UNIVERSIDADE DA REGIÃO DE JOINVILLE - UNIVILLE

Bacharelado em Engenharia de Software (BES)

Estatística para computação

Professora Priscila Ferraz Franczak

Engenheira Ambiental - UNIVILLE

Mestre em Ciência e Engenharia de Materiais - UDESC

Doutora em Ciência e Engenharia de Materiais - UDESC

priscila.franczak@gmail.com

Plano de Aula

Testes Estatísticos

1. Introdução
2. Teste t
   - 2.1 Intervalo de confiança para média populacional.
   - 2.2 Intervalo de confiança para a diferença de médias de duas populações.
3. Teste F
4. Teste para a Normalidade
5. Exercícios

1. Introdução

- O R, em sua gama de utilidades, é urna poderosa ferramenta de análise estatística.

- Dentre os procedimentos incluídos no R podemos destacar os testes de média, amplamente usados em várias áreas do conhecimento.

- Ao tratarmos da análise de experimentos, é necessária a realização de testes estatísticos para verificação de determinadas hipóteses.

- Assim, é preciso que se faça uma breve revisão sobre alguns conceitos relacionados à inferência estatística, ou, mais especificamente, aos testes de hipóteses.

O teste de hipóteses é uma técnica estatística utilizada para avaliar alguma afirmação feita sobre uma população de interesse através de dados amostrais.

Exemplo: um engenheiro pode estar interessado em avaliar a hipótese de que o tempo de duração de um fusível seja de 1.000 horas, contra a hipótese de que tal valor seja diferente de 1.000 horas.

Essa seria uma afirmação sobre uma média, uma vez que a variável de interesse – tempo de duração – é quantitativa.

Nesse caso, o objetivo é testar se a hipotética média de 1.000 horas é verdadeira.

- No exemplo em questão, seria impraticável observar o tempo de duração de todos os fusíveis fabricados, ou seja, da população de interesse.

- De forma que é necessário a utilização de dados amostrais.

- O engenheiro poderia selecionar alguns fusíveis, calcular o valor da média e comparar com o valor proposto de 1.000 horas

A construção e o significado de uma hipótese estatística

- Uma hipótese estatística pode ser construída a partir de alguma teoria sobre determinado assunto, ou através de alguma afirmação sobre certo parâmetro da população em análise.

- No caso do engenheiro interessado em testar se o tempo médio de duração de um fusível é 1.000 horas, a hipótese não se deu através de uma teoria, mas possivelmente em função da experiência dele com o assunto.

Um teste estatístico tem como objetivo o fornecimento de evidências para subsidiar a decisão de rejeitar ou não rejeitar uma hipótese sobre algum parâmetro de uma população através de dados obtidos por uma amostra.

A afirmação sobre a média populacional é tida como a hipótese nula.

Damos o nome de hipótese alternativa à afirmação contrária à da hipótese nula.

Hipótese nula: Refere-se a uma afirmação do que queremos provar sobre algum parâmetro. Geralmente representada por Ho.

Hipótese alternativa: Refere-se a uma afirmação contrária ao que queremos provar. Geralmente representada por H_1 ou Ha.

Normalmente a hipótese nula é formulada sob a forma de uma igualdade, ou seja, é uma hipótese simples.

Exemplo: Um fabricante afirma que o tempo médio de secagem da tinta de sua marca é de 30 minutos.

Uma pessoa decide testar se essa afirmação é verdadeira.

Para isso, marca o tempo de secagem de 40 paredes e depois calcula a média. Quais seriam as hipóteses nula e alternativa?

A hipótese nula é o tempo de secagem, igual a 30 minutos.

A hipótese alternativa é o contrário (ou o complemento): o tempo de secagem é diferente de 30 minutos. As hipóