Inferência Estatística

Estatística

Engenharias
2013/14

Rui Santos e Helena Ribeiro
Escola Superior de Tecnologia e Gestão
Instituto Politécnico de Leiria

- **Estatística Descritiva** — recolha, apresentação, análise e interpretação de dados:
  - Quadros de frequências;
  - Representação gráfica;
  - Redução dos dados.
- **Estatística Indutiva ou inferência estatística** — técnicas que permitem, a partir da informação contida na amostra, tirar conclusões sobre características desconhecidas da população.
  - Estimação pontual e por intervalos;
  - Testes de hipóteses paramétricos e não paramétricos.

- **População** (representada pela variável aleatória X) é caracterizada por uma determinada distribuição de probabilidade.

- **Amostra aleatória simples** $(X_1, X_2, \ldots, X_n)$ – todos os elementos da população têm igual probabilidade de pertencerem à amostra.
    - $X_1, X_2, \ldots, X_n$ são n variáveis aleatórias independentes;
    - $X_i,\ i = 1, \ldots, n$, têm a mesma distribuição que a população.

  Assim sendo $X_i,\ i = 1, \ldots, n$, são n variáveis aleatórias independentes e identicamente distribuídas a X.

- **Amostra concreta** $(x_1, x_2, \ldots, x_n)$ – corresponde à realização das variáveis aleatórias $(X_1, X_2, \ldots, X_n)$.

- **Parâmetros** – características da população que, embora sejam habitualmente desconhecidas, são consideradas fixas.

$$\mu \Rightarrow \text{média da população;}$$
$$\sigma^2 \Rightarrow \text{variância da população;}$$
$$p \Rightarrow \text{proporção de sucessos na população (Bernoulli).}$$

- **Estatísticas** – variáveis aleatórias obtidas através de uma função de uma amostra aleatória $(X_1, X_2, \cdots, X_n)$ e que não dependem de nenhum parâmetro desconhecido: $T\,(X_1, X_2, \cdots, X_n)$.

$$\overline{X} \Rightarrow \text{média da amostra;}$$
$$S^2 \Rightarrow \text{variância da amostra;}$$
$$\widehat{p} \Rightarrow \text{proporção de sucessos numa amostra (Bernoulli).}$$

- O objetivo da **estimação pontual** consiste em obter uma estatística [função da amostra $T(X_1, X_2, \cdots, X_n)$] que "melhor" aproxima o valor do parâmetro desconhecido θ.

- Um **estimador** de θ, representado por $\widehat{\theta}(X_1, X_2, \cdots, X_n)$ ou simplesmente por $\widehat{\theta}$, é uma estatística que usa a informação contida na amostra com o objetivo de estimar o valor de parâmetros desconhecidos da população.

- Uma **estimativa** $\widehat{\theta}(x_1, x_2, \cdots, x_n)$, $\widehat{\theta}_{obs}$ ou simplesmente $\widehat{\theta}$, é o valor assumido por um estimador numa amostra concreta.

- Os métodos mais comuns de obtenção de estimadores são o **método dos momentos** e o **método da máxima verosimilhança**.

- $\widehat{\theta}$ é um estimador **centrado** (ou não enviesado) de θ se:
$$\mathbb{E}\left(\widehat{\theta}\right) = \theta.$$
- Sejam $\widehat{\theta_1}$ e $\widehat{\theta_2}$ dois estimadores centrados para θ.
$\widehat{\theta_1}$ **é mais eficiente do que** $\widehat{\theta_2}$ se:
$$\text{Var}\left(\widehat{\theta_1}\right) < \text{Var}\left(\widehat{\theta_2}\right).$$
- $\widehat{\theta}$ é um estimador de θ **consistente** se:
$$\lim_{n\to\infty} \text{P}\left(\left|\widehat{\theta} - \theta\right| < \varepsilon\right) = 1, \ \forall \varepsilon > 0.$$
Condições suficientes para um estimador de θ ser consistente:
$$\lim_{n\to\infty} \mathbb{E}\left(\widehat{\theta}\right) = \theta \quad \text{e} \quad \lim_{n\to\infty} \text{Var}\left(\widehat{\theta}\right) = 0.$$

Propriedades da média amostral $\overline{X}$ para estimar a média da população μ:

- $\mathbb{E}\left(\overline{X}\right) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \mu$;
- $\text{Var}\left(\overline{X}\right) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{\sigma^2}{n} \qquad$ (e $\lim_{n\to\infty} \text{Var}\left(\overline{X}\right) = 0$).

Assim, $\widehat{\mu} = \overline{X}$ é um estimador centrado e consistente para μ (logo $\widehat{p}$ é também um estimador centrado e consistente para p).

Nota: $\widehat{\sigma^2} = S^2$ é um estimador centrado e consistente para σ^2.

Uma vez que $\overline{X}$ é uma v.a., podemos utilizar a sua distribuição para determinar probabilidades recorrendo, por exemplo, a:

- Se a população for normal então $\overline{X} \sim \text{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ pela Estabilidade da Lei Normal; caso contrário, se a amostra for grande, então $\overline{X} \overset{\bullet}{\sim} \text{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ pelo Teo. Limite Central.

- A **estimação pontual** utiliza estimadores para fornecer um valor (estimativa) para um parâmetro desconhecido θ. No entanto, **não é avaliada a precisão da estimativa**.

- Esta precisão pode ser avaliada utilizando a teoria da estimação por intervalos, pois, neste caso, em vez de se indicar um valor concreto para o parâmetro desconhecido θ, **constrói-se um intervalo que, com determinada "probabilidade" previamente definida, contém o verdadeiro valor do parâmetro θ.**

Intervalo aleatório para θ — Sejam T_1, T_2, T_3 e T_4 estatísticas (funções da amostra).

- $]T_1, T_2[$ é um **intervalo aleatório bilateral** para θ com probabilidade $1-\alpha$ se:

$$\mathrm{P}\left[T_1\left(X_1, \cdots, X_n\right) < \theta < T_2\left(X_1, \cdots, X_n\right)\right] = 1-\alpha.$$

- $]-\infty, T_3[$ é um **intervalo aleatório unilateral inferior** para θ com probabilidade $1-\alpha$ se:

$$\mathrm{P}\left[\theta < T_3\left(X_1, \cdots, X_n\right)\right] = 1-\alpha.$$

- $]T_4, +\infty[$ é um **intervalo aleatório unilateral superior** para θ com probabilidade $1-\alpha$ se:

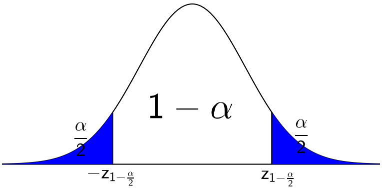
$$\mathrm{P}\left[\theta > T_4\left(X_1, \cdots, X_n\right)\right] = 1-\alpha.$$

Exemplo Intervalo aleatório para μ numa população normal com σ conhecido. Tendo em conta que $\overline{X} \sim \mathrm{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ então:

$$Z = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathrm{N}\left(0, 1\right).$$

Um intervalo para Z, com probabilidade $1-\alpha$, é dado por:

$$\mathrm{P}\left(-z_{1-\frac{\alpha}{2}} < \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < z_{1-\frac{\alpha}{2}}\right) = 1-\alpha,$$



onde $z_{1-\frac{\alpha}{2}}$ representa o quantil $1-\frac{\alpha}{2}$ da v.a. Z.
Esta equação pode ser resolvida em ordem a μ, obtendo-se:

$$\mathrm{P}\left(\overline{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha.$$

Desta forma concluímos que $\mu \in \left]\overline{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right[$ com probabilidade $1-\alpha$.

Exemplo Intervalo aleatório para μ numa população normal com σ desconhecido. Neste caso, pelas propriedades das distribuições, pode-se demonstrar que:

$$T = \frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t_{(n-1)}.$$

Um intervalo aleatório para T com probabilidade $1-\alpha$ é dado por:

$$\mathrm{P}\left(-t_{1-\frac{\alpha}{2}} < \frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} < t_{1-\frac{\alpha}{2}}\right) = 1-\alpha,$$

onde $t_{1-\frac{\alpha}{2}}$ representa o quantil $1-\frac{\alpha}{2}$ da v.a. T.

Esta equação pode ser resolvida em ordem a μ, obtendo-se:

$$\mathrm{P}\left(\overline{X} - t_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right) = 1-\alpha.$$

Portanto, $\mu \in \left]\overline{X} - t_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}, \overline{X} + t_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right[$ com probabilidade $1-\alpha$.

Intervalo de confiança para θ — Se num intervalo aleatório para θ com probabilidade $1-\alpha$ substituir-se nas funções $T_i\left(X_1,\cdots,X_n\right), i=1,2,3,4$, as variáveis aleatórias por valores de uma amostra concreta, obtêm-se os intervalos com $(1-\alpha)\times 100$ por cento de confiança para θ.

- $\left]T_1\left(x_1,\cdots,x_n\right), T_2\left(x_1,\cdots,x_n\right)\right[\;\rightarrow$ **intervalo bilateral** com $(1-\alpha)\times 100\%$ de confiança para θ.
- $\left]-\infty, T_3\left(x_1,\cdots,x_n\right)\right[\;\rightarrow$ **intervalo unilateral inferior** com $(1-\alpha)\times 100\%$ de confiança para θ.
- $\left]T_4\left(x_1,\cdots,x_n\right), +\infty\right[\;\rightarrow$ **intervalo unilateral superior** com $(1-\alpha)\times 100\%$ de confiança para θ.

Interpretação de um intervalo de confiança

- Um intervalo com 95 por cento de confiança significa que, quando se concretiza os intervalos para um grande número de amostras (passando de intervalo aleatório para intervalo de confiança ao substituir por valores concretos de cada amostra), **os intervalos resultantes irão conter o verdadeiro valor em 95 por cento da vezes**.
- Se aplicar este método infinitas vezes, em **95 por cento dos intervalos de confiança obtidos o verdadeiro valor do parâmetro estará lá contido**.

Propriedades de um intervalo de confiança

- Quanto maior for o nível de confiança maior será a amplitude do intervalo;
- Quanto maior for a dimensão da amostra menor será a amplitude do intervalo;
- Quanto maior for a variância da população maior será a amplitude do intervalo.

Nota As amplitudes dos intervalos de confiança deduzidos nas páginas 10 e 11 são dadas respetivamente por:

$$\mathrm{I}_1 = 2z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \quad \text{e} \quad \mathrm{I}_2 = 2t_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}.$$

- **Hipóteses**: população com distribuição normal.[1]
- R: Statistics → Means → Single-sample t-test → escolher a variável, o nível de significância e o tipo de intervalo ...

```
                One Sample t-test
data: DRH_F_Operacional$Altura
t = 289.7111, df = 91, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
    158.2769 160.4623
sample estimates:
mean of x
    159.3696
```

Logo $\widehat{\mu} = 159.3696$ e $\mu \in]158.2769, 160.4623[$ com 95% de confiança (o resultado corresponde à aplicação da fórmula apresentada na pág. 11).

[1] Deve ser efetuado previamente um teste de normalidade (cf. pág. 32).

- **Hipóteses**: população de Bernoulli.
- R: Statistics → Proportions → Single-sample proportion test → escolher a variável binária, o nível de significância, o tipo de intervalo e o tipo de teste (exato ou aproximado).[2]

Exact binomial test

data: rbind(.Table)
number of successes = 86, number of trials = 92, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
99 percent confidence interval:
0.8385484 0.9829707
sample estimates:
probability of success
0.9347826

Logo $\widehat{p} = 0.9347826$ e $p \in]0.8385484, 0.9829707[$ com 99% de confiança.

[2] O IC calculado corresponde à primeira categoria, para alterar utilizar: Data → Manage Variable in Active data set → Reorder variable factors.

- Confrontar duas hipóteses acerca de uma característica da população:
  - **Hipótese nula** — H_0;
  - **Hipótese alternativa** — H_1.
- Vamos optar por uma das hipóteses, sendo esta decisão tomada de acordo com a **informação contida na amostra**.
- Existem dois tipos de erros que podem ocorrer:
  - **Erro do Tipo I** — Rejeitar H_0 quando H_0 é verdadeira;
  - **Erro do Tipo II** — Não Rejeitar H_0 quando H_0 é falsa.

- **Hipótese nula** — H_0, é uma **hipótese simples**, e.g. $\mathrm{H}_0 : \theta = \theta_0$ onde θ representa o parâmetro em análise e θ_0 um valor particular desse parâmetro ou H_0 : a população é caracterizada pela distribuição normal (i.e., $X \sim \mathrm{N}(\mu, \sigma)$).

- **Hipótese alternativa** — H_1, é uma **hipótese composta**, ou seja, nela é especificada mais do que um valor para o parâmetro/distribuição, por exemplo:

$\mathrm{H}_1 : \theta \neq \theta_0 \quad \longrightarrow$ hipótese alternativa **bilateral**;

$\mathrm{H}_1 : \theta > \theta_0 \quad \longrightarrow$ hipótese alternativa **unilateral superior**;

$\mathrm{H}_1 : \theta < \theta_0 \quad \longrightarrow$ hipótese alternativa **unilateral inferior**;

ou H_1 : a população não é caracterizada pela distribuição normal.

Exemplo: A empresa *DelFonte* comercializa garrafas de água de 1500 mililitros. As garrafas são enchidas, através de um processo automático, com uma quantidade de água com valor médio igual a μ mililitros.

Observem-se três situações distintas para as hipóteses a testar:

- Considere-se que a empresa pretende testar se o processo de enchimento das garrafas está a funcionar devidamente, ou seja, se de facto o processo enche as garrafas, em média, com 1500 mililitros. Neste caso as hipóteses a testar seriam:

$$\mathrm{H}_0 : \mu = 1500 \quad \text{versus} \quad \mathrm{H}_1 : \mu \neq 1500.$$

- Considere-se que a empresa pretende controlar os custos do processo produtivo. Assim, para averiguar se o processo de enchimento está a encher as garrafas com uma quantidade superior à devida, dever-se-ía testar:

$$\mathrm{H}_0 : \mu = 1500 \quad \text{versus} \quad \mathrm{H}_1 : \mu > 1500.$$

- Considere-se que um conjunto de consumidores afirmam que as garrafas de água DelFonte possuem menos quantidade de água que a devida. Neste caso, as hipóteses a testar seriam:

$$\mathrm{H}_0 : \mu = 1500 \quad \text{versus} \quad \mathrm{H}_1 : \mu < 1500.$$

Decisão	**Situação** H_0 é verdadeira	**Situação** H_0 é falsa
Rejeitar H_0	Erro do tipo I $P\left(\text{Erro do tipo I}\right) = \alpha$	Decisão correta
Não Rejeitar H_0	Decisão correta	Erro do tipo II $P\left(\text{Erro do tipo II}\right) = \beta$

- O **erro do tipo I** ou **erro de primeira espécie** é o erro que se comete quando se rejeita a hipótese nula $\left(H_0\right)$ e esta é verdadeira.
- Denomina-se por **nível de significância** α a probabilidade de se cometer um erro do tipo I, ou seja,

$$\alpha = P\left(\text{rejeitar } H_0 | H_0 \text{ verdadeira}\right).$$

Decisão	**Situação** H_0 é verdadeira	 H_0 é falsa
Rejeitar H_0	Erro do tipo I $P\left(\text{Erro do tipo I}\right) = \alpha$	Decisão correta
Não Rejeitar H_0	Decisão correta	Erro do tipo II $P\left(\text{Erro do tipo II}\right) = \beta$

- O **erro do tipo II** ou **erro de segunda espécie** é o erro que se comete quando não se rejeita a hipótese nula (H_0) e esta é falsa.
- Representa-se por β a probabilidade de se cometer um erro do tipo II, ou seja, $\beta = P\left(\text{não rejeitar } H_0 | H_0 \text{ falsa}\right)$.
- Denomina-se por **potência do teste** a probabilidade de não se cometer um erro do tipo II, isto é, $1 - \beta$.

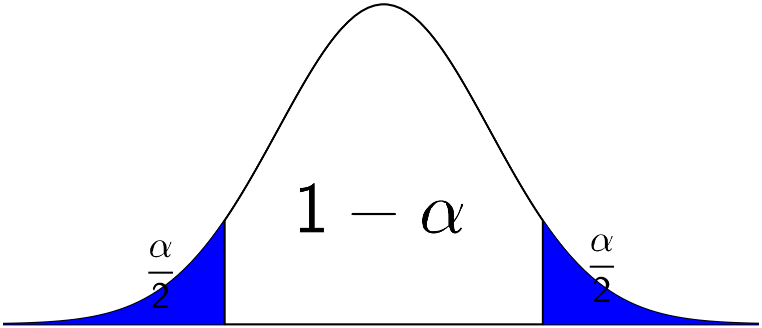
1. Se **diminuirmos** a probabilidade de um tipo de erro a probabilidade do outro tipo de erro **aumenta**;
2. Pode-se controlar a probabilidade de ocorrência de um erro da primeira espécie (α - nível de significância) mas não se consegue controlar a probabilidade de ocorrência de um erro da segunda espécie (β);
3. Os testes de hipóteses são feitos considerando a **probabilidade de um erro do tipo I fixa** (**nível de significância** – α), logo H_0 é protegida (p. ex. $\alpha = 0.05$).
4. Como H_0 é protegida devem-se usar as expressões "rejeitar H_0" e "não rejeitar H_0" em vez de "aceitar H_1" ou "aceitar H_0".
5. Existem duas formas de fazer um teste de hipótese: **Região crítica** ou **Valor de prova** (**p-value**).

- A região na qual a decisão é rejeitar H_0 denomina-se por **região crítica**, sendo os valores limites da região crítica denominados por **valores críticos**.
- Quando se pretende testar se a média de uma população é igual a μ_0 ($\mathrm{H}_0 : \mu = \mu_0$), como o parâmetro em análise é a média da população, o estimador que será utilizado no teste será a média da amostra, medindo-se o seu afastamento relativamente a μ_0.
- O afastamento máximo (para não rejeitar H_0) será definido de forma a garantir que a probabilidade de erro do tipo I (nível de significância) seja igual ao valor α definido, isto é,

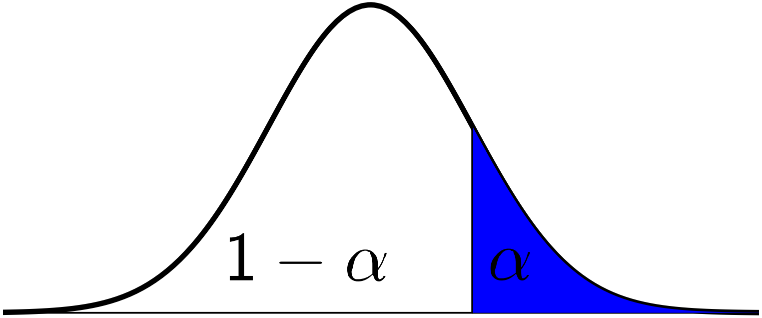
$$\alpha = \mathrm{P}\left(\text{rejeitar } \mathrm{H}_0 | \mathrm{H}_0 \text{ verdadeira}\right).$$

- A região crítica é definida pela hipótese alternativa.

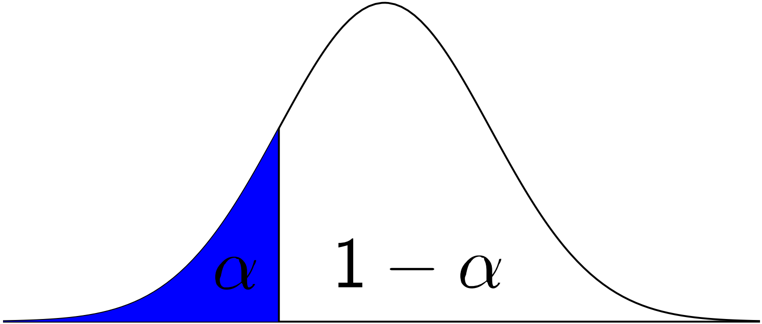
- No caso de a hipótese alternativa consistir na média da população ser diferente de μ_0 ($\mathrm{H}_1 : \mu \neq \mu_0$), deve-se rejeitar H_0 se o valor da média amostral for suficientemente distante de μ_0 (quer seja inferior ou superior).



- No caso de a hipótese alternativa ser a média superior a μ_0 ($H_1 : \mu > \mu_0$) só se deve rejeitar H_0 se o valor da média amostral for suficientemente distante e superior a μ_0.



- Pelo mesmo raciocínio, se a hipótese alternativa for a média ser inferior a μ_0 ($H_1 : \mu < \mu_0$) deve-se rejeitar H_0 nos casos em que o valor da média amostral seja suficientemente inferior a μ_0.



- Na prática a utilização da região crítica corresponde a construir um intervalo de confiança para realizar o teste de hipóteses: se μ_0 pertencer ao intervalo de confiança então não rejeitamos H_0, caso contrário rejeitamos H_0. O tipo de intervalo utilizado depende da hipótese alternativa.

  - Se $H_1 : \mu \neq \mu_0$, então rejeita-se H_0 se μ_0 não pertencer ao intervalo de confiança bilateral.

  - Se $H_1 : \mu > \mu_0$, então rejeita-se H_0 se μ_0 não pertencer ao intervalo de confiança unilateral superior.

  - Se $H_1 : \mu < \mu_0$, então rejeita-se H_0 se μ_0 não pertencer ao intervalo de confiança unilateral inferior.

- O ***p*-value** (**valor p** ou **valor de prova**) é a probabilidade de observar uma amostra mais desfavorável para a hipótese nula (H_0) do que aquela que foi observada, considerando que a hipótese nula é verdadeira.

  - Nos casos em que o *p-value* assume um valor pequeno significa que a probabilidade de haver uma amostra mais desfavorável que a observada, sob a hipótese de H_0 ser verdadeira, é pequena, logo deve-se rejeitar H_0.

  - A definição de uma probabilidade pequena para rejeitar H_0 é feita pelo nível de significância, logo

$$p\text{-}value < \alpha \quad \Longrightarrow \quad \text{Rejeitar } H_0$$

$$p\text{-}value \geq \alpha \quad \Longrightarrow \quad \text{Não rejeitar } H_0$$

- Num tribunal um arguido é considerado inocente até prova de contrário, logo a inocência do arguido é protegida (num teste de hipóteses consideramos que H_0 é verdadeira ao longo de todo o processo, logo H_0 **é protegida**).

  - Se um arguido é condenado significa que existem provas suficientes para o condenar (se H_0 **é rejeitada** significa que a **amostra contém suficiente informação para a rejeitar**).

  - Se um arguido não é condenado, não significa necessariamente que seja inocente mas antes que não foram apresentadas provas suficientes para o condenar (se H_0 **não é rejeitada** não implica que esta seja verdadeira, mas antes que a **amostra utilizada não contém informação suficiente para a rejeitar**).

Testes paramétricos *versus* testes não paramétricos

- Testes **paramétricos** — supõem que a distribuição da população é conhecida (p. ex. a população ser caracterizada pela distribuição normal) e as hipóteses a testar referem-se a valores para os parâmetros.

- Testes **não paramétricos** — não impõem qualquer restrição à distribuição da população e as hipóteses a testar podem referir-se a valores para os parâmetros ou à própria distribuição que caracteriza a população.

Teste de normalidade de Shapiro-Wilks

- H_0 : a população é caracterizada pela distribuição normal (i.e. $X \sim N(\mu, \sigma)$)

  versus

  H_1 : a população não é caracterizada pela distribuição normal

- **Hipóteses do teste**: Nenhuma.

- R: Statistics → Summaries → Shapiro-Wilks test of normality… escolher a variável.

