 79

Assim, a menor e a maior observação é 12 e 79, respectivamente. O número de observações é 57. O primeiro *quartil* é a observação

$$Q_1 = \frac{57+1}{4} = 14.5 = y_{(14,5)} = 25 \text{ g},$$

e o terceiro *quartil*

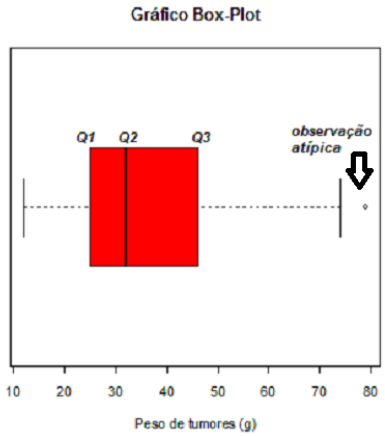
$$Q_3 = \frac{3(57+1)}{4} = 43.5 = y_{(43,5)} = 46{,}5 \text{ g}$$

Script no R para os cálculos acima

```
> # entrando com os dados
> p.tumor <- c(68, 63, 42, 27, 30, 36, 28, 32, 79, 27,
+                22, 23, 24, 25, 44, 65, 43, 25, 74, 51,
+                36, 42, 28, 31, 28, 25, 45, 12, 57, 51,
+                12, 32, 49, 38, 42, 27, 31, 50, 38, 21,
+                16, 24, 69, 47, 23, 22, 43, 27, 49, 28,
+                23, 19, 46, 30, 43, 49, 12)

> min.ptumor <- min(p.tumor)  # mínmo do vetor p.tumor
> min.ptumor
[1] 12

> max.ptumor <- max(p.tumor) # máximo do vetor p.tumor
> max.ptumor
[1] 79

> amplitude<-max.ptumor-min.ptumor   #  amplitude
```

```
> amplitude
[1] 67

> q.20 <- quantile(p.tumor,0.20)            # quantil 0.20
> q.20
20%
 24

> q1 <- quantile(p.tumor,0.25)              # quartil 0.25
> q1
25%
 25
 > q2 <- quantile(p.tumor,0.50)             # 2º quartil 0.50
> q2
50%
 32

> q3<- quantile(p.tumor,0.75)              # 3º quartil 0.75
> q3
75%
 46

> mediana<- median(p.tumor)                # mediana
> mediana
[1] 32
# reparem que a mediana é igual ao segundo quartil
# cálculo dos 3 quartis (0.25, 0.50, 0.75) de uma única vez
> quartis <- c(0.25,0.50,0.75)
> quantile(p.tumor,quartis)
25% 50% 75%
 25  32  46

> summary(p.tumor)                          # função summary( )
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  12.00   25.00   32.00   36.72   46.00   79.00

# 2 gráficos pela função boxplot()
> boxplot(p.tumor)  # gráfico default

# incrementando o gráfico
> boxplot(p.tumor,
+           col=2,                    # colocando cor no gráfico
+           horizontal= T,            # na posição horizontal
+           main= "Gráfico Box-Plot")  # colocando título principal
```

Gráfico produzido pela última função *boxplot( )*




O exame deste Gráfico revela que 50% das observações estão no retângulo entre os valores do Q1=25 e Q3=46,5. A linha vertical dentro da caixa representa o valor da mediana, $Q_2$, a qual é 32. A longa cauda a direita do gráfico indica que a distribuição de peso de tumores é levemente assimétrica à direita. O símbolo da bolinha indica que existe uma observação atípica neste conjunto de dados, observação cujo valor é 79, com uma probabilidade de ocorrência muito baixa.

---

### 3.3.5 Medidas da forma da distribuição

As medidas da forma de uma distribuição são os coeficientes de *assimetria* (skewness) e *curtosis* (kurtosis).

Assimetria é uma medida da assimetria da distribuição de freqüência. Ela mostra se os desvios da média são maiores de um lado do que do outro lado da distribuição. Ela é dada por

$$ass = \frac{n}{(n-1)\ (n-2)} \sum_{i=1}^{n} \left( \frac{y_i - \bar{y}}{s} \right)^3$$

Para uma distribuição simétrica o coeficiente de assimetria é *zero*. Ela é positiva quando a cauda da direita é mais alongada e negativa quando a cauda da esquerda é mais alongada.

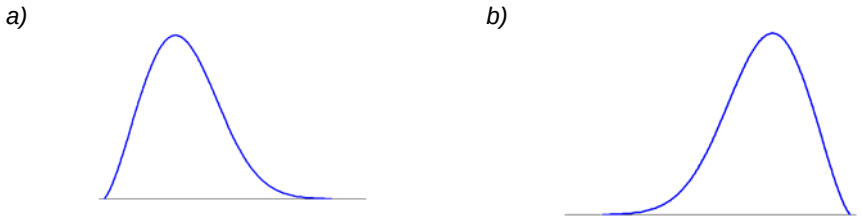



Figura 3.3 Ilustrações da assimetria a) negativa e b) positiva

*Curtosis* é uma medida da forma das caudas de uma distribuição. Ela é dada por

$$ct = \frac{n\ (n+1)}{(n-1)\ (n-2)\ (n-3)} \sum_{i=1}^{n} \left( \frac{y_i - \bar{y}}{s} \right)^4 - \frac{3\ (n-1)^2}{(n-2)\ (n-3)}$$

Para variáveis, tais como, peso, altura ou produção de leite, espera-se que a distribuição de freqüência seja simétrica em torno da média e tenha a forma de um sino. Estas são as distribuições normais. Se as observações têm distribuição normal então a curtosis é igual a zero ($ct = 0$). Uma distribuição com curtosis positiva tem uma grande freqüência de observações próximas da

média e caudas finas. Uma distribuição com curtosis negativa tem as caudas mais grossas e uma baixa freqüência de dados perto da média.

No R pode-se definir funções para executar tarefas que requer mais de um passo. Funções pode simplificar a digitação, organizar nossos pensamentos, e salvar o nosso trabalho para reutilização. No R, uma função tem um nome (usualmente), uma regra (o corpo da função), um amaneira de definir as entradas (os argumentos da função), e uma saída ( 0 último comando avaliado)

Funções no R são criadas com palavra-chave **function ( )** . Por exemplo: o script no R para os cálculos dos coeficientes de assimetria e curtosis, podem ser feitos por meio de funções, como se segue:

```
# definindo uma função para o cálculo do coef. de assimetria (item 3.3.5)
> ass<-function(x){                  # definindo a função
+ m3<-sum((x-mean(x))^3)            # inicio do corpo da função
+ s3<-sd(x)^3
+ n <- length(x)
+ coef<- n/((n-1)*(n-2))
+ coef*m3/s3 }                      # término do corpo da função

> ass(p.tumor)                      # saída da função ass( )
[1] 0.7612649

# definindo uma função para o cálculo do coef. de curtosis
> ct <-function(x) {
+ m4<-sum((x-mean(x))^4)            # inicio do corpo da função
+ s4<-sd(x)^4
+ n<-length(x)
+ coef1<-n*(n+1)/((n-1)*(n-2)*(n-3))
+ coef2<- 3*(n-1)^2/((n-2)*(n-3))
+ coef1*m4/s4 - coef2}              # término do corpo da função
> ct(p.tumor)                       # saída da função
[1] 0.1301841
```

A seguir definimos uma função **ed( )** a qual calcula grande parte das estatísticas descritivas de uma variável

```
> # definindo uma função ed( ) que calcula todas as estatísticas descritivas
> ed<-function (x) {                 # inicio da funçao
+ media<-mean(x)                    # cálculo da média
+ dp<-sd(x)                         # cálculo do desvio padrão
+ minimo<-min(x)                    # cálculo do mínimo
+ maximo<-max(x)                    # cálculo do máximo
+ q1<-quantile(x,0.25)              # cálculo do 1 quartil
+ mediana<-median(x)                # cálculo da mediana q2
+ q3<-quantile(x,0.75)              # cálculo do terceiro quartil
+ cv<-sd(x)/mean(x)*100             # cálculo do coef. variação

+ m3<-sum((x-mean(x))^3)            # cálculo do coef. de assimetria
+ s3<-sd(x)^3
```

```
+ n <- length(x)
+ coef<- n/((n-1)*(n-2))
+ ass<-coef*m3/s3

+ m4<-sum((x-mean(x))^4)          # cálculo do coef. curtosis
+ s4<-sd(x)^4
+ n<-length(x)
+ coef1<-n*(n+1)/((n-1)*(n-2)*(n-3))
+ coef2<- 3*(n-1)^2/((n-2)*(n-3))
+ ct<-coef1*m4/s4 - coef2

+ # definindo a saída
+ c(mínimo=minimo,Q1=q1,média=media,mediana=mediana,desv_pad=dp,
+ Q3=q3,máximo=maximo,CV=cv,Assimetria=ass,Curtosis=ct)
+ }                        # final da função ed( )

# aplicando a função ed( ) aos dados de p.tumor
round(ed(p.tumor),1)    # a função round( ) controla as casas decimais
```

Abaixo estão estas estatísticas calculadas pela função **ed( )** aplicada aos dados do tumor armazenados no objeto

```
mínimo    Q1.25%   mediana   desv_pad    Q3.75%    máximo        CV
    12.0      25.0      32.0      15.9      46.0      79.0      43.2
Assimetria   Curtosis
       0.8       0.1
```

Alternativamente,vários pacotes disponibilizam o cálculo das estatísticas descritivas, dentre estes destacamos o pacote **pastecs**, e dentro deste pacote temos a função **stat.dec ( )**. Uma aplicação é apresentada a seguir:

```
> library(pastecs)                    # requrendo o pacote pastecs
> # usando o commando round ( ) junto com stat.desc( )
> # para arredondar para 1 casa decimal
> round(stat.desc(p.tumor,basic=FALSE,norm=T),1)
     median        mean     SE.mean  CI.mean.0.95         var     std.dev
       32.0        36.7         2.1          4.2       251.7        15.9
   coef.var    skewness     skew.2SE    kurtosis    kurt.2SE    normtest.W
        0.4         0.7          1.1        -0.1        -0.1         0.9
normtest.p
        0.0
```

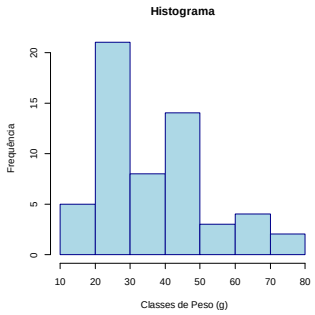
Para uma melhor compreensão desta função basta acionar a ajuda no R, digitando **?stat.dec( ),** ou **help(stat.dec)**.

### 3.3.6 **Histograma e Box-Plot**

O gráfico do histograma é outro recurso visual muito usado para a análise da forma da distribuição, enquanto que o gráfico Box-Plot mostra as cinco medidas resumo, as contém o mínimo, o 1º quartil (Q1), a mediana (Q2), o 3º Quartil e a observação máxima . No script do R abaixo são apresentados alguns exemplos da função **hist( )** e sua correspondência com o gráfico Box-Plot para os dados p.tumor.

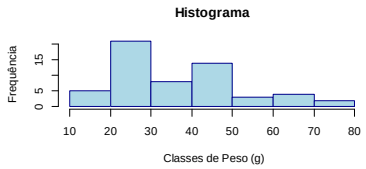
```
> hist(p.tumor)                          # gráfico default do histograma

# histograma com mais opções
> hist(p.tumor,
+      col="light blue",                 # colocando a cor azul
+      xlab=" Classes de Peso (g)",      # título do eixo x
+      ylab="Frequência",                # título do eixo y
+      nclass=8,                          # número de colunas
+      border="dark blue")                # colocando bordas no gráfico
```

Saída fornecida pelo script acima




Apresentação do histograma e do Box-Plot juntos em uma mesma janela gráfica

```
par(mfrow=c(2,1))        #dividindo a janela gráfica em 2 linhas e 1 coluna

# histograma
> hist(p.tumor,
+      col="light blue",                 # colocando a cor azul
+      xlab=" Classes de Peso (g)",      # título do eixo x
+      ylab="Frequência",                # título do eixo y
+      nclass=8,                          # número de colunas
+      border="dark blue",                # colocando bordas no gráfico
+      main="Histograma")                 # título principal

# gráfico box plot
> boxplot(p.tumor,
+           col=2,                       # colocando cor no gráfico
+           horizontal= T,               # na posição horizontal
+  main="Gráfico Box-Plot")              # colocando título principal
```



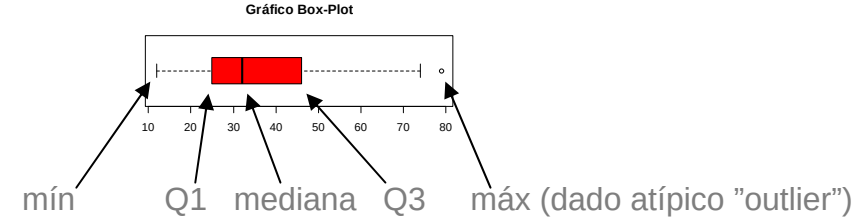
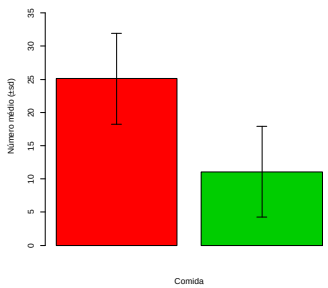




Ilustração das 5 medidas resumo ilustradas no gráfico Boxplot.

Gráficos para dados com uma classificação é uma ferramenta muito útil na ánalise exploratória de dados. Considere a questão nº 1 da 1ª Lista de exercícios, apresentada ao final da Aula 1. Nesta questão é solicitado a construção do gráfico de barras para cada tipo de comida. O script no R para construir estes gráficos é:

```
> # entrando com as observações por tipo de comida
> moscas.cr <- c(15,20,31,16,22,22,23,33,38,28,25,20,21,23,29,26,
                 40,20,19,31)
> moscas.su<-c(6,19,0,2,11,12,13,12,5,16,2,7,13,20,18,19,19,9,9,9)
>
> #calculando a média para cada tipo de comida
> media.cr<-mean(moscas.cr)
> media.cr
[1] 25.1
> media.su<-mean(moscas.su)
> media.su
[1] 11.05
>
> # reunindo as duas médias em um único vetor
> media<-c(media.cr,media.su)
> #calculando o desvio-padrão para cada tipo de comida
> dp.cr<-sd(moscas.cr)
> dp.cr
[1] 6.84336
> dp.su<-sd(moscas.su)
> dp.su
[1] 6.194012
> # reunindo os dois dp em um único vetor
> dp<-c(dp.cr,dp.cr)
> # gráfico de barras do valor médio de cada tipo de comida
> bar.moscas<-barplot(media, cex.names=0.7, xlab="Comida",col=c(2,3),
+ ylab="Número médio (±sd)", ylim=c(0,max(media+2*dp)))
>
> # colocando os eixos do desvio-padrão no gráfico de barras
> arrows(bar.moscas,media-dp, bar.moscas,media+dp, length=0.1,
angle=90, code=3)
```

O script acima fornecendo o seguinte gráfico:




Existem outros tipos de gráficos que ajudam a entender a distribuição dos dados. Uma forma útil de se fazer isto para um conjunto de dados é com o

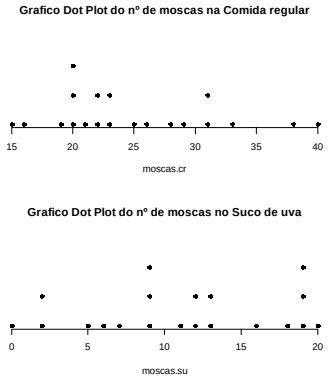
gráfico caule-e-folha (stem-and-leaf plot), o qual é uma maneira de codificar um conjunto de valores que minimiza a escrita e dá uma idéia de como a amplitude e a distribuição dos dados. Cada observação é representada de uma maneira bem compacta. A função que faz o gráfico caule-e-folha no R é a **stem( )**. Aplicando esta função aos dados das moscas temos:

```
> stem(moscas.cr)

  The decimal point is 1 digit(s) to the right of the |

  1 | 569
  2 | 000122335689
  3 | 1138
  4 | 0

> stem(moscas.su)

  The decimal point is 1 digit(s) to the right of the |

  0 | 022
  0 | 567999
  1 | 12233
  1 | 68999
  2 | 0
```

Um número como 16 será escrito como como **1** para o caule e **6** para a folha. Um alternativa ao gráfico caule-e-folha é o gráfico de pontos (**dot plot**), o qual é feito pela função DOTplot ( ) avaliável no pacote **UsingR**. Usando esta função nos dados das moscas temos:

```
> library(UsingR)
> par(mfrow=c(2,1))    # dividindo a janela gráfica em duas linas e 1 coluna
> DOTplot(moscas.cr,main="Grafico Dot Plot do nº de moscas na Comida regular")
> DOTplot(moscas.su,main="Grafico Dot Plot do nº de moscas no Suco de uva")
```

Cuja saída fornecida pelo R é




## 3.7 Coeficiente de variação *(CV)*

O desvio-padrão é útil como medida de variação dentro de um conjunto de dados. Quando desejamos comparar a dispersão de dois conjuntos de dados, a comparação dos desvios-padrões dos dois conjuntos de dados pode nos levar a conclusões falsas. Pode acontecer que as duas variáveis envolvidas estão medidas em unidades diferentes. Por exemplo, podemos estar interessados em saber se os níveis do soro de colesterol, medido em miligramas por 100 ml são mais variáveis do que o peso corporal, medido em kilograma.

O que é necessário nesta situação é o uso de uma medida de variação relativa do que uma medida absoluta. Tal medida é o ***COEFICIENTE DE VARIAÇÃO (CV)***, a qual expressa o desvio padrão como uma porcentagem da média, e sua fórmula é

$$cv = \frac{s}{\bar{y}}(100)\%,$$

a qual é uma medida independente da unidade.

Exemplo: considere os valores abaixo de média e desvio-padrão de dois grupo de cães, identificados pelas suas idades

| | *Amostra 1* | *Amostra 2* |
|---|---|---|
| *Grupo* | *10 anos* | *4 anos* |
| *Peso médio* | *145* | *80* |
| *Desvio-padrão* | *10* | *10* |

Uma comparação dos seus respectivos desvios-padrões leva a uma conclusão de que as duas amostras têm a mesma variabilidade. Se calcularmos os coeficientes de variação, para o grupo 1

$$cv = \frac{10}{145}(100) = 6{,}9\%$$

e para o grupo 2,

$$cv = \frac{10}{80}(100) = 12{,}5\%,$$

e comparando estes resultados temos uma impressão bem diferente. O grupo 2 tem uma variabilidade de 1,8 vezes maior em relação ao grupo 1. O coeficiente de variação é muito útil na comparação de resultados obtidos por diferentes pesquisadores que investigam a mesma variável. Visto que o coeficiente de variação é independente da unidade, ele é útil para comparar a variabilidade de duas ou mais variáveis medidas em diferentes unidades. No R é necessário definir uma função para o cálculo do CV. Nos dados das moscas temos:

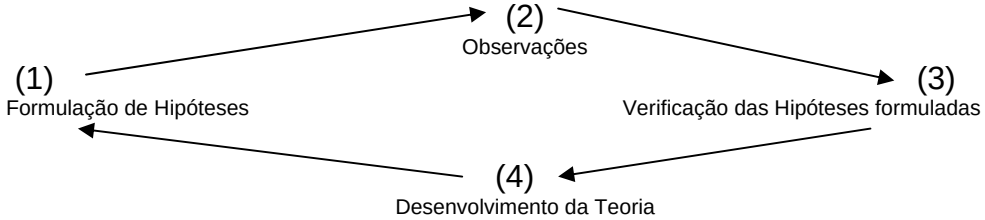
```
> # definindo uma função para o cálculo do coeficiente de variação
> cv <- function(x) sd(x)/mean(x)*100
>
> # aplicando a função aos dados das moscas temos
> cv(moscas.cr)
[1] 27.26438
> cv(moscas.su)
[1] 56.05441
```

Assim, pelos resultados acima vemos que o conjunto de dados do nº de moscas da comida regular tem menor variabilidade do que o conjunto do nº de moscas do suco de uva.

# 4.ESTATÍSTICA EXPERIMENTAL

## 4.1 INTRODUÇÃO

Numa pesquisa científica o procedimento geral é formular hipóteses e verificá-las diretamente ou por suas conseqüências. Para isto é necessário um conjunto de observações e o planejamento de experimentos é então essencial para indicar o esquema sob o qual as hipóteses possam ser verificadas com a utilização de métodos de análise estatística que dependem da maneira sob a qual as observações foram obtidas. Portanto, planejamento de experimentos e análises dos resultados estão intimamente ligados e devem ser utilizados em uma seqüência nas pesquisas científicas das diversas áreas do conhecimento. Isto pode ser visto por meio da seguinte representação gráfica da circularidade do método científico.




Fica evidente nesta ilustração que as técnicas de planejamento devem ser utilizadas entre as etapas (1) e (2) e os métodos de análise estatística devem ser utilizados na etapa (3).

Desenvolvendo um pouco mais está idéia podemos dizer que uma pesquisa científica estatisticamente planejada consiste nas seguintes etapas:

1. Enunciado do problema com formulação de hipóteses.
2. Escolha dos fatores (variáveis independentes) que devem ser incluídos no estudo.
3. Escolha da unidade experimental e da unidade de observação.
4. Escolha das variáveis que serão medidas nas unidades de observação.
5. Determinação das regras e procedimentos pelos quais os diferentes tratamentos são atribuídos às unidades experimentais (ou vice-versa).
6. Análise estatística dos resultados.
7. Relatório final contendo conclusões com medidas de precisão das estimativas, interpretação dos resultados com possível referência a outras pesquisas similares e uma avaliação dos itens de 1 a 6 (desta pesquisa) com sugestões para possíveis alterações em pesquisas futuras.

Ilustrações destas etapas com exemplos.

1. Enunciado do problema.

Como vimos uma pesquisa científica se inicia sempre com a formulação de hipóteses. Essas hipóteses são primeiramente formuladas em termos científicos dentro da área de estudo (hipótese científica) e em seguida em termos estatísticos (hipótese estatística) . Deve haver uma correspondência perfeita entre as hipóteses científica e estatística para evitar ambigüidade.

Portanto, no enunciado do problema, a hipótese científica deve ser formulada de maneira precisa e objetiva.

Exemplo:Um pesquisador está interessado em estudar o efeito de vários tipos de ração que diferem pela quantidade de potássio no ganho de peso de determinado tipo de animal.

Este objetivo pode ser atingido se planejarmos a pesquisa com uma das seguintes finalidades:

*a) comparar as médias dos aumentos de peso obtidas com cada uma das rações (igualdade das médias);*

b) *Estabelecer uma relação funcional entre o aumento do peso médio e a quantidade de potássio.*

2. Escolha dos fatores e seus respectivos níveis.

No exemplo apresentado em 2.1, a variável independente “ração” é um fator e os tipos de rações são os níveis deste fator, ou tratamentos. Assim, em um experimento para se estudar o efeito de 4 rações e 3 suplementos no ganho de peso de animais, temos dois fatores: ração com quatro níveis e suplementos com 3 níveis. Podemos dizer que este experimento envolve 12 tratamentos, correspondentes às combinações dos níveis dos dois fatores.

Pelo próprio conceito de fator, temos que em um experimento, a escolha dos fatores e seus respectivos níveis é basicamente um problema do pesquisador. No entanto é importante para o planejamento e análise distinguirmos as duas situações, descritas a seguir:

a) uma fazenda de inseminação adquiriu 5 touros de uma determinada raça para a produção de sêmen, e está interessada em realizar um experimento para verificar se os cinco touros são homogêneos quanto a produção de sêmen.

b) A mesma fazenda de inseminação está interessada em realizar um experimento para verificar se a produção de sêmen de touros, de uma determinada raça, é homogênea. Como a população de touros da fazenda é muito grande o pesquisador decidiu realizar um experimento com uma amostra de touros (5 touros), mas as conclusões devem ser estendidas para a população de touros.

Na situação descrita em a) dizemos que o fator “touro” é fixo e na situação em b) o fator “touro” é aleatório. A diferença fundamental entre estes dois tipos de fatores é, então, que no caso de fatores fixos, as conclusões se referem apenas aos níveis do fator que estão presentes no experimento. No caso de fatores aleatórios as conclusões devem ser estendidas para a população de níveis.

3. Escolha da unidade experimental.

Em um grande número de situações práticas a unidade experimental é determinada pela própria natureza do material experimental. Por exemplo, experimentos com animais, em geral a unidade experimental é um animal. Em outras situações a escolha de outras unidades experimentais não é tão evidente, exigindo do pesquisador juntamente com o estatístico algum estudo, no sentido de escolher a unidade experimental mais adequada. A escolha de uma unidade experimental, de um modo geral, deve ser orientada no sentido de minimizar o erro experimental, isto é, as unidades devem ser as mais homogêneas possíveis, para, quando submetidas a dois tratamentos diferentes, seus efeitos, sejam facilmente detectados.

4. Escolha das variáveis a serem medidas.

As medidas realizadas nas unidades experimentais após terem sido submetidas aos tratamentos constituem os valores da variável dependente. A variável dependente, em geral, é pré-determinada pelo pesquisador, isto é, ele sabe qual variável que ele quer medir. O que constitui problema, às vezes, é a maneira como a variável é medida, pois disto dependem a precisão das observações, e a distribuição de probabilidade da variável a qual é essencial para a escolha do método de análise. Assim, por exemplo, se os valores de uma variável são obtidos diretamente por meio de um aparelho de medida (régua, termômetro, etc.) a precisão das observações vai aumentar se, quando possível, utilizarmos como observação a média de três medidas da mesma unidade experimental. Com relação à distribuição de probabilidade em muitas situações as observações não são obtidas diretamente e sim por expressões matemáticas que as ligam a outros valores obtidos diretamente. Neste caso, a distribuição de probabilidade das observações vai depender da distribuição de probabilidade da variável obtida diretamente e da expressão matemática que as relaciona.

Portanto, as variáveis, necessariamente presentes em um experimento são: a variável dependente, medida nas unidades experimentais, e o conjunto de fatores (variáveis independentes) que determinam as condições sob as quais os valores da variável dependente são obtidos.

Qualquer outra variável que possa influir nos valores da variável dependente deve ser mantida constante.

5. Regras segundo as quais os tratamentos são atribuídos às unidades experimentais.

Nas discussões apresentadas em cada um dos itens anteriores a colaboração da estatística é bem limitada exigindo-se a essencial colaboração do pesquisador. Porém, o assunto discutido neste item é o que poderíamos denominar de planejamento estatístico de experimento. Trata-se das regras que associam as unidades experimentais aos tratamentos e que praticamente determinam os diferentes planos experimentais, ou seja, a ***Aleatorização*** ou ***Casualização***. Lembramos, neste ponto, que os tratamentos são cada uma das combinações entre os níveis de todos os fatores envolvidos no experimento.

Para que a metodologia estatística possa ser aplicada aos resultados de um experimento é necessário que em alguma fase do experimento, o principio a ser obedecido é o da ***Repetição***, segundo o qual devemos ter repetições do experimento para que possamos ter uma medida da variabilidade necessária aos testes da presença de efeitos de tratamentos ou a estimação desses efeitos.

**Aleatorização**

Aleatorização é a designação dos tratamentos às unidades experimentais, tal que estas têm a mesma chance (mesma probabilidade) de receber um tratamento. Sua função é assegurar estimativas *não-viesadas* das médias dos tratamentos e do erro experimental. Nesta fase do planejamento de um experimento já sabemos quais fatores serão estudados e o número de níveis de cada fator que estarão presentes no experimento. Sabemos ainda qual é a unidade experimental escolhida e a variável dependente. Podemos imaginar que de um lado temos um conjunto

- **U** de unidades experimentais, e de outro,
- **T** um conjunto de tratamentos, que podem ser as combinações dos níveis de todos os fatores envolvidos. Precisamos estabelecer esquemas que associam subconjuntos de elementos de **U** a cada elemento de **T**. Vamos apresentar o esquema mais simples. Para efeito de notação vamos supor que o conjunto **U** tem **n** elementos, o conjunto **T** tem **a** elementos, e o número de elementos de **U** submetidos ao tratamento $\mathbf{T_i}$ é $n_i$, com i=1, 2, ..., a, de tal modo que $\sum_{i=1}^{k} n_i = n$ .

O número de unidades experimentais $n_i$ para cada tratamento $\mathbf{T_i}$ é determinado a partir de informações sobre a variabilidade das unidades experimentais em termos da variabilidade da variável dependente.

O **plano *completamente aleatorizado*** é um esquema em que as unidades experimentais que vão ser submetidas a cada tratamento são escolhidas completamente ao acaso. Isto significa que cada unidade experimental tem igual probabilidade de receber qualquer um dos tratamentos.

Por exemplo, um pesquisador quer realizar um experimento para estudar o efeito de um resíduo industrial que é adicionado em rações de animais. Ele suspeita que este resíduo contenha uma substância tóxica, cuja presença no organismo, produz um aumento relativo de alguns órgãos, como o fígado, por exemplo. Após uma entrevista com o pesquisador conseguimos as seguintes informações

- O experimento irá envolver um único fator, ração, com três níveis: $t_1$ - ração normal, sem resíduo industrial (grupo controle); $t_2$ - ração normal com o resíduo tratado, e $t_3$ - ração normal com resíduo não tratado. Portanto, o conjunto T tem três tratamentos
- Um conjunto **U,** é formado por um grupo de 18 camundongos todos, recém nascidos, com o mesmo peso inicial e homogêneos com relação às características genéticas gerais. Por isto foi decidido distribuir completamente ao acaso 6 animais para cada tratamento.
- A variável dependente (resposta) é o peso relativo do fígado após 90 dias do início do experimento.

Uma maneira de se proceder ao sorteio é a seguinte:

- enumera-se as unidades experimentais de 1 a 18.
- coloca-se os tratamentos em seqüência , por exemplo:

  $T_1$ $T_1$ $T_1$ $T_1$ $T_1$ $T_1$ , $T_2$ $T_2$ $T_2$ $T_2$ $T_2$ $T_2$ , $T_3$ $T_3$ $T_3$ $T_3$ $T_3$ $T_3$
- sorteia-se uma sequência de 18 números aleatórios. Pode-se obter, por exemplo, a sequência :

  3, 1, 11, 15, 18, 16, 4, 5, 9, 12, 8, 7, 17, 14, 2, 6, 13, 10

Gerando uma seqüência de números aleatórios no R:

```
> # gerando uma sequencia de números de 1 a 18
> x<-seq(1:18)
> x
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18
>
> # sequencia aleatória de tamanho 18 de x
```

```
> sample(x,18,replace=F)
 [1]  4  5  7 10 17  8 16  3  2 14 11 13  9  1 18  6 15 12
```

- Distribuição das unidades experimentais aos tratamentos de acordo com a seqüência gerada no R

| *Trat.* | *Repetições* | | | | | |
|---|---|---|---|---|---|---|
| $T_1$ | $u_4$ | $u_5$ | $u_7$ | $u_{10}$ | $u_{17}$ | $u_8$ |
| $T_2$ | $u_{16}$ | $u_3$ | $u_2$ | $u_{14}$ | $u_{11}$ | $u_{13}$ |
| $T_3$ | $u_9$ | $u_1$ | $u_{18}$ | $u_6$ | $u_{15}$ | $u_{12}$ |

Este plano experimental é mais eficiente quanto maior for o grau de homogeneidade entre as unidades experimentais em termos da variável dependente. Se as unidades experimentais são heterogêneas o número n de unidades experimentais necessárias para uma boa precisão pode ser muito grande. Algumas alterações no planejamento descrito, tal como, a introdução de blocos, ou simplesmente a utilização de uma co-variável medida nas unidades experimentais, a qual é correlacionada com à variável dependente, podem reduzir consideravelmente o erro experimental.

Observações:

1) o plano experimental completamente aleatorizado não depende do numero de fatores envolvidos e nem da maneira pela qual os fatores são combinados.
2) Existem alguns fatores que pela própria natureza, impõe restrições na aleatorização, porém para efeito de análise, o experimento é considerado completamente aleatorizado.

**Plano experimental em blocos**

Quando o conjunto **U** de unidades experimentais for muito heterogêneo (em termos da variável independente), o plano experimental completamente aleatorizado torna-se pouco preciso, pois o erro experimental fica muito grande. Em algumas situações dispomos de informações segundo as quais, antes da realização do experimento, é possível agruparmos as unidades experimentais mais ou menos homogêneas, em que **a** é o número de tratamentos envolvidos no experimento. Estes subconjuntos são denominados de blocos. Assim, a maior parte da heterogeneidade interna do conjunto **U** é expressa pela heterogeneidade entre blocos. A distribuição das unidades experimentais entre os tratamentos obedece a uma restrição imposta pelos blocos, isto é, as **a** unidades de cada bloco são distribuídas aleatoriamente entre os tratamentos.

Na análise de um experimento em blocos, além dos fatores de interesse, deve-se levar em conta o fator experimental bloco, diminuindo desta forma o erro experimental. Quanto maior for a heterogeneidade entre blocos, maior é a eficiência deste plano experimental em relação ao completamente aleatorizado.

Exemplo: Um pesquisador deseja testar o efeito de três tratamentos ($T_1$, $T_2$, $T_3$ ) no ganho de peso de ovelhas . Antes do inicio do experimento as ovelhas foram pesadas e ordenadas de acordo com o peso e atribuídas a 4 blocos. Em cada bloco tinham 3 animais aos quais os tratamentos foram sorteados. Portanto, 12 animais foram usados.

**Repetição**

Repetição significa que o mesmo tratamento é aplicado sobre duas ou mais unidades experimentais. Sua função é fornecer uma estimativa do “erro experimental” e dar uma medida mais precisa dos efeitos dos tratamentos. O

número de repetições requeridas em um particular experimento depende da magnitude das diferenças que o pesquisador deseja testar e da variabilidade da variável dependente em que se esta trabalhando.

# 1º EXERCÍCIO PRÁTICO ESTATÍSTICA EXPERIMENTAL

1) Em um estudo genético, uma alimentação regular era colocada em 20 frascos e o número moscas de um determinado genótipo era contado em cada frasco. O número de moscas também era contado em outros 20 frascos que continham suco de uva. O número de moscas contados foram:

| Número de moscas | |
|---|---|
| Comida regular | Suco de uva |
| 15 20 31 16 22 22 23 33 38 28<br>25 20 21 23 29 26 40 20 19 31 | 6 19 0 2 11 12 13 12 5 16<br>2 7 13 20 18 19 19 9 9 9 |

(a) Calcule a média amostral, a variância amostral, o desvio padrão amostral e o coeficiente de variação de cada conjunto de dados. Comente. Qual destes dois conjuntos de dados tem maior variabilidade?

(b) Calcule a média amostral, a variância amostral, o desvio padrão amostral de cada conjunto de dados utilizando os recursos imediatos de sua calculadora.

(c) Para cada conjunto de dados utilize o R para calcular a média, a mediana, o $Q_1$, o $Q_3$, a observações mínima e máxima, construa os gráficos do Histograma, do Box-Plot, do caule-e-folha (stem-plot) , do Dot-plot e o gráfico de barras com os desvio-padrões para cada tipo de comida. Comente os resultados.

2) Demonstre sua familiaridade com a notação da somatória, desdobrando-as e calculando as seguintes expressões com

$x_1 = 1$, $x_2 = -2$, $x_3 = 4$, e $x_4 = 5$:

Dica para o item (a) $\sum_{i=1}^{4} x_i = x_1 + x_2 + x_3 + x_4 = 1 + (-2) + 4 + 5 = 8$

(a) $\sum_{i=1}^{4} x_i$ (b) $\sum_{i=1}^{4} 4x_i$ (c) $\sum_{i=1}^{4} (x_i - 3)$ (d) $\sum_{i=2}^{4} (x_i - 4)$ (e) $\sum_{i=1}^{3} (x_i - 4)^2$

(f) $\sum_{i=1}^{4} x_i^2$ (g) $\sum_{i=1}^{4} (x_i - 2)^2$ (h) $\sum_{i=1}^{4} (x_i^2 - 4x_i + 4)$

3) Uma observação qualquer do conjunto de dados abaixo pode ser representada por $y_{ij}$ , com o índice ***i=1, 2, 3*** controlando as linhas e ***j=1, 2, 3, 4, 5, 6*** controlando as colunas. Por exemplo, $y_{23} = 100$. Calcule as seguintes expressões (fazendo o desdobramento):

a) $\sum_{i=1}^{3} y_{i2}$ b) $\sum_{j=1}^{6} y_{2j}$ c) $\sum_{i=1}^{3}\sum_{j=1}^{6} y_{ij}$ d) $\sum_{i=1}^{3}\sum_{j=1}^{6} y_{ij}^2$ e) $\left(\sum_{i=1}^{3}\sum_{j=1}^{6} y_{ij}\right)^2$

| | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $C_6$ |
|---|---|---|---|---|---|---|
| $L_1$ | 550 | 950 | 950 | 750 | 650 | 700 |
| $L_2$ | 350 | 500 | 100 | 550 | 350 | 350 |
| $L_3$ | 600 | 450 | 150 | 500 | 100 | 250 |

4) Os dados a seguir referem-se ao nível de glicose em sangue de 10 cães

56 62 63 65 65 65 65 68 70 72

Calcule manualmente e depois utilize o R para calcular: a) média; b) a mediana; c) mínimo e máximo; d) os quartis $Q_1$ e $Q_3$. Construa o histograma e gráfico tipo Box – Plot. Comente a respeito da dispersão dos dados.

7) Determinações de açúcar no sangue ( mg/ 100ml ) foram feitas em 5 raças de animais experimentais, sendo 10 amostras por raça. Os resultados foram:

| *Raças* | | | | |
|---|---|---|---|---|
| *A* | *B* | *C* | *D* | *E* |
| 124 | 111 | 117 | 104 | 142 |
| 116 | 101 | 142 | 128 | 139 |
| 101 | 130 | 121 | 130 | 133 |
| 118 | 108 | 123 | 103 | 120 |
| 118 | 127 | 121 | 121 | 127 |
| 120 | 129 | 148 | 119 | 149 |
| 110 | 122 | 141 | 106 | 150 |
| 127 | 103 | 122 | 107 | 149 |
| 106 | 122 | 139 | 107 | 120 |
| 130 | 127 | 125 | 115 | 116 |

Utilize o R para calcular para cada raça: a) média; b) a mediana; c) desvio padrão; d) o erro padrão; e) mínimo e máximo; f) os quartis $Q_1$ e $Q_3$. Construa o histograma, o gráfico tipo Box-Plot e o gráfico de barras para cada raça. Comente a respeito da dispersão dos dados em cada raça.

**Somatório e Algebrismo**

c) Seja Y a variável tempo de recuperação da anestesia de tilápias, com 10 observações:

Y = { 17,0; 8,9; 28,7; 20,5; 8,9; 26,1; 43,9 }

Calcular passo-a-passo:

a) $\sum_{i=1}^{7} y_i$ b) $\frac{\sum_{i=1}^{7} y_i}{7}$ c) Quadrado da Soma $\left(\sum_{i=1}^{7} y_i\right)^2$;

d) Soma de Quadrados $\sum_{i=1}^{7} y_i^2$; e) Suponha k = 15, calcule $\sum_{i=1}^{7} k y_i$;

f) Considerando-se $\bar{Y}$ como uma constante, desenvolva algebricamente o seguinte quadrado:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2, \text{ lembre-se que } \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

g) Reescreva a expressão $s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (y_i - \bar{y})^2$ em função do desenvolvimento do item f.

h) Considere a variável X tempo (segundos) de indução da anestesia para as mesmas 7 tilápias, respectivamente:

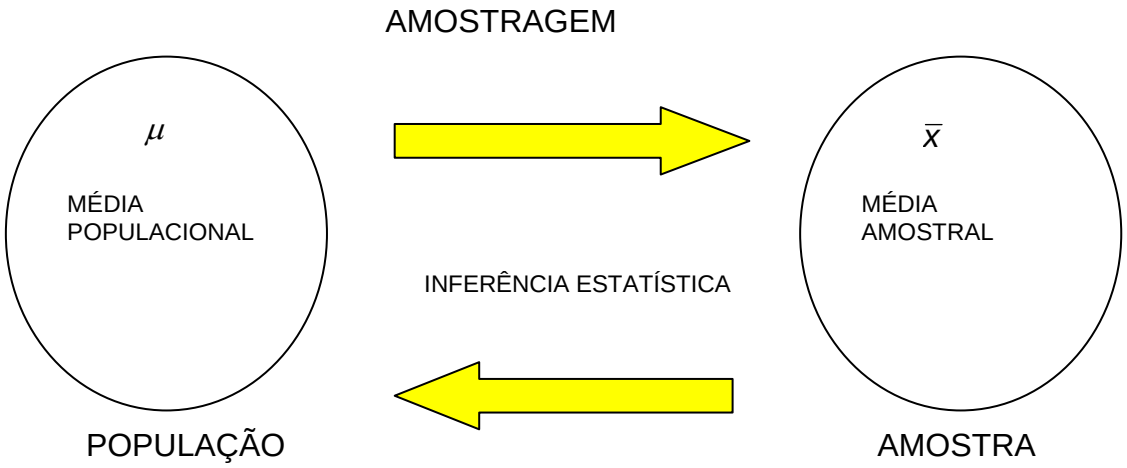
X = {165; 183; 161; 147; 146; 152; 174}

Calcule: $\sum_{i=1}^{7} x_i y_i$

**Aula 2 – Testes de significância**

**1 Introdução**

Um dos principais objetivos da estatística é a tomada de decisões a respeito de parâmetros da população com base nas observações de amostras.




Ao tomarmos decisões, é conveniente a formulação de Hipóteses relativas às populações, as quais podem ser ou não verdadeiras.

Exemplo: Um veterinário está interessado em estudar o efeito de 4 tipos de rações que diferem pela quantidade de potássio no aumento de peso de coelhos.

*$H_0$ : Não existe diferença entre as rações, ou seja, quaisquer diferenças observadas são devidas a fatores não controlados*

*$H_1$ : As rações propiciam aumentos de pesos dist **int** os*

***$H_0$*** é denominada de hipótese de nulidade, a qual assume que não existe efeito dos tratamentos e ***$H_1$*** é a contra hipótese.

**Testes de hipóteses ou testes de significância**

São os processos que nos permitem decidir se aceitamos ou rejeitamos uma determinada hipótese, ou se os valores observados na amostra diferem significativamente dos valores esperados (População)

**2 Tipos de erros nos testes de significância**

QUADRO RESUMO: condições sobre as quais os erros Tipo I e Tipo II podem ser cometidas

| | | **Condição da Hi pótese nula** | |
|---|---|---|---|
| | | **$H_0$ Verdadeiro** | **$H_0$ Falsa** |
| **Possível ação** | **Rejeição de $H_0$** | **Erro Tipo I (α)** | **Decisão correta** |
| | **Não rejeição de $H_0$** | **Decisão correta** | **Erro Tipo II (β)** |

**Erro Tipo I**: é o erro cometido ao rejeitar $H_0$, quando $H_0$ é verdadeira.
**Erro Tipo II**: é o erro cometido ao aceitar $H_0$, quando ela é falsa. E

$\alpha = P[ErroTipoI]$; *e* $\beta = P[ErroTipoII]$

Esses dois erros estão de tal forma associados que, se diminuirmos a probabilidade de ocorrência de um deles, automaticamente aumentamos a probabilidade de ocorrência do outro. Em geral, controlamos somente o *Erro Tipo I*, por meio do **nível de significância** (daí vem a denominação de Testes de Significância) do teste representado por $\alpha$ , o qual é a probabilidade máxima com que nos sujeitamos a correr um risco de cometer um erro do Tipo I, ao testar a hipótese $H_0$. Dado que rejeitar uma hipótese nula, $(H_0)$, verdadeira constitui um erro, parece razoável fixarmos esta probabilidade de rejeitar uma hipótese nula, $(H_0)$, verdadeira pequena, e de fato, é isto que é feito. Na prática é comum fixarmos $\alpha$ = 0,05 (5%) ou $\alpha$ = 0,01 (1%).

Se, por exemplo, foi escolhido $\alpha$ = 0,05, isto indica que temos 5 possibilidades em 100 de rejeitarmos a hipótese de nulidade ($H_0$), quando na verdade ela deveria ser não rejeitada, ou seja, existe uma confiança de 95% de que tenhamos tomado uma decisão correta, esta confiabilidade é denominada grau de confiança do teste e é representada por $\boldsymbol{1-\alpha}$ e expressa em porcentagem. Nunca saberemos qual tipo de erro estamos cometendo ao rejeitarmos ou ao não rejeitarmos uma hipótese nula ($H_0$), dado que a verdadeira condição é desconhecida. Se o teste nos leva à decisão de rejeitar $H_0$, podemos ficar tranqüilos pelo fato de que fizemos $\alpha$ pequeno e, portanto, a probabilidade de cometer o erro Tipo I é bem pequena.

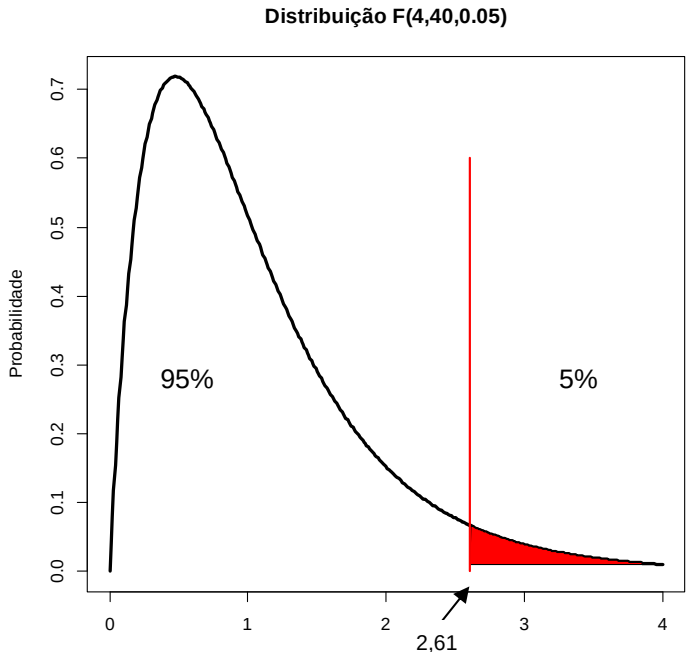
**3 Teste F para a Análise de Variância (ANOVA)**

O teste F é a razão entre duas variâncias e é usado para determinar se duas estimativas independentes da variância podem ser assumidas como estimativas da mesma variância. Na análise de variância, o teste F é usado para testar a igualdade de médias, isto é, para responder a seguinte questão, é razoável supor que as médias dos tratamentos são amostras provenientes de populações com médias iguais? Considere o seguinte exemplo de cálculo da estatística F; vamos supor que de uma população normal $N(\mu,\sigma^2)$ foram retiradas, aleatoriamente, 5 (n=5) amostras de tamanho 9 (r=9).

- Calcule as médias das 5 amostras e $s_i^2 = \dfrac{\sum_{i=1}^{9} (y_i - \bar{y})^2}{(9-1)}$
- Estime $\sigma^2$ por meio da fórmula $s^2 = \dfrac{(s_1^2 + \ldots + s_5^2)}{5}$, a qual é uma média das variâncias das amostras e será denominada de variabilidade dentro das amostras ($s_D^2$).
- Estime a variância populacional das médias $\sigma_{\bar{y}}^2$, por meio das médias das 5 amostras: $s_{\bar{y}}^2 = \dfrac{\sum_{i=1}^{5} (\bar{y}_{i+}^2 - \bar{y}_{++})^2}{5-1}$
- De $s_{\bar{y}}^2$, estime novamente $\sigma^2$, usando a relação $s_{\bar{y}}^2 = \dfrac{s^2}{r}$, *ou* $s^2 = r s_{\bar{y}}^2$ , denominada de variabilidade entre as amostras ($s_E^2$).

- Calcule $F_c = \frac{s_E^2}{s_D^2}$

A estimativa de $s_E^2$ do numerador foi feita com base em n - 1 = 4 graus de liberdade (n é o número de amostras) e a estimativa de $s_D^2$ do denominador foi feita com base em *n(r – 1) = 5(9-1) = 40*. A repetição deste procedimento amostral muitas vezes gera uma população de valores de F, os quais quando colocados em um gráfico de distribuição de freqüência tem o seguinte formato




O valor de F = 2,61 é o valor acima do qual, 5% dos valores de F calculados têm valor acima dele. Este é o valor para um $\alpha = 5\%$ encontrado na Tabela F para 4 e 40 graus de liberdade (veja Tabela F). Dado que as estimativas da variância utilizadas no cálculo da estatística F são estimativas da mesma variância $\sigma^2$ **,** espera-se que o valor de F seja bem próximo de 1, a menos que um conjunto de amostras não usual foi retirado. Para qualquer conjunto de amostras retiradas de n = 5 e r = 9 a probabilidade (ou a chance) de um valor de F calculado ser maior ou igual a 2,61 é 0,05 (5%) ($P[F > 2{,}61] = 0{,}05$ ).

As hipóteses estatísticas que testamos quando aplicamos o teste F são

$$H_0 : \sigma_1^2 = \sigma_2^2 \qquad \text{ou} \qquad H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$$
$$H_1 : \sigma_1^2 > \sigma_2^2 \qquad\qquad H_1 : \frac{\sigma_1^2}{\sigma_2^2} > 1$$

A hipótese $H_0$ estabelece que as duas variâncias populacionais são iguais, o que equivale a admitir que as amostras foram retiradas da mesma população. A hipótese $H_1$ (contra hipótese, ou hipótese alternativa) estabelece que as variâncias são provenientes de populações diferentes e, mais ainda, a variância da primeira é maior que a variância da segunda. Os valores de F são tabelados em função dos graus de liberdade das estimativas de $\boldsymbol{s^2}$ do

numerador ($n_1$) e do denominador ($n_2$) no cálculo da estatística F e para diferentes valores de níveis de significância (*5%, 1%*, etc.). Também podem ser fornecidos por comandos do programa **R**. Os valores teóricos da distribuição F podem ser facilmente obtidos no **R.** Por exemplo, os valores de

$$F_{(0,05,4,40)} \; ; F_{(0,01,4,40)},$$

são obtidos com os seguintes comandos:

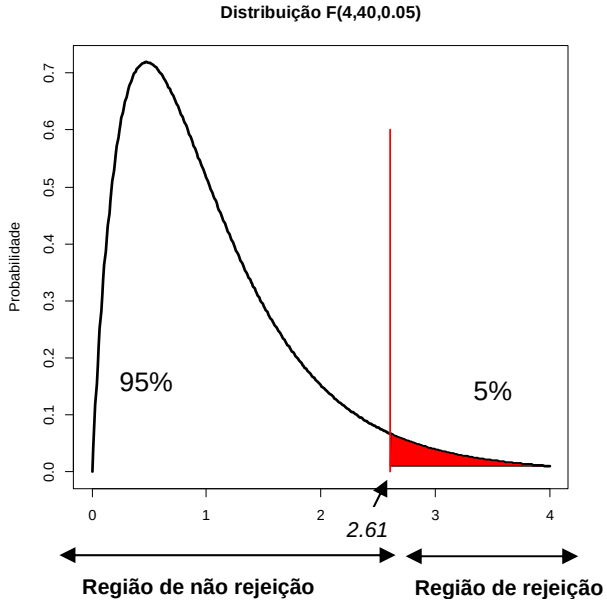
```
> # valores teóricos das distribuições F(0,05,4,40)
> qf(1-0.05,4,40)
[1] 2.605975
> qf(1-0.01,4,40)
[1] 3.828294
```

(fazer os gráficos destas distribuições com estes valores).

**Regra de decisão para o teste da estatística F**

Todos os possíveis valores que o teste estatístico pode assumir são pontos no eixo horizontal do gráfico da distribuição do teste estatístico e é dividido em duas regiões; uma região constitui o que denominamos de *região de rejeição* e a outra região constitui o que denominamos de *região de não rejeição*. Os valores do teste estatístico que formam a *região de rejeição* são aqueles valores menos prováveis de ocorrer se a hipótese nula é verdadeira, enquanto que os valores da *região de aceitação* são os mais prováveis de ocorrer se a hipótese nula é verdadeira. *A regra de decisão nos diz para* ***rejeitar $H_0$*** *se o valor do teste estatístico calculado da amostra é um dos valores que está na região de rejeição e para* ***não rejeitar $H_0$*** *se o valor calculado do teste estatístico é um dos valores que está na região de não rejeição.* O procedimento usual de teste de hipóteses é baseado na adoção de um critério ou regra de decisão, de tal modo que $\alpha$ *= P(Erro tipo I)* não exceda um valor pré-fixado. Porém, na maioria das vezes, a escolha de $\alpha$ é arbitrária. Um procedimento alternativo consiste em calcular o *"menor nível de significância para o qual a hipótese $H_0$ é rejeitada, com base nos resultados amostrais"*. Este valor, denominado de *nível descritivo do teste ou nível mínimo de significância do teste,* será denotado por ***valor de p*** ( "***p-value***"). Todos os modernos programas computacionais fornecem este valor nos testes estatísticos.

A representação gráfica a seguir mostra uma ilustração da regra de decisão do teste F, visto anteriormente,




Outro exemplo: Amostras aleatórias simples e independentes do nível de glicose no plasma de ratos após uma experiência traumática forneceram os seguintes resultados para dois tipos de esforços:

Esforço 1: 54 99 105 46 70 87 55 58 139 91

Esforço 2: 93 91 93 150 80 104 128 83 88 95 94 97

Estes dados fornecem suficiente evidência para indicar que a variância é maior na população de ratos submetidos ao esforço 1 do que nos ratos submetidos ao esforço 2. Quais as suposições necessárias para se aplicar o teste?

Solução:

- As variâncias amostrais são $s_1^2 = 852,9333$ e $s_2^2 = 398,2424$, respectivamente.
- Suposições: Os dados constituem amostras aleatórias independentes retiradas, cada uma, de uma população com distribuição normal. *(Esta é a suposição geral que deve ser encontrada para que o teste seja válido).*
- Hipóteses estatísticas

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$$

$$H_1 : \frac{\sigma_1^2}{\sigma_2^2} > 1$$

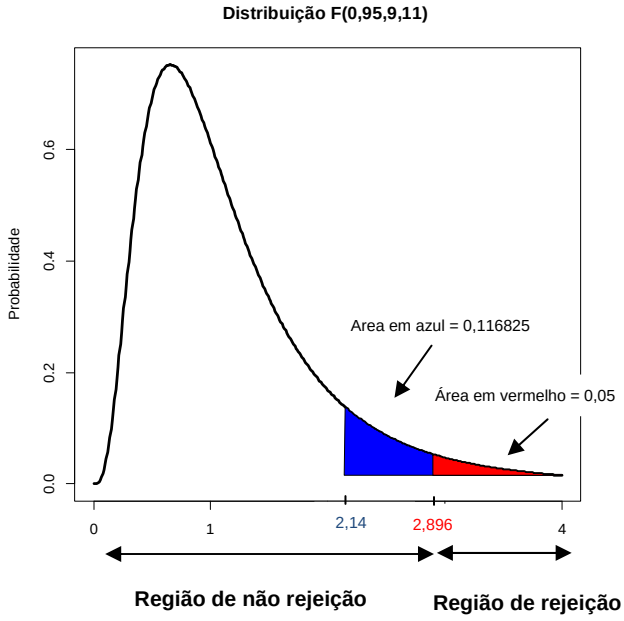
- Cálculo da estatística

$$F_c = \frac{s_1^2}{s_2^2} = \frac{852,9333}{398,2424} = 2,1417$$

Distribuição do Teste Estatístico: quando $H_0$ é verdadeira a estatística F tem distribuição F com $n_1 - 1$ e $n_2 - 1$ graus de liberdade, ou seja,

Regra de Decisão: o valor teórico da distribuição

- Existe duas maneiras de tomar a decisão estatística, ou seja, rejeitar ou não rejeitar $H_0$:
  1) **Pelo procedimento clássico: não rejeitar $H_0$**, dado que 2,1417<2,896; isto é, o valor de $F_c=2,141$ é menor que o valor teórico, $F_t=2,896$, portanto o valor de $F_c$ obtido com base nos dados da amostra esta na região de não rejeição;
  2) **Usando o procedimento do valor- p: não rejeitar $H_0$**, o valor-p é 0,116825 é maior que $\propto = 0,05$ fixado anteriormente da realização do experimento. O valor de 0,116825 refere-se a cauda à direita da distribuição $F_{(-0.05,9,11)}$ como mostrado na figura abaixo




- Conclusão: os dados não fornecem suficiente evidências para rejeitarmos $H_0$, ou seja, as variâncias das observações dos esforços 1 e 2 não podem ser consideradas diferentes.

Script no R para o teste F

```
> # entrando com os dados
> esf1<-c(54,99,105,46,70,87,55,58,139,91)
> esf2<-c(93,91,93,150,80,104,128,83,88,95,94,97)
> var.esf1<-var(esf1)       # calculo da variância dos dados do esforço1
> var.esf1
[1] 852.9333
> var.esf2<-var(esf2)       # calculo da variância dos dados do esforço2
> var.esf2
[1] 398.2424
> fc<-var.esf1/var.esf2          # calculo da estatística F
> fc
[1] 2.141744
> # valor teórico desta distribuição para alfa=0,05
```

```
> ft<-qf(1-0.05,length(esf1)-1,length(esf2)-1)
> ft
[1] 2.896223
> # valor de p associado a estatística calculada fc
> valor.p<-1-pf(fc,length(esf1)-1,length(esf2)-1)
> valor.p
[1] 0.116825
```

Uma outra forma de se testar esta hipótese é usar a função **var.test ( )**, ou seja,

```
> var.test(esf1,esf2,alternative="greater")

        F test to compare two variances

data:  esf1 and esf2
F = 2.1417, num df = 9, denom df = 11, p-value = 0.1168
alternative hypothesis: true ratio of variances is greater than 1
95 percent confidence interval:
 0.7394956       Inf
sample estimates:
ratio of variances
          2.141744
```

Exemplo da construção do gráfico de uma distribuição $F_{(9,11)}$ com a demarcação do valor teórico de para $\propto = 0{,}05$ .

```
> # gráfico da distribuição
> xv<-seq(0,4,0.01)        # gerando uma sequência de números de 0 a 4
> yv<-df(xv,9,11)          # gerando os valores de
>                          # distr. F(9,11) com a sequencia xv
> plot(xv,yv,type="l",main="Distribuição F(0,95,9,11) ",
+ ylab="Probabilidade",xlab="",lwd=3)  # gráfico da distribuição F
> fcr=qf(0.95,9,11)
> fcr                                  # valor crítico para alfa=5%
[1] 2.605975
> lines(c(fcr,fcr),c(0,0.6),
+ col=2,lwd=2,pch=2)                  # linha sinalizando o valor crítico

> # preenchimento da área sob a curva acima do valor crítico
> polygon(c(xv[xv>=2.605975],2.605975),
+ c(yv[xv>=2.60975],yv[xv==4]),col="red")
```

## 4 Análise de variância

Embora o teste F possa ser aplicado independentemente, a sua maior aplicação é na análise de variância dos Delineamentos Experimentais. Vamos considerar os seguintes dados de Delineamento Inteiramente Casualizado, (DIC).

| Tratamentos | Repetições 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| **A** | 12,4 | 15,2 | 14,3 | 12,6 |
| **B** | 13,2 | 16,2 | 14,8 | 12,9 |
| **C** | 12,1 | 11,3 | 10,8 | 11,4 |
| **D** | 10,9 | 9,8 | 9,4 | 8,3 |

$\Rightarrow \sigma_e^2$

$\sigma_e^2+\sigma_T^2$

Dentro de um mesmo tratamento o valor observado nas diferentes repetições não é o mesmo, pois estes valores estão sujeitos à variação ao acaso ($\sigma_e^2$). Quando passamos de um tratamento para outro, os dados também não são iguais, pois estes estão sujeitos a uma variação do acaso acrescida de uma variação devida ao efeito do tratamento, i.é, $\sigma_e^2+\sigma_T^2$

**Quadro da análise de variância do DIC**

Considere os dados do exemplo anterior, onde tínhamos 4 tratamentos (k=4) e 4 repetições. A Tabela da Análise de variância fica sendo

| *Fonte de variação* | *G.L.* | *Soma de Quadrados* | *Quadrado médio* | *Estatística F* |
|---|---|---|---|---|
| *Entre* | ***k - 1*** | $\sum_{i=1}^{k}\frac{y_{i+}^2}{r}-\frac{(y_{++})^2}{kr}$ | $\frac{S.Q.Trat.}{k-1}$ | $\frac{Q.M.Trat.}{Q.M.\mathrm{Re}s.}$ |
| *Dentro* | ***n - k*** | $\sum_{i=1}^{k}\sum_{j=1}^{r}y_{ij}^2-\sum_{i=1}^{k}\frac{(y_{i+})^2}{r}$ | $\frac{S.Q.\mathrm{Re}s.}{kr-k}$ | |
| *Total* | ***n - 1*** | $\sum_{i=1}^{k}\sum_{j=1}^{r}y_{ij}^2-\frac{(y_{++})^2}{kr}$ | | |

Deste quadro notamos que o Quadrado médio do resíduo estima a variação casual (do resíduo) $\sigma_e^2$ . Enquanto que o quadrado médio dos tratamentos estima a variação casual (resíduo) acrescida de uma possível variância devido ao efeito dos tratamentos ($\sigma_e^2+\sigma_T^2$ ), então

$$F=\frac{\sigma_e^2+\sigma_T^2}{\sigma_e^2}$$

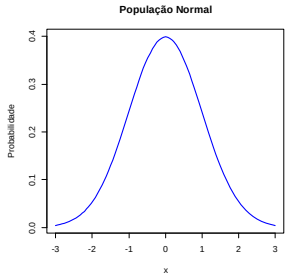
Se não houver efeito dos tratamentos os dois quadrados médios *(Quadrado médio dos tratamentos e quadrado médio do resíduo)* estimam a mesma variância, o que implica o valor de F ≅ 1,0, e qualquer diferença que ocorra entre os valores médios dos tratamentos é meramente casual.

**5 Teste *t – Student*.**

Considere uma outra retirada de amostras repetidas de um determinado tamanho, por exemplo, r = 5 de uma população normal. Para cada amostra calcule a média $\bar{y}$ o desvio padrão, $s$ , o erro padrão da média $s_{\bar{y}}$ e uma outra estatística

$$t_c=\frac{\bar{y}-\mu}{s_{\bar{y}}}$$

Graficamente temos

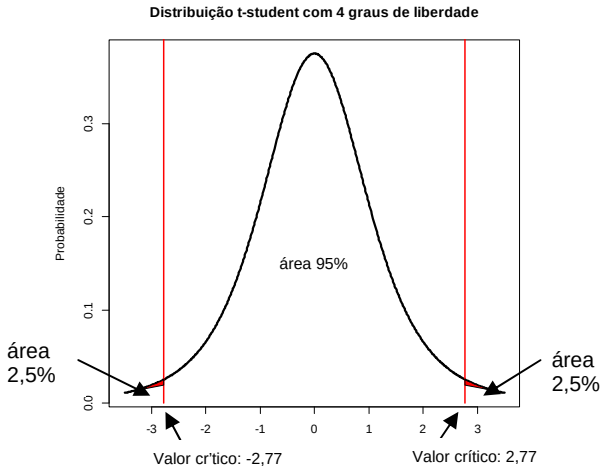



$$amostra\ 1 \rightarrow s_1^2 = \frac{\sum (y_i - \bar{y})^2}{5-1}; s_{\bar{y}_1} = \sqrt{\frac{s_1^2}{5}};\ t_1 = \frac{\bar{y}_1 - \mu}{s_{\bar{y}_1}}$$

$$amostra\ 2 \quad \text{..... ..... ..... ..... ..... .....}$$

$$amostra\ m \quad s_m^2 = \frac{\sum (y_i - \bar{y})^2}{5-1}; s_{\bar{y}_m} = \sqrt{\frac{s_1^2}{5}};\ t_m = \frac{\bar{y}_m - \mu}{s_{\bar{y}_1}}$$

Organizando estes milhares de valores da estatística *t* em distribuição de probabilidade teremos a seguinte forma



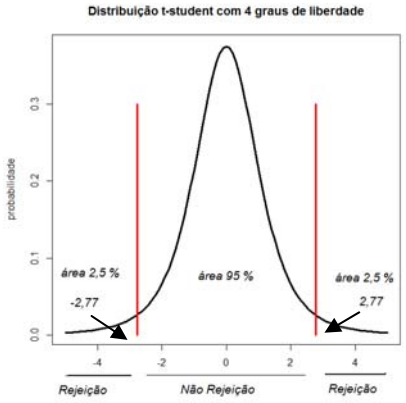

Existe uma única distribuição *t* para cada tamanho de amostra. Neste exemplo em que r = 5 (repetições = 5), 2,5 % dos valores de *t* serão maiores ou iguais do que 2,776 e 2,5% serão menores ou iguais do que -2,776. Os valores da estatística ***t – student*** são apresentados em tabelas (ver Tabela da distribuição *t* ). Por exemplo, no exemplo acima, para 4 graus de liberdade, o valor tabelado esperado para $|t|$ com probabilidade de 0,05 (5%) é 2,77.

No **R** a obtenção dos valores teóricos da distribuição *t-student* é dado pela seguinte função

```
> #valor teórico da distribuição t-student pela função qt( ) para alfa=0.05 e
4 #graus de liberdade
> alfa<-0.05
> qt(1-alfa/2,4)
[1] 2.776445
```

A distribuição ***t – student*** converge rapidamente para a distribuição normal. Quanto maior for a amostra maior é aproximação da distribuição *t – student* com a distribuição normal. Quando os valores de t são calculados em amostras de tamanho r = 60, estes são bem próximos dos valores da distribuição normal.

**Regra de decisão para a estatística *t-student***




Todos os possíveis valores que o teste estatístico pode assumir são pontos no eixo horizontal do gráfico da distribuição do teste estatístico e é dividido em duas regiões; uma região constitui o que denominamos de *região de não rejeição* e a outra região constitui o que denominamos de *região de rejeição*. Os valores do teste estatístico que formam a *região de rejeição* são aqueles valores menos prováveis de ocorrer se a hipótese nula é verdadeira, enquanto que os valores da *região de não rejeição* são os mais prováveis de ocorrer se a hipótese nula é verdadeira. *A regra de decisão nos diz para* ***rejeitar $H_0$*** *se o valor do teste estatístico calculado da amostra ($t_c$) é um valor que está na região de rejeição e para* ***não rejeitar $H_0$*** *se o valor calculado do teste estatístico é um dos valores que está na região de não rejeição.* Em particular, no caso do teste *t – student* a regra de decisão fica sendo: rejeita-se $H_0$ se $|t_c| \geq t_{(n-1,\frac{\alpha}{2})}$. Outra forma de se tomar decisão sobre rejeitar ou não rejeitar $H_0$ é pelo valor de p associado ao valor calculado da estatística $t_c$ . Se $p \geq 0{,}05$ não se rejeita $H_0$ , caso contrário ( $p < 0{,}05$ ) rejeita-se $H_0$ . Neste caso não há necessidade de se consultar a Tabela teórica da distribuição *t-student*.

Exemplo: Em um hospital veterinário amostras de soro de amilase de 15 animais sadios e 22 animais hospitalizados foram colhidas. Os resultados da média e dos desvios-padrões foram os seguintes:

$$\bar{y}_1 = 120 \text{ unidades/ml}, \quad s_1 = 40 \text{ unidades/ml}$$

$$\bar{y}_2 = 96 \text{ unidades/ml}, \quad s_2 = 35 \text{ unidades/ml}$$

Neste exemplo, o erro padrão amostral $s_{\bar{y}}$ da fórmula da estatística t, será substituído pelo erro padrão da média “pooled”, ou seja,

$$s_P^2 = \frac{(r_1 - 1)s_1^2 + (r_2 - 1)s_2^2}{(r_1 - 1) + (r_2 - 1)}$$

Cálculos:

- Suposições: os dados constituem duas amostras independentes, cada uma, retirada de uma população normal. As variâncias populacionais são desconhecidas e assumidas iguais;

- Hipóteses: $H_0 : \mu_1 = \mu_2$; $H_1 : \mu_1 \neq \mu_2$
- Teste estatístico: $t_c = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{r_1} + \frac{s_p^2}{r_2}}}$;
- Distribuição do teste estatístico: quando $H_0$ for verdadeira, o teste segue uma distribuição *t – Student* com $r_1 + r_2 – 2$ graus de liberdade;
- Regra de decisão: Rejeita-se $H_0$ se $|t_c| \geq t_{(r_1+r_2-2;\frac{\alpha}{2})}$, neste exemplo, $|t_c| \geq 2{,}030$;
- Cálculo do teste estatístico: primeiro o cálculo da variância amostral

$$s_p^2 = \frac{14(40^2) + 21(35^2)}{14+21} = 1375 \quad e$$

$$t_c = \frac{(120-96)-0}{\sqrt{\frac{1375}{15} + \frac{1375}{22}}} = \frac{24}{12{,}42} = 1{,}93$$

- Decisão estatística: não se rejeita $H_0$, visto que $-2{,}030 < 1{,}88 < 2{,}030$, ou seja, 1,88 está na região de não rejeição;
- Conclusão: com base nestes dados não podemos concluir que as médias das duas populações são diferentes. Neste teste o nível mínimo de significância do teste é p= 0,069 (p>0,05).

Script no R para resolver o exemplo acima

```
> # definição das médias, dos desvio padrões e do tamanho das amostras
> m.y1<-120;m.y2<-96;sd.y1<-40;sd.y2<-35;n.y1<-15;n.y2<-22
>
> # calculo da variância “pooled”
> v.pool<-((n.y1-1)*sd.y1^2+(n.y2-1)*sd.y2^2)/((n.y1-1)+(n.y2-1))
> v.pool
[1] 1375
> # calculo da estatistica t
> tc<-(m.y1-m.y2)/sqrt(v.pool/n.y1+v.pool/n.y2)
> tc
[1] 1.932929
> # valor de t tabelado a 5% e 35 graus de liberdade
> alfa<-0.05
> t.tab <- qt(1-alfa/2,35)
> t.tab
[1] 2.030108
> # valor de p correspondente a este valor de t
> # multiplica-se o valor de p por 2 pois o teste é bi-lateral
> valor.p <- 2*(1-pt(tc,35))
> valor.p
[1] 0.06136792
```

## 2º EXERCÍCIO PRÁTICO DE ESTATÍSTICA EXPERIMENTAL

**1-** A tabela abaixo mostra a porcentagem de gordural corporal para vários homens e mulheres. Estas pessoas participaram de um programa de controle de peso de três vezes por semana por um ano. As medidas referem-se a porcentagem de gordura de seus corpos.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Homens* | *13,3* | *19,0* | *20,0* | *8,0* | *18,06* | *22,0* | *20,0* | *31,0* | *21,0* |
| | *12,0* | *16,0* | *12,0* | *24,0* | | | | | |
| *Mulheres* | *22,0* | *26,0* | *16,0* | *12,0* | *21,7* | *23,2* | *21,0* | *28,0* | *30,0* |
| | *23,0* | | | | | | | | |

a) Faça um gráfico de barras e um gráfico boxplot para cada grupo
a) Quais as suposições sob as quais o teste F pode ser aplicado.
b) Podemos concluir que a variabilidade do grupo das mulheres seja maior que o do grupo homens. (Use $\alpha = 0{,}05 \ e \ 0{,}01$).

**2-** Em um estudo, a seguintes contagens de linfócitos foi obtido em vacas de dois anos da raça Holstein e de vacas de dois anos da raça Guernseys. Os resultados estão na Tabela abaixo:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Holstein* | *5166* | *6080* | *7290* | *7031* | *6700* | *8908* | *4214* | *5135* | *5002* |
| | *4900* | *8043* | *6205* | *3800* | | | | | |
| *Guernseys* | *6310* | *6295* | *4497* | *5182* | *4273* | *6591* | *6425* | *4600* | *5407* |
| | *5509* | | | | | | | | |

Calcular:
a)- a média geral, um gráfico de barras para cada raça e um gráfico boxplot para cada raça, a média de cada raça, a variância amostral e o desvio-padrão de cada raça;
b)- declare as suposições sob as quais o teste *t –student,* para amostras independentes, pode ser aplicado;
c)- teste se as variâncias das duas populações são iguais. (Teste F)
d)- em função do resultado do teste do item c) podemos concluir que a contagem de linfócitos nas duas raças diferem assumindo que as variâncias são desconhecidas e iguais? Considere $\alpha = 5\%$.

**3-** Retirou-se 5 amostras de tamanho 5 de uma população $N(\mu, \sigma^2)$. Para cada amostra foi aplicado um antiparasitário (tratamentos). Em seguida os pesos dos animais foram analisados para cada tratamento. Teste se existe efeito de antiparasitário no peso dos animais, ou seja, teste a hipótese estatística,

$$H_0 = \mu_1 = \mu_2 = \ldots = \mu_5$$
$$H_1 = \mu_i \neq \mu_j \ \ para \ i \neq j$$

Os tratamentos (antiparasitários) e os pesos, em quilogramas, dos animais estão dados na tabela abaixo:

| | *Tratamentos* | | | | |
|---|---|---|---|---|---|
| | *Neguvon* | *Methiridim* | *TH* | *Haloxon* | *Controle* |
| | *330* | *315* | *298* | *286* | *279* |
| | *314* | *304* | *289* | *273* | *240* |
| | *331* | *307* | *273* | *269* | *266* |
| | *311* | *320* | *240* | *278* | *269* |
| | *320* | *305* | *121* | *274* | *250* |
| *(Média)* ( $\bar{y}_{+j}$ ) | | | | | |
| (Variância) $S_i^2$ | | | | | |
| *(Desvio padrão)* $S_i$ | | | | | |

Roteiro dos cálculos:
a)- Faça uma estimativa da $\sigma^2$ utilizando $s_D^2$ e $s_E^2$ pelas fórmulas:

$$s_D^2 = \frac{s_1^2 + s_2^2 + \ldots + s_5^2}{5} \quad e \ de \quad s_E^2 = r s_{\bar{y}}^2, \ sendo \ s_{\bar{y}}^2 = \frac{\sum_{j=1}^{5} (\bar{y}_{+j} - \bar{y})^2}{5-1}$$

Calcule a estatística $F = \frac{s_E^2}{s_D^2}$ e compare com o valor teórico da distribuição F a 5%, sendo $S_D$ = variância dentro dos tratamentos e $S_E$ = variância entre os tratamentos.

**4-** Obter por meio das tabelas das distribuições F e t os valores de

a)- $F_{(5,\ 6,\ 0{,}05)}$ ; $F_{(5,\ 6,\ 0{,}01)}$; $F_{(10,\ 15,\ 0{,}05)}$; $F_{(10,\ 15,\ 0{,}01)}$; $F_{(11,\ 12,\ 0{,}05)}$; $F_{(11,\ 12,\ 0{,}01)}$.

b)- $t_{(7,\ 0{,}05)}$; $t_{(7,\ 0{,}01)}$; $t_{(15,\ 0{,}025)}$; $t_{(16,\ 0{,}05)}$; $t_{(10,\ 0{,}10)}$; $t_{(18,\ 0{,}20)}$;

(faça os desenhos das distribuições com os respectivos valores). Finalmente, obtenha os mesmos valores e os mesmos gráficos no **R**

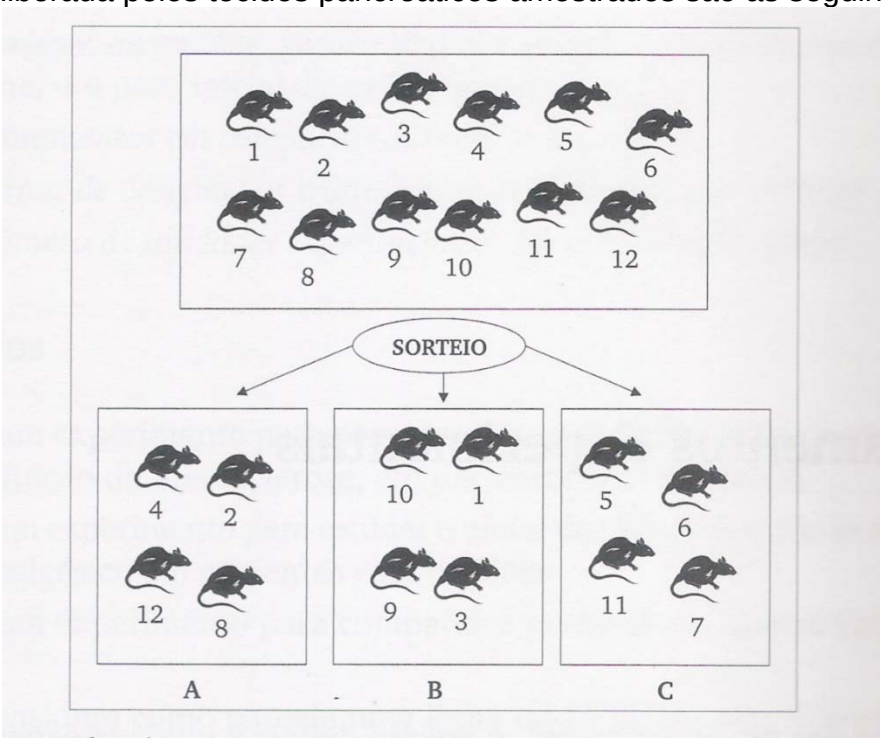
**Delineamento inteiramente casualizado (DIC)**

**1 Introdução**

O DIC é mais simples dos delineamentos. Os tratamentos se distribuem ao acaso em todas as unidades experimentais e o número de repetições por tratamento pode ser igual ou diferente. O DIC é muito utilizado para estudos de métodos, técnicas de trabalhos em laboratório, ensaios de vegetação e em experimentos com animais. Para sua aplicação, há necessidade que o meio atue de forma uniforme em todas as unidades experimentais e que estas sejam facilmente identificadas para receber o tratamento.

Vamos começar com um exemplo:

- Em um estudo do efeito da glicose na liberação de insulina, 12 espécies de tecido pancreático idênticas foram subdivididas em três grupos de 4 espécies cada uma. Três níveis (baixo - tratamento 1, médio tratamento - 2 e alto tratamento - 3) de concentrações de glicose foram aleatoriamente designados aos três grupos, e cada espécie dentro de cada grupo foi tratado com o nível de concentração de glicose sorteado a eles. A quantidade de insulina liberada pelos tecidos pancreáticos amostrados são as seguintes:




| *Tratamento* | *Repetições* | | | | | *Total* | *Média* | *Variância* |
|---|---|---|---|---|---|---|---|---|
| | *1* | *2* | *3* | *4* | $r_i$ | | | |
| *T1* | *1,59* | *1,73* | *3,64* | *1,97* | *4* | *8,93* | *2,23* | *0,91* |
| $T_2$ | *3,36* | *4,01* | *3,49* | *2,89* | *4* | *13,75* | *3,44* | *0,21* |
| $T_3$ | *3,92* | *4,82* | *3,87* | *5,39* | *4* | *18,00* | *4,50* | *0,54* |
| ***Total*** | | | | | ***12*** | ***40,68*** | | |

Este é um estudo experimental com 12 unidades experimentais (amostras de tecido pancreático) e *k=3* tratamentos. Cada tratamento é um nível de fator simples: concentração de glicose. Existem 4 repetições para cada tratamento. Os dados, quantidade de insulina liberada pelo tecido pancreático podem ser considerados como três amostras aleatórias, cada uma com *r=4* repetições, ou de tamanho *r=4* sorteadas de três populações.

Dado que os tratamentos são designados às unidades experimentais completamente ao acaso, este delineamento é denominado de *DELINEAMENTO INTEIRAMENTE AO ACASO* (DIC). Em geral, em um DIC, um número fixo de k tratamentos são sorteados às *N* unidades experimentais de tal forma que o *i-ésimo* tratamento é sorteado a exatamente $r_i$ unidades experimentais. Assim, $r_i$ é o número de repetições do *i-ésimo* tratamento e $r_1 + r_2 + r_3 + ... + r_k = N$. No caso em que $r_i$ são iguais, i.é., $r_1 = r_2 = r_3 = ... = r_k = r$, então $N = rk$ e o delineamento é *balanceado.*

**Notação:**

| | Repetições | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Tratamento** | **1** | **2** | **3** | **...** | **j** | **...** | **r** | **Total** | **Média** |
| 1 | $y_{11}$ | $y_{12}$ | $y_{13}$ | ... | ... | ... | $y_{1r}$ | $y_{1+}$ | $\bar{y}_{1+}$ |
| 2 | $y_{21}$ | $y_{22}$ | $y_{23}$ | ... | ... | ... | $y_{2r}$ | $y_{2+}$ | $\bar{y}_{2+}$ |
| 3 | $y_{31}$ | $y_{32}$ | $y_{33}$ | ... | ... | ... | $y_{3r}$ | $y_{2+}$ | $\bar{y}_{2+}$ |
| . | . | . | . | . | ... | . | . | . | . |
| . | . | . | . | . | ... | . | . | . | . |
| . | . | . | . | . | ... | . | . | . | . |
| i | . | . | . | . | $y_{ij}$ | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . |
| k | $y_{k1}$ | $y_{k2}$ | $y_{k3}$ | ... | ... | ... | $y_{kr}$ | $y_{k+}$ | $\bar{y}_{k+}$ |
| | | | | | | | N=rk | $y_{++}$ | $\bar{y}_{++}$ |

Convenções:

- $y_{i+}$ *e* $\bar{y}_{i+}$ representam, respectivamente, o total e a média do i-ésimo tratamento, respectivamente,
- $y_{++}$ *e* $\bar{y}_{++}$ representam, respectivamente, o total geral (soma de todas as observações) e a média geral de todas as observações.

## 2 Quadro da Análise de Variância (ANOVA)

O método da análise de variância pode ser visto como uma extensão do teste *t de student* para amostras independentes. Como no teste *t* de amostras independentes, o método da ANOVA compara uma medida da magnitude da variabilidade observada dentro das k amostras com uma medida da variabilidade entre as médias das k amostras.

## 3 Modelo matemático do DIC com efeitos de tratamentos fixos

O modelo associado ao DIC com efeitos fixos é

$$y_{ij} = \mu + \tau_i + e_{ij},$$

sendo,

- $y_{ij}$ é a observação na unidade experimental que recebeu o i-ésimo tratamento na j-ésima repetição;
- $\mu$ é a média geral comum a todas as observações definida como

$\mu = \dfrac{\sum_{i=1}^{k} r_i \mu_i}{N}$, com $\mu_i$ a média populacional do i-ésimo tratamento;

- $\tau_i$ o efeito do i-ésimo tratamento na variável dependente Y e mede o afastamento da média $\mu_i$ em relação a $\mu$, isto é, $\tau_i = \mu_i - \mu$; e
- $e_{ij}$ é um erro casual não observável.

Pela definição de $\mu$ e $\tau_i$ acima, temos que este modelo possui a restrição $\sum_{i=1}^{k} n_{i_i} \tau_i = 0$, pois, $\sum_{i=1}^{k} n_i \tau_i = \sum_{i=1}^{k} n_i (\mu_i - \mu) = \sum_{i=1}^{k} n_i \mu_i - n_i \mu = 0$.

## 4 Suposições associadas ao modelo

As suposições usualmente associadas aos componentes do modelo do DIC são que os $e_{ij}$ são variáveis aleatórias independentes e identicamente distribuídas com distribuição $N(0, \sigma^2)$. Como os $y_{ij}$ são funções lineares dos $e_{ij}$, das suposições sobre os erros decorre que:

- $E(y_{ij}) = \mu + \tau_i = \mu_i$;
- $Var(y_{ij}) = \sigma^2$;
- $y_{ij}$ são normalmente distribuídos e independentes, ou, resumidamente que $y_{ij} \sim N(\mu_i, \sigma^2)$.

Portanto, estamos supondo que as observações do experimento a ser analisado correspondem a amostras aleatórias de k populações normais com a mesma variância e que podem ou não ter médias diferentes. A figura abaixo representa graficamente esse fato, considerando, no caso, três tratamentos.

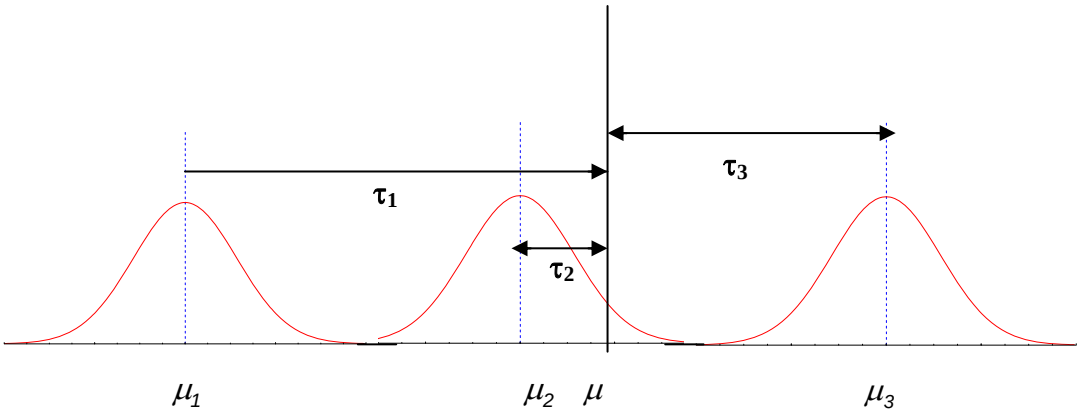



Figura: Ilustrações das suposições do modelo matemático associado ao DIC com um fator fixo.

## 5 Hipóteses estatísticas

A Hipótese geral é:

$$H_0 : \tau_1 = \tau_2 = ... = \tau_k = 0,$$

ou seja, vamos testar a não existência de efeito do fator (tratamento).

## 6 Partição da soma de quadrados

Voltemos ao quadro de representação das observações no DIC na página 30. Podemos identificar os seguintes desvios:

- $y_{ij} - \bar{y}_{++}$, como o desvio de uma observação em relação a média amostral geral;

- $y_{ij} - \bar{y}_{i+}$, como o desvio da observação em relação à média de seu grupo ou do i-ésimo tratamento;
- $\bar{y}_{i+} - \bar{y}_{++}$, como o desvio da média do i-ésimo tratamento em relação á média geral.

Consideremos a identidade

$$(y_{ij} - \bar{y}_{++}) = (y_{ij} - \bar{y}_{i+}) + (\bar{y}_{i+} - \bar{y}_{++}),$$

a qual diz que a " *a variação de uma observações em relação à média geral amostral é igual à soma variação desta observação em relação à média de seu grupo com a variação da média do i-ésimo tratamento em que se encontra esta observação em relação à média geral amostral ".* Elevando-se ao quadrado os dois membros da identidade acima e somando em relação aos índices i e j, obtemos:

$$\sum_{i=1}^{k}\sum_{j=1}^{r_i} (y_{ij} - \bar{y}_{++})^2 = \sum_{i=1}^{k}\sum_{j=1}^{r_i} (y_{ij} - \bar{y}_{i+})^2 + \sum_{i=1}^{k} r_i (\bar{y}_{i+} - \bar{y}_{++})^2,$$

os duplos produtos são nulos.

O termo

$$\sum_{i=1}^{k}\sum_{j=1}^{r_i} (y_{ij} - \bar{y}_{++})^2,$$

é denominado de ***Soma de Quadrados Total*** e vamos denotá-lo por ***SQT***.O número de graus de liberdade associado à ***SQT*** é kr - 1, ou N – 1, pois temos N observações e a restrição

$$\sum_{i=1}^{k}\sum_{j=1}^{r_i} (y_{ij} - \bar{y}_{++}) = 0.$$

A componente:

$$\sum_{i=1}^{k}\sum_{j=1}^{r_i} (y_{ij} - \bar{y}_{i+})^2,$$

é denominada de ***Soma de Quadrados Residual***, representada por ***SQR,*** e é uma medida da homogeneidade interna dos tratamentos. Quanto mais próximas estiverem as observações dentro de cada grupo (tratamento), menor é a ***SQR***. Notem que a magnitude da ***SQR*** não depende da diferença entre as médias dos tratamentos. Considerando apenas o i-ésimo tratamento, temos que

$$\sum_{j=1}^{r_i} (y_{ij} - \bar{y}_{i+})^2$$

possui $r_i - 1$ graus de liberdade. Assim, o número de graus de liberdade associado à ***SQR*** é:

$$\sum_{i=1}^{k} (r_i - 1) = kr - k = N - k.$$

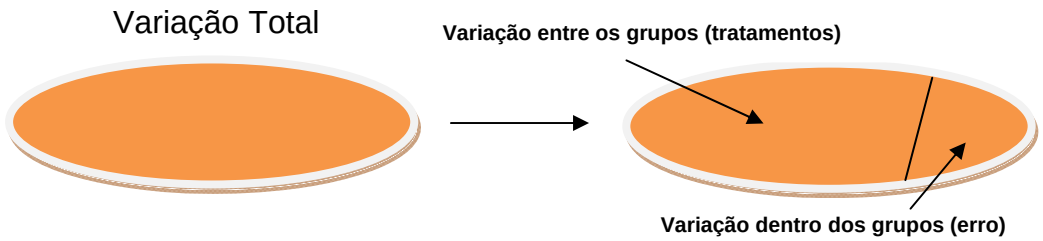
A componente $\sum_{i=1}^{k} r_i (\bar{y}_{i+} - \bar{y}_{++})^2$, mede a variabilidade entre as médias dos tratamentos e por isso é denominada de ***Soma de Quadrados Entre Tratamentos***, representada por ***SQTr.*** Quanto mais diferentes entre si forem as médias dos tratamentos, maior será a ***SQTr.*** Desde que temos k tratamentos e a restrição de que

$$\sum_{i=1}^{k} r_i \left(\bar{y}_{i+} - \bar{y}_{++}\right) = 0 ,$$

A ***SQTr*** possui k - 1 graus de liberdade. Com esta notação, podemos escrever que:

$$\boldsymbol{SQT = SQR + SQTr.}$$

A ilustração abaixo mostra como a variação total é particionada em variação entre grupos (tratamentos) e variação dentro dos grupos (erro).




---

## 6 Quadrados médios

Dividindo a ***SQR*** e ***SQTr*** pelos seus correspondentes graus de liberdade, obtemos, respectivamente o ***Quadrado Médio Residual*** (***QMR***) e o ***Quadrado Médio Entre Tratamentos*** (***QMTr***), isto é,

$$QMR = \frac{SQR}{N-k} \quad e \quad QMTr = \frac{SQTr}{k-1}$$

---

## 7 Estatística e região crítica do teste

A estatística para o teste é

$$F_c = \frac{QMTr}{QMR} ,$$

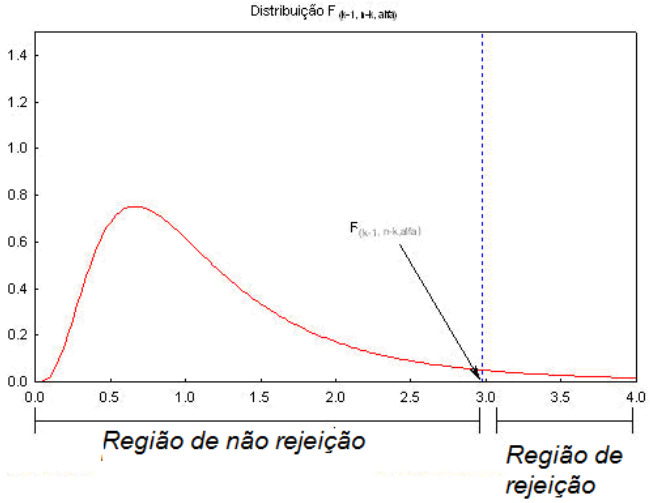
a qual, deve ser próximo de 1 se $H_0$ for verdadeira, enquanto que valores grandes dessa estatística são uma indicação de que $H_0$ é falsa. A teoria nos assegura que $F_c$ tem, sob $H_0$ distribuição *F – Snedecor* com (k -1) e (N – k) graus de liberdade. Resumidamente, indicamos:

$$F_c \sim F_{(k-1,N-K)}, \; sob \; H_0 .$$

Rejeitamos $H_0$ para o nível de significância $\alpha$ se

$$F_c > F_{(k-1,N-K,\alpha)},$$

sendo, $F_{(k-1,N-K,\alpha)}$ o quantil de ordem $(1-\alpha)$ da distribuição F-Snedecor com (k -1) e (N – k) graus de liberdade. Graficamente temos:




## 8 Quadro da análise de variância (ANOVA)

Dispomos as expressões necessárias ao teste na Tabela abaixo denominada de Quadro de Análise de Variância (ANOVA).

| Fonte de variação | g.l. | SQ | QM | $F_c$ |
|---|---|---|---|---|
| ***Tratamentos*** ***(Entre)*** | ***k - 1*** | $\sum_{i=1}^{r_i} \frac{Y_{i+}^2}{r_i} - \frac{(Y_{++})^2}{N}$ | $QMTr = \frac{SQTr}{k-1}$ | $\frac{QMTr}{QMR}$ |
| ***Resíduo*** ***(dentro dos trat.)*** | ***N - k*** | $\sum_{i=1}^{k}\sum_{j=1}^{r} Y_{ij}^2 - \sum_{i=1}^{k} \frac{(Y_{i+})^2}{r}$ | $QMR = \frac{SQR}{N-k}$ | |
| ***TOTAL*** | ***N - 1*** | $\sum_{i=1}^{k}\sum_{j=1}^{r} Y_{ij}^2 - \frac{(Y_{++})^2}{N}$ | | |

Pode-se provar que:

- $E(QMR) = \sigma^2$, ou seja, QMR é um estimador não viesado da variância $\sigma^2$;
- $E(QMTr) = \sigma^2 + \frac{r}{(k-1)}\sum_{i=1}^{k} \tau_i$, ou seja, QMTr é um estimador não viesado da variância $\sigma^2$ se a hipótese $H_0 : \tau_1 = \tau_2 = ... = \tau_k = 0$ é verdadeira.

## 9 Detalhes dos cálculos

Apresentaremos alguns passos que facilitam os cálculos das somas de quadrados da ANOVA.

- Calcule a correção para a média $CM = \frac{(Y_{++})^2}{N}$;
- Calcule a *Soma de Quadrados dos Totais* (***SQT***)

  $SQT = \sum_{i=1}^{k}\sum_{j=1}^{r_i} y_{ij}^2 - CM$;

- Calcule a *Soma de Quadrados Entre os Tratamentos* (***SQTr***)
  $SQTr = \sum_{i=1}^{r_i} \frac{Y_{i+}^2}{r_i} - CM$ ;
- Calcule a *Soma de Quadrados Residual* (***SQR***) pela diferença, isto é, $SQR = SQT - SQTr$ ;
- Calcule os *Quadrados Médios Entre os Tratamentos* (***QMTr***) e o *Quadrado Médio Residual* (***QMR***) $QMTr = \frac{SQTr}{k-1}$ *e* $QMR = \frac{SQR}{N-k}$
- Calcule $F_c$ para tratamentos $F_c = \frac{QMTr}{QMR}$

Notem que estas fórmulas computacionais assumem que existe $r_i$ repetições para o *i-ésimo* tratamento; consequentemente, para um experimento balanceado com *r* repetições para cada tratamento, $r_i$ deve ser substituído por *r*. Estas várias soma de quadrados obtidas nestes cinco passos podem ser resumidas no quadro da ANOVA apresentado no item 8.

**Exemplo 1**

Vamos considerar os dados apresentados na pg 49. Desejamos testar a hipótese nula

$H_0 : \mu_1 = \mu_2 = \mu_3$

$H_1 : \mu_i \neq \mu_j$ *para pelo menos um par* $i \neq j$

Os cálculos para montarmos o quadro da ANOVA são: temos k = 3, r = 4, e N = 3 x 4 =12. Então

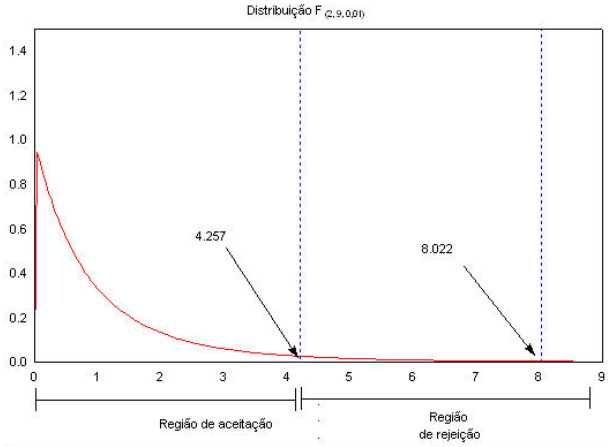
- Graus de liberdade:
  $Total = N - 1 = 12 - 1 = 11; Trat. = k - 1 = 3 - 1 = 2$
  $\text{Re}siduo = N - k = 12 - 3 = 9$
- $CM = \frac{(40{,}68)^2}{12} = 137{,}91$
- $SQT = (1{,}59)^2 + (1{,}73)^2 + \ldots + (5{,}39)^2 - CM = 153{,}18 - 137{,}18 = 15{,}28$
- $SQTr = \frac{(8{,}93)^2}{4} + \frac{(13{,}75)^2}{4} + \frac{(18{,}00)^2}{4} - CM = 148{,}20 - 137{,}91 = 10{,}30$
- $SQR = SQT - SQTr = 15{,}28 - 10{,}30 = 4{,}98$
- $QMTr = \frac{10{,}30}{2} = 5{,}15 \quad e \quad QMR = \frac{4{,}98}{9} = 0{,}55$
- $F_c = \frac{QMTr}{QMR} = \frac{4{,}98}{0{,}55} = 9{,}31$

O quadro da ANOVA para a variável insulina liberada é o seguinte:

| Fonte de var. | g.l. | SQ | QM | $F_c$ |
|---|---|---|---|---|
| ***Tratamentos*** *(Entre)* | ***2*** | ***10,30*** | ***5,15*** | ***9,31*** |
| ***Resíduo*** *(dentro dos tratamentos)* | ***9*** | ***4,98*** | ***0,55*** | |
| ***TOTAL*** | ***11*** | ***15,28*** | | |

Das tabelas das distribuições F, temos que
$F_{(2,9,0,05)} = 4{,}257$ *e* $F_{(2,9,0,01)} = 8{,}022$ . O valor $F_c$=9,31 é maior do que estes

valores tabelados, então rejeitamos a hipótese nula $H_0$ para $\alpha = 0{,}01$, ou 1% de probabilidade (se é significativo a 1%, logo também é significativo a 5%).




Podemos concluir que, para um nível de $\alpha = 0{,}01$, ou 1%, que a quantidade de insulina liberada é diferente para pelo menos dois níveis de glicose.
Script da resolução do exemplo da pg 48 no R

```
> #
> # exemplo 1 (DIC) pg 48
> #
>
> # entrando com o número de repetições
> r <- 4
> # entrando com os dados
> insulina <- c(1.59, 1.73, 3.64, 1.97, 3.36, 4.01, 3.49, 2.89, 3.92, 4.82, 3.87,
5.39)
>
> # entrando com os níveis da insulina (Tratamentos)
> trat <- c(rep("Baixo", r), rep("Medio", r), rep("Alto", r))
> media.geral<- mean(insulina)   # calculando a média geral
> media.geral
[1] 3.39
>
> # aplicando o comando tapply ao objeto insulina para o cálculo dos
> # totais dos tratamentos
> total.trat <- tapply(insulina, trat, sum)
> total.trat
 Alto Baixo Medio
18.00  8.93 13.75
>
> # aplicando o comando tapply ao objeto insulina para o cálculo das
> # médias dos tratamentos
> media.trat <- tapply(insulina, trat, mean)
> media.trat
  Alto  Baixo  Medio
4.5000 2.2325 3.4375
> # aplicando o comando tapply ao objeto insulina para o cálculo dos
> # desvio-padrões dos tratamentos
> desvio.trat <- tapply(insulina, trat, sd)
```

```
> desvio.trat
    Alto        Baixo        Medio
0.7366139  0.9513631   0.4605341
> # fazendo a análise de variância
> insulina.av <- aov(insulina~factor(trat))
>
> #imprimindo o quadro da anova
> summary(insulina.av)
#imprimindo o quadro da anova
summary(insulina.av)
```

Quadro da anova fornecido pelos comandos básicos do R

| | *Df* | *Sum Sq* | *Mean Sq* | *F value* | *Pr(>F)* |
|---|---|---|---|---|---|
| *trat* | *2* | *10.2967* | *5.1483* | *9.3054* | *0.006445 *** |
| *Residuals* | *9* | *4.9794* | *0.5533* | | |

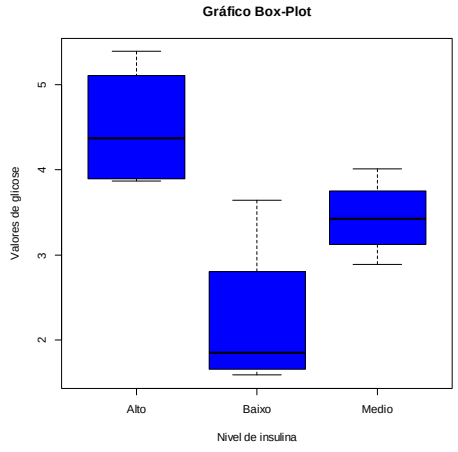
*---*

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Notem que os comandos do script acima não fornecem os resultados do da fonte de variação do Total e por outro lado fornecem uma coluna a mais referente a Pr(>Fr) com o valor de p ***(p-value)*** da estatística F. Este valor deve ser comparado com o valor de probabilidade referência ( $\propto$ ) selecionado antes de fazer a análise. Geralmente este valor de $\propto$ é fixado em 5% ou 1%.

Um gráfico muito prático no DIC é o Box-Plot para cada tratamento (nível de insulina), o qual pode ser obtido pelos comandos

```
> # mostrando o s gráficos box plot para cada tratamento
> boxplot(insulina~factor(trat),xlab="Nivel de insulina",col="blue",
+         ylab="Valores de glicose",
+         main="Gráfico Box-Plot")
```




Outra forma de obter o quadro da ANOVA é pela função ***crd( )*** do pacote ***ExpDes***. Pacotes (*packages*) ou bibliotecas (*library*) são os nomes mais usados para designar conjuntos de funções, exemplos, e documentações desenvolvidas para determinadas tarefas. Os comandos básicos do R, por exemplo, estão em uma biblioteca chamada **base**. Existem inúmeras bibliotecas, algumas já inclusas na instalação do R. No R pode-se encontrar pacotes desenvolvidos pelos responsáveis pelo R ou implementados por

usuários. A seguir apresentamos o script da instalação e do uso do pacote ExpDes.

```
> # instalando o pacote ExpDes (Experimental Designs)
> #install.packages("ExpDes")
>
> # requerendo o ExpDes
> require(ExpDes)
Carregando pacotes exigidos: ExpDes
>
> # sintaxe do comando que faz a ANOVA no ExpDes
> # crd(treat, resp, quali = TRUE, mcomp = "tukey", sigT = 0.05, sigF =
0.05)
> crd(trat,insulina,mcomp=F)

------------------------------------------------------------------------
Analysis of Variance Table
------------------------------------------------------------------------
             DF     SS       MS       Fc        Pr>Fc
Treatament    2  10.2967  5.1483   9.3054   0.0064452
Residuals     9   4.9794   0.5533
Total        11  15.2760
------------------------------------------------------------------------
CV = 21.94 %
------------------------------------------------------------------------
Shapiro-Wilk normality test
p-value:  0.08657144
According to Shapiro-Wilk normality test at 5% of significance, residuals
can be considered normal.
------------------------------------------------------------------------
```

Podemos chegar a mesma conclusão anteriormente, simplesmente analisando o valor de p (***Pr>Fc, (p=0,006445)***), o qual é bem menor que 0,01. Assim, sem recorrer à tabela F, concluímos que o teste F é significativo pelo valor de p (*p=0,006445*) fornecido pela função **crd()** do ***ExpDes***, rejeitamos $H_0$ e concluímos que a quantidade de glicose é diferente para pelo menos dois níveis de glicose.

A função ***aov( )*** armazena os valores da tabela da anova acima na forma matricial (2 x 5). Neste exemplo, ***insulina.av*** é o objeto que recebeu os resultados do quadro da análise de variância no script R listado anteriormente.

O esquema das posições de armazenamento dos resultados do quadro da anova do DIC no R é

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| **trat** | **[1,1]** | **[1,2]** | **[1,3]** | **[1,4]** | **[1,5]** |
| **Residuals** | **[2,1]** | **[2,2]** | **[2,3]** | | |

Para obter o valor do quadrado médio de resíduo basta digitar e executar o comando ***anova(insulina.av)[2,3]***.

Uma forma de se obter ajuda de alguma função no R é por meio da execução do comando *??**nome da função**()*. Por exemplo, para se obter uma ajuda da sintaxe da função ***mean( )*** deve-se executar o comando *??mean()*.

Informações sobre o pacote *ExpDes()*, basta executar o comando ***??ExpDes()***, e seguir a sequência de passos indicados abaixo

***ExpDes::ExpDes-package/***
***Index/***
***Documentation for package 'ExpDes' version 1.1.2/ExpDes-package/***
***crd***

E a seguinte exlicação aparecerá

**crd(treat,resp,quali=TRUE,mcomp="tukey",sigT=0.05,sigF=0.05)**

em que:

***treat*** *Vetor numérico contendo os tratamentos;*

***resp*** *Vetor numérico contendo a variável resposta;*

***quali*** *Lógico. Se TRUE (default), os tratamentos são assumidos qualitativos, se FALSE, quantitativos;*

***mcomp*** *Permite a escolha do teste de comparação múltiplo; o "***default***"é o teste de Tukey, entretanto, as opções são: o teste LSD ('lsd'), o teste LSD com a proteção de Bonferroni ('lsdb'), o teste de Duncan ('duncan'), o teste de Student-Newman-Keuls ('snk'), o teste de Scott-Knott ('sk') e o teste de comparação múltipla bootstrap ('ccboot');*

***sigT*** *A significância a ser usada para o teste de comparação múltipla; o "default" é 5%;*

***sigF*** *A significância a ser usada no teste F da ANOVA; o "default" é 5%* ,

são os argumentos desta função.

---

**Exemplo 2**

Em um experimento em que se mediu o peso corporal (kg), 19 porcos foram distribuídos aleatoriamente a 4 grupos. Cada grupo foi alimentado com dietas diferentes. Deseja-se testar se os pesos dos porcos são os mesmos para as 4 dietas.

Desejamos testar a hipótese nula

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$H_1 : \mu_i \neq \mu_j \text{ para pelo menos um par } i \neq j$$

As observações obtidas foram:

| *Tratamento* | *Repetições* | | | | |
|---|---|---|---|---|---|
| | *1* | *2* | *3* | *4* | *5* |
| *Dieta 1* | *60,8* | *57,7* | *65,0* | *58,6* | *61,7* |
| *Dieta 2* | *68,7* | *67,7* | *74,0* | *66,3* | *69,8* |
| *Dieta 3* | *102,6* | *102,1* | *100,2* | *96,5* | * |
| *Dieta 4* | *87,9* | *84,2* | *83,1* | *85,7* | *90,3* |

Temos um experimento desbalanceado com número de repetições desigual para os tratamentos. Então, os cálculos para montarmos o quadro da ANOVA são:

- Graus de liberdade:
  $Total = N - 1 = 19 - 1 = 18; Trat. = k - 1 = 4 - 1 = 3$
  $\text{Res} = N - k = 19 - 4 = 15$
- $CM = \frac{(1482,9)^2}{19} = 115736,44$

- $SQT = (60{,}8)^2 + ... + (90{,}3)^2 - CM = 120062{,}19 - 115736{,}44 = 4325{,}75$
- $SQTr = \frac{(303{,}8)^2}{5} + \frac{(346{,}5)^2}{5} + \frac{(401{,}4)^2}{4} + \frac{(431{,}2)^2}{5} - CM = 119938{,}52 - 115736{,}44 = 4202{,}07$
- $SQR = SQT - SQTr = 4325{,}75 - 4202{,}07 = 123.67$
- $QMTr = \frac{4202{,}07}{3} = 1400{,}69 \quad e \quad QMR = \frac{123{,}67}{15} = 8{,}24$
- $F_c = \frac{QMTr}{QmR} = \frac{1400{,}69}{8{,}24} = 169{,}89$

O quadro da ANOVA para a variável peso (kg) é o seguinte:

| *Fonte de var.* | *g.l.* | *SQ* | *QM* | $F_c$ |
|---|---|---|---|---|
| *Tratamentos* | *3* | *4202,07* | *1400,69* | *169,89* |
| *Resíduo* | *15* | *123,67* | *8,24* | |
| *TOTAL* | *18* | *4325,75* | | |

Script no R para resolver o exemplo 2

Atenção, antes de rodar este script é necessário remover todos os objetos definidos no script do exemplo 1 com o comando ***rm(list=ls(all=TRUE))***, ou pelo atalho na aba do menu da janela da console clicar em ***Misc/Remover todos os objetos***

Neste exemplo vamos organizar os dados na forma de arquivo que será a forma padrão a ser seguida a partir de agora em diante. O primeiro passo é montar o arquivo no Bloco de Nota ("notepad") e gravá-lo no diretório de trabalho com o nome de ***ex2.txt*** na seguinte forma:

```
Dieta  Peso
Dieta1  60,8
Dieta1  57,7
Dieta1  65,0
Dieta1  58,6
Dieta1  61,7
Dieta2  68,7
Dieta2  67,7
Dieta2  74,0
Dieta2  66,3
Dieta2  69,8
Dieta3  102,6
Dieta3  102,1
Dieta3  100,2
Dieta3  96,5
Dieta3  NA
Dieta4  87,9
Dieta4  84,2
Dieta4  83,1
Dieta4  85,7
Dieta4  90,3
```

"Este arquivo tem as seguintes características: é no formato linhas e colunas, ou seja, duas colunas: uma denominada de ***Dieta***, a qual recebe os níveis dos tratamentos e outra , denominada de **Peso**, com os valores do peso corporal dos suínos e 21 linhas, sendo 5 repetições para cada tratamento e uma para o cabeçalho. A última repetição da Dieta3 é uma obeservação perdida "**missing value**" . Por isso, foi colocado o código **NA** , iniciais de "**not avaliable**" para dados faltantes. Uma observação final é sobre o separador das casas decimais, neste caso a virgula".

Script no R para análise de variância

```
> # exemplo 2 da Aula 3 (DIC) pg 59
```

```
> # entrando com os dados com o comando read.table( )
> dados.peso<-read.table("ex2.txt",header=TRUE,dec=",",
                                 na.strings="NA")
> head(dados.peso)          # imprimindo as 6 primeiras linhas do arquivo
   Dieta Peso
1 Dieta1 60.8
2 Dieta1 57.7
3 Dieta1 65.0
4 Dieta1 58.6
5 Dieta1 61.7
6 Dieta2 68.7
> attach(dados.peso)        # colocando o arquivo no caminho de procura
# comando tapply() para o cálculo das médias dos tratamentos
> media.dieta <- tapply(Peso,Dieta,mean,na.rm=T)
> media.dieta
Dieta1 Dieta2 Dieta3 Dieta4
 60.76  69.30 100.35  86.24
# comando by() para o cálculo das médias dos tratamentos
> by(Peso,Dieta,mean,na.rm=T)
Dieta: Dieta1
[1] 60.76
------------------------------------------------------------
Dieta: Dieta2
[1] 69.3
------------------------------------------------------------
Dieta: Dieta3
[1] 100.35
------------------------------------------------------------
Dieta: Dieta4
[1] 86.24
> peso.av<-aov(Peso~factor(Dieta))# anova pelo comando aov()
> summary(peso.av)                  # imprimindo o quadro da anova
                Df  Sum Sq     Mean Sq       F value     Pr(>F)
factor(Dieta)  3     4202         1400.7        169.9     8.45e-12 ***
Residuals     15      124              8.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
1 observation deleted due to missingness
```

O símbolo *** na frente do valor de p (p=8.45e-12) é o código de significância fornecido pelo R. No caso indica que o teste é significativo para um valor de $\alpha$ muito pequeno em torno de 0,01%.

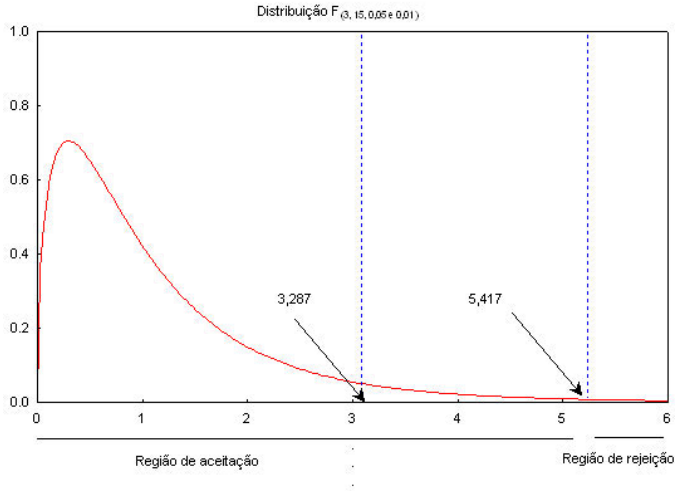
```
> require(ExpDes)                   # requerendo o pacote ExpDes
Carregando pacotes exigidos: ExpDes
> crd(Dieta,Peso,mcomp=F)        # usando o comando crd()
------------------------------------------------------------------------
Analysis of Variance Table
------------------------------------------------------------------------
                DF     SS         MS         Fc         Pr>Fc
Treatament  3    4202.1   1400.69    169.88    8.4501e-12
Residuals  15      123.7       8.24
Total          18    4325.7
------------------------------------------------------------------------
CV = NA %
```

```
---------------------------------------------------------------------
Shapiro-Wilk normality test
p-value:  0.3775306
According to Shapiro-Wilk normality test at 5% of significance, residuals can be
considered normal.
---------------------------------------------------------------------
```

A forma tradicional de interpretar o resultado do teste F da anova é consultar as tabelas das distribuições F. Desta consulta, temos que $F_{(3,15,0,05)} = 3,287$ e $F_{(3,15,0,01)} = 5,417$. O valor da estatística $F_c$=169,89 é bem superior que estes valores tabelados, assim, o valor desta estatística fornecida pelos dados esta na região de rejeição de $H_0$, logo rejeitamos a hipótese nula $H_0$ a um nível $\alpha = 0,01$, ou 1% de probabilidade (se é significativo a 1%, logo também é significativo a 5%).

Atenção! Pode-se chegar a esta mesma conclusão analisando somente pelo valor de p associado à estatística F calculada, o qual é apresentado na forma exponecial *p=8,45 e-12* ou $p=8,45 \times 10^{-12}$, bem menor que 0,001, portanto significativo a 0,1%.

Graficamente a regra de decisão fica




Evidentemente que o valor 189,88 esta bem a direita do valor crítico 5,417, assim podemos concluir que, para um nível de $\alpha = 0,001$, ou 0,01%, que os pesos dos porcos são diferentes para pelo menos duas dietas.

**12 Estimadores de mínimos quadrados.**

Nesta seção mostraremos os estimadores dos termos do modelo matemático do DIC $y_{ij} = \mu + \tau_i + e_{ij}$, os quais são obtidos minimizando-se a expressão do erro deste modelo

$$\sum_{i=1}^{k}\sum_{j=1}^{r_i} (y_{ij} - \hat{y}_{ij})^2,$$

em relação a $\mu$ e $\tau_i$, i=1, 2, ...k, sujeito a restrição $\sum_{i=1}^{k} r_i\tau_i = 0$. Assim procedendo, obtemos os estimadores de $\mu$, $\tau_i$ e $\mu_i$, dados por $\hat{\mu} = \bar{y}_{++}$, $\hat{\tau}_i = \bar{y}_{i+} - \bar{y}_{++}$ e de $\hat{\mu}_i = \hat{\mu} - \hat{\tau}_i = \bar{y}_{i+}$, $i = 1,2,...,k$.

Para construir um intervalo de confiança para a média de cada tratamento, devemos notar que a estatística:

$$\frac{\bar{y}_{i+} - \mu_i}{\sqrt{\frac{QMR}{r_i}}} \sim t_{(n-k)},$$

i.é., tem distribuição *t – Student* com (n – k) graus de liberdade. Um intervalo de confiança para $\mu_i$ com um coeficiente de confiança $(1-\alpha)$ é dado pela expressão

$$IC(\mu_i; 1-\alpha) = \bar{y}_{i+} \pm t_{\left(\frac{\alpha}{2}; N-k\right)} \sqrt{\frac{QM\,\mathrm{Res}}{r}},$$

sendo, $t_{\left(\frac{\alpha}{2}, N-k\right)}$ o quantil de ordem $\left(1-\frac{\alpha}{2}\right)$ da distribuição *t* – Student com (n – k) graus de liberdade, os mesmos graus de liberdade do resíduo da ANOVA.

Como exemplo, vamos considerar os dados do experimento apresentado no item 1, cujos cálculos foram mostrados no item 10. As médias destes dados são:

- $\bar{y}_{1+} = \frac{8{,}93}{4} = 2{,}23$ ; $\bar{y}_{2+} = \frac{13{,}75}{4} = 3{,}44$; $\bar{y}_{3+} = \frac{18{,}00}{4} = 4{,}50$ *e* ; $\bar{y}_{++} = 3{,}39$
- do quadro da ANOVA temos os valores de SQR para calcular $\sqrt{\frac{QMR}{r}} = \sqrt{\frac{0{,}553}{4}} = 0{,}372$ ;
- o valor de $t_{(0{,}025, 9)} = 2{,}262$ .

Assim, os intervalos são dados por:

$$IC(\mu_i; 95\%) = \bar{y}_{i+} \pm 2{,}262\sqrt{\frac{0{,}553}{4}} = \bar{y}_{i+} \pm 0{,}841$$

Resumindo temos o quadro a seguir

| | ***Nível baixo de glicose*** | ***Nível médio de glicose*** | ***Nível alto de glicose*** |
|---|---|---|---|
| $\bar{y}_i$ | *2,23* | *3,44* | *4,50* |
| $IC(\mu_i, 95\%)$ | *(1,389; 3,071)* | *(2,599; 4,281)* | *(3,659; 5,341)* |

Problema: identificar quais os níveis de glicose (tratamentos) que tiveram efeitos não nulos sobre a liberação de insulina dos tecidos.

Script no R para o calculo dos intervalos de confiança do exemplo 1. Atenção! É necessário executar novamente o script das páginas 55 e 56.

```
> # definindo os objetos para o cálculo dos IC´s
> # obtenção dos gl do residuo no quadro da anova
> glr <- anova(insulina.av)[2,1]
> glr
[1] 9
>
> # obtenção da QMR no quadro da anova
> qmr <- anova(insulina.av)[2,3]
> qmr
[1] 0.5532611
```

```
> # intervalo de confiança para o nível baixo de glicose
> ic.baixo <- media.trat[2] + qt(c(0.025, 0.975), df = glr) * sqrt(qmr/r)
> ic.baixo
[1] 1.391187 3.073813ic.baixo

> # intervalo de confiança para o nível médio de glicose
> ic.medio <- media.trat[3] + qt(c(0.025, 0.975), df = glr) * sqrt(qmr/r)
> ic.medio
[1] 2.596187 4.278813

> # intervalo de confiança para o nível baixo de glicose
> ic.alto <- media.trat[1] + qt(c(0.025, 0.975), df = glr) * sqrt(qmr/r)
> ic.alto
[1] 3.658687 5.341313
```

Como segundo exemplo, vamos considerar os dados do experimento apresentado no item 11. As médias destes dados são:

- $\bar{y}_{1+} = \frac{303,8}{5} = 60,76$ ; $\bar{y}_{2+} = \frac{346,50}{5} = 69,30$; $\bar{y}_{3+} = \frac{401,4}{4} = 100,35$; $e\ \ \bar{y}_{4+} = \frac{431,2}{5} = 86,24$ *e a média geral é* $\bar{y}_{++} = 79,13$
- do quadro da ANOVA temos o valor do QMR para calcular desvio padrão médio para os tratamentos 1, 2 e 4 é $\sqrt{\frac{QMR}{r_i}} = \sqrt{\frac{8,557}{5}} = 1,31$. Para o terceiro tratamento o erro padrão médio é $\sqrt{\frac{QMR}{r_i}} = \sqrt{\frac{8,557}{4}} = 1,46$
- o valor de $t_{(0,025;\,15)} = 2,1314$ .

Assim, os intervalos são dados por:

$$IC(\mu_i; 95\%) = \bar{y}_{i+} \pm 2,1314\sqrt{\frac{8,557}{5}} = \bar{y}_{i+} \pm 1,31,\ para\ i = 1,2,e\ 4$$

$$e\ IC(\mu_i; 95\%) = \bar{y}_{i+} \pm 2,1314\sqrt{\frac{8,557}{4}} = \bar{y}_{i+} \pm 1,46,\ para\ i = 3$$

Resumindo temos o quadro abaixo

| | ***Dieta 1*** | ***Dieta 2*** | ***Dieta 3*** | ***Dieta 4*** |
|---|---|---|---|---|
| $\bar{y}_i$ | *60,76* | *69,30* | *100,35* | *86,24* |
| $IC(\mu_i, 95\%)$ | *(58,02; 63,48)* | *(66,56; 72,04)* | *(97,29; 103,41)* | *(83,50; 88,98)* |

Problema: identificar quais as Dietas (tratamentos) que tiveram efeitos não nulos sobre o peso dos suínos.

Script no R para calcular os IC´s do exemplo 2. Antes porém execute novamente o script do R descritos na página 59 e 60.

```
# definindo os objetos para o cálculo dos IC´s
#  definindo o vetor de repetições dos tratamentos
r<- c(5,5,4,5)
```

```
> # definindo os objetos para o cálculo dos IC´s
> #  definindo o vetor de repetições dos tratamentos
> r<- c(5,5,4,5)
>
> # obtenção dos gl do residuo no quadro da anova
> glr <- anova(peso.av)[2,1]
> glr
[1] 15
>
> # obtenção da QMR no quadro da anova
> qmr <- anova(peso.av)[2,3]
> qmr
[1] 8.244933
>
> # intervalo de confiança para a Dieta1
> ic.dieta1 <- media.dieta[1] + qt(c(0.025, 0.975), df = glr) * sqrt(qmr/r[1])
> ic.dieta1
[1] 58.02294 63.49706
>
> # intervalo de confiança para a Dieta2
> ic.dieta2 <- media.dieta[2] + qt(c(0.025, 0.975), df = glr) * sqrt(qmr/r[2])
> ic.dieta2
[1] 66.56294 72.03706
>
> # intervalo de confiança para a Dieta3
> ic.dieta3 <- media.dieta[3] + qt(c(0.025, 0.975), df = glr) * sqrt(qmr/r[3])
> ic.dieta3
[1]  97.28988 103.41012
>
> # intervalo de confiança para a Dieta4
> ic.dieta4 <- media.dieta[4] + qt(c(0.025, 0.975), df = glr) * sqrt(qmr/r[4])
> ic.dieta4
[1] 83.50294 88.97706
```

---

## 13 Coeficientes de determinação ($R^2$) e de variação *(CV)*

A parte da *Soma de Quadrados Total* (***SQT***), a variação total nas observações , que pode ser explicada pelo modelo matemático do DIC, é denominada de *coeficiente de determinação.* Assim, o coeficiente de determinação para modelo do DIC, $y_{ij} = \mu + \tau_i + e_{ij}$, é definido como

$$R^2 = \frac{SQTr}{SQT} x\, 100\,\%.$$

Pode ser verificado que $0 \leq R^2 \leq 100$ e que $R^2 = 100\%$ quando toda variabilidade nas observações esta sendo explicada pelo modelo matemático do DIC.

A variabilidade entre as unidades experimentais de experimentos envolvendo diferentes unidades de medidas e/ou tamanhos de parcelas pode ser comparada pelos *coeficientes de variação*, os quais expressam o desvio padrão por unidade experimental como uma porcentagem da média geral do experimento, ou seja,

$$CV = \frac{S}{\bar{y}_{++}} x\,100\,\%\,.$$

Da ANOVA sabemos que $S = \sqrt{QMR}$ , daí resulta que

$$CV = \frac{\sqrt{QMR}}{\bar{y}_{++}} * 100\,.$$

Como exemplo vamos considerar os dados do experimento apresentado no item 1, cujos cálculos foram mostrados no item 10. Neste exemplo temos:

$SQT = 15{,}28$ *e* $SQTr = 10{,}30$, *então*

- $R^2 = \frac{SQTr}{SQT} = \frac{10{,}30}{15{,}28} x\,100 = 67{,}4\,\%$
- $CV = \frac{\sqrt{QMR}}{\bar{y}_{++}} * 100 = \frac{\sqrt{0{,}55}}{3{,}39} * 100 = 21{,}88\,\%$

Concluímos que 67,4% da variabilidade que existe nas observações deste experimento em torno de seu valor médio é explicada pelo modelo matemático do DIC e este experimento apresenta um coeficiente de variação de aproximadamente 22%.

Script no R para calcular os coeficientes de determinação ($R^2$ ) e de variação (*CV)*

```
> # calculo do CV
> cv <- sqrt(qmr)/mean(Peso,na.rm=T)*100
> cv
[1] 3.679047
>
> sqtr <- anova(peso.av)[1,2]  # obtenção da SQTr da anova
> sqtr
[1] 4202.073
> sqr <- anova(peso.av)[2,2]   # obtenção da SQR da anova
> sqr
[1] 123.674
```

---

## 14 Checando as violações das suposições de normalidade dos dados e da homogeneidade das variâncias dos tratamentos Anova

De um modo geral, o teste F da ANOVA não é muito sensível às violações da suposição de distribuição normal. Ele também é moderadamente insensível às violações de variâncias iguais, se os tamanhos das amostras são iguais e não muito pequenas em cada tratamento. Entretanto, variâncias desiguais podem ter um efeito marcante no nível do teste, especialmente se amostras pequenas estão associadas com tratamentos que têm as maiores variâncias. Existe uma série de procedimentos para se testar se as suposições da ANOVA são violadas. Entre estes temos o *teste de Anderson-Darling, teste de Shapiro-Wilks e teste de Kolmogorov-Smirnov,* que testam a normalidade da população. A igualdade das variâncias (homocedasticidade) pode ser testada pelo *teste de Bartlett*. Com o advento dos modernos computadores, métodos gráficos são ferramentas muito populares para a visualização das violações das suposições teóricas da ANOVA. Alguns destes métodos gráficos mais

comumente usados para checar as suposições da ANOVA são baseados em gráficos dos resíduos.

**Resíduos**. O resíduo correspondente a uma observação $y_{ij}$ é definido como:

$$e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \hat{\mu} - \hat{\tau}_i = y_{ij} - \bar{y}_{i+},$$

ou seja, o resíduo corresponde á parte da observação que não foi explicada pelo modelo. Calculando os resíduos correspondentes a todas as observações de um experimento e analisando-os descritivamente de forma apropriada, podemos ter alguma indicação, graficamente, se as suposições da ANOVA estão sendo satisfeitas.

**Gráfico dos resíduos para testar a normalidade.** Técnicas gráficas para checar se uma amostra de resíduos é provenientes de uma população normal incluem os gráficos do Histograma, do Box – Plot, etc. Outra importante técnica é o *gráfico* ***q-q normal*** *(quantile-quantile normal plot)*. O gráfico ***q-q normal***, é um gráfico entre os resíduos e um conjunto de percentis devidamente escolhidos da normal padronizada. Sob a hipótese de normalidade este gráfico ***q-q normal*** deve se aproximar de uma reta. Se o gráfico é sigmóide é uma indicação de que a população tem as caudas pesadas ou leves. A assimetria é indicada por gráficos côncavos (assimetria a esquerda) e convexos (assimetria a direita).

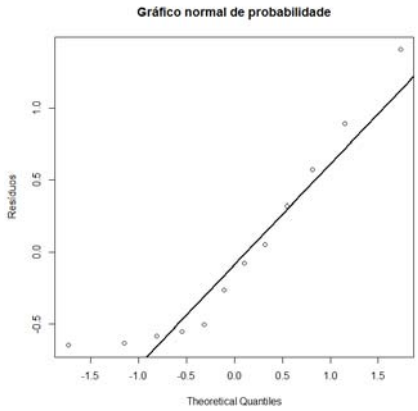
O primeiro passo na construção de um gráfico ***q-q normal*** é o cálculo de $p_{ij} = \frac{n^o\, de\ resíduos \le e_{ij}}{N+1}$, a qual é denominada de probabilidade empírica acumulada, e está associada a todo $e_{ij}$, de tal forma que $p_{ij} = \frac{posto\ de\ e_{ij}}{N+1}$. Como exemplo, a probabilidade empírica acumulada associada ao resíduo, cujo posto é o sexto (seu rank=6) em um conjunto de N=10 resíduos é *p=6/11 = 0.545.* O gráfico ***q-q normal*** de um conjunto de resíduos é obtido com o gráfico dos resíduos $e_{ij}$ *vs* $q_{ij} = z_\alpha(1 - p_{ij})$ sendo que: $z_\alpha$ é o valor critico de nível $\alpha$ de uma distribuição normal padronizada

Vamos considerar os dados apresentados no item 1 e construir um *gráfico q-q normal* para ver se a suposição de normalidade parece razoável para a quantidade de insulina liberada do exemplo 1.

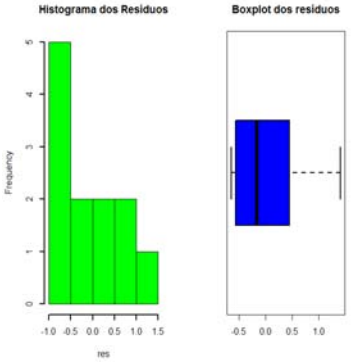
O Quadro abaixo apresenta os dados, o valor estimado pelo modelo, os resíduos e os percentis associados:

| *i* | *j* | $Y_{ij}$ | $Y_{est}$ | $e_{ij}$ | $R(e_{ij})$ | $P_{ij}$ | $Q_{ij}$ |
|---|---|---|---|---|---|---|---|
| *1* | *1* | *1.59* | *2.23* | *-0.64* | *1* | *0.077* | *-1.426* |
| *1* | *2* | *1.73* | *2.23* | *-0.50* | *5* | *0.385* | *-0.293* |
| *1* | *3* | *3.64* | *2.23* | *1.41* | *12* | *0.923* | *1.426* |
| *1* | *4* | *1.97* | *2.23* | *-0.26* | *6* | *0.462* | *-0.097* |
| *2* | *1* | *3.36* | *3.44* | *-0.08* | *7* | *0.538* | *0.097* |
| *2* | *2* | *4.01* | *3.44* | *0.57* | *10* | *0.769* | *0.736* |
| *2* | *3* | *3.49* | *3.44* | *0.05* | *8* | *0.615* | *0.293* |
| *2* | *4* | *2.89* | *3.44* | *-0.55* | *4* | *0.308* | *-0.502* |
| *3* | *1* | *3.92* | *4.50* | *-0.58* | *3* | *0.231* | *-0.736* |
| *3* | *2* | *4.82* | *4.50* | *0.32* | *9* | *0.692* | *0.502* |
| *3* | *3* | *3.87* | *4.50* | *-0.63* | *2* | *0.154* | *-1.020* |
| *3* | *4* | *5.39* | *4.50* | *0.89* | *11* | *0.846* | *1.020* |

e o gráfico *q-q normal* ( $e_{ij}$ *x* $q_{ij}$ ) fica sendo:




e os gráficos do Histograma e do Box – Plot dos resíduos ficam:




Pelo gráfico ***qq normal***, pelo histograma e pelo Box-Plot é razoável supor a normalidade para os dados de liberação de insulina.

O script do R que fornece os resultados acima são:

```
> # extraindo os resíduos do objeto pc.av
> residuo <-insulina.av$res
> residuo
     1      2      3      4      5      6      7      8      9     10     11     12
-0.6425 -0.5025  1.4075 -0.2625 -0.0775  0.5725  0.0525 -0.5475 -0.5800  0.3200 -0.6300  0.8900
>
> # fazendo o gráfico q-q plot
> qqnorm(residuo, ylab ="Resíduos",main="Gráfico normal de
probabilidade")
> qqline(residuo,lwd=2)
> # dividindo a tela gráfica em 2 colunas e uma linha
> par(mfrow=c(1,2))
>
> # histograma dos resíduos
> hist(residuo, main="Histograma dos Resíduos",lwd=2,col="green")
>
> #   gráfico boxplot dos resíduos
> boxplot(residuo,  horizontal=T,main="Boxplot dos resíduos",
+ col="blue",lwd=2)
```

Estes recursos gráficos não são quantitativos. É necessário um teste. O script no R que fornece o teste de normalidade de Shapiro-Wilks para testar as hipóteses:

$H_0$ : *a população amostrada tem distruição normal*

$H_1$ : *a população amostrada não tem distruição normal*

*ou*

$H_0 : e_{ij} \sim N(0, \sigma^2)$

$H_1 : e_{ij}$ *não tem* $N(0, \sigma^2)$

é dado a seguir

```
> # teste de normalidade de Shapiro-Wilks dos dados do exemplo 1
> shapiro.test(residuo)

        Shapiro-Wilk normality test

data:  residuo
W = 0.8796, p-value = 0.08657
```

No resultado fornecido pelo R e pelo valor de p (*p=0,08657*) associado a estatística *W=0,8796* do teste de *Shapiro-Wilks* é não significativo,portanto não rejeitamos $H_0$, logo é razoável supor a normalidade para os dados de liberação de insulina. O teste de *Bartlett* testa as hipóteses

$$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$$

$$H_1 : \sigma_i^2 \neq \sigma_j^2 \quad i \neq j \text{ ,}$$

ou seja, a homogeneidade das variâncias dos tratamentos.

O script no R que fornece este teste é

```
> # teste de homogeneidade das variâncias dos tratamentos dos dados do
> #exemplo 1 (Teste de Bartlett)
> bartlett.test(insulina ~ trat)

        Bartlett test of homogeneity of variances

data:  insulina by trat
Bartlett's K-squared = 1.27, df = 2, p-value = 0.5299
```

Pelo valor de p=0,5299 do teste de Bartlett, este teste não é significativo, portanto não rejeitamos $H_0$. Concluímos, então, que a homogeneidade das variâncias é uma suposição plausível para os dados da liberação da insulina. Assim é razoável supor que este conjunto de dados suporta as suposições básicas de normalidade e homogeneidade da variância para a correta aplicação da ANOVA.

---

## 14 Vantagens e desvantagens do DIC

As principais vantagens do DIC são:

- é fácil de ser planejado e é flexível quanto ao número de tratamento e de repetições tendo como única limitação o número de unidades experimentais disponíveis para o experimento;

- o número de repetições pode variar de tratamento para tratamento, embora o desejável é ter o mesmo número de unidades experimentais em todos os tratamentos;
- o DIC proporciona o número máximo de graus de liberdade para o resíduo;
- a análise estatística é simples mesmo que se perca algumas unidades experimentais.

Algumas desvantagens são:

- é mais apropriado para um pequeno número de tratamentos e para um material experimental homogêneo;
- todas as fontes de variação não associadas aos tratamentos farão parte do **resíduo**, podendo comprometer a precisão das análises;
- super-estima a variância residual.

## 15 Resumo

O DIC é mais útil onde não existe nenhuma fonte de variação identificável entre as unidades experimentais, exceto às dos efeitos dos tratamentos. É o mais flexível com respeito ao arranjo físico das unidades experimentais. Ele maximiza os graus de liberdade para a estimação da variância por unidade experimental (erro experimental ou erro residual) e minimiza o valor da estatística F requerida para a significância estatística.

## 3º EXERCÍCIO PRÁTICO DE ESTATÍSTICA EXPERIMENTAL

1- Para avaliar o efeito de altos níveis de cobre na alimentação de pintinhos, seis pintinhos foram alimentados com uma dieta basal padrão às quais foram adicionadas três níveis de cobre (0, 400, e 800 ppm). Os dados abaixo mostram a razão da eficiência da dieta (g dieta/ g ganho de peso) ao final de 3 semanas
Use o R para apresentar os resultados.

| *Tratamentos* | *Pintinhos* | | | | | |
|---|---|---|---|---|---|---|
| *(nível de cobre)* | *1* | *2* | *3* | *4* | *5* | *6* |
| *0* | *1,57* | *1,54* | *1,65* | *1,57* | *1,59* | *1,58* |
| *400* | *1,91* | *1,71* | *1,55* | *1,67* | *1,64* | *1,67* |
| *800* | *1,88* | *1,62* | *1,75* | *1,97* | *1,78* | *2,20* |

(extraído de Statistical Research Methods in the Life Science, P. V. Rao, pg. 287).

(a) Calcular os totais dos tratamentos $y_{i+}$ , i=1,2,3, as médias dos tratamentos $\bar{y}_{i+}$ , os desvios padrões dos tratamentos $s_i$ , *i=1,2,3,* o total geral $y_{++}$, e a média geral $\bar{y}_{++}$.

(b) Estabelecer as hipóteses estatísticas $H_0$ e $H_1$ e as suposições básicas para se testar estas hipóteses.

(c) Monte o quadro da anova

(d) Com base nos resultados do teste F da anova faça as conclusões pertinentes sobre as hipóteses do item (b).

(e) Calcular os intervalos de confiança das médias dos tratamentos $IC(\mu_i\ ;\ 95\%)$. Apresente os resultados. (Siga o modelo da tabela pg 62 da apostila).

(f) Calcular os coeficientes: de determinação $\boldsymbol{R^2}$ e o de variação do experimento (***CV***). Comente os resultados.

(g) Verifique as suposições básicas da ANOVA. Apresente e comente os resultados.

2- Num experimento inteiramente casualizado com 5 tratamentos e 4 repetições, estudou-se o efeito de 5 carrapaticidas (tratamentos) no controle de carrapatos em bovinos.
Analisando- se o número de carrapatos que cairam por animal, obtiveram-se as seguintes somas de quadrados:
S.Q. Tratamentos = 41,08
S.Q. Total = 57,46
Estabelecer as hipóteses estatísticas $H_0$ e $H_1$, montar o quadro de análise de variância, concluir e calcular o coeficiente de determinação $R^2$

3- Cite as vantagens e as desvantagens do delineamento inteiramente casualizado.
4- Escreva o modelo matemático do delineamento inteiramente casualizado para os dados apresentados na 2ª questão.
5- Descreva os procedimentos de um experimento ***cego***, e dos experimentos duplamente ***cego***.
6- Quando um experimento será considerado planejado (Descreva as etapas).
7- Quais os princípios básicos da experimentação.

## Aula 4 Testes de comparações múltiplas

### 1 Introdução

Os testes de comparações múltiplas também conhecidos como testes de comparações de médias servem como um complemento ao teste F da análise de variância quando este é significativo e são usados para detectar diferença entre médias. Considere o exemplo a seguir

**Exemplo 1.** Em um experimento de alimentação de porcos, foram utilizados quatro rações (A, B, C e D), cada uma fornecida a 5 animais. Os ganhos de peso, kg, foram:

| ***Rações*** | | | |
|---|---|---|---|
| ***A*** | ***B*** | ***C*** | ***D*** |
| *35* | *40* | *39* | *27* |
| *19* | *35* | *27* | *12* |
| *31* | *46* | *20* | *13* |
| *15* | *41* | *29* | *28* |
| *30* | *33* | *45* | *30* |

Calculando-se as somas de quadrados podemos construir o seguinte quadro de análise de variância:

| *F.V.* | *g.l.* | *S.Q.* | *QM* | $F_c$ |
|---|---|---|---|---|
| *Rações* | *3* | *823,75* | *274,58* | *3,99* |
| *Resíduo* | *16* | *1100,00* | *68,75* | |
| *Total* | *19* | *1923,75* | | |

- Das tabelas das distribuições F, temos que $F_{(3,16,0,05)} = 3,24$ e $F_{(3,16,0,01)} = 5,29$. O valor $F_c$=3,99 é maior que o valor do F tabelado a 5%, então, rejeitamos a hipótese nula $H_0$ a 5% de probabilidade.
- Dúvida: Qual é a ração que tem o melhor desempenho no ganho de peso?

Para responder a questão, conheceremos alguns ***PROCEDIMENTOS DE COMPARAÇÕES DE MÚLTIPLAS*** ou ***MÉTODOS DE COMPARAÇÕES DE MÉDIAS***, como por exemplo, os testes *t-Student* , *Scheffé*, *Tukey*, *Duncan*, *Dunnett* e *Bonferroni*, dentre outros.

### 2 Definições básicas

Consideremos um experimento com k tratamentos, cujas médias populacionais são $\mu_1, \mu_2, ..., \mu_K$ e seus estimadores $\bar{y}_1, \bar{y}_2, ..., \bar{y}_k$ foram obtidas de amostras de tamanhos $r_1, r_2, ..., r_K$.

**Definição 1** Um contraste de médias é qualquer função do tipo

$$Y = c_1\mu_1 + c_2\mu_2 + ... + c_k\mu_k ,$$

com $\sum_{i=1}^{k} c_i = c_1 + c_2 + ... + c_k = 0$ e $\mu_i$, é a média do tratamento *i = 1, 2, ..., k*

**Definição 2** Dizemos que dois contrastes são ortogonais se $\sum_{i=1}^{k} \frac{a_i b_i}{r_i} = 0$ . Quando o experimento é balanceado ($r_i = r$) a condição de ortogonalidade é que a soma dos produtos de seus coeficientes é nula, i.é., $\sum_{i=1}^{k} a_i b_i = 0$ .

- Quando um experimento envolve k tratamentos, podemos definir diversas comparações entre as k médias, mas somente $(k - 1)$ são ortogonais;
- Nos contrastes envolvendo duas médias podemos definir $\frac{k(k-1)}{2}$ contrastes possíveis, os quais não são ortogonais.

Supondo que os tratamentos têm variância constante $\sigma^2$ e que uma estimativa não viesada desta variância é o QMR da ANOVA, tem-se que:

- $\hat{Y} = c_1\bar{x}_1 + c_2\bar{x}_2 + c_3\bar{x}_3 + \cdots + c_n\bar{x}_k$ é um estimador não viesado do contraste $Y = c_1\mu_1 + c_2\mu_2 + \ldots + c_k\mu_k$;
- $V(\hat{Y}) = (c_1^2 + c_2^2 + \cdots + c_n^2)\frac{\sigma^2}{r_i} = \sum_{i=1}^{n} c_i^2 \frac{\sigma^2}{r_i}$ e um estimador não viesado é dado por $\hat{V}(\hat{Y}) = (c_1^2 + c_2^2 + \cdots + c_n^2)\frac{QMR}{r_i} = \sum_{i=1}^{n} c_i^2 \frac{QMR}{r_i}$, se o experimento é balanceado $r_1 = r_2 = \ldots = r_K = r$, as expressões acima ficam, respectivamente,

$V(\hat{Y}) = (c_1^2 + c_2^2 + \cdots + c_n^2)\frac{\sigma^2}{r} = \sum_{i=1}^{n} c_i^2 \frac{\sigma^2}{r}$ e

$\hat{V}(\hat{Y}) = (c_1^2 + c_2^2 + \cdots + c_n^2)\frac{QMR}{r} = \sum_{i=1}^{n} c_i^2 \frac{QMR}{r}$

**Exemplo 1**. Em um experimento dois antibióticos em duas dosagens cada um para a cura da mastite em bovinos. A variável resposta é tempo de cura em dias

| *Tratamento* | *Descrição* |
|---|---|
| *T1* | *Dose baixa da droga A* |
| *T2* | *Dose alta da droga A* |
| *T3* | *Dose baixa da droga B* |
| *T4* | *Dose alta da droga B* |

Podemos definir os seguintes contrastes:

- $Y_1 = \mu_1 + \mu_2 - \mu_3 - \mu_4$ : compara as doses da droga A com as doses da droga B;
- $Y_2 = \mu_1 - \mu_2$: compara as doses da droga A;
- $Y_3 = \mu_3 - \mu_4$: compara as doses da droga B.

A afirmação de que o contraste $Y_1$ é nulo ($Y_1 = 0$) é o mesmo que afirmar que: $\mu_1 + \mu_2 = \mu_3 + \mu_4$, *ou que*, $\frac{\mu_1 + \mu_2}{2} = \frac{\mu_3 + \mu_4}{2}$, ou ainda, que a média dos tratamentos 1 e 2 é igual à média dos tratamentos 3 e 4.

Para verificarmos se estes contrastes são ortogonais é aconselhável uma tabela com os coeficientes dos *(k – 1*) contrastes e a partir daí, verificar que a soma dos produtos dos coeficientes, aos pares, é nula.

| *Contraste* | $\mu_1$ | $\mu_2$ | $\mu_3$ | $\mu_4$ |
|---|---|---|---|---|
| $Y_1$ | *+1* | *+1* | *-1* | *-1* |
| $Y_2$ | *+1* | *-1* | *0* | *0* |
| $Y_3$ | *0* | *0* | *+1* | *-1* |

Portanto estes contrastes são ortogonais 2 a dois e ortogonais entre si.

**3 Teste *t - student*** O teste *t – student* pode ser utilizado para comparar médias de tratamentos. Os requisitos básicos para sua utilização são:

- as comparações devem ser determinadas *a priori*, ou seja, antes de serem examinados os dados.
- não existe limite para o número de contrastes envolvendo as médias de tratamentos, porém, o número de contrastes ortogonais é, no máximo, igual ao número de graus de liberdade dos tratamentos.
- A ortogonalidade entre os contrastes de médias garante independência entre as conclusões.

O objetivo é testar a hipótese

$$H_0 : Y_i = 0$$
$$H_1 : Y_i \neq 0 \text{ ,}$$

Usamos a estatística $t = \frac{\hat{Y}_i}{\sqrt{\hat{V}(\hat{Y})}} = \frac{\hat{Y}_i}{\sqrt{\frac{QMR}{r}\sum_{i=1}^{k} c_i^2}} \sim t_{(gl\ res,\ \alpha)}$, a qual sob $H_0$ verdadeira tem distribuição *t-student* com o mesmo número de graus de liberdade do resíduo, no DIC é *( n-k )*. Para um valor fixado de nível de significância $\alpha$, devemos buscar o valor de *t* tabelado (arquivo Tab_tstudent, disponibilizado na página ou nos livros indicados na bibliografia) e compará-lo com o valor da estatística $t_c$, calculada para o contraste $Y_i$ e aplicar a regra de decisão:

- Se $|t_c| \geq t_{Tabelado}$ rejeitamos $H_0$ para um determinado valor de $\alpha$, geralmente 5% ou 1%, caso contrário ($|t_c| < t_{Tabelado}$), não rejeitamos $H_0$.

*(veja o esquema gráfico desta regra de decisão apresentado no item 6 da 2ª Aula).*

**Exemplo 2**: Num experimento inteiramente casualizado com 4 tratamentos e 4 repetições, estudaram-se os efeitos de Bacitracina de zinco(BDZ) e Anti-stress sobre frangos de corte alimentados com rações à base de sorgo, desde a fase inicial até a final. A resposta medida foi conversão alimentar. Foram utilizados os seguintes tratamentos:

| Tratamento | Descrição | Média(kg) |
|---|---|---|
| 1 | Concentrado Comercial + Milho | 2,03 |
| 2 | Concentrado Comercial + Sorgo | 2,24 |
| 3 | Concentrado Comercial + Sorgo + BDZ | 2,04 |
| 4 | Concentrado Comercial + Sorgo + Anti-stress | 2,22 |

Sabendo-se que da ANOVA o valor do $QMR = 0,0044375$ , com 12 graus de liberdade. Pode - se estabelecer os contrastes de médias dos tratamentos para cada componente do desdobramento:

- *Milho vs. sorgos*, o qual é expresso pela combinação linear $Y_1 = 3\mu_1 - \mu_2 - \mu_3 - \mu_4$, *estimado por* $\hat{Y}_1 = 3\bar{y}_1 - \bar{y}_2 - \bar{y}_3 - \bar{y}_4$ ;
- *Sorgo vs. Sorgo + Aditivos*, o qual é expresso pela combinação linear $Y_2 = 2\mu_2 - \mu_3 - \mu_4$, *estimado por* $\hat{Y}_2 = 2\bar{y}_2 - \bar{y}_3 - \bar{y}_4$ ;
- *Bacitracina vs. Anti-stress*, o qual é expresso por $Y_3 = \mu_3 - \mu_4$, *etimado por* $\hat{Y}_3 = \bar{y}_3 - \bar{y}_4$ ;

A verificação se os contrastes são ortogonais pode ser feita facilmente no quadro abaixo:

| Contraste | $\mu_1$ | $\mu_2$ | $\mu_3$ | $\mu_4$ | $\hat{Y}_l$ | $\sum_{l=1}^{4} c_i^2$ | $t_c$ |
|---|---|---|---|---|---|---|---|
| $Y_1$ | +3 | -1 | -1 | -1 | -0,41 | 12 | -3,55 (p=0,00198) |
| $Y_2$ | 0 | +2 | -1 | -1 | 0,22 | 6 | 2,70 (p=0,0097) |
| $Y_3$ | 0 | 0 | +1 | -1 | -0,18 | 2 | -3,82 (p=0,0012) |

p< 0,01 significativo a 1% e a 5%; p< 0,05 significativo a 5% e p> 0,05 não-significativo a 5%.

O objetivo é testar a hipótese $\begin{matrix} H_0 : Y_i = 0 \\ H_1 : Y_i \neq 0 \end{matrix}$ , para *i = 1,2,3*.

Assim, para o contraste $Y_1$ , temos que:

- $H_0 : Y_1 = 0$
- $H_1 : Y_1 \neq 0$

  $\hat{Y}_1 = 3(2,03) - (2,24) - (2,04) - (2,22) = -0,41$ *e*
- $\hat{V}(\hat{Y}_1) = \frac{QMR}{r}\sum_{i=1}^{4} c_i^2 = \frac{0,0044375}{4} 12 = 0,0133$
- $t_c = \frac{\hat{Y}_1}{\sqrt{\frac{QMR}{r}\sum_{i=1}^{4} c_i^2}} = \frac{-0,41}{\sqrt{0,0133}} = -3,55$
- $t_{(12,\ 0,025)} = 2,179$ . Como $|t_c| > t_{Tab}$ , então rejeitamos $H_0$ (0,005<p<0,001). *(Repetir estes passos para os contrastes $Y_2$ e $Y_3$ )*.

Script do R para o cálculo dos resultados apresentados acima

```
> # é necessário fornecer os valores
> # definindo o número de repetições
> r <- 4
>
> # definindo os graus de liberdade do resíduo
> glr <-  12
> # quadrado médio do resíduo
```

```
> qmr <- 0.0044375
>
> # definindo as médias dos tratamentos
> m.trat <- c(  2.03, 2.24, 2.04, 2.22)
>
> # definindo os coeficientes do contraste
> c <- c( 3, -1,-1,-1)
>
> #calculo da variância do contraste
> var.c<- qmr/r*sum(c^2)
>
> # cálculo da estatística tc da estatística t-student
> tc <- sum(c*m.trat)/sqrt(qmr/r*sum(c^2))
> tc
[1] -3.553481
>
> # cálculo do valor de p associado à estatistica t calculada anteriormente
> valor.p<- 1-pt(abs(tc),glr)
> valor.p
[1] 0.001985572
```

(repita este procedimento adaptando-o aos demais contrastes)

Com base nos resultados dos testes de hipóteses, concluímos que:

- os animais tratados com o concentrado comercial + milho têm uma conversão alimentar melhor do que os animais tratados com concentrado comercial + sorgo;
- os animais tratados com o concentrado comercial + sorgo+aditivos têm uma conversão alimentar melhor do que os animais tratados com concentrado comercial + sorgo, ou seja, os aditivos BDZ e anti-stress quando adicionados ao concentrado comercial não melhoram a conversão alimentar;
- os animais tratados com o concentrado comercial + sorgo+BDZ têm uma conversão alimentar melhor do que os animais tratados com concentrado comercial + sorgo+anti-stress.

---

**4 Teste de Scheffé**

O teste de Scheffé pode testar qualquer contraste envolvendo médias de tratamentos do tipo $Y = c_1\mu_1 + c_2\mu_2 + ... + c_k\mu_k$ definido *a priori* ou não, sendo baseado na estatística ***S***, definida como:

| $r_i = r$ *para todo* $i$ *(Experimento balanceado)* | $r_i \neq r_j$ *para* $\forall\ i \neq j$ *(Experimento desbalanceado)* |
|---|---|
| $S = \sqrt{(k-1)F_{(k-1,gl\ res,\ \alpha)}\hat{V}(\hat{Y}_i)} = \sqrt{(k-1)F_{(k-1,gl\ res,\ \alpha)}QMR\sum_{i=1}^{k}\frac{c_i^2}{r}};$ | $S = \sqrt{(k-1)F_{(k-1,gl\ res,\ \alpha)}QMR\sum_{i=1}^{k}\frac{c_i^2}{r_i}};$ |

Sendo: *k – 1* o número de graus de liberdade de tratamentos; $F_{(k-1,\ gl\ res,\ \alpha)}$ é o valor crítico da Tabela *F-Snedecor*, a qual depende dos graus de

liberdade de tratamentos e do resíduo; $c_i$ são os coeficientes do contraste e $r_i$ é o número de repetições do *i-ésimo* tratamento. A Regra de Decisão do teste de Scheffé para rejeitarmos ou não se o contraste é diferente de zero é comparar a estimativa do contraste $\hat{Y}$ com o valor de ***S***:

- se $|Y_i| \geq S$, rejeitamos a hipótese $H_0 : Y_i = 0$, e concluímos que o contraste de médias é diferente de zero;
- se $|Y_i| < S$, não rejeitamos a hipótese $H_0 : Y_i = 0$, e concluímos que o contraste de médias não é diferente de zero

Aplicando o teste de Scheffé ao exemplo anterior do teste de *t-student*, temos

| *Contraste* | $\mu_1$ | $\mu_2$ | $\mu_3$ | $\mu_4$ | $\hat{Y}_I$ | $\sum_{I=1}^{4} c_i^2$ | $S$ |
|---|---|---|---|---|---|---|---|
| $Y_1$ | *+3* | *-1* | *-1* | *-1* | *-0,41* | *12* | *0,3733* ** |
| $Y_2$ | *0* | *+2* | *-1* | *-1* | *0,22* | *6* | *0,2640* ** |
| $Y_3$ | *0* | *0* | *+1* | *-1* | *-0,18* | *2* | *0,1524* ***ns*** |

***** significativo a 5%; ***ns*** não significativo a 5%.

O objetivo é testar a hipótese $\begin{matrix} H_0 : Y_i = 0 \\ H_1 : Y_i \neq 0 \end{matrix}$, para *i = 1,2,3*.

Assim, para o contraste $Y_1$, temos que:

- $H_0 : Y_1 = 0$
- $H_1 : Y_1 \neq 0$
- $\hat{Y}_1 = 3(2{,}03) - (2{,}24) - (2{,}04) - (2{,}22) = -0{,}41$ *e*
- $\hat{V}(\hat{Y}_1) = \frac{QMR}{r}\sum_{i=1}^{4} c_i^2 = \frac{0{,}0044375}{4}\,12 = 0{,}0133$
- $S = \sqrt{(4-1)F_{(3,12,0,05)}\frac{0{,}0044375}{4}\,(3^2 + 1^2 + 1^2 + 1^2)} =$
- $\sqrt{(4-1)\,(3{,}49)\,\left(\frac{0{,}0044375}{4}\right)\,(12)} = \sqrt{0{,}1394} = 0{,}3733$
- Pela regra de decisão $|Y_i| > S$, logo rejeitamos $H_0$ a 5% de probabilidade e concluímos que a ração comercial com milho tem uma conversão alimentar melhor do que a que a ração comercial com sorgo.

O script no R para o cálculo da estatística de Scheffé é

```
> # é necessário fornecer os valores
>  # definindo o número de repetições
>  r <- 4
>
>  # definindo os graus de liberdade dos tratamentos
>  gltr<- 3
>  # definindo os graus de liberdade do resíduo
>  glr <-  12
>
>  # quadrado médio do resíduo
```

```
>
>  qmr <- 0.0044375
>
>  # definindo as médias dos tratamentos
>  m.trat <- c(  2.03, 2.24, 2.04, 2.22)
>
>  # definindo os coeficientes do contraste
>  c <- c( 3, -1,-1,-1)
>
>  # cálculo da estimativa do contraste
>  y.est<-sum(m.trat*c)
>  y.est
[1] -0.41
>
>  # cálculo da variância do contraste
>  var.c<- qmr/r*sum(c^2)
>
>  # cálculo da estatística S de Scheffé
>  s<- sqrt(gltr*qf(0.95,gltr,glr)*var.c)
>  s
[1] 0.3733546
```

*(Repetir esse procedimento para os contrastes $Y_2$ e $Y_3$ e tirar as conclusões).*

---

**5 Teste de Tukey**

O Teste de Tukey é baseado na amplitude total *"estudentizada"* (*studentized range*) e pode ser usado para comparar todo contraste entre duas médias de tratamentos do tipo

- Hipóteses: $H_0 : Y_i = \mu_i - \mu_j = 0 \; para \; i \neq j$ / $H_1 : Y_i = \mu_i - \mu_j \neq 0$
- Calcular o valor da *diferença mínima significativa* (d.m.s):

| $r_i = r$ *para todo* $i$ **(Experimento balanceado)** | $r_i \neq r_j$ *para* $\forall \; i \neq j$ **(Experimento desbalanceado)** |
|---|---|
| $dms = q_{(k;\, gl\ res,\, \alpha)} \sqrt{\frac{QMR}{r}}$ | $dms = q_{(k;\, gl\ res,\, \alpha)} \sqrt{\frac{QMR}{2} \left(\frac{1}{r_i} + \frac{1}{r_j}\right)}$ |

sendo: $q_{(k,\, gl\ res,\, \alpha)}$ é o valor da *amplitude total "estudentizada"* e é obtido de tabela própria, e depende do número de tratamentos (k) e do número de graus de liberdade para o resíduo, o qual neste exemplo é (n - k). Após calcular o *d.m.s.*, calculamos a estimativa dos contrastes entre os pares de médias $\hat{Y}_i = \bar{x}_i - \bar{x}_j$ e comparamos esses valores com o valor do *d.m.s.*, aplicando a seguinte regra de decisão:

- se $\left|\hat{Y}_i\right| \geq d.m.s.$ **rejeitamos $H_0$**, ao nível a de significância, e concluímos que as médias dos tratamentos envolvidos são diferentes;

- se $\left|\hat{Y}_i\right| < d.m.s.$ **não rejeitamos $H_0$** e concluímos que as médias dos tratamentos envolvidos são iguais.

**Exemplo 3:** usaremos os dados do exemplo 1 apresentado no início desta aula, o quadro da anova fornece

- $k = 4,\ QMR = 68{,}75$ com 16 graus de liberdade e $q_{(5,16,0{,}05)} = 4{,}046$ e $dms = q_{(k;\, n-k,\, \alpha)}\sqrt{\frac{QMR}{r}} = 4{,}046\sqrt{\frac{68{,}75}{5}} = 15{,}00$

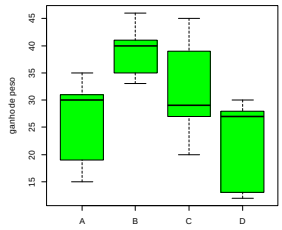
Assim, toda estimativa de contraste do tipo $\left|\hat{Y}_i\right| = \left|\bar{y}_i - \bar{y}_j\right|$ que exceder o valor do *d.m.s.= 15,00* é significativo a 5%.

| *Estimativa do contraste* |
|---|
| $\left\|\hat{Y}_1\right\| = \left\|\bar{y}_B - \bar{y}_A\right\| = \left\|39 - 26\right\| = 13$ *ns* |
| $\left\|\hat{Y}_2\right\| = \left\|\bar{y}_C - \bar{y}_A\right\| = \left\|32 - 26\right\| = 6$ *ns* |
| $\left\|\hat{Y}_3\right\| = \left\|\bar{y}_D - \bar{y}_A\right\| = \left\|22 - 26\right\| = 4$ *ns* |
| $\left\|\hat{Y}_4\right\| = \left\|\bar{y}_B - \bar{y}_C\right\| = \left\|39 - 32\right\| = 7$ *ns* |
| $\left\|\hat{Y}_5\right\| = \left\|\bar{y}_B - \bar{y}_D\right\| = \left\|39 - 22\right\| = 17$ * |
| $\left\|\hat{Y}_6\right\| = \left\|\bar{y}_C - \bar{y}_D\right\| = \left\|32 - 22\right\| = 10$ *ns* |

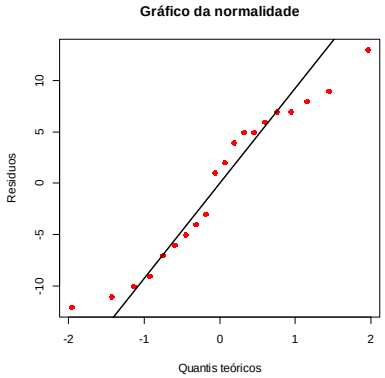
** - significativo a 5%; ns – não significativo a 5%*

Script no R para o cálculo da anova e o teste de Tukey

```
> # entrando com os dados de ganho de peso
> gp <- c(35,19,31,15,30,
+         40,35,46,41,33,
+         39,27,20,29,45,
+         27,12,13,28,30)
>
> # entrando com o número de repetições dos tratamentos
> r <- 5
>
> # entrando com os níveis dos tratamentos
> trat <- c(rep("A",r),rep("B",r),rep("C",r),rep("D",r))
> trat
 [1] "A" "A" "A" "A" "A" "B" "B" "B" "B" "B" "C" "C" "C" "C" "C" "D" "D" "D" "D"
[20] "D"
>
> # cálculo das medias dos tratamentos
> m.trat <- tapply(gp,trat,mean)
> m.trat
 A  B  C  D
26 39 32 22
> # Gráfico Box-Plot
> boxplot(gp~trat, vertical=T,ylab="ganho de peso",col="green")
>
```

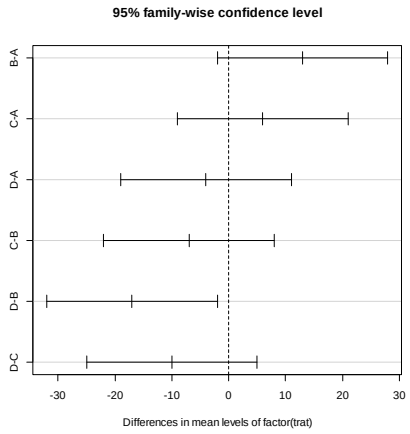



```
> # análise da variância - ANOVA
> gp.av <- aov(gp~factor(trat))
> summary(gp.av)
             Df  Sum Sq   Mean Sq    F value   Pr(>F)
factor(trat)  3   823.8   274.58      3.994    0.0267 *
Residuals    16 1100.0    68.75
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
>
> # obtendo os residuos
> residuo <- aov(gp.av)$res
> residuo
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
 9 -7  5 -11  4  1 -4  7  2 -6  7 -5 -12 -3 13  5 -10 -9  6  8
>
> # gerando o gráfico normal de probabilidade
> qqnorm(residuo,ylab="Residuos", main="Gráfico Normal de
Probabilidade dos Resíduos",
+        xlab="Quantiles Teóricos",pch=16,col=2)
>
> # colocando a reta da distribuição teórica normal
> qqline(residuo,lwd=2)
>
```




```
> # testando a normalidade dos resíduos "Teste de Shapiro-Wilks"
> shapiro.test(residuo)

        Shapiro-Wilk normality test

data:  residuo
W = 0.9387, p-value = 0.227
>
```

```
> # teste da homogeneidade das variâncias "Teste de Bartllet"
> bartlett.test(gp ~ trat)

        Bartlett test of homogeneity of variances

data:  gp by trat
Bartlett's K-squared = 1.5284, df = 3, p-value = 0.6757

>
> # Teste de Tukey
> compara.tu <- TukeyHSD(gp.av)
> compara.tu
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = gp ~ factor(trat))

$`factor(trat)`
       diff       lwr         upr          p adj
B-A   13   -2.003315   28.003315  0.1018285
C-A    6   -9.003315   21.003315  0.6687032
D-A   -4  -19.003315   11.003315  0.8698923
C-B   -7 -22.003315     8.003315  0.5553529
D-B  -17 -32.003315   -1.996685   0.0237354
D-C  -10 -25.003315    5.003315   0.2640642
>
> # grafico do teste de Tukey
> plot(compara.tu)
```




Conclusões desta análise:

- *o teste F é significativo (p=0,0267), rejeitamos $H_0$. Assim existe pelo menos dois tratamentos que diferem entre si.*

- *o teste de Shapiro-Wilks de normalidade* é não significativo (p=0,227), não rejeitamos $H_0$ e concluímos que os dados deste experimento suportam a suposição de normalidade.

- *o teste de Bartlett* é não significativo (p=0,6757), não rejeitamos $H_0$ e concluímos que os dados deste experimento suportam a suposição de homogeneidade das variâncias populacionais dos tratamentos.

- o teste de Tukey é significativo (p=0,0237) para o contraste entre as médias dos tratamentos D e B. As outras comparações de pares de médias populacionais dos tratamentos não são significativas.

Uma forma simples de apresentação destes resultados é a seguinte:

- coloque as médias em ordem decrescente;
- una as médias que não diferem entre si por meio de uma linha

No exemplo temos:

| $\bar{y}_D$ | $\bar{y}_A$ | $\bar{y}_C$ | $\bar{y}_B$ * |
|---|---|---|---|
| 22 | 26 | 32 | 39 |

** médias seguidas pela mesma linha não diferem entre si pelo teste de Tukey a 5% de probabilidade.*

Outra forma, muito utilizada pelos pesquisadores é a que substitui a linha por letras, ou seja,

| $\bar{y}_D$ | $\bar{y}_A$ | $\bar{y}_C$ | $\bar{y}_B$ |
|---|---|---|---|
| 22a | 26ab | 32ab | 39b, |

*médias seguidas pela mesma letra minúscula não diferem entre si pelo teste de* Tukey a 5% de probabilidade

ou ainda, na forma

| *Tratamentos* | *Médias* |
|---|---|
| *D* | *22 a* |
| *A* | *26 ab* |
| *C* | *32 ab* |
| *B* | *39 b* |

*médias seguidas pela mesma letra minúscula nas colunas não diferem entre si pelo teste de* Tukey a 5% de probabilidade.

A saída do pacote ***ExpDes*** para o teste de Tukey já contempla esta facilidade das médias seguidas pelas letras.

O script do R usando os recursos deste pacote é dado por:

```
> # usando ao função crd( ) do ExpDes
> # requerendo o ExpDes
> library(ExpDes)
> crd(trat,gp,quali=T,mcomp="tukey")
------------------------------------------------------------------------
Analysis of Variance Table
------------------------------------------------------------------------
             DF     SS       MS       Fc       Pr>Fc
Treatament  3  823.75   274.58  3.9939   0.026711
Residuals  16 1100.00    68.75
Total      19 1923.75
------------------------------------------------------------------------
CV = 27.87 %
```

```
---------------------------------------------------------------------------
Shapiro-Wilk normality test
p-value:  0.2270061
According to Shapiro-Wilk normality test at 5% of significance, residuals
can be considered normal.
---------------------------------------------------------------------------

Tukey's test
---------------------------------------------------------------------------
Groups Treatments Means
a      B          39
ab     C          32
ab     A          26
 b     D          22
---------------------------------------------------------------------------
```

**6 Teste de Dunnet (comparando todas as médias com um controle)**

É um teste no qual as únicas comparações de interesse são aquelas entre os tratamentos e um determinado tratamento padrão, geralmente a testemunha (controle), não havendo interesse na comparação dos demais tratamentos entre si. Para testarmos o contraste $H_0 : \mu_i - \mu_c \; vs \; H_1 : \mu_i \neq \mu_c$, o qual envolve a média do tratamento "*i*" e do tratamento controle "*c",* usamos a estatística:

$$D = d_{(k,\, gl\ res,\, \alpha)} \sqrt{\left(\frac{1}{r_i} + \frac{1}{r_c}\right) QMR},$$

sendo: "$d_{(k,gl\ res,\alpha)}$" o valor tabelado para $\alpha$ fixado freqüentemente em 5%, que depende do número total de tratamentos (*k*) (excluindo o controle), do número de graus de liberdade do resíduo *(gl res)*, o qual neste exemplo é (*n-k*) e de $\alpha$; $r_i$ e $r_c$ correspondem ao número de repetições dos tratamentos "*i"* e "*c"*. A seguir, calculamos uma estimativa para cada um dos contrastes $\hat{Y}_i = \bar{y}_i - \bar{y}_c$ e comparamos o valor da estatística *D'* e aplicamos a seguinte regra de decisão:

- se $\left|\hat{Y}_i\right| \geq D$ rejeitamos $H_0$ e concluímos que a média do tratamento "*i*" difere significativamente da média do tratamento *"c"* o padrão;
- se $\left|\hat{Y}_i\right| < D$ não rejeitamos $H_0$ e concluímos que a média do tratamento "*i*" é igual ao do tratamento padrão "*c*".

Considere os dados do exemplo 1, com o tratamento A sendo o controle, então, *k = 3, QMR = 68,75* com 16 graus de liberdade e consultando a Tabela do Teste de Dunnett (*VIEIRA, S. pg 183 e 184*) a 5% de probabilidade, temos que $d_{(3,16,0,05)} = 2{,}63$ e a estatística $D' = 2{,}63\sqrt{\frac{2\,(68{,}75)}{5}} = 13{,}79$

| Rações | *média* |
|---|---|
| *A (controle)* | *26 a* |
| *B* | *39 a* |
| *C* | *32 a* |
| *D* | *22 a* |

*Médias com a mesma letra do controle não diferem deste pelo teste de Dunnett a 5% de probabilidade.*

- é fácil verificar que os tratamentos B, C e D não diferem significativamente do controle A, ou seja, apresentam resultados semelhantes ao do controle.

O teste de Dunnett no R está implementado no pacote ***multcomp***. Um exemplo de sua utilização para os dados do exemplo 1 desta aula, considerando o tratamento A como controle, é dado pelo script abaixo

```
>
> # instalando o pacote multicomp
> # install.packages("multcomp")
>
> # requerendo o pacote multcomp
> require(multcomp)
>
> #  teste de Dunnett
> gp.dunnet<- glht(gp.av, linfct = mcp(trat = "Dunnet"))
> summary(gp.dunnet)

     Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts

Fit: aov(formula = gp ~ trat)

Linear Hypotheses:
            Estimate  Std. Error  t value  Pr(>|t|)
B - A == 0   13.000       5.244    2.479    0.0622 .
C - A == 0    6.000       5.244    1.144    0.5402
D - A == 0   -4.000       5.244   -0.763    0.7883
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```

Conclusão: todos os valores de p do teste t-student são superiores a 0,05 (p > 0,05) logo nenhum dos tratamentos B, C e D diferem do controle A.

---

**7 Teste de Duncan**

A aplicação do teste de *Duncan* é bem mais trabalhosa que o teste de *Tukey*, mas chega-se a resultados mais detalhados e se discrimina com mais facilidade entre os tratamentos. Geralmente, o Teste de *Duncan* indica resultados significativos em casos em que o Teste de *Tukey* não permite obter significância estatística. Para a aplicação do teste é importante ordenarmos as médias dos tratamentos em ordem crescente ou decrescente de tamanho. A seguir, calculamos o valor da amplitude total mínima significativa (*shortest significant range*) para o contraste entre a **maior** e a **menor** das médias dos tratamentos, usando a fórmula:

$$d.m.s. = z_{(p,gl\ res,\alpha)}\sqrt{\frac{QMR}{r}},$$

sendo: *p=i-j+1* ( nº de médias abrangidas pelo intervalo delimitado pelas médias comparadas), $z_{(p,gl\ res,\alpha)}$ *é o nível* $\alpha$ *da amplitude mínima estudentizada*

*de Duncan (obtido da Tabela de Duncan – arquivo Tab_Duncan_5%.pdf)*, neste exemplo os graus de liberdade do resíduo é n-k.

A regra de decisão é:

- se $\left|\hat{Y}_i\right| \geq d.m.s.$ rejeita-se $H_0$ , ou seja, se o valor absoluto da diferença entre as médias em comparação é igual ou maior que a *d.m.s.*
- se $\left|\hat{Y}_i\right| < d.m.s.$ não rejeitamos $H_0$

Considere os dados do exemplo 1. O ordenamento dos tratamentos, segundo a grandeza das médias, é:

| *Tratamentos* | *D* | *A* | *C* | *B* |
|---|---|---|---|---|
| *Médias* | *22* | *26* | *32* | *39* |

Os valores do *d.m.s*. para comparar as médias é:

| *nº de tratamentos (k)* | *2* | *3* | *4* |
|---|---|---|---|
| $z_{(p,16,0,05)}$ | *3,0* | *3,15* | *3,23* |
| *d.m.s* | *11,12* | *11,68* | *11,98* |

Teste as diferenças na seguinte ordem: a maior média menos a menor, a maior média menos a segunda maior, a segunda maior média menos a menor, segunda maior média menos a segunda menor média, e assim por diante. Do quadro de médias acima, a diferença entre as médias dos tratamentos D e B envolve 4 médias, assim, esta diferença para ser significativa $\left|\hat{Y}_1\right| = \left|\bar{y}_B - \bar{y}_D\right| = \left|39 - 22\right| = \left|17\right| = 17$ , a qual é > do que o dms=11,98, o que pela regra de decisão nos leva a rejeitar a $H_0 : Y_1 = 0$ e concluímos que a média da Droga B é significativamente maior que a média da Droga D, a 5% de probabilidade. Em detalhes temos que as outras diferenças são dados por:

As comparações entre a média da Droga B e Drogas A e D, envolvem intervalos de cinco médias e o calculo do *d.m.s.* do teste de Duncan fica:

- os contrastes $Y_2 = \mu_B - \mu_A$ e $Y_3 = \mu_C - \mu_D$ envolvem 3 médias, e suas estimativas em módulo são

  $\left|\hat{Y}_2\right| = \left|39 - 26\right| = 13 > 11{,}68;$ ***significativo***

  $\left|\hat{Y}_3\right| = \left|32 - 22\right| = 10 < 11{,}68;$ ***não significativo***

  ;

portanto, ***rejeitamos*** a hipótese $H_0 : Y_2 = \mu_B - \mu_A = 0$ ***e não rejeitamos*** $H_0 : Y_3 = \mu_c - \mu_D = 0$ e concluímos que o contraste $Y_2$ é significativo a 5% de probabilidade e o contraste $Y_3$ é não significativo a 5% de probabilidade.

Da mesma forma para comparar o controle e as médias das Drogas B vs C, de C vs A e A vs D, todas elas envolvendo 2 médias, temos,

- os contrastes $Y_4 = \mu_B - \mu_C$ , $Y_5 = \mu_C - \mu_A$ *e* $Y_6 = \mu_A - \mu_D$ e seus valores de suas estimativas em módulo são $|\hat{Y}_4| = |39 - 32| = 7$, $|Y_5| = |32 - 26| = 6$ *e* $|Y_6| = |26 - 22| = 4$ todos menores que 11,12. Portanto não rejeitamos as hipóteses $H_0 : Y_4 = \mu_B - \mu_C = 0$; $H_0 : Y_5 = \mu_C - \mu_A = 0$ *e* $H_0 : Y_6 = \mu_A - \mu_D = 0$ e concluímos que estes contrastes não são significativos a 5% de probabilidade.

O resultado da aplicação do teste de Duncan é representado da seguinte maneira:

| *Tratamentos* | *D* | *A* | *C* | *B* |
|---|---|---|---|---|
| *médias* | *(2)b* | *(8)b* | *(10)ab* | *(13)a* |

*Médias seguidas pela mesma letra minúscula não diferem entre si pelo teste de Duncan a 5% de probabilidade.*

```
> # usando ao função crd( ) do ExpDes
> # requerendo o ExpDes (atenção !!! se o ExpDes já foi requerido não é
> # necessário requerê-lo novamente
> # library(ExpDes)
> crd(trat,gp,quali=T,mcomp="duncan")
------------------------------------------------------------------------
Analysis of Variance Table
------------------------------------------------------------------------
             DF     SS      MS      Fc      Pr>Fc
Treatament  3  823.75  274.58  3.9939  0.026711
Residuals  16 1100.00   68.75
Total      19 1923.75
------------------------------------------------------------------------
CV = 27.87 %
------------------------------------------------------------------------
Shapiro-Wilk normality test
p-value:  0.2270061
According to Shapiro-Wilk normality test at 5% of significance, residuals
can be considered normal.
------------------------------------------------------------------------
Duncan's test
------------------------------------------------------------------------
Groups  Treatments  Means
a       B          39
ab      C          32
 b      A          26
 b      D          22
------------------------------------------------------------------------
```

Conclusão (somente para o teste de Duncan): O tratamento B difere significativamente dos tratamentos A e D. Reparem que o teste de Duncan indicou uma diferença a mais, entre os tratamentos B e A, a qual não foi indicada pelo teste de Tukey.

A função *crd( )* do pacote ExpDes fornece outras opções de testes de comparações múltiplas para serem colocadas no comando ***mcomp* = " "**. O teste *default* é o teste de Tukey (**"tukey"**). As outras opções são o teste LSD equivalente ao teste *t-student* (**"lsd"**); o teste LSD com proteção Bonferroni

(**"lsdb"**); o teste de Duncan (**"duncan"**); o teste de Student-Newman-Kews (**"snk"**) e o teste de Scott-Knott (**"sk"**).

## 8 ALGUMAS CONSIDERAÇÕES SOBRE O USO DE PROCEDIMENTOS DE COMPARAÇÕES MÚLTIPLAS.

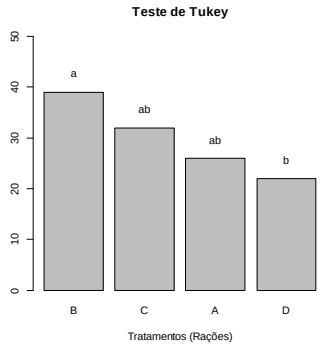
Quando desejamos comparar os diversos tratamentos com um tratamento controle ou padrão (testemunha), o teste de Dunnett é o mais indicado. Os testes de Duncan e de Tukey têm fundamentos muito semelhantes, mas o teste de Duncan é menos conservador e menos exigente que o teste de Tukey, isto é, indica diferenças significativas com mais facilidade. Vale lembrar também que o teste de Duncan é um teste seqüencial e a sua aplicação é mais trabalhosa. Ambos os testes são exatos quando os números de repetições por tratamento forem iguais; caso contrário os testes são apenas aproximados. O teste t-Sudent é pouco rigoroso quando usado indiscriminadamente, devendo ser usado com cautela para testar contrastes ortogonais definidos *a priori*. Já o teste de Scheffé é bastante rigoroso e seu uso é desaconselhável (como o teste t-Student) para a comparação entre duas médias de tratamentos, sendo mais indicado para testar contrastes que envolvem mais de duas médias.

O pacote **"agricolae "** também pode ser a utilizado para as comparações múltiplas. A seguir é fornecido um script utilizando este pacote

```
> # entrando com os dados de ganho de peso
>
> gp <- c(35,19,31,15,30,
+           40,35,46,41,33,
+           39,27,20,29,45,
+           27,12,13,28,30)
>
>  # entrando com o número de repetições dos tratamentos
>  r <- 5
>
>  # entrando com os níveis dos tratamentos
>  trat <- c(rep("A",r),rep("B",r),rep("C",r),rep("D",r))
>
>  # análise da variância - ANOVA
>  gp.av <- aov(gp~trat)
>  summary(gp.av)
             Df  Sum Sq Mean Sq  F value   Pr(>F)
trat          3   823.8  274.58    3.994  0.0267 *
Residuals    16  1100.0   68.75
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> # instatlando o pacote "agricolae"
> # install.packages("agricolae")
>
> # requerendo o pacote “agricolae”
> require("agricolae")
>
> #  teste de Tukey
> compara.tukey <- HSD.test(gp.av,"trat",group=T)
```

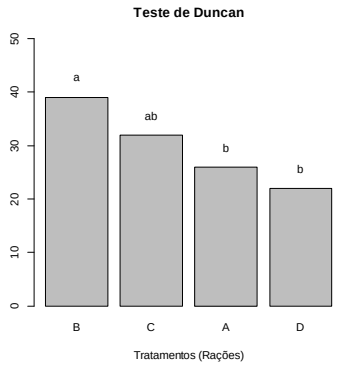
```
> compara.tukey
$statistics
   Mean      CV      MSerror      HSD
  29.75    27.8708   68.75      15.00331

$parameters
  Df ntr StudentizedRange
  16   4        4.046093

$means
  gp      std       r  Min Max
A 26 8.544004  5  15   35
B 39 5.147815  5  33   46
C 32 9.949874  5  20   45
D 22 8.746428  5  12   30

$comparison
NULL

$groups
   trt  means  M
1   B    39      a
2   C    32      ab
3   A    26      ab
4   D    22       b
```

```
# bar.group(compara.tukey,main="Teste de Tukey", ylim=c(0,50),
            xlab="Tratamentos (Rações)")
```




```
> # teste de Duncan
> compara.duncan <- duncan.test(gp.av,"trat")
> compara.duncan
$statistics
   Mean      CV      MSerror
  29.75    27.8708   68.75

$parameters
  Df ntr
  16   4
```

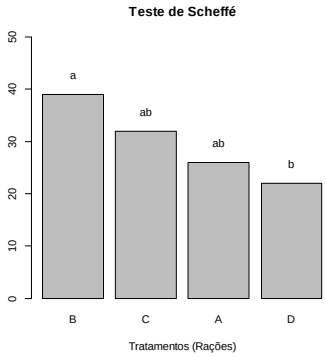
```
$Duncan
    Table CriticalRange
2 2.997999      11.11688
3 3.143802      11.65753
4 3.234945      11.99550

$means
   gp       std       r  Min Max
A 26  8.544004   5  15    35
B 39  5.147815   5  33    46
C 32  9.949874   5  20    45
D 22  8.746428   5  12    30
$comparison
NULL

$groups
  trt means  M
1   B    39  a
2   C    32  ab
3   A    26    b
4   D    22    b

#  gráfico de barras das médias com as letras segundo o teste de Duncan
# bar.group(compara.duncan,main="Teste de Duncan",ylim=c(0,50),
#         xlab="Tratamentos (Rações)")
```




```
> compara.scheffe <- scheffe.test(gp.av,"trat")
> compara.scheffe
$statistics
   Mean      CV        MSerror    CriticalDifference
  29.75    27.8708   68.75          16.34646

$parameters
  Df   ntr       F          Scheffe
  16   4      3.238872   3.117148

$means
   gp      std       r  Min  Max
A 26   8.544004  5  15     35
B 39   5.147815  5  33     46
```

```
C 32   9.949874  5  20    45
D 22   8.746428  5  12    30

$comparison
NULL

$groups
   trt  means  M
1   B    39     a
2   C    32     ab
3   A    26     ab
4   D    22      b
```

```
#  gráfico de barras das médias com as letras segundo o teste de Scheffé
#bar.group(compara.scheffe, main="Teste de Scheffé",ylim=c(0,50),
 #                xlab="Tratamentos (Rações)")
```



# 4º EXERCÍCIO PRÁTICO DE ESTATÍSTICA EXPERIMENTAL

1. Três extratos de origem vegetal foram fornecidos a 20 cães por via oral com a finalidade de testar o possível efeito sobre a pressão arterial sistólica desses animais. Os cães foram divididos em grupos de cinco animais, recebendo cada grupo um tipo de extrato, ao acaso, B, C ou D, além de um grupo controle – A, tratado com placebo. Os dados obtidos foram:

| *Trat.(extratos)* | *Cães* | | | | | *Totais* | *Médias* | $S_i$ |
|---|---|---|---|---|---|---|---|---|
| *(Controle) A* | *74,0* | *71,0* | *73,0* | *79,0* | *68,0* | | | |
| *B* | *99,0* | *91,0* | *94,0* | *101,0* | *97,0* | | | |
| *C* | *100,0* | *95,0* | *97,0* | *99,0* | *98,0* | | | |
| *D* | *78,0* | *74,0* | *75,0* | *86,0* | *72,0* | | | |
| | | | | | | *Total Geral* | | |

(a) Escreva o *script* da linguagem R para ler os dados da tabela acima e calcular os totais dos tratamentos, as médias dos tratamentos, os desvios padrões dos tratamentos, o total geral, e a média geral. Apresente os resultados na mesma tabela acima.
(b) Escrever o modelo matemático do experimento, estabelecer as hipóteses estatísticas $H_0$ e $H_1$ e as suposições básicas para se testar estas hipóteses.
(c) Escreva o script para os cálculos do quadro da análise de variância e apresente monte o quadro da anova. Apresente as conclusões.
(d) Aplique o teste de Tukey para comparar as médias 2 a 2. Apresente um quadro e um gráfico de barras das médias juntamente com as letras explicando as diferenças. Tire as conclusões.
(e) Aplique o teste de Duncan para comparar as médias 2 a 2. Apresente um quadro e um gráfico de barras das médias juntamente com as letras explicando as diferenças. Tire as conclusões.
(f) Aplique o teste de Dunnett para comparar as médias com o controle A. Comente os resultados.

2- A redução da pressão sangüínea sistólica (RPS) depois da administração de drogas para hipertensão é um dos indicadores de como os pacientes estão respondendo às drogas. No tratamento da hipertensão, os efeitos colaterais associados com as drogas têm um particular interesse. Neste estudo, duas drogas X e Y para a redução dos efeitos colaterais de uma droga padrão (P) de hipertensão foi avaliada. O estudo foi conduzido em um delineamento inteiramente casualizado com cinco tratamentos, assim definidos:

- $T_1$ – Droga padrão (P)
- $T_2$ – P combinada com uma dose baixa de X (P+DBX)
- $T_3$ – P combinada com uma dose alta de X (P+DAX)
- $T_4$ – P combinada com uma dose baixa de Y (P+DBY)
- $T_5$ – P combinada com uma dose alta de Y (P+DAY)

A redução na pressão sangüínea (mm Hg) em um período de quatro semanas observadas em cães experimentais está tabulada abaixo:

| *Tratamentos* | *Repetição* | | | | *Total* | *Média* |
|---|---|---|---|---|---|---|
| | *1* | *2* | *3* | *4* | | |
| $T_1$ | *27* | *26* | *21* | *26* | | |
| $T_2$ | *19* | *13* | *15* | *16* | | |
| $T_3$ | *15* | *10* | *10* | *11* | | |
| $T_4$ | *22* | *15* | *21* | *18* | | |
| $T_5$ | *20* | *18* | *17* | *16* | | |

Pede-se:
a) A análise de variância para testar a hipótese geral de igualdade das médias dos tratamentos;
b) Aplique os testes t-student e Scheffé nos contrates abaixo:
b.1 Existe efeito das drogas combinadas ($T_2$ $T_3$ $T_4$ $T_5$) na RPS?.
b.2 Existe diferença entre os efeitos médios das doses baixa e alta da droga Y?.
b.3 Existe diferença entre a resposta média esperada das duas doses de X?.
(extraído de Statistical Research Methods in the Life Science, P. V. Rao, pg. 327).

**Aula 5 Testes F planejados**

No planejamento de um experimento, frequentemente pode-se utilizar o teste F para responder algumas questões mais específicas. Isto implica na decomposição dos graus de liberdade e da soma de quadrados do efeito dos tratamentos em componentes de comparações. Estes componentes podem ser classes de comparações ou tendência das respostas. Eles podem ser testados pela partição dos graus de liberdade e da soma de quadrados dos efeitos dos tratamentos em contrastes simples e específicos e suas soma de quadrados associadas. O número de contrastes independentes e ortogonais que podem ser definidos é igual ao número de graus de liberdade do efeito do tratamento. O poder e a simplicidade deste método não são muito apreciados e compreendidos pelos pesquisadores com deveria ser. Esta metodologia envolve a definição de contrastes ortogonais, e talvez este termo, cria a impressão de que ele é complicado e difícil. Isto esta longe de ser verdade. Atualmente este método tem três grandes vantagens:

- permite responder a questões específicas e importantes a respeito dos efeitos dos tratamentos;
- os cálculos são simples; e,
- proporciona uma checagem útil da soma de quadrados dos tratamentos.

Esta metodologia também é denominada de "**desdobramento, ou a decomposição dos graus de liberdade de tratamentos".**

**2 Soma de quadrados de um contraste**.

Quando utilizamos contrastes na decomposição dos graus de liberdade dos efeitos dos tratamentos usamos a seguinte definição para o cálculo da soma de quadrados:

- Definição: a soma de quadrados de um contraste é calculada pela fórmula $SQ(Y_i) = \frac{\left(\sum_{i=1}^{k} c_i Y_{i+}\right)^2}{r\sum_{i=1}^{k} c_i^2}$ *ou* $SQ(Y_i) = \frac{\left(\hat{Y}_i\right)^2}{r\sum_{i=1}^{k} c_i^2}$, sendo: $c_i$ os coeficientes do contraste; $Y_{i+}$ os totais dos tratamentos e *r* o numero de repetições *(neste caso* $r = r_1 = \ldots = r_k$ *)* e $\hat{Y}_i = \left(\sum_{i=1}^{k} c_i Y_{i+}\right)^2$ é uma estimativa do contraste com base nos totais.

Observações importantes:

- todo contraste tem sempre 1 grau de liberdade, assim $QM(Y_i) = SQ(Y_i)$.
- geralmente, testamos $H_0: Y_i = 0$ *vs.* $H_1: Y_i \neq 0$ e para tanto usamos a estatística *F-Snedecor* tendo como denominador o quadrado médio do erro experimental *(QMR)*.
- os contrastes devem ser planejados *a priori* e podem ser tão numerosos quanto acharmos necessário.
- o número de *contrastes ortogonais* entre os totais dos tratamentos é igual ao número de graus de liberdade associados a essa fonte de variação, isto é, se o fator tratamento tem k níveis então conseguiremos definir *somente* (k-1) contrastes ortogonais.

- se $Y_1, Y_2,...,Y_{k-1}$ são contrastes ortogonais envolvendo os totais dos k níveis do fator, então $SQ(Y_1) + SQ(Y_2) + ... + SQ(Y_{k-1}) = SQTr$
- a *ortogonalidade* dos contrastes garante a *independência* entre as conclusões.

**Exemplo 1**: Foram comparados os efeitos de cinco tratamentos no crescimento de alevinos de carpas (mediu-se o comprimento em cm aos dois meses de idade) em um DIC.

*$T_1$ – ração comum. (rc)*
*$T_2$ – ração comum + esterco. (rce)*
*$T_3$ – ração comum + esterco de porco + vitamina $B_{12}$. (rce$B_{12}$)*
*$T_4$ – ração comum + farinha de osso. (rcfo)*
*$T_5$ – ração comum + farinha de osso + vitamina $B_{12}$. (rcfo$B_{12}$)*

Dados

| | *Repetições* | | | |
|---|---|---|---|---|
| *Trat.* | *1* | *2* | *3* | *4* |
| $T_1$ | 4,6 | 5,1 | 5,8 | 5,5 |
| $T_2$ | 6,0 | 7,1 | 7,2 | 6,8 |
| $T_3$ | 5,8 | 7,2 | 6,9 | 6,7 |
| $T_4$ | 5,6 | 4,9 | 5,9 | 5,7 |
| $T_5$ | 5,8 | 6,4 | 6,8 | 6,8 |

Análise de variância usual.

| *Causas da Variação* | *G.L.* | *S.Q.* | *Q.M.* | *F* |
|---|---|---|---|---|
| *Tratamentos* | 4 | 7,72 | 1,91 | 7,19 |
| *Resíduo* | 15 | 4,03 | 0,27 | |
| *Total* | 19 | 11,75 | | |

$F_{(4,15;0,05)} = 3,06$ e $F_{(4,15;0,01)} = 4,89$

Conclusão: o teste é significativo a 1% de probabilidade, portanto rejeitamos $H_0$, os tratamentos apresentam efeitos distintos sobre o crescimento de alevinos de carpas. Esta é uma informação geral sobre os efeitos dos tratamentos. Para obtermos informações detalhadas devemos decompor os 4 graus de liberdade dos efeitos dos tratamentos em quatro contrastes ortogonais.

Comparações objetivas:

- ***rc* vs demais.**
$$\hat{Y}_1 = 4T_1 - T_2 - T_3 - T_4 - T_5 = 4Y_{1+} - Y_{2+} - Y_{3+} - Y_{4+}$$
- ***rce* vs rcfo**
$$\hat{Y}_2 = T_2 + T_3 - T_4 - T_5 = Y_{2+} + Y_{3+} - Y_{4+} - Y_{5+}$$
- ***rce* vs *rce$B_{12}$*.**
$$\hat{Y}_3 = T_2 - T_3 = Y_{2+} - Y_{3+}$$
- ***rcfo* vs *rcfo$B_{12}$***
$$\hat{Y}_4 = T_4 - T_5 = Y_{4+} - Y_{5+}$$

| Contraste | $Y_{1+}$ | $Y_{2+}$ | $Y_{3+}$ | $Y_{4+}$ | $Y_{5+}$ |
|---|---|---|---|---|---|
| $Y_1$ | 4 | -1 | -1 | -1 | -1 |
| $Y_2$ | 0 | +1 | +1 | -1 | -1 |
| $Y_3$ | 0 | +1 | -1 | 0 | 0 |
| $Y_4$ | 0 | 0 | 0 | +1 | -1 |

Usando a fórmula definida acima para o cálculo da soma de quadrados dos contrastes temos**:**

**1) *rc* vs *demais*.**

$$\hat{Y}_1 = 4(21{,}0) - 27{,}1 - 26{,}6 - 22{,}1 - 25{,}6 = -17{,}6 \ cm$$

$$\sum_{i=1}^{5} c_i^2 = (4)^2 + (-1)^2 + (-1)^2 + (-1)^2 + (-1)^2 = 20$$

$$S.Q.(\hat{Y}_1) = \frac{(\hat{Y}_1)^2}{r\sum_{1}^{5} c_i^2} = \frac{(-17{,}4)^2}{4(20)} = 3{,}78$$

*(A obtenção das S.Q. dos outros contrastes são deixadas como exercícios)*

A anova com os testes *F planejados* ou com os desdobramentos dos graus de liberdade do efeito dos tratamentos fica:

| *Causas da Variação* | *G.L.* | *S.Q.* | *Q.M.* | *F* | *Pr(>F)* |
|---|---|---|---|---|---|
| *rc vs demais ($Y_1$)* | *1* | *3,87* | *3,87* | *14,43* | *0.00175* |
| *rce vs rcfo ($Y_2$)* | *1* | *2,10* | *2,10* | *7,83* | *0.55014* |
| *rce vs rce$B_{12}$ ($Y_3$)* | *1* | *0,03* | *0,03* | *0,12* | *0.00200* |
| *Rcfo vs rcfo$B_{12}$ ($Y_4$)* | *1* | *1,72* | *1,72* | *6,38* | *0.87434* |
| *Tratamentos* | *(4)* | *(7,72)* | *1,91* | *7,19* | *0.00193* |
| *Resíduo* | *15* | *4,03* | *0,27* | | |
| *Total* | *19* | *11,75* | | | |

$F_{(4,15;0,05)} = 3{,}06$ ; $F_{(4,15;0,01)} = 4{,}89$; $F_{(1,15;0,05)} = 4{,}54$ e $F_{(1,15;0,05)} = 8{,}68$

Conclusões:

- ***rc vs demais*** – o contraste é significativo (*p<0,01)* e pelo resultado do contraste devemos utilizar *rce* ou *rce$B_{12}$*, ou ainda, *rcfo* ou *rcfo$B_{12}$*, quando comparada com a *rc*.
- ***rce vs rcfo*** – o contraste é significativo (p<0,05) e pelo resultado do contraste verificamos que *rce* tem um efeito superior no crescimento dos alevinos, quando comparada com *rcfo.*
- ***rce vs rce$B_{12}$*** - o contraste é não significativo (*p>0,05),* portanto o acréscimo de vitamina $B_{12}$ à *rce (rce$B_{12}$)* não afeta significativamente, o crescimento dos alevinos, quando comparada com a *rce.*
- ***rcfo vs rcfo$B_{12}$*** – o contraste é significativo *(p<0,05)* e pelo resultado do contraste devemos adicionar vitamina $B_{12}$ à ração comum com farinha de osso,quando comparada com a *rcfo*.

Script no R para os cálculos descritos acima

```
> # lendo os dados do arquivo compdados.txt
> dadoscomp<-read.table("compdados.txt",h=T,dec=",")
> head(dadoscomp)
  trat comp
1   T1  4.6
2   T1  5.1
3   T1  5.8
4   T1  5.5
5   T2  6.0
6   T2  7.1
> attach(dadoscomp)
>
> # fazendo a análise da variância - ANOVA
> comp.av <- aov(comp~factor(trat))
> # imprimindo o quadro da ANOVA
> summary(comp.av)
                Df  Sum Sq    Mean Sq   F value  Pr(>F)
factor(trat)     4      7.717      1.9292    7.19       0.00193 **
Residuals    15      4.025      0.2683
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


> # definição do nº de repetições
> r<-4
> # Definição do contraste
> c <- c(4,-1,-1,-1,-1)     #  contraste  rc vs demais
>
> # obtenção do QMR no quadro da anova
> qmr <- anova(comp.av)[2,3]
> qmr
[1] 0.2683333
>
> # obtenção dos gl do residuo no quadro da anova
> glr <- anova(comp.av)[2,1]
> glr
[1] 15
>
> # cálculo dos totais por tratamento
> t.trat <- tapply(comp,trat,sum)
> t.trat
  T1   T2   T3   T4   T5
21.0 27.1 26.6 22.1 25.8
>
> # estimativa do contraste Y com base nos totais
> y.est <-sum(c*t.trat)
> y.est
[1] -17.6
> # cálculo da soma de quadrados
> sqy <- (y.est^2)/(r*sum(c^2))
> sqy
```

```
[1] 3.872
>
> # Cálculo da estatística F
> fc <- sqy/qmr
> fc
[1] 14.42981
> # Cálculo do valor de p associado à estatística fc
> valor.p <- 1-pf(fc,1,glr)
> valor.p
[1] 0.001748013
```

Para obter os resultados referentes aos outros contrastes basta substituir o objeto c na linha # Definição do contraste pelo contraste correspondente definido abaixo e executar todo o script novamente

```
#c <- c(0, 1, 1,-1,-1)    #  contraste rce vs rcfo
#c <- c(0, 1,-1, 0, 0)    #  contraste rce vs rceB12
#c <- c(0, 0, 0, 1,-1)    #  contraste rcfo vs rcfoB12
```

Estes resultados podem ser obtidos facilmente com o pacote "***gmodels"***

```
> # instalando o pacote gmodels
> # install.packages("gmodels")
>
> # requerendo o pacote para o ambiente R
> require(gmodels)
>
> # juntando os 4 contrastes no objeto cte
> cte <-rbind(c(4,-1,-1,-1,-1), c(0, 1, 1,-1,-1),
+             c(0, 1,-1, 0, 0), c(0, 0, 0, 1,-1))
>
> # calculando a anova com desdobramento dos gl dos tratamentos
> comp.av  <- aov(comp ~ trat,contrast = list(trat = make.contrasts(cte)))
>
> # imprimindo o quadro da ANOVA
> summary(comp.av, split = list(trat = 1:4))
             Df  Sum Sq  Mean Sq  F value  Pr(>F)
trat          4   7.717    1.929    7.190 0.00193 **
  trat: C1    1   3.872    3.872   14.430 0.00175 **
  trat: C2    1   2.102    2.102    7.835 0.01348 *
  trat: C3    1   0.031    0.031    0.116 0.73764
  trat: C4    1   1.711    1.711    6.377 0.02331 *
Residuals  15   4.025    0.268
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Tratamentos com níveis quantitativos**

Quando os tratamentos e/ou fatores utilizados num experimento são de natureza **qualitativa** (raça, sexo, cultivares, tratos culturais etc.) os testes de comparações de médias (teste t-Student, testes de Tukey, Duncan, Scheffé

etc.) se aplicam sem restrições. A esses casos se equiparam os fatores ou tratamentos **quantitativos** (doses de uma droga, tempo, etc.) *quando há só dois níveis* (presença e ausência, por exemplo). O mesmo não acontece, porém, quando o tratamento ou fator quantitativo *tem mais de dois níveis*, por exemplo:

- doses crescentes de cobre na alimentação de galinhas (0, 400 e 800 ppm);
- doses crescentes de uma droga;
- 0%, 20%, 40% e 60% de substituição de um ingrediente da ração por farelo de soja.

Em tais situações é essencial avaliar o *comportamento da variável resposta ao longo dos níveis do fator*, através de uma *equação de regressão*. Por exemplo: a equação que associa a freqüência cardíaca em função de doses de uma droga é quase sempre desconhecida, mas em geral, pode ser bem *estimada* por meio de uma *equação polinomial* do tipo:

$$Y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \ldots,$$

sendo *Y*, a resposta avaliada e *x* os níveis quantitativos do fator (tratamentos).

O ajuste e a interpretação da equação de regressão quando o polinômio é de grau muito elevado são tarefas bastante complexas. Porém, quando os níveis do fator quantitativo são igualmente espaçados, o estudo do comportamento das médias pode ser feito utilizando o método dos polinômios ortogonais, que será apresentado a seguir através de um exemplo.

Exemplo: os efeitos de quatro tratamentos no ganho de peso (g) de alevinos de carpas foram comparados em um DIC.

*T1 – ração comum.*
*T2 – ração comum + 10 mg de B12.*
*T3 – ração comum + 20 mg de B12.*
*T4 – ração comum + 30 mg de B12.*

Dados

| | *Repetições* | | | |
|---|---|---|---|---|
| *Trat.* | *1* | *2* | *3* | *4* |
| *0* | *6,80* | *6,50* | *6,40* | *6,50* |
| *10* | *7,90* | *6,60* | *6,80* | *6,20* |
| *20* | *8,30* | *8,40* | *8,60* | *9,20* |
| *30* | *9,50* | *9,80* | *10,00* | *10,70* |

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

$H_1 :$ *∃ pelo menos duas médias* $\neq$

Análise de variância usual.

| *Causas da Variação* | *G.L.* | *S.Q.* | *Q.M.* | *F* |
|---|---|---|---|---|
| *Tratamentos* | *3* | *31,03* | *10,02* | *41,3* |
| *Resíduo* | *12* | *2,95* | *0,25* | |
| *Total* | *15* | *33,98* | | |

$F_{(3,12;0,05)} = 3,49$ $F_{(3,12;0,01)} = 5,95$

O teste F é significativo a 1% de probabilidade, portanto rejeita-se Ho, os tratamentos apresentam efeitos distintos sobre o crescimento dos alevinos de carpas. Como os níveis são eqüidistantes, 0, 10, 20 e 30 mg a decomposição dos graus de liberdade pode ser feita com uso de polinômios ortogonais,

usando-se os coeficientes dos contrastes encontrados em tabelas. As tabelas são construídas em função do número de tratamentos, denominados níveis. Assim, como temos 4 tratamentos, temos 4 níveis e o polinômio máximo é o de grau 3. Consultando as tabelas dos coeficientes dos polinômios ortogonais (Gomes P., 1966, p. 314, Sampaio, I.B.M, 1998, p. 215) , podemos montar a seguinte tabela

| Tratamentos (Totais) | Coeficientes para 4 níveis | | |
|---|---|---|---|
| | *1º grau* | *2º grau* | *3º grau* |
| $T_1$=26,20 | -3 | +1 | -1 |
| $T_2$=27,50 | -1 | -1 | +3 |
| $T_3$=34,50 | +1 | -1 | -3 |
| $T_4$=40,00 | +3 | +1 | +1 |
| $\sum_i c_i^2$ | 20 | 4 | 20 |

Assim, para o *efeito linear* temos:

$$\hat{Y}_{Linear} = (-3)T_1 + (-1)T_2 + (1)T_3 + (3)T_4 =$$
$$= (-3)(26,20) + (-1)(27,50) + (1)(34,50) + (3)(40,00) =$$
$$= 48,40$$

$$S.Q.(Y_{Linear}) = \frac{(48,40)^2}{(4)(20)} = 29,28$$

*(A obtenção das SQ dos efeitos quadráticos e cúbicos são deixados como exercício)*

A análise de variância com desdobramento dos graus de liberdade dos tratamentos por polinômios ortogonais.

| *Causas da Variação* | *G.L.* | *S.Q.* | *Q.M.* | *F* |
|---|---|---|---|---|
| ***Regressão linear*** | *1* | *29,28* | *29,28* | *119,32* |
| ***Regressão Quadrática*** | *1* | *1,10* | *1,10* | *4,49* |
| ***Regressão Cúbica*** | *1* | *0,65* | *0,65* | *2,64* |
| *Tratamentos* | *(3)* | *(31,03)* | *10,34* | *42,15* |
| *Resíduo* | *12* | *2,95* | *0,25* | |
| *Total* | *15* | *33,98* | | |

$F_{(3,12;0,05)} = 3,49$; $F_{(3,12;0,01)} = 5,95$; $F_{(1,12;0,05)} = 4,75$ e $F_{(1,12;0,01)} = 9,33$

Conclusão: somente a componente do 1º grau foi significativa *(p<0,01)*, ou seja, a diferença entre os valores médios dos tratamentos está sendo explicada por uma equação linear, $Y = a + bx$ , cujos parâmetros a e b são estimados por:

$$\hat{b} = \frac{\sum_{i=1}^{k} X_i Y_i - \frac{\sum_{i=1}^{k} X_i \sum_{i=1}^{k} Y_i}{k}}{\sum_{i=1}^{k} X_i^2 - \frac{\left(\sum_{i=1}^{k} X_i\right)^2}{k}} \quad \text{e} \quad \hat{a} = \bar{Y} - \hat{b} * \bar{X},$$

sendo: $\hat{b}$ *e* $\hat{a}$ , os estimadores de mínimos – quadrados de *b* e *a*, respectivamente, $x_i$ = 0, 10, 20 e 30 as doses de vitamina $B_{12}$; $\bar{y}_{i+}$ = *6,55, 6,80,*

*8,63 e 10,00* são os comprimentos médios dos alevinos, para *i = 1, 2, 3, 4.*
Utilizando essas fórmulas, obtemos a equação

$$\hat{Y} = 6{,}168 + 0{,}122\,X$$

Script no R para os cálculos acima

```
> # entrando com os dados do arquivo gpdadosquantitativos.txt
> dados.gpq<-read.table("gpdadosquantitativos.txt",h=T,dec=",")
>
> head(dados.gpq)
   Dose g.peso
1    0    6.8
2    0    6.5
3    0    6.4
4    0    6.5
5   10    7.9
6   10    6.6
> attach(dados.gpq)
>
> # calculando o quadro da  ANOVA
> gpq.av <- aov(g.peso~factor(Dose))
> summary(gpq.av)
             Df Sum Sq  Mean Sq  F value  Pr(>F)
factor(Dose)  3 31.033   10.344    42.15 1.2e-06 ***
Residuals    12  2.945    0.245
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> # imprimindo o quadro da anova com o comando anova( )
> anova(gpq.av)
Analysis of Variance Table
Response: g.peso
             Df  Sum Sq  Mean Sq  F value    Pr(>F)
factor(Dose)  3  31.032  10.3442   42.149 1.196e-06 ***
Residuals    12   2.945   0.2454
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> # obtenção do QMR no quadro da anova
> qmr <- anova(gpq.av)[2,3]
> qmr
[1] 0.2454167
>
> # obtenção dos gl do residuo no quadro da anova
> glr <- anova(gpq.av)[2,1]
> glr
[1] 12
>
> # cálculo dos totais por tratamento
> t.trat <- tapply(g.peso,Dose,sum)
> t.trat
  0  10  20  30
26.2 27.5 34.5 40.0
```

```
>
> # definindo o nº de repetições
> r<-4
>
> # Definição do contraste
> c <- c(-3,-1,1,3)     # efeito linear
>
> #  estimativa do contraste linear com base nos totais
> y.est <-sum(c*t.trat)
> y.est
[1] 48.4
>
> # cálculo da soma de quadrados
> sqy<- (y.est^2)/(r*sum(c^2))
> sqy
[1] 29.282
>
> # calculo da estatística F
> fc <- sqy/qmr
> fc
[1] 119.3154
>
> # calculo do valor de p da estatística fc
> valor.p <- 1-pf(fc,1,glr)
> valor.p
[1] 1.368487e-07
```

Para obter os resultados referentes aos outros contrastes basta substituir o objeto c na linha # Definição do contraste pelo contraste correspondente definido abaixo e executar todo o script novamente

```
#c <- c(1,-1,-1,1)   # efeito quadrático
#c <- c(-1,3,-3,1)   # efeito cúbico
```

Este quadro da anova pode ser obtido facilmente com o pacote "*gmodels".* Não há necessidade de instalar o pacote *gmodels* novamente, dado que ele já foi instalado no script anterior. Basta requerê-lo.

Script no R utilizando o pacote *gmodels*

```
> # requerendo o pacote para o ambiente R
> # require(gmodels)
>
> # juntando os 3 contrastes no objeto cte
> cte<-rbind(c(-3, -1,  1, 3), c(1, -1, -1, 1), c(-1,  3, -3, 1))
> Dose<-factor(Dose)
> # cálculando a anova com desdobramento dos gl dos tratamentos
> gpeso.av  <- aov(g.peso ~ Dose,contrast = list(Dose =
make.contrasts(cte)))
> # imprimindo o quadro da ANOVA
> summary(gpeso.av, split = list(Dose = 1:3))
```

```
             Df   Sum  Sq Mean Sq F value   Pr(>F)
Dose          3   31.033  10.344      42.149  1.20e-06 ***
  Dose: C1    1   29.282  29.282     119.315  1.37e-07 ***
  Dose: C2    1    1.102   1.102       4.492   0.0556 .
  Dose: C3    1    0.648   0.648       2.640   0.1301
Residuals    12    2.945  0.245
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

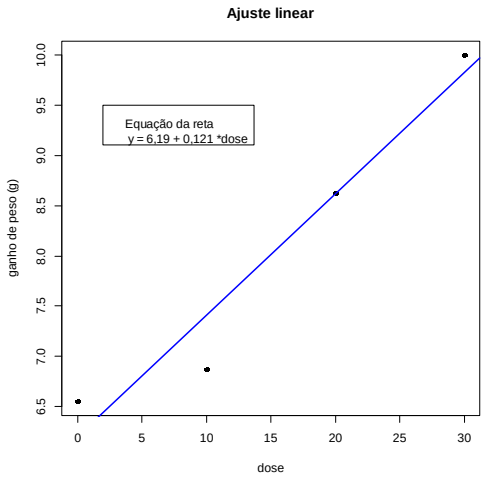
Este quadro da anova com os desdobramentos dos graus de liberdade dos tratamentos junto com as equações linear, quadrática e cúbica são facilmente obtidos com o pacote ***ExpDes***.

Script no R utilizando o pacote ***ExpDes***

```
> #requrendo o pacote ExpDes
> require(ExpDes)
>
> #quadro da anova com o desdobramento dos graus de liberdade dos
trat
> crd(Dose,g.peso,quali=F)
------------------------------------------------------------------------
Analysis of Variance Table
------------------------------------------------------------------------
               DF     SS      MS         Fc        Pr>Fc
Treatament  3  31.033  10.3442   42.149   1.1963e-06
Residuals  12   2.945   0.2454
Total      15  33.978
------------------------------------------------------------------------
CV = 6.18 %
------------------------------------------------------------------------
Shapiro-Wilk normality test
p-value:  0.1290953
According to Shapiro-Wilk normality test at 5% of significance, residuals
can be considered normal.
------------------------------------------------------------------------
Adjustment of polynomial models of regression
------------------------------------------------------------------------
$`Linear Model\n------------------------------------------------------------------------`
   Estimate   Standard.Error      tc      p.value
b0   6.1975       0.20724        29.90512      0
b1   0.1210       0.01108        10.92316      0

$`R2 of linear model`
[1] 0.9435914

$`Analysis of Variance of linear model`
                 DF      SS        MS        Fc       p.value
Linear Effect    1    29.2820  29.28200  119.32       0
Lack of fit      2     1.7505   0.87525   3.57      0.06087
Residuals       12     2.9450   0.24542
------------------------------------------------------------------------
```

```
$`Quadratic Model\n----------------------------------------------------------------------------`
  Estimate Standard.Error       tc      p.value
b0 6.460000       0.24143  26.75769  0.00000
b1 0.042250       0.03877   1.08974  0.29723
b2 0.002625       0.00124   2.11952  0.05558

$`R2 of quadratic model`
[1] 0.9791187

$`Analysis of Variance of quadratic model`
                    DF     SS       MS        Fc      p.value
Linear Effect        1  29.2820  29.28200  119.32      0
Quadratic Effect     1   1.1025   1.10250   4.49     0.05558
Lack of fit          1   0.6480   0.64800   2.64     0.13013
Residuals           12   2.9450   0.24542

-----------------------------------------------------------------------
$`Cubic Model\n----------------------------------------------------------------------------`
      Estimate   Standard.Error      tc          p.value
b0  6.550000       0.24770          26.44352    0.00000
b1 -0.098750       0.09504          -1.03903    0.31927
b2  0.016125       0.00840           1.91968    0.07898
b3 -0.000300       0.00018          -1.62493    0.13013

$`R2 of cubic model`
[1] 1

$`Analysis of Variance of cubic model`
                    DF     SS       MS        Fc      p.value
Linear Effect        1  29.2820  29.28200  119.32      0
Quadratic Effect     1   1.1025   1.10250    4.49    0.05558
Cubic Effect         1   0.6480   0.64800    2.64    0.13013
Lack of fit          0   0.0000   0.00000    0       1
Residuals           12   2.9450   0.24542
-----------------------------------------------------------------------
   Levels   Observed Means
1       0          6.550
2      10          6.875
3      20          8.625
4      30         10.000
-----------------------------------------------------------------------
>
> # entrando com os valores da dose (x)
> dose<-c(0,10,20,30)
> # cálculo das médias dos tratamentos (y)
> m.trat<-tapply(g.peso,Dose,mean)
> m.trat
    0     10     20     30
 6.550  6.875  8.625 10.000
>
> # gráfico de dispersão (dose x ganho de peso)
```

```
> plot(dose,m.trat,pch=16, col="black",ylab="ganho de peso (g)")
>
> # ajustando a reta de regressão
> reg.lin<-lm(m.trat~dose)
>
> #imprimindo os resultados do ajuste
> summary(reg.lin)
Call:
lm(formula = m.trat ~ dose)

Residuals:
     0      10      20      30
 0.3525 -0.5325  0.0075  0.1725

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.19750    0.39137  15.835  0.00396 **
dose         0.12100    0.02092   5.784  0.02861 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4678 on 2 degrees of freedom
Multiple R-squared:  0.9436,    Adjusted R-squared:  0.9154
F-statistic: 33.46 on 1 and 2 DF,  p-value: 0.02861
>
> # colocando a reta estimada no gráfico de dispersão
> abline(reg.lin,col="blue",lwd=2)
legend(2,9.5,"Equação da reta\n y = 6,19 + 0,121 *dose")
```




```
detach(dados.gpq)
```

## 5º EXERCÍCIO PRÁTICO DE ESTATÍSTICA EXPERIMENTAL

1) Num experimento estudou-se a adição de triguilho, a uma dieta básica de milho e farelo de soja na alimentação se suínos, mestiços ( Landrace x Large White), com peso inicial de 10,5 kg durante um período experimental de 40 dias, mantidos em gaiolas metálicas de 1,90 x 0,74 m. O delineamento experimental foi o inteiramente casualizado com 5 tratamentos e 8 repetições e a parcela experimental representada por 4 animais (dois machos castrados e duas fêmeas). Os tratamentos consistiram na inclusão de 0; 7,5; 15,0; 22,5; e 30% de triguilho em dietas à base de milho e soja.

Os ganhos de peso médio diário em gramas (média dos 4 animais na parcela) foram:

| *Tratamentos* | | | | *Repetições* | | | | | *Total* |
|---|---|---|---|---|---|---|---|---|---|
| *% de triguilho* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | |
| *0,0* | *340* | *320* | *310* | *350* | *320* | *340* | *330* | *340* | *2650* |
| *7,5* | *360* | *350* | *350* | *360* | *370* | *380* | *340* | *350* | *2860* |
| *15,0* | *370* | *370* | *380* | *390* | *360* | *370* | *360* | *380* | *2980* |
| *22,5* | *380* | *390* | *380* | *390* | *360* | *360* | *360* | *390* | *3010* |
| *30,0* | *400* | *390* | *410* | *420* | *380* | *390* | *410* | *420* | *3220* |
| | | | | | | | | | 14720 |

A análise de variância preliminar é a seguinte:

| *Causa da variação* | *GL* | *S. Quadrados* | *Q. M* | *F* |
|---|---|---|---|---|
| *Tratamentos* | *4* | *21915.00* | *5478,55* | *31,30*** |
| *Resíduo* | *35* | *6125,00* | *175,00* | |
| *Total* | *39* | | | |

** Significativo p<0,01

a- Escrever o script na linguagem do R para reproduzir o quadro da anova acima.

b- Escrever também o script para montar a tabela de análise de variância com desdobramento dos graus de liberdade de tratamentos por polinômios ortogonais.

| *Causa da variação* | *GL* | *S. Q.* | *Q. M.* | *F* | *Vapor de p* |
|---|---|---|---|---|---|
| *Tratamentos* | | *21915.00* | *5478,55* | *31,30*** | |
| *$Y_1$ (Linear)* | | | | | |
| *$Y_2$ (Quadrático)* | | | | | |
| *$Y_3$ (Cúbico)* | | | | | |
| *$Y_4$ (4ª grau)* | | | | | |
| *Resíduo* | | *6125,00* | *175,00* | | |
| *Total* | | *0,3426* | | | |

c- Tirar as conclusões práticas possíveis para este experimento.

d- Calcular as médias e os erros padrões das médias dos tratamentos e o coeficiente de determinação e de variação do experimento.

*Coeficientes dos polinômios ortogonais para 5 tratamentos:*

| | | | | | |
|---|---|---|---|---|---|
| *Linear:* | *-2* | *-1* | *0* | *1* | *2* |
| *Quadrático. :* | *2* | *-1* | *-2* | *-1* | *2* |
| *Cúbico:* | *-1* | *2* | *0* | *-2* | *1* |
| *4º Grau :* | *1* | *-4* | *6* | *-4* | *1* |

2) Num experimento inteiramente casualizado de competição de linhagens de aves visando o ganho de peso aos 60 dias de idade, foram utilizados 4 tratamentos e 6 repetições. Os tratamentos, com as respectivas médias de ganho de peso foram as seguintes:

*1- ARBOR ACRES* $\bar{y}_{1+} = 1{,}81$ kg

*2- KIMBER 44* $\bar{y}_{2+} = 1{,}59$ kg

*3- PILCH* $\bar{y}_{3+} = 1{,}61$ *kg*

*4- COBBS* $\bar{y}_{4+} = 1{,}71$ kg

Para a análise de variância dos ganhos de peso, obteve-se: S.Q. Tratamentos = 0,2266 e S.Q. Total = 0,3426

a) Sejam os contrastes:

$y_1 = \mu_1 + \mu_2 - \mu_3 - \mu_4$ ; $y_2 = \mu_1 - \mu_2$ ; $y_3 = \mu_3 - \mu_4$

Verificar se estes contrastes são ortogonais entre si.

b) Preencher o quadro da anova abaixo:

| F.V. | GL | S. Q. | Q. M. | F | Vapor de p |
|---|---|---|---|---|---|
| Tratamentos | | 0,2266 | | | |
| $Y_1$ | | | | | |
| $Y_2$ | | | | | |
| $Y_3$ | | | | | |
| Resíduo | | | | | |
| Total | | 0,3426 | | | |

c) Apresente as conclusões destes testes.

d) Calcular $R^2$ e o C.V. deste experimento e concluir.

3- Num experimento inteiramente casualizado, com 5 tratamentos e 6 repetições, estudou-se o efeito da infestação de ovinos e caprinos por larvas de ***Gaigeria pachyscelis*** (Nematoda: Ancylostomatoidea).

Os tratamentos aplicados foram:

***$T_1$*** *- infestação com 150 larvas por animal*
***$T_2$*** *- infestação com 300 larvas por animal*
***$T_3$*** *- infestação com 600 larvas por animal*
***$T_4$*** *- infestação com 1200 larvas por animal*
***$T_5$*** *- infestação com 2400 larvas por animal.*

A análise de variância do número de semanas decorridas até a morte do animal apresentou os seguintes resultados.

S.Q. Tratamentos = 5,7204

S.Q. Total =13,1829

Sabendo-se, também que as médias do número de semanas, decorridas até a morte do animal, por tratamento foram:

$\overline{y}_{1+} = 4{,}28$ $\overline{y}_{2+} = 4{,}16$ $\overline{y}_{3+} = 3{,}55$ $\overline{y}_{4+} = 3{,}22$ $\overline{y}_{5+} = 2{,}71$

Pede-se:

a) Montar a análise de variância e concluir.

| F.V. | GL | S. Q. | Q. M. | F | Vapor de p |
|---|---|---|---|---|---|
| Tratamentos | | 5,7204 | | | |
| Resíduo | | | | | |
| Total | | 13,1829 | | | |

b) Verificar pelo teste de “Tukey”, “Duncan” e “Scheffé” ao nível de 5% de probabilidade, quais as médias de tratamentos que estão diferindo significantemente entre si.

**Aula 6 Delineamento em blocos casualizados (DBC)**

Suponha que um experimentador esteja interessado em estudar os efeitos de 3 diferentes dietas. A primeira providência do pesquisador foi a de se inteirar a respeito da natureza do material experimental disponível. Feito isto, constatou que ele disporia de 12 animais com aproximadamente o mesmo peso. Entretanto, estes 12 animais eram provenientes de 4 ninhadas, cada uma contendo três animais. Dentro de uma ninhada, os três animais foram sorteados às três dietas. Os animais foram colocados em 12 baias idênticas e alimentados com as dietas sorteadas, em idênticas condições. Mediu-se, então, o ganho de peso desses animais depois de 12 semanas. Os dados obtidos são apresentados no quadro abaixo:

| *Dieta* | *Ninhada* | | | | *Total* |
|---|---|---|---|---|---|
| | *1* | *2* | *3* | *4* | |
| *A* | *28,7* | *29,3* | *28,2* | *28,6* | *114,8* |
| *B* | *30,7* | *34,9* | *32,6* | *34,4* | *132,6* |
| *C* | *31,9* | *34,2* | *34,9* | *35,3* | *136,3* |
| *Total* | *91,3* | *98,4* | *95,7* | *98,3* | *383,7* |

Organizando as observações em arquivos com extensão ***.xls*** ou ***.txt***

dieta.xls

| dieta | ninhada | gpeso |
|---|---|---|
| A | Ninhada1 | 28.7 |
| A | Ninhada2 | 29.3 |
| A | Ninhada3 | 28.2 |
| A | Ninhada4 | 28.6 |
| B | Ninhada1 | 30.7 |
| B | Ninhada2 | 34.9 |
| B | Ninhada3 | 32.6 |
| B | Ninhada4 | 34.4 |
| C | Ninhada1 | 31.9 |
| C | Ninhada2 | 34.2 |
| C | Ninhada3 | 34.9 |
| C | Ninhada4 | 35.3 |

dieta.txt

| *dieta* | *ninhada* | *gpeso* |
|---|---|---|
| *A* | *Ninhada1* | *28.7* |
| *A* | *Ninhada2* | *29.3* |
| *A* | *Ninhada3* | *28.2* |
| *A* | *Ninhada4* | *28.6* |
| *B* | *Ninhada1* | *30.7* |
| *B* | *Ninhada2* | *34.9* |
| *B* | *Ninhada3* | *32.6* |
| *B* | *Ninhada4* | *34.4* |
| *C* | *Ninhada1* | *31.9* |
| *C* | *Ninhada2* | *34.2* |
| *C* | *Ninhada3* | *34.9* |
| *C* | *Ninhada4* | *35.3* |

*(Dica: primeiro digite os dados no excel, para depois colocá-lo no bloco de notas)*

O delineamento experimental para este ensaio de dietas é um exemplo de um ***Delineamento em Blocos Casualizados*** com três tratamentos e quatro blocos. Os tratamentos são níveis de um fator experimental, as três dietas; os blocos são os níveis do fator confundido, as ninhadas. Dado que os animais em diferentes ninhadas respondem diferentemente a uma dada dieta, a ninhada é considerada, um fator de confundimento. As 12 unidades experimentais (animais) são agrupados em 4 blocos, de tal forma que, dentro de cada grupo, três unidades são afetadas pelo mesmo nível do fator de confundimento. Por causa da porção das características inerentes aos animais dentro de uma mesma ninhada (bloco), suas respostas serão muito similares, enquanto que as respostas dos animais pertencentes a diferentes ninhadas irão variar muito; isto é, as unidades experimentais são mais homogêneas dentro dos blocos do que entre os blocos. Assim, resumidamente, podemos definir que um ***DBC*** é um delineamento no qual as unidades (unidades experimentais) às quais os tratamentos são aplicados são subdivididos em grupos homogêneos, denominados de blocos, tal que o número de unidades experimentais em um bloco é igual ao número (ou algum múltiplo do número) de tratamentos estudados. Os tratamentos são então sorteados às unidades experimentais

dentro de cada bloco. Deve-se ressaltar que cada tratamento aparece em cada bloco, e todo bloco recebe todos os tratamentos. Quando se usa o ***DBC,*** o objetivo é isolar e remover do termo de erro (resíduo) a variação atribuída ao bloco, garantindo assim, que as médias dos tratamentos estão livres do efeito dos blocos. A efetividade deste delineamento depende da habilidade em se obter blocos homogêneos de unidades experimentais. A habilidade para formar blocos homogêneos depende do conhecimento que o pesquisador tem do material experimental. Quando os blocos são usados adequadamente, o QMR (quadrado médio do resíduo) no quadro da ANOVA será reduzido, a estatística F aumentará, e a chance de se rejeitar $H_0$ (hipótese de nulidade) será maior.

Em experimentos com animais, quando suspeita-se que diferentes raças de animais responderá diferentemente ao mesmo tratamento, a raça do animal pode ser usada como um fator a ser considerado na formação dos blocos. O ***DBC*** pode, também, ser empregado efetivamente quando um experimento deve ser conduzido em mais de um laboratório (bloco) ou quando vários dias (blocos) são requeridos para a realização do experimento. No DBC temos os três princípios básicos da experimentação: **repetição**, **casualização** e **controle local.**

Vantagens do DBC

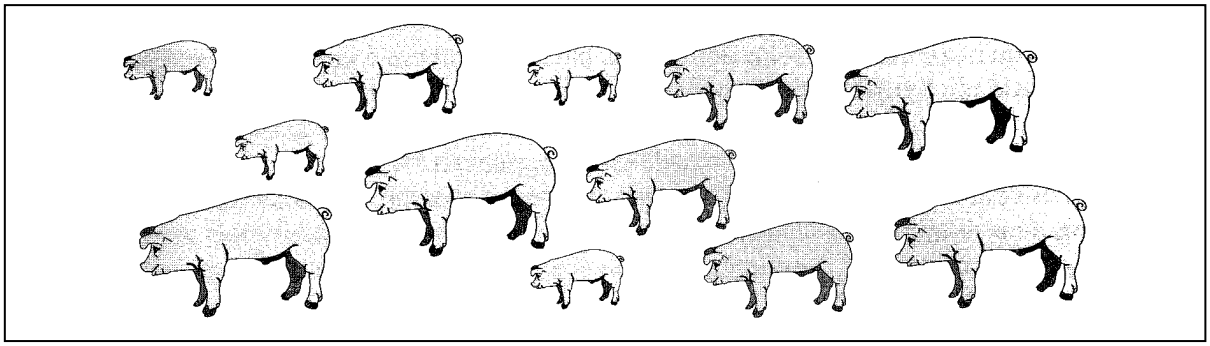
- Com o agrupamento das parcelas, geralmente se obtém resultados mais precisos que aqueles obtidos num DIC.
- Desde exista material experimental suficiente, o delineamento será sempre balanceado, podendo-se incluir qualquer número de tratamentos.
- A análise estatística é bastante simples.
- Se a variância do erro experimental é maior para alguns tratamentos que para outros, pode-se obter um erro não viesado para testar qualquer combinação específica das médias dos tratamentos.

Principal desvantagem

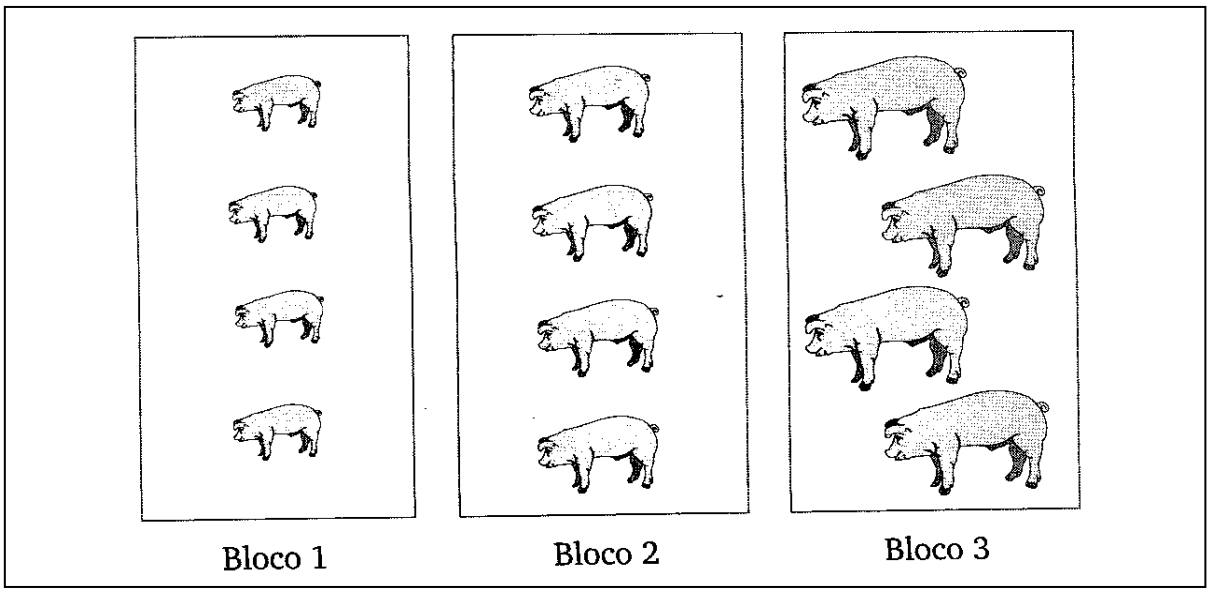
Ocorre quando da perda de parcela(s) em algum tratamento. Apesar de existir um método apropriado de estimação desses valores, há a perda de eficiência na comparação de médias envolvendo esses tratamentos.

Esquematicamente para um DBC com 4 tratamentos e 3 blocos (classes de idade) temos:

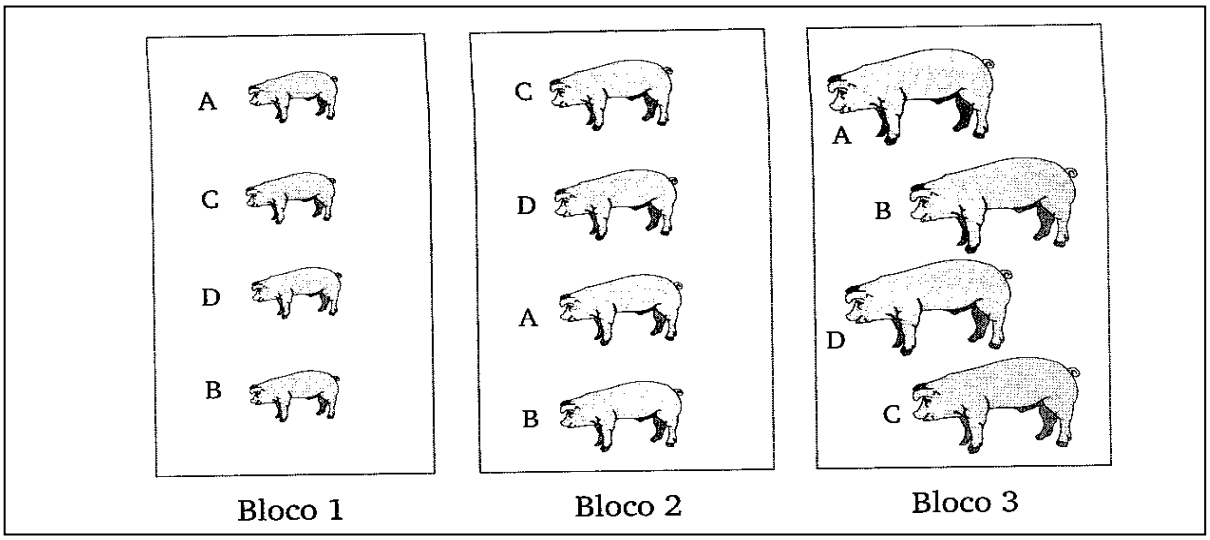
1) Unidades experimentais heterogêneas (Fonte: Vieira, 2006, pag. 15).



2) Constituição dos 3 blocos. ( 3 classes de idades ).




3) Delineamento de um experimento em blocos casualizados.




**2 Organização dos dados no *DBC*.**

Vamos considerar *k* -tratamentos; *r* – blocos e $y_{ij}$ é o valor observado na parcela que recebeu o tratamento *i* e se encontra no bloco *j.* Assim, um quadro para representar os valores amostrais de um ***DBC*** pode ser da forma abaixo:

| Trat. | Blocos 1 | 2 | 3 | ... | j | ... | r | Total | Média |
|---|---|---|---|---|---|---|---|---|---|
| 1 | $Y_{11}$ | $Y_{12}$ | $Y_{13}$ | ... | ... | ... | $Y_{1r}$ | $Y_{1+}$ | $\overline{Y}_{1+}$ |
| 2 | $Y_{21}$ | $Y_{22}$ | $Y_{23}$ | ... | ... | ... | $Y_{2r}$ | $Y_{2+}$ | $\overline{Y}_{2+.}$ |
| 3 | $Y_{31}$ | $Y_{32}$ | $Y_{33}$ | ... | ... | ... | $Y_{3r}$ | $Y_{3+}$ | $\overline{Y}_{3+}$ |
| . | . | . | . | . | ... | . | . | . | . |
| . | . | . | . | . | ... | . | . | . | . |
| . | . | . | . | . | ... | . | . | . | . |
| i | . | . | . | . | $Y_{ij}$ | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . |
| k | $Y_{k1}$ | $Y_{k2}$ | $Y_{k3}$ | ... | ... | ... | $Y_{kr}$ | $Y_{k+}$ | $\overline{Y}_{k+}$ |
| TOTAL | $Y_{+1}$ | $Y_{+2}$ | $Y_{+3}$ | ... | $Y_{+j}$ | ... | $Y_{+r}$ | $Y_{++}$ | |

**3 Modelo matemático**

$$Y_{ij} = \mu + \beta_j + \tau_i + \varepsilon_{ij} \quad i = 1, 2, \cdots, k \quad e \quad j = 1, 2, \cdots, r$$

- sendo:

  $y_{ij}$ *a observação que recebeu o* $i-$*ésimo tratamento no* $j-$*ésimo bloco*;

  $\mu$ *é média geral comum a todas as observações*;

  $\beta_j$ *é o efeito do* $j-$*ésimo bloco*, *com* $\sum_{j=1}^{r} \beta_j = 0$;

  $\tau_i$ *é efeito do* $i-$*ésimo tratamento com* $\sum_{j=1}^{r} \tau_i = 0$;

  $\varepsilon_{ij}$ *é o efeito do erro aleatório*.

**4 Suposições do modelo.** Neste modelo,

- cada $y_{ij}$ observado constitui uma amostra aleatória independente de tamanho 1 de cada uma das *kr* populações
- os $\varepsilon_{ij}$ são independentes e normalmente distribuídos com média 0 e variância $\sigma^2$, ou seja, $\varepsilon_{ij} \sim N(0, \sigma^2)$. Isto implica em que as *kr* populações são normalmente distribuídas com média $\mu_{ij}$ e a mesma variância $\sigma^2$, ou seja, $y_{ij} \sim N(\mu_{ij}, \sigma^2)$;
- os efeitos de blocos e tratamentos são aditivos. Esta suposição pode ser interpretada como *não existe interação entre tratamentos e blocos.* Em outras palavras, uma particular combinação bloco-tratamento não produz um efeito que é maior que ou menor que a soma dos efeitos individuais.

**4 Hipótese estatística**

Podemos testar

$H_0 : \tau_i = 0$, *com* $i = 1,2,...,k$
$H_1$ : *nem todos os* $\tau_i = 0$

ou

$H_0 : \mu_1 = \mu_2 = ... = \mu_k = 0$
$H_1 : \mu_i \neq \mu_j \quad i \neq j$

Geralmente o teste de hipótese com relação aos efeitos de blocos não é feito por dois motivos: primeiro o interesse principal é testar os efeitos de tratamento, o propósito usual dos blocos é eliminar fontes estranhas de variação. Segundo, embora as unidades experimentais sejam distribuídas aleatoriamente aos tratamentos, os blocos são obtidos de uma maneira não aleatória.

**6 Partição da soma de quadrados**

Voltemos ao quadro de representação das observações no DBC no item 2

Podemos identificar os seguintes desvios:

- $y_{ij} - \bar{y}_{++}$, como o desvio de uma observação em relação a média amostral geral;
- $y_{ij} - \bar{y}_{i+}$, como o desvio da observação em relação à média de seu grupo ou do i-ésimo tratamento;
- $\bar{y}_{i+} - \bar{y}_{++}$, como o desvio da média do i-ésimo tratamento em relação á média geral.
- $\bar{y}_{+j} - \bar{y}_{++}$ como o desvio da média do j-ésimo bloco em relação á média geral.

Consideremos a identidade

$$(y_{ij} - \bar{y}_{++}) = (\bar{y}_{i+} - \bar{y}_{++}) + (\bar{y}_{+j} - \bar{y}_{++}) + (y_{ij} - \bar{y}_{i+} - \bar{y}_{+j} + \bar{y}_{++}),$$

a qual representa a " *a variação de uma observações em relação à média geral amostral como uma soma da variação desta observação em relação à média de seu grupo, com a variação desta observaçãoem relação à média do j-ésimo bloco em que se encontra esta observação, com a variação do erro experimental ".* Elevando-se ao quadrado os dois membros da identidade acima e somando em relação aos índices i e j, obtemos:

$$\sum_{i=1}^{k}\sum_{j=1}^{r}(y_{ij} - \bar{y}_{++})^2 = \sum_{i=1}^{k}\sum_{j=1}^{r}(y_{ij} - \bar{y}_{i+})^2 + \sum_{i=1}^{k}\sum_{j=1}^{r}(\bar{y}_{+j} - \bar{y}_{++})^2 +$$

$$+ \sum_{i=1}^{k}\sum_{j=1}^{r}(y_{ij} - y_{i+} - \bar{y}_{+j} + \bar{y}_{++})^2,$$

Descrição de cada termo da expressão acima. O termo

$$\sum_{i=1}^{k}\sum_{j=1}^{r_i}(y_{ij} - \bar{y}_{++})^2,$$

é denominado de ***Soma de Quadrados Total*** e vamos denotá-lo por ***SQT***.O número de graus de liberdade associado à ***SQT*** é kr - 1, ou N – 1, pois temos N observações e a restrição

$$\sum_{i=1}^{k}\sum_{j=1}^{r_i}(y_{ij} - \bar{y}_{++}) = 0.$$

O termo:

$$\sum_{i=1}^{k}\sum_{j=1}^{r_i}(\bar{y}_{i+} - \bar{y}_{++})^2,$$

é denominado de ***Soma de quadrados de tratamentos***, representada por ***SQTr,*** e é uma medida da variabilidade entre os tratamentos. Quanto mais

diferentes entre si forem as médias dos tratamentos, maior será a ***SQTr.*** Desde que temos k tratamentos e a restrição de que

$$\sum_{i=1}^{k}(\bar{y}_{i+} - \bar{y}_{++}) = 0 ,$$

a ***SQTr*** está associada a *k-1* graus de liberdade.

O termo

$$\sum_{i=1}^{k}\sum_{j=1}^{r_i}(\bar{y}_{+j} - \bar{y}_{++})^2 ,$$

é denominado de ***Soma de quadrados de blocos,*** representada por ***SQB,*** e é uma medida da variabilidade entre os blocos. Quanto mais diferentes entre si forem as médias dos blocos, maior será a ***SQB***, justificando assim, a utilização do delineamento em blocos. Desde que temos r blocos e a restrição

$$\sum_{j=1}^{r}(\bar{y}_{+j} - \bar{y}_{++}) = 0 ,$$

a ***SQB*** está associada a *r-1* graus de liberdade.

Finalmente, o termo

$$\sum_{i=1}^{k}\sum_{j=1}^{r}(y_{ij} - y_{i+} - \bar{y}_{+j} + \bar{y}_{++})^2 ,$$

é denominado ***SQR***. Notem que a magnitude da ***SQR*** não depende da diferença entre as médias dos tratamentos. Os graus de liberdade associada à ***SQR*** é *(k-1)(r-1)*, isto é, o produto dos graus de liberdade dos tratamentos e blocos. Assim,

$$SQT = SQB + SQTr + SQR ,$$

e os graus de liberdade associados a cada membro da equação acima fica

| total | | blocos | | tratamentos | | resíduo |
|---|---|---|---|---|---|---|
| *kr-1* | = | *(r-1)* | + | *(k-1)* | + | *(k-1)(r-1)* |

**7 Quadrado médios.**

Dividindo a *SQB*, *SQTr e SQR* pelos correspondentes graus de liberdade, obtemos, respectivamente o ***Quadrado Médio Blocos*** (***QMB***), o ***Quadrado Médio Entre Tratamentos*** (***QMTr***) e o ***Quadrado Médio Resíduo***, isto é,

$$QMB = \frac{SQB}{r-1} \quad e \quad QMTr = \frac{SQTr}{k-1} \; e \; QMR = \frac{SQR}{(k-1)\;(r-1)}$$

**8 Estatística e região crítica do teste**

A estatística para o teste é

$$F_c = \frac{QMTr}{QMR} ,$$

a qual, deve ser próximo de 1 se $H_0$ for verdadeira, enquanto que valores grandes dessa estatística são uma indicação de que $H_0$ é falsa. A teoria nos assegura que $F_c$ tem, sob $H_0$ distribuição *F – Snedecor* com (*k -1) e (k-1)(r-1)* graus de liberdade no numerador e no denominador, respectivamente.

Resumidamente, indicamos:

$$F_c \sim F_{(k-1,\,(k-1)\;(r-1),\,\alpha)}, \; sob \; H_0 .$$

Rejeitamos $H_0$ para o nível de significância $\alpha$ se

$$F_c > F_{(k-1,(k-1)(r-1),\alpha)},$$

sendo, $F_{(k-1,(k-1)(r-1),\alpha)}$ o quantil de ordem $(1-\alpha)$ da distribuição F-Snedecor com $(k-1)$ e $(k-1)(r-1)$ graus de liberdade no numerador e no denominador.

**9 Quadro de análise de variância (anova)**

Dispomos as expressões necessárias ao teste na Tabela abaixo, denominada de Quadro de Análise de Variância (ANOVA).

| Fonte de variação | *gl* | *SQ* | *QM* | *F* |
|---|---|---|---|---|
| *Blocos* | $r-1$ | $\sum_{j=1}^{r}\frac{Y_{+j}^2}{k}-\frac{(Y_{++})^2}{kr}$ | $\frac{SQB}{r-1}$ | |
| *Tratamentos* | $k-1$ | $\sum_{i=1}^{k}\frac{Y_{i+}^2}{r}-\frac{(Y_{++})^2}{kr}$ | $\frac{SQTr}{k-1}$ | $\frac{QMTr}{QMR}$ |
| *Resíduo* | $(k-1)(r-1)$ | $\neq$ | $\frac{SQR}{(k-1)(r-1)}$ | |
| *TOTAL* | $kr-1$ | $\sum_{i=1}^{k}\sum_{J=1}^{r}Y_{ij}^2-\frac{(Y_{++})^2}{kr}$ | | |

Pode-se provar que:

- $E(QMR)=\sigma^2$, ou seja, QMR é um estimador não viesado da variância $\sigma^2$;
- $E(QMTr)=\sigma^2+\frac{r}{(k-1)}\sum_{i=1}^{k}\tau_i$, ou seja, QMTr é um estimador não viesado da variância $\sigma^2$ se a hipótese $H_0: \tau_1=\tau_2=...=\tau_k=0$ é verdadeira.
- $E(QMB)=\sigma^2+\frac{k}{(r-1)}\sum_{j=1}^{r}\beta_j$

**10 Detalhes computacionais**

Apresentaremos alguns passos que facilitam os cálculos das somas de quadrados da ANOVA.

- Calcule a correção para a média $CM=\frac{(y_{++})^2}{N}$;
- Calcule a *Soma de Quadrados dos Totais* (***SQT***)
  $SQT=\sum_{i=1}^{k}\sum_{j=1}^{r}y_{ij}^2-CM$;
- Calcule a *Soma de Quadrados Entre os Tratamentos* (***SQTr***)
  $SQTr=\sum_{i=1}^{r}\frac{Y_{i+}^2}{r}-CM$;
- Calcule a *Soma de Quadrados de blocos* (***SQB***)
  $SQB=\sum_{j=1}^{r}\frac{Y_{+j}^2}{k}-CM$;
- Calcule a *Soma de Quadrados Residual* (***SQR***) pela diferença, isto é, $SQR=SQT-SQTr-SQB$;

- Calcule o *Quadrado Médio entre os Tratamentos* (***QMTr***) e o *Quadrado Médio Residual* (***QMR***)

$$QMB = \frac{SQB}{r-1},\ QMTr = \frac{SQTr}{k-1} \text{ e } QMR = \frac{SQR}{(k-1)(r-1)}$$

- Calcule $F_c$ para tratamentos $F_{cTr} = \frac{QMTr}{QMR}$ e $F_{cBl} = \frac{QMB}{QMR}$

**11 Exemplo 1**

Vamos considerar os dados apresentados no item1.

Os cálculos para montar-mos o quadro da ANOVA são:

$k = 3, r = 4,$ e $kr = N = (3)(4) = 12$. Então

- Graus de liberdade:

$$Total = kr - 1 = N - 1 = (3)(4) - 1 = 12 - 1 = 11; Trat. = k - 1 = 3 - 1 = 2$$

$$Blocos = r - 1 = 4 - 1 = 3 \text{ e } Res = (k-1)(r-1) = (3)(2) = 6$$

$$CM = \frac{(383,80)^2}{12} = 12275,20$$

- $SQT = (28,7)^2 + (29,3)^2 + ... + (35,3)^2 - CM =$
  $12353,35 - 12268,81 = 84,54$
- $SQTr = \frac{(114,8)^2}{4} + \frac{(132,6)^2}{4} + \frac{(136,3)^2}{4} - CM =$
  $12334,87 - 12268,81 = 66,06$
- $SQB = \frac{(91,3)^2}{3} + \frac{(98,4)^2}{3} + \frac{(95,7)^2}{3} + \frac{(98,3)^2}{3} - CM =$
  $12279,88 - 12268,81 = 11,07$
- $SQR = SQT - SQTr - SQB = 84,54 - 66,06 - 11,07 = 7,41$
- $QMTr = \frac{66,06}{2} = 33,03,\ QMB = \frac{11,07}{3} = 3,69$ e $QMR = \frac{7,41}{6} = 1,24$

$$F_{cTr} = \frac{QMTr}{QMR} = \frac{33,03}{1,24} = 26,64 \text{ e } F_{cBl} = \frac{QMB}{QMR} = \frac{3,69}{1,24} = 2,99$$

Organizando estes resultados no Quadro da ANOVA, temos:

| ***Fonte de variação*** | ***g.l.*** | ***SQ*** | ***QM*** | ***$F_c$*** |
|---|---|---|---|---|
| *Dietas* | *2* | *66,06* | *33,03* | *26,75* |
| *Ninhadas* | *3* | *11,07* | *3,69* | *2,99* |
| *Resíduo* | *6* | *7,41* | *1,235* | |
| *Total* | *11* | *84,54* | | |

Das tabelas das distribuições F, temos que $F_{(2,6,0,05)} = 5,14$ e $F_{(2,6,0,01)} = 10,92$. O valor $F_{cTr}$ = 26,75 é maior do que estes valores tabelados, então rejeitamos a hipótese nula $H_0$ para um nível $\alpha = 0,01$, ***ou 1%*** de probabilidade (se é significativo a 1%, também é significativo a 5%), e concluímos que existe uma diferença entre as três dietas. As conclusões sobre as diferenças entre os efeitos de ninhadas (blocos) podem ser baseadas no $F_c$ para blocos ($F_{cBl}$ *= 2,98 com p=0,118)*. Os resultados indicam que não existe uma variação significativa entre as ninhadas nos ganhos de peso.

O teste F da ANOVA para os blocos é um teste aproximado mesmo quando as suposições são satisfeitas. Alguns pesquisadores sugerem que não se considere o efeito colocado nos blocos em futuros estudos similares, somente se o valor mínimo significativo *(valor de p)* associado à estatística calculada for maior ou igual a 0,25 $(p \geq 0{,}25)$. Para estes dados, $F_{cBl} = 2{,}99$ tem um $p = 0{,}118$. Portanto, mesmo que existe insuficientes evidências para rejeitar $H_0 : \beta_j = 0$, ou seja, não existe efeito de ninhada, não é uma boa idéia ignorar os efeitos de ninhada em futuros estudos.
O script no R para obter os resultados acima é apresentado abaixo