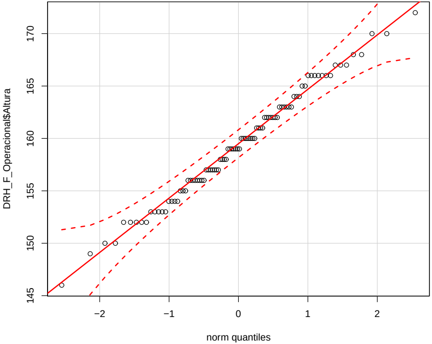
- Pode ser representado o gráfico Q-Q (quantil-quantil), que compara os quantis da distribuição normal com os da amostra. No R: Graphs → Quantile-comparison plot… escolher a variável e a distribuição normal.

Shapiro-Wilk normality test

data: DRH_F_Operacional$Altura

W = 0.9909, p-value = 0.784

Como p-value $= 0.784 \geq \alpha = 0.05 \Rightarrow$ não rejeitar H_0, logo não há evidência estatística de que a população não seja caracterizada pela distribuição normal (logo, podemos supor $X \sim \mathrm{N}(\mu, \sigma)$).

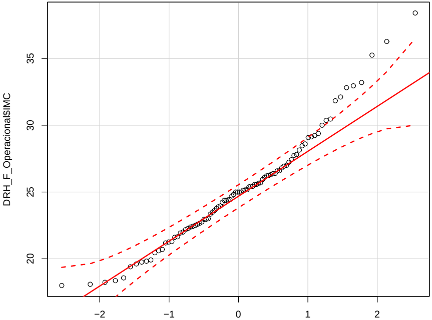


Shapiro-Wilk normality test

data: DRH_F_Operacional$IMC

W = 0.9635, p-value = 0.01122

Como p-value $= 0.01122 < \alpha = 0.05 \Rightarrow$ rejeitar H_0, logo há evidência estatística de que a população não é caracterizada pela distribuição normal (logo, os resultados obtidos em testes de hipóteses que supõem distribuição normal não são tão fiáveis).

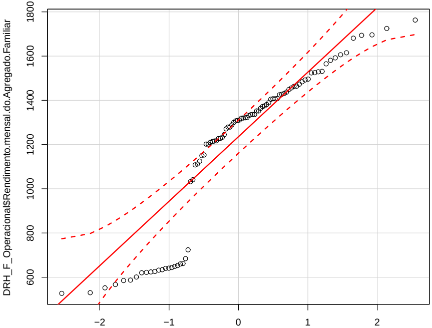


Shapiro-Wilk normality test

data: DRH_F_Operacional$Rendimento.mensal.do.Agregado.Familiar

W = 0.8728, p-value = 2.328e-07

Como p-value $= 2.328 \times 10^{-7} < \alpha = 0.05 \Rightarrow$ rejeitar H_0, logo há evidência estatística de que a população não é caracterizada pela distribuição normal (logo, os resultados obtidos em testes de hipóteses que supõem distribuição normal não são tão fiáveis).



Uma amostra — teste para a média da população

- $\mathrm{H}_0 : \mu = \mu_0$ *versus* $\mathrm{H}_1 : \begin{cases} \mu \neq \mu_0 \\ \mu < \mu_0 \\ \mu > \mu_0 \end{cases}$

- **Hipóteses do teste**: população normal.[3]

- R: Statistics → Means → Single-sample t-test → ... escolher a variável e a forma da hipótese alternativa.

[3] Deve ser efetuado previamente um teste de normalidade (cf. pág. 32).

$$H_0 : \mu = 160 \textit{ versus } H_1 : \mu \neq 160$$

One Sample t-test
data: DRH_F_Operacional$Altura
t = -1.146, df = 91, p-value = 0.2548
alternative hypothesis: true mean is not equal to 160
95 percent confidence interval:
158.2769 160.4623
sample estimates:
mean of x
159.3696

Como p-value $= 0.2548 > \alpha = 0.05 \Rightarrow$ não rejeitar H_0, logo não há evidência estatística de que a média seja distinta de 160.

Duas amostras independentes[4] – comparação de médias

- $\text{H}_0 : \mu_1 = \mu_2$ *versus* $\text{H}_1 : \begin{cases} \mu_1 \neq \mu_2 & & \mu_1 - \mu_2 \neq 0 \\ \mu_1 < \mu_2 & \text{ou} & \mu_1 - \mu_2 < 0 \\ \mu_1 > \mu_2 & & \mu_1 - \mu_2 > 0 \end{cases}$
- **Hipóteses do teste**: amostras independentes, populações normais.[5]
- R: Statistics → Means → Independent samples t-test → … escolher a variável em teste e a variável que define os grupos, a forma da hipótese alternativa e indicar se o teste deve considerar variâncias iguais ou distintas.[6]

[4] Duas amostras referentes a duas populações disjuntas.

[5] Deve ser efetuado previamente um teste de normalidade (cf. pág. 32).

[6] Deve ser efetuado previamente um teste de igualdade de variâncias (cf. pág. 53).

$$H_0 : \mu_1 - \mu_2 = 0 \textit{ versus } H_1 : \mu_1 - \mu_2 \neq 0$$

Two Sample t-test data: Altura by Mais43
t = 0.2433, df = 90, p-value = 0.8083
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-1.939531 2.480899
sample estimates:
mean in group 0 mean in group 1
159.4902 159.2195

Como p-value $= 0.8083 > \alpha = 0.05 \Rightarrow$ não rejeitar H_0, logo não há evidência estatística de que as médias sejam distintas.

Duas amostras emparelhadas[7] — comparação de médias

- $\mathrm{H}_0 : \mu_1 = \mu_2$ *versus* $\mathrm{H}_1 : \begin{cases} \mu_1 \neq \mu_2 & & \mu_1 - \mu_2 \neq 0 \\ \mu_1 < \mu_2 & \text{ou} & \mu_1 - \mu_2 < 0 \\ \mu_1 > \mu_2 & & \mu_1 - \mu_2 > 0 \end{cases}$
- **Hipóteses do teste**: Amostras emparelhadas, população normal.[8]
- R: Statistics → Means → Paired Samples t-test → … escolher a variável em teste e a variável que define os grupos e a forma da hipótese alternativa.

[7] Nas amostras emparelhadas é observada a mesma característica (variável) nos mesmos indivíduos em duas ocasiões distintas.