```
> # lendo o arquivo dieta.txt e armazenando no objeto dados
> dados.ex1 <- read.table("dieta.txt",h=T)
> head(dados.ex1)
  dieta  ninhada gpeso
     1   A Ninhada1   28.7
     2    A Ninhada2  29.3
     3    A Ninhada3  28.2
     4   A Ninhada4   28.6
     5   B Ninhada1   30.7
     6   B Ninhada2   34.9

> #  anexando o objeto dados.ex1 no caminho de procura
> attach(dados.ex1)
>
> # cálculo da média da coluna com dados de ganho de peso (gpeso)
> mean(gpeso)
[1] 31.975
>
> # mostra o caminho agora com o objeto dados.ex1 incluído
> search()
 [1] ".GlobalEnv"        "dados.ex1"         "package:ExpDes"
 [4] "package:stats"     "package:graphics"  "package:grDevices"
 [7] "package:utils"     "package:datasets"  "package:methods"
[10] "Autoloads"         "package:base"
>
> # gráficos box-plot para cada dieta com a cor 5
> boxplot(gpeso~dieta,col=5)
>
> # estatisticas descritivas do box-plot de cada dieta
> e.des<- tapply(gpeso,dieta,summary)
> e.des
$A
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  28.20   28.50   28.65   28.70   28.85   29.30
$B
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  30.70   32.12   33.50   33.15   34.52   34.90
$C
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  31.90   33.62   34.55   34.08   35.00   35.30
```

```
> # média do ganho de peso de cada dieta
> m.gpeso <- tapply(gpeso,dieta,mean)
> m.gpeso
    A         B        C
28.700  33.150 34.075
>
> # desvio padrão do ganho de peso de cada dieta
> sd.gpeso <- tapply(gpeso,dieta,sd)
> sd.gpeso
       A              B              C
0.4546061    1.9087518 1.5195942
>
> # análise de variância
> gpeso.av <- aov(gpeso~factor(ninhada) + factor(dieta))
> summary(gpeso.av)
                    Df  Sum Sq   Mean Sq  F value    Pr(>F)
factor(ninhada)  3    11.07        3.69       2.988    0.11772
factor(dieta)       2    66.06       33.03      26.753   0.00103 **
Residuals           6     7.41        1.23
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # outra forma de se obter as médias do gpeso das dietas e das ninhadas
> model.tables(gpeso.av,type="means")
Tables of means
Grand mean
31.975

factor(ninhada)
factor(ninhada)
Ninhada1 Ninhada2 Ninhada3 Ninhada4
  30.43    32.80    31.90    32.77

factor(dieta)
factor(dieta)
   A     B     C
28.70 33.15 34.08
>
> # efeitos das dietas e das ninhadas
> model.tables(gpeso.av,type="effects")
Tables of effects

factor(ninhada)
factor(ninhada)
Ninhada1 Ninhada2 Ninhada3 Ninhada4
 -1.5417   0.8250  -0.0750   0.7917

factor(dieta)
factor(dieta)
    A      B      C
-3.275  1.175  2.100
```

```
> # obtendo os resíduos de cada observação
> residuos <- resid(gpeso.av)
> residuos
       1            2            3            4            5            6            7
 1.5416667 -0.2250000 -0.4250000 -0.8916667 -0.9083333  0.9250000 -0.4750000
       8            9           10           11           12
 0.4583333  -0.6333333  -0.7000000  0.9000000  0.4333333
>
> # gráfico Q-Q da normalidade
> qqnorm(residuos,pch=16,col=1)
> qqline(residuos,lwd=2,col=2)
>
> # teste de normalidade de Shapiro-Wilks para os resíduos
> shapiro.test(residuos)

        Shapiro-Wilk normality test

data:  residuos
W = 0.9014, p-value = 0.1652

> # teste de Bartlett para a igualdade das variância populacionais das
dietas
> bartlett.test(gpeso~factor(dieta))

        Bartlett test of homogeneity of variances

data:  gpeso by factor(dieta)
Bartlett's K-squared = 4.1933, df = 2, p-value = 0.1229
>
> # requerendo o pacote ExpDes
> require(ExpDes)
>
> # anova pelo ExpDes
> rbd(dieta,ninhada,gpeso,quali=T,mcomp=F)
------------------------------------------------------------------------
Analysis of Variance Table
------------------------------------------------------------------------
               DF     SS      MS      Fc      Pr>Fc
Treatament      2 66.065  33.032 26.7530  0.001025
Block           3 11.069   3.690  2.9883  0.117724
Residuals       6  7.408   1.235
Total          11 84.542
------------------------------------------------------------------------
CV = 3.48 %
------------------------------------------------------------------------
Shapiro-Wilk normality test
p-value:  0.1651555
According to Shapiro-Wilk normality test at 5% of significance, residuals
can be considered normal.
------------------------------------------------------------------------
> detach(dados.ex1)
```

```
> # mostra o caminho agora com o objeto dados.ex1 excluído
> search()
 [1] ".GlobalEnv"        "package:ExpDes"    "package:stats"
 [4] "package:graphics"  "package:grDevices" "package:utils"
 [7] "package:datasets"  "package:methods"   "Autoloads"
[10] "package:base"
```

NOTA IMPORTANTE: Sempre use detach () antes de anexar um novo arquivo de dados, especialmente se as colunas dos dois arquivos tem nomes idênticos, se não haverá problemas!

**12 Parcela perdida**

Um problema relativamente comum neste tipo de delineamento ocorre quando perdemos uma (ou mais) parcela(s) durante o desenvolvimento do experimento. Vamos considerar o seguinte exemplo:

Exemplo:

| | *Classe de idade (Blocos)* | | | | |
|---|---|---|---|---|---|
| *Trat.* | *1* | *2* | *3* | *4* | *Total* |
| *A* | *15* | *11* | *20* | *18* | *64* |
| *B* | *22* | *31* | *45* | *26* | *124* |
| *C* | *33* | *37* | * | *30* | *100* |
| *D* | *44* | *31* | *49* | *34* | *158* |
| *E* | *37* | *30* | *36* | *21* | *124* |
| *Total* | *151* | *140* | *150* | *129* | *570* |

A generalização destes dados pode ser representada no quadro abaixo

| | *Blocos* | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *Trat.* | *1* | *2* | *3* | *...* | *j* | *...* | *r* | *Total* |
| *1* | $Y_{11}$ | $Y_{12}$ | $Y_{13}$ | ... | ... | ... | $Y_{1r}$ | $Y_{1+}$ |
| *2* | $Y_{21}$ | $Y_{22}$ | $Y_{23}$ | ... | ... | ... | $Y_{2r}$ | $Y_{2+}$ |
| . | . | . | . | . | ... | . | . | . |
| . | . | . | . | . | ... | . | . | . |
| . | . | . | . | . | ... | . | . | . |
| *i* | . | . | . | . | $\hat{Y}_{ij}$ | . | . | $Y'_{i+}$ |
| . | . | . | . | . | | . | . | |
| . | . | . | . | . | . | . | . | |
| . | . | . | . | . | . | . | . | . |
| | | | | | . | | | . |
| *K* | $Y_{k1}$ | $Y_{k2}$ | $Y_{k3}$ | ... | ... | ... | $Y_{kr}$ | $Y_{i+}$ |
| *Total* | $Y_{+1}$ | $Y_{+2}$ | ... | | $Y'_{+j}$ ... | | $Y_{+r}$ | $Y'_{++}$ |

sendo:

$\hat{Y}_{ij}$ *a estimativa da parcela perdida*;

*k o número de tratamentos e r o número de blo*cos;

$Y'_{++}$ *o total das parcelas disponíveis*;

$Y'_{i+}$ *o total das parcelas res*tan*tes no tratamento onde ocorreu a parcela perdida*;

$Y'_{+j}$ *o total das parcelas res*tan*tes no bloco onde ocorreu a parcela perdida*.

Atualmente a perda de parcela não causa problemas de estimabilidade, ou seja todos os os efeitos são estimáveis, bem como todos os contrastes.

Rode a análise comum aov(y~factor(blocos)+factor(trat)) com o código ***NA*** para a o dado da parcela perdida. O que acontece são precisões diferentes nas estimativas dos efeitos e isto impossibilita o uso de testes de médias com base no ***dms*** constante, como os testes de Tukey, t-student, etc.

Um script no R para o exemplo acima é dado por:

```
> # lendo o arquivo dieta.txt e armazenando no objeto dados
> dados.expp <- read.table("ex ppdbc.txt",h=T)
>
> # imprimindo os dados
> dados.expp
   trat blocos resp
1     A      I   15
2     A     II   11
3     A    III   20
4     A     IV   18
5     B      I   22
6     B     II   31
7     B    III   45
8     B     IV   26
9     C      I   33
10    C     II   37
11    C    III   NA
12    C     IV   30
13    D      I   44
14    D     II   31
15    D    III   49
16    D     IV   34
17    E      I   37
18    E     II   30
19    E    III   36
20    E     IV   21
>
> #  anexando o objeto dados.ex1 no caminho de procura
> attach(dados.expp)
>
> # análise de variância
> resp.av <- aov(resp~factor(blocos) + factor(trat))
> summary(resp.av)
                Df  Sum Sq  Mean Sq  F value    Pr(>F)
factor(blocos)  3   333.4    111.13    3.521   0.05241 .
factor(trat)    4  1253.4    313.35    9.928   0.00119 **
Residuals      11   347.2     31.56
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
1 observation deleted due to missingness
```

**Observação:** Nesta análise, a soma de quadrados, bem como o quadrado médio dos tratamentos e dos resíduos está corretamente estimada, mas a soma de quadrados dos blocos esta subestimada.

Fazendo esta mesma análise no MiniTab, com asterisco no lugar da parcela perdida temos o seguinte resultado

**General Linear Model: Y versus Bloco; Trat**

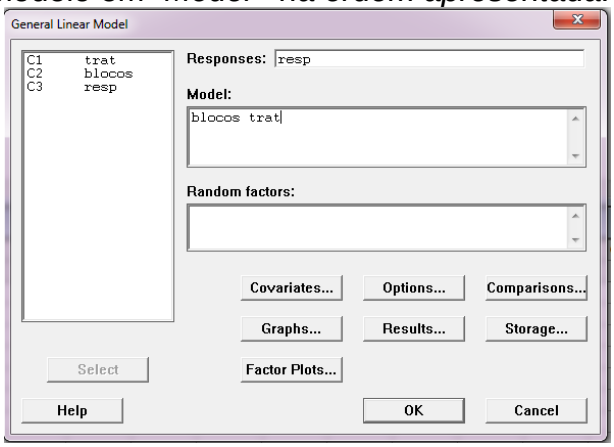
*Factor Type Levels Values*
*Bloco fixed 4 1 2 3 4*
*Trat fixed 5 A B C D E*

*Analysis of Variance for Y, using Adjusted SS for Tests*

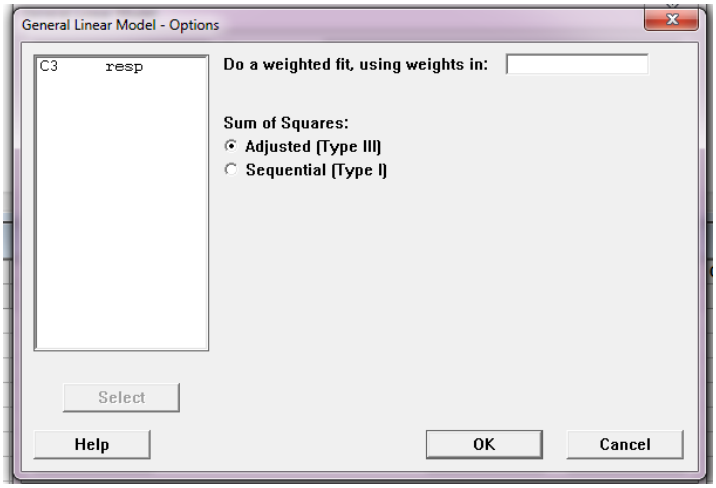
| Source | DF | Seq SS | Adj SS | Adj MS | F | P |
|---|---|---|---|---|---|---|
| Bloco | 3 | 333.40 | 400.48 | 133.49 | 4.23 | 0.032 |
| Trat | 4 | 1253.42 | 1253.42 | 313.35 | 9.93 | 0.001 |
| Error | 11 | 347.18 | 347.18 | 31.56 | | |
| Total | 18 | 1934.00 | | | | |

Reparem que a ***SQTr*** e de ***Bloco*** esta corrigida, ou seja, quando se usa o MiniTab ou o SAS. Nestes programas a correção da *SQTr* é feita automaticamente. No MiniTab é necessário seguir os seguintes passos:

- *Stat/ANOVA/General Linear Models* e nesta janela colocar os termos do modelo em "*Model" na ordem apresentada.*




- Antes de acionar o *OK* nesta janela vá à janela "*General Model – Options"* marcar na *Sum of Square* a opção *Adjusted (Type III) e OK*



## 13 Análise de variância de medidas repetidas

Um delineamento experimental de medidas repetidas é aquele, no qual várias medidas são feitas na mesma unidade experimental (geralmente animal), e estas medidas repetidas constituem as repetições. Para ilustrar melhor esta característica vamos considerar o exemplo 2, item 11 da Aula 3 , pg 38. Neste exemplo tínhamos 4 amostras independentes de animais e todos os animais de cada grupo foram alimentados, depois do sorteio, com uma das 4 dietas. Nos delineamentos de medidas repetidas não existe amostras independentes de animais, ao contrário, cada um dos 5 animais terão seus pesos medidos depois que foram submetidos a uma determinada dieta, depois de um certo período de tempo, os mesmos cinco animais terão seus pesos avaliados depois de terem sidos submetidos a outra dieta, e assim sucessivamente, até serem submetidos a todas as dietas. A tabulação dos dados pode ser bem parecida com a representação dos dados do ***DBC***. Neste exemplo podemos ter:

| | *Dietas* | | | | |
|---|---|---|---|---|---|
| *Animais* | *1* | *2* | *3* | *4* | *Total* |
| *1* | $Y_{11}$ | $Y_{12}$ | $Y_{13}$ | $Y_{14}$ | $Y_{1+}$ |
| *2* | $Y_{21}$ | $Y_{22}$ | $Y_{23}$ | $Y_{24}$ | $Y_{2+}$ |
| *3* | $Y_{31}$ | $Y_{32}$ | $Y_{33}$ | $Y_{34}$ | $Y_{3+}$ |
| *4* | $Y_{41}$ | $Y_{42}$ | $Y_{43}$ | $Y_{44}$ | $Y_{4+}$ |
| *5* | $Y_{51}$ | $Y_{52}$ | $Y_{53}$ | $Y_{54}$ | $Y_{5+}$ |
| *Total* | $Y_{+1}$ | $Y_{+2}$ | $Y_{+3}$ | $Y_{+4}$ | $Y_{++}$ |

Os resultados dos cálculos da ANOVA de um delineamento de medidas repetidas são os mesmos de uma análise de um ***DBC*.** A grande vantagem deste tipo de delineamento é o seu econômico requerimento de unidades experimentais (animais). Este delineamento tem desvantagens se existe um efeito por causa da seqüência em que os tratamentos são administrados (dietas no presente exemplo) aos animais. Outra desvantagem surge se o tempo entre a aplicação de diferentes tratamentos é insuficiente para evitar a sobreposição de efeitos do tratamento anterior.

**Exemplo 3** Considere o conjunto de dados abaixo os quais se referem a níveis de concentração de colesterol (mg/dl) em sangue de 7 animais experimentais, depois que foram tratados cada um com uma das três drogas, com suficiente tempo entre as aplicações das drogas para que seu efeito desaparecesse do animal.

| | *Drogas* | | | |
|---|---|---|---|---|
| *Animal* | *A* | *B* | *C* | *Total* |
| *1* | *164* | *152* | *178* | *494* |
| *2* | *202* | *181* | *222* | *605* |
| *3* | *143* | *136* | *132* | *411* |
| *4* | *210* | *194* | *216* | *620* |
| *5* | *228* | *219* | *245* | *692* |
| *6* | *173* | *159* | *182* | *514* |
| *7* | *161* | *157* | *165* | *483* |
| *Total* | *1281* | *1198* | *1340* | *3819* |

A hipótese de interesse é que a média do nível de colesterol no sangue é a mesma independente da droga (tratamento). (Extraído de *ZAR, J. H. Biostatistical Analysis, pg. 255, 1999*)

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

$$H_1 : \exists \text{ pelo menos duas médias diferentes}$$

Script no R para obter os resultados do exemplo 3

```
> #  Entrando com os dados pelo comando read.table( )
> dados.ex3 <- read.table("ex3dbc.txt",h=T)
> head(dados.ex3)
  droga animal colesterol
1     A      1        164
2     A      2        202
3     A      3        143
4     A      4        210
5     A      5        228
6     A      6        173
>
> # anexando o objeto dados.ex3 no caminho de procura
> attach(dados.ex3)
> # quadro da anova pela função aov( )
> colesterol.av <- aov(colesterol~factor(animal)+factor(droga))
> summary(colesterol.av)
                  Df  Sum Sq    Mean Sq    F value      Pr(>F)
factor(animal)  6    18731      3121.9       53.88      5.53e-08 ***
factor(droga)   2     1454       727.0        12.55     0.00115 **
Residuals      12      695        57.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> shapiro.test(resid(colesterol.av)) # teste de normalidade

        Shapiro-Wilk normality test

data:  resid(colesterol.av)
W = 0.9555, p-value = 0.4302
>
> bartlett.test(colesterol~droga) #teste de homogeneidade das variâncias

        Bartlett test of homogeneity of variances

data:  colesterol by droga
Bartlett's K-squared = 0.5751, df = 2, p-value = 0.7501
>
> #Utilizando os recursos do pacote ExpDes
> # requerendo o pacote ExpDes
> require(ExpDes)
>
> # anova pelo ExpDes
> rbd(droga,animal,colesterol,quali=T,mcomp="tukey")
```

```
------------------------------------------------------------------------
Analysis of Variance Table
------------------------------------------------------------------------
             DF       SS       MS       Fc       Pr>Fc
Treatament    2   1454.0   727.00   12.547   0.00114640
Block         6  18731.2  3121.87   53.877   0.00000006
Residuals    12    695.3    57.94
Total        20  20880.6
------------------------------------------------------------------------
CV = 4.19 %
------------------------------------------------------------------------
Shapiro-Wilk normality test
p-value:  0.4302259
According to Shapiro-Wilk normality test at 5% of significance, residuals
can be considered normal.
------------------------------------------------------------------------
Tukey's test
------------------------------------------------------------------------
Groups Treatments Means
a       C       191.4286
a       A       183
 b      B       171.1429
------------------------------------------------------------------------
>
> # retirando o objeto dados.ex3 do caminho de procura
> detach(dados.ex3)
```

# 6º EXERCÍCIO PRÁTICO DE ESTATÍSTICA EXPERIMENTAL

1 - Contagens médias de linfócitos de células de ratos (1000/mm$^3$) foram comparadas dando uma de duas drogas ou um placebo (controle). Ninhadas de ratos do mesmo sexo foram usadas para formar ***blocos*** homogêneos de 3 ratos cada; dentro de cada bloco, 3 tratamentos foram sorteados ao acaso. Parece razoável assumir que os efeitos dos três tratamentos deve ser relativamente constante para vários genótipos de ratos para diferentes ninhadas.

| | *Blocos* | | | | | | |
|---|---|---|---|---|---|---|---|
| *Tratamentos* | *I* | *II* | *III* | *IV* | *V* | *VI* | *VII* |
| *Placebo* | *5,4* | *4,0* | *7,0* | *5,8* | *3,5* | *7,6* | *5,5* |
| *Droga 1* | *6,0* | *4,8* | *6,9* | *6,4* | *5,5* | *9,0* | *6,8* |
| *Droga 2* | *5,1* | *3,9* | *6,5* | *5,6* | *3,9* | *7,0* | *5,4* |

a) Escrever o modelo matemático deste experimento e estabelecer as hipóteses estatísticas $H_0$ e $H_1$ para testar os efeitos dos tratamentos.
b) Montar o quadro da análise de variância para testar as hipóteses do item a).
c) Fazer o gráfico de barras das médias dos tratamentos com o desvio padrão.
d) Calcular as médias dos tratamentos e o erro padrão das médias com base na variância conjunta do experimento (QMR da ANOVA).
e) Faça um gráfico dos itens c) e d).
f) Verificar pelo teste de Dunnett se os efeitos de cada droga diferem do controle (trat1).
g) Calcular os coeficientes de variação (CV) e de determinação ($R^2$) do experimento.

2- A Tabela abaixo mostra os dados da produção de leite, de vacas da raça Gir, filhas de 3 touros, na 1ª, 2ª e 3ª parições, em 305 dias de lactação, delineados segundo um ***DBC***. com amostragem na parcela.

| *Touros* | *Parições (Blocos)* | | | | | | | | | *Total* |
|---|---|---|---|---|---|---|---|---|---|---|
| | *I* | | | *II* | | | *III* | | | |
| *1* | *1750* | *1650* | *1600* | *2250* | *2200* | *2220* | *2400* | *2650* | *2610* | *19330* |
| *2* | *1250* | *1150* | *1120* | *1750* | *1600* | *1350* | *1800* | *1900* | *1710* | *13630* |
| *3* | *1600* | *1700* | *1900* | *2300* | *2400* | *2200* | *2700* | *2750* | *2680* | *20230* |
| *Total* | *13720* | | | *18270* | | | *21200* | | | *53190* |

Pede-se:
a) Escrever o modelo matemático deste experimento e estabelecer as hipóteses estatísticas $H_0$ e $H_1$
b) Montar o quadro da análise de variância e testar as hipóteses do item a).
c) Calcular as médias dos tratamentos e o erro padrão das médias com base na variância comum (QMR da ANOVA).
d) Fazer o gráfico de barras das médias dos tratamentos com o erro padrão.
e) Verificar, pelo teste de Tukey, se existem diferenças entre as médias dos touros.
f) Calcular o coeficiente de variação e de determinação do experimento $R^2$ do experimento.

3 - Num experimento objetivando verificar a influência da suplementação concentrada de enzimas amilolíticas, celulolíticas e proeolíticas sobre o ganho de peso em ovinos da raça ideal (POLWARTH), criados a pasto, foram utilizados os seguintes tratamentos:

1 - Pasto de ***Cynodon dactylon*** + ração concentrada
2 - Pasto de ***Cynodon dactylon*** + ração concentrada + BIOVITASE
3 - Pasto de ***Cynodon dactylon*** + ração concentrada + PANASE-S
4 - Pasto de ***Cynodon dactylon*** ( Testemunha**)**

O experimento foi em ***blocos ao acaso***, com 5 blocos e 4 tratamentos, e os resultados obtidos para o ganho de peso médio, em kg, durante o experimento foram:

| | ***Blocos*** | | | | |
|---|---|---|---|---|---|
| *Tratamentos* | *I* | *II* | *III* | *IV* | *V* |
| *1- **Cynodon dactilon (testemunha)*** | *6,10* | *5,80* | *3,60* | *5,30* | *6,30* |
| *2- Ração Concentrada (RC)* | *10,90* | *13,75* | *14,50* | *11,70* | *13,10* |
| *3- RC + BIOVITASE* | *11,70* | *16,28* | *14,40* | *15,50* | *11,60* |
| *4- RC + PANASE-S* | *16,80* | *14,10* | *8,60* | *16,10* | *14,30* |

Pede-se:

a) Estabelecer as hipóteses estatísticas $H_0$ e $H_1$
b) Montar o quadro da análise de variância e testar as hipóteses do item a).
c) Calcular as médias dos tratamentos e erros padrões das médias.
d) Use o teste de Dunnett para testar os tratamentos que diferem da testemunha (RC).
e) Definir 3 contrastes ortogonais de interesse entre as médias dos tratamentos e testá-los através da análise de variância. (decomposição dos graus de liberdade).
f) Calcular os coeficientes de variação e de determinação do experimento.

4 - Num experimento estudou-se o efeito do farelo de arroz desengordurado (FAD) ) como fatores de retardamento da maturidade sexual de frangas. O ensaio, organizado em blocos completos casualizados, abrangeu duas fases distintas e foi constituído de 5 tratamentos e 5 repetições com 8 aves por unidade experimental.
A 1ª fase iniciada quando as aves atingiram 9 semanas de idade, teve duração de 12 semanas. As pesagens eram efetuadas com intervalos de duas semanas, e o consumo de ração era registrado também com intervalo de duas semanas.
Os tratamentos, na 1ª fase eram formados por rações que continham 0, 15, 30, 45, 60 % de FAD em substituição ao milho. Os resultados obtidos na 1ª fase do ensaio, para conversão alimentar foram os seguintes:

| *Tratamentos* | *1º Bloco* | *2º Bloco* | *3º Bloco* | *4º Bloco* | *5º Bloco* |
|---|---|---|---|---|---|
| *A - 0% de FAD* | *6,5* | *6,4* | *6,2* | *5,8* | *7,3* |
| *B - 15% de FAD* | *7,1* | *7,4* | *6,9* | *7,3* | *7,0* |
| *C - 30% de FAD* | *7,5* | *8,1* | *6,7* | *7,4* | *7,7* |
| *D - 45% de FAD* | *8,4* | *8,5* | *8,7* | *8,3* | *7,9* |
| *E - 60% de FAD* | *9,3* | *9,9* | *9,5* | *8,5* | *8,9* |

Fazer a análise de variância e caso haja significância entre os tratamentos fazer a decomposição dos graus de liberdade dos tratamentos por meio da técnica dos polinômios ortogonais (regressão linear, quadrática, etc.). Ajuste a equação de regressão linear às médias dos tratamentos.

5 - No estudo do ganho de peso de porcos *guinea,* quatro dietas foram testadas. Vinte animais foram usados neste experimento, 5 animais para cada dieta. Entretanto o pesquisador acreditou que alguns fatores ambientais podem afetar o ganho de peso. Não foi possível reunir os 20 animais em uma mesma condição ambiental. Portanto, foram estabelecidos 5 blocos de unidades experimentais sob idênticas condições de temperatura, luz, etc.

| | *Dietas* | | | |
|---|---|---|---|---|
| *Blocos* | *1* | *2* | *3* | *4* |
| *1* | *7,0* | *5,3* | *4,9* | *8,8* |
| *2* | *9,9* | *5,7* | *7,6* | *8,9* |
| *3* | *8,5* | *4,7* | *5,5* | *8,1* |
| *4* | *5,1* | *3,5* | *2,8* | *3,3* |
| *5* | *10,3* | *7,7* | *8,4* | *9,1* |

a) Estabelecer as hipóteses estatísticas $H_0$ e $H_1$
b) Montar o quadro da análise de variância e testar as hipóteses do item a).
c) Fazer o gráfico de barras das médias dos tratamentos com o erro padrão.
d) Verificar, pelo teste de Tukey, se existem diferenças entre as médias das dietas. Qual foi a dieta que proporcionou o melhor ganho de peso?
e) Calcular o coeficiente de variação e de determinação do experimento $R^2$ do experimento.

6- Os resultados apresentados pelo programa R a uma análise de dados de um experimento foram:

*Response: dados*

| | *Df* | *Sum Sq* | *Mean Sq* | *F value* | *Pr(>F)* |
|---|---|---|---|---|---|
| *blocos* | *3* | *37.35* | *12.45* | *2.7978* | *0.08549 .* |
| *tratamentos* | *4* | *2530.20* | *632.55* | *142.1461* | *5.361e-10 ****** |
| *Residuals* | *12* | *53.40* | *4.45* | | |

*---*
*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

a) Interprete estes resultados
b) As médias dos tratamentos são apresentadas abaixo:

| *Trat A* | *Trat B* | *Trat C* | *Trat D* | *Trat E* |
|---|---|---|---|---|
| *12.75* | *23.50* | *37.00* | *45.00* | *34.50* |

Analisando os resultados da saída do Teste de Tukey preencha a tabela abaixo com as médias seguidas das letras.

*$tratamentos*

| | *diff* | *lwr* | *upr* | *p adj* |
|---|---|---|---|---|
| *Trat B-Trat A* | *10.75* | *5.995488* | *15.504512* | *0.0000864* |
| *Trat C-Trat A* | *24.25* | *19.495488* | *29.004512* | *0.0000000* |
| *Trat D-Trat A* | *32.25* | *27.495488* | *37.004512* | *0.0000000* |
| *Trat E-Trat A* | *21.75* | *16.995488* | *26.504512* | *0.0000000* |
| *Trat C-Trat B* | *13.50* | *8.745488* | *18.254512* | *0.0000085* |
| *Trat D-Trat B* | *21.50* | *16.745488* | *26.254512* | *0.0000001* |
| *Trat E-Trat B* | *11.00* | *6.245488* | *15.754512* | *0.0000689* |
| *Trat D-Trat C* | *8.00* | *3.245488* | *12.754512* | *0.0012997* |
| *Trat E-Trat C* | *-2.50* | *-7.254512* | *2.254512* | *0.4820549* |
| *Trat E-Trat D* | *-10.50* | *-15.254512* | *-5.745488* | *0.0001086* |

| ***Tratamentos*** | ***Médias*** |
|---|---|
| | |
| | |
| | |
| | |
| | |