[8] Deve ser efetuado previamente um teste de normalidade (cf. pág. 32).

$$H_0 : \mu_1 - \mu_2 = 0 \textit{ versus } H_1 : \mu_1 - \mu_2 > 0$$

Paired t-test

data: DRH_F_Operacional$Peso and DRH_F_Operacional$Peso2
t = 58.3956, df = 91, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
1.239011 Inf
sample estimates:
mean of the differences
1.275302

Como p-value $< 2.2 \times 10^{-16} < \alpha = 0.01 \Rightarrow$ rejeitar H_0, logo há evidência estatística de que a média μ_1 é superior à média μ_2.

ANOVA a um fator (One-way ANOVA) Comparação de k médias

- $\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ *versus* $\mathrm{H}_1 : \exists_{i \neq j} : \mu_i \neq \mu_j$
- **Hipóteses do teste**: k amostras independentes, população normal, variâncias iguais.[9]
- R: Statistics → Means → One-Way ANOVA → … escolher a variável em teste e a variável que define os grupos.
- Pode-se ainda utilizar o seguinte gráfico nesta análise: Graphs → Plot of means…

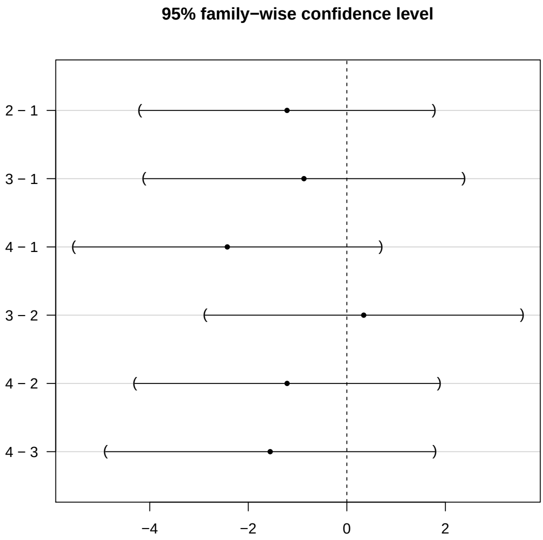
[9] Devem ser efetuados previamente o teste de normalidade (cf. pág. 32) e o teste de igualdade das variâncias (cf. pág. 55).

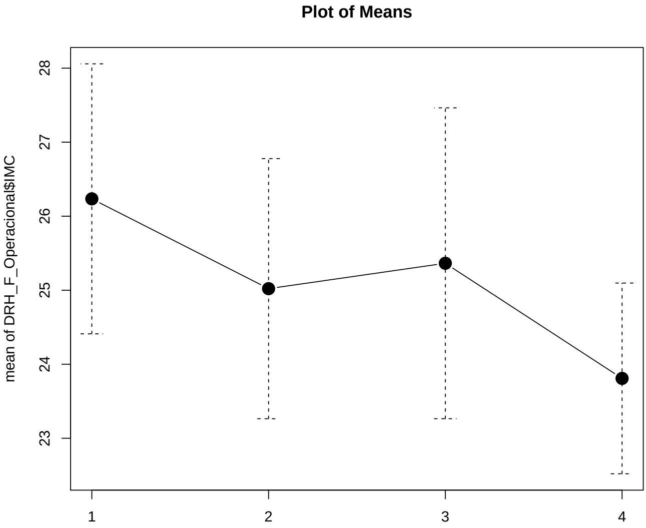
$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \ \textit{versus} \ H_1 : \exists_{i \neq j} : \mu_i \neq \mu_j$$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Grupos	3	70.3	23.42	1.411	0.245
Residuals	88	1460.6	16.60		

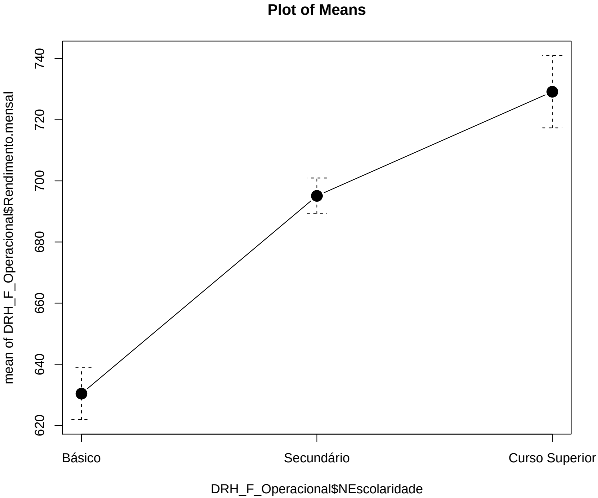
	mean	sd	%	data:n
1	26.23431	4.418623	0	25
2	25.02099	4.350515	0	26
3	25.36386	4.356370	0	19
4	23.80851	2.905211	0	22

Como p-value $= 0.254 > \alpha = 0.05 \Rightarrow$ não rejeitar H_0, logo não há evidência estatística de que as médias sejam distintas.


95% family−wise confidence level
2 − 1
3 − 1
4 − 1
3 − 2
4 − 2
4 − 3
−4
−2
0
2


Plot of Means
mean of DRH_F_Operacional$IMC
23
24
25
26
27
28
1
2
3
4

Exemplo com diferenças significativas na média dos grupos



Teste para uma proporção

- $H_0 : p = p_0$ *versus* $H_1 : \begin{cases} p \neq p_0 \\ p < p_0 \\ p > p_0 \end{cases}$
- **Hipóteses do teste**: população de Bernoulli.
- R: Statistics → Proportions → Single-sample proportion test → ... escolher a variável binária, a forma da hipótese alternativa e o tipo de teste (exato ou aproximado).
- A proporção testada corresponde à primeira categoria da variável. Para alterar utilizar R: Data → Manage Variable in Active data set → Reorder variable factors.

$$H_0 : p = 0.65 \textit{ versus } H_1 : p < 0.65$$

Exact binomial test

data: rbind(.Table)
number of successes = 51, number of trials = 92, p-value = 0.03641
alternative hypothesis: true probability of success is less than 0.65
95 percent confidence interval:
0.0000000 0.6426998
sample estimates:
probability of success
0.5543478

Como p-value $= 0.03641 < \alpha = 0.05 \Rightarrow$ rejeitar H_0, logo há evidência estatística de que a proporção é inferior a 0.65.

Comparação de duas proporções

- $\mathrm{H}_0 : p_1 = p_2$ *versus* $\mathrm{H}_1 : \begin{cases} p_1 \neq p_2 & & p_1 - p_2 \neq 0 \\ p_1 < p_2 & \text{ou} & p_1 - p_2 < 0 \\ p_1 > p_2 & & p_1 - p_2 > 0 \end{cases}$
- **Hipóteses do teste**: amostras independentes, populações de Bernoulli.
- R: Statistics → Proportions → Two-sample proportion test → ... escolher as duas variáveis binárias (a que define o grupo e a que define a proporção a testar), a forma da hipótese alternativa e o tipo de teste (tipo de aproximação).
- A proporção testada corresponde à primeira categoria da variável resposta. Para alterar utilizar R: Data → Manage Variable in Active data set → Reorder variable factors.

$$H_0 : p_1 = p_2 \ \textit{versus} \ H_1 : p_1 \neq p_2$$

2-sample test for equality of proportions with continuity correction
data: .Table
X-squared = 1.041, df = 1, p-value = 0.3076
alternative hypothesis: two.sided
95 percent confidence interval:
-0.09492297 0.34552077
sample estimates:
prop 1 prop 2
0.6862745 0.5609756

Como p-value $= 0.3076 \geq \alpha = 0.05 \Rightarrow$ não rejeitar H_0, logo não há evidência estatística de que as proporções sejam diferentes.

Teste do Qui-Quadrado para igualdade de k proporções

- $\mathrm{H}_0 : p_1 = \phi_1, p_2 = \phi_2, \cdots, p_k = \phi_k$
  (com $\phi_1 + \phi_2 + \cdots + \phi_k = 1$)

  versus

  H_1 : pelo menos uma das proporções não corresponde
- **Hipóteses do teste**: variável dividida em k categorias (com $k \geq 2$).
- Statistics → Summaries → Frequency Distribution→ escolher a opção "Chi-square goodness-of-fit test"... indicar as proporções pretendidas para cada categoria (valores de $\phi_1, \phi_2, \ldots, \phi_k$).

$$H_0 : p_1 = \frac{1}{4}, \quad p_2 = \frac{1}{4}, \quad p_3 = \frac{1}{4}, \quad p_4 = \frac{1}{4}$$

versus

$$H_1 : \text{Pelo menos umas das proporções é diferente de } \frac{1}{4}$$

1	2	3	4
27.17	28.26	20.65	23.91

Chi-squared test for given probabilities

data: .Table

X-squared = 1.3043, df = 3, p-value = 0.7281

Como p-value $= 0.7281 \geq \alpha = 0.05 \Rightarrow$ não rejeitar H_0, logo não há evidência estatística de que as proporções sejam diferentes das indicadas em H_0 (proporção igual a $\frac{1}{4}$ em cada um dos 4 grupos).

Teste F para comparar duas variâncias

- $\mathrm{H}_0 : \sigma_1 = \sigma_2$ *versus* $\mathrm{H}_1 : \begin{cases} \sigma_1 \neq \sigma_2 & & \dfrac{\sigma_1}{\sigma_2} \neq 1 \\ \sigma_1 < \sigma_2 & \text{ou} & \dfrac{\sigma_1}{\sigma_2} < 1 \\ \sigma_1 > \sigma_2 & & \dfrac{\sigma_1}{\sigma_2} > 1 \end{cases}$

- **Hipóteses do teste**: amostras independentes, população normal.[10]

- R: Statistics → Variances → Two-variances F-test... escolher a variável em teste, a variável que define os grupos e a forma da hipótese alternativa.

[10] Deve ser efetuado previamente o teste de normalidade (cf. pág. 32).

$$\mathrm{H}_0 : \sigma_1 = \sigma_2 \; \textit{versus} \; \mathrm{H}_1 : \sigma_1 < \sigma_2 \quad \left(\frac{\sigma_1}{\sigma_2} < 1\right)$$

```
            F test to compare two variances
data: Altura by Mais43
F = 0.9445, num df = 50, denom df = 40, p-value = 0.4205
alternative hypothesis: true ratio of variances is less than 1
95 percent confidence interval:
   0.000000    1.543024
sample estimates:
ratio of variances
   0.9445074
```

Como p-value $= 0.4205 \geq \alpha = 0.05 \Rightarrow$ não rejeitar H_0, logo não há evidência estatística de que a variância da primeira população seja inferior à variância da segunda população.

Teste de Bartlett — Igualdade de k variâncias ($k \geq 2$)

- $\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$ *versus* $\mathrm{H}_1 : \exists_{i \neq j} : \sigma_i^2 \neq \sigma_j^2$
- **Hipóteses do teste**: k amostras independentes, população normal.[11]
- R: Statistics → Variances → Bartlett's test... escolher a variável em teste, a variável que define os grupos e se o teste deve ser baseado na média ou na mediana.

[11] Deve ser efetuado previamente o teste de normalidade (cf. pág. 32).

$$\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$
$$\textit{versus}$$
$$\mathrm{H}_1 : \exists_{i \neq j} : \sigma_i^2 \neq \sigma_j^2$$

1	2	3	4
27.34000	26.81385	16.78363	40.41775

Bartlett test of homogeneity of variances
data: Altura by Grupos
Bartlett's K-squared = 3.5667, df = 3, p-value = 0.3122

Como p-value $= 0.3122 \geq \alpha = 0.05 \Rightarrow$ não rejeitar H_0, logo não há evidência estatística de que haja qualquer variância distinta das restantes.

Teste de Levene — Igualdade de k variâncias ($k \geq 2$)

- $\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$ *versus* $\mathrm{H}_1 : \exists_{i \neq j} : \sigma_i^2 \neq \sigma_j^2$
- **Hipóteses do teste**: k amostras independentes.
- R: Statistics → Variances → Levene's test... escolher a variável em teste, a variável que define os grupos e se o teste deve ser baseado na média ou na mediana.

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$
$$versus$$
$$H_1 : \exists_{i \neq j} : \sigma_i^2 \neq \sigma_j^2$$

1	2	3	4
15.506667	2.364615	1.251462	18.928571

Levene's Test for Homogeneity of Variance (center = median)

	Df	F value	Pr(>F)
group	3	3.3839	0.02167
	88		

Como p-value $= 0.02167 < \alpha = 0.05 \Rightarrow$ rejeitar H_0, logo há evidência estatística de que pelo menos uma das variâncias é distinta das restantes.

Quando se utilizam os testes não paramétricos de localização?

- quando a população não tem distribuição normal (dados rejeitados pelo teste de normalidade);
- quando a amostra é pequena (não sendo fácil analisar a sua normalidade);
- os dados estão definidos numa escala ordinal.

Nota 1: Os testes de localização não paramétricos baseiam-se, regra geral, na mediana (Me) enquanto que os paramétricos na média (μ).

Nota 2: Alguns dos testes apresentados anteriormente são também testes não paramétricos, contudo esta apresentação segue a ordem com que os comandos aparecem no R (p. ex. o teste de normalidade ou o teste de igualdade de variâncias de Levene são testes não paramétricos).

Duas amostras independentes
Comparação de medianas

- $\mathrm{H}_0 : \mathrm{Me}_1 = \mathrm{Me}_2$ *versus* $\mathrm{H}_1 : \begin{cases} \mathrm{Me}_1 \neq \mathrm{Me}_2 \\ \mathrm{Me}_1 < \mathrm{Me}_2 \\ \mathrm{Me}_1 > \mathrm{Me}_2 \end{cases}$

- **Hipóteses do teste**: amostras independentes.

- R: Statistics → Nonparametric Tests → Two-sample Wilcoxon Test… escolher a variável em teste, a variável que define os grupos, a forma da hipótese alternativa e o tipo de teste (exato ou aproximação).

$$H_0 : Me_1 = Me_2 \textit{ versus } H_1 : Me_1 \neq Me_2$$

0	1
25.42612	24.99893

Wilcoxon rank sum test
data: IMC by Mais43
W = 1188, p-value = 0.2661
alternative hypothesis: true location shift is not equal to 0

Como p-value $= 0.2661 \geq \alpha = 0.05 \Rightarrow$ não rejeitar H_0, logo não há evidência estatística de que as medianas sejam distintas

Duas amostras emparelhadas
Comparação de medianas

- $H_0 : Me_1 = Me_2$ *versus* $H_1 : \begin{cases} Me_1 \neq Me_2 \\ Me_1 < Me_2 \\ Me_1 > Me_2 \end{cases}$

- **Hipóteses do teste**: amostras emparelhadas.

- R: Statistics → Nonparametric Tests → Paired-samples Wilcoxon Test... escolher a variável em teste, a variável que define os grupos, a forma da hipótese alternativa e o tipo de teste (exato ou aproximação).

$$H_0 : \mathrm{Me}_1 = \mathrm{Me}_2 \;\textit{versus}\; H_1 : \mathrm{Me}_1 > \mathrm{Me}_2$$

Wilcoxon signed rank test with continuity correction

data: DRH_F_Operacional$Rendimento.mensal and
DRH_F_Operacional$Rendimento.mensal.há.dois.anos
V = 4278, p-value < 2.2e-16
alternative hypothesis: true location shift is greater than 0

Como p-value $< 2.2 \times 10^{-16} < \alpha = 0.05 \Rightarrow$ rejeitar H_0, logo há evidência estatística de que a mediana Me_1 é superior à mediana Me_2.

Comparação de k medianas em populações independentes

- $\mathrm{H_0} : \mathrm{Me}_1 = \mathrm{Me}_2 = \cdots = \mathrm{Me}_k$
  versus
  $\mathrm{H_1} : \exists_{i \neq j} : \mathrm{Me}_i \neq \mathrm{Me}_j$

- **Hipóteses do teste**: k amostras independentes.

- **R**: Statistics → Nonparametric Tests → Kruskal-Wallis test... escolher a variável em teste e a variável que define os grupos.

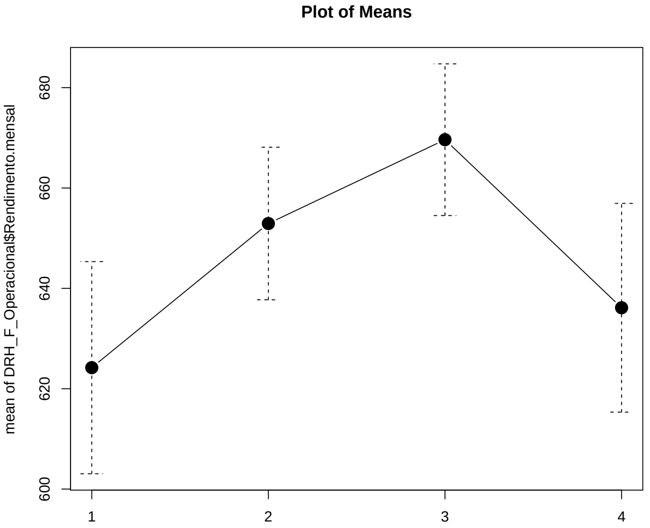
- Pode-se ainda utilizar o seguinte gráfico nesta análise (compara as médias entre as k populações):
  Graphs → Plot of means...

$$\mathrm{H}_0 : \mathrm{Me}_1 = \mathrm{Me}_2 = \mathrm{Me}_3 = \mathrm{Me}_4 \textit{ versus } \mathrm{H}_1 : \exists_{i \neq j} : \mathrm{Me}_i \neq \mathrm{Me}_j$$

1	2	3	4
632.0	647.5	672.0	636.0

```
Kruskal-Wallis rank sum test
data: Rendimento.mensal by Grupos
Kruskal-Wallis chi-squared = 12.7379,    df = 3,    p-value = 0.005239
```

Como p-value $= 0.005239 < \alpha = 0.05 \Rightarrow$ rejeitar H_0, logo há evidência estatística de que pelo menos uma das medianas é distinta das restantes.


Plot of Means
mean of DRH_F_Operacional$Rendimento.mensal
600
620
640
660
680
1
2
3
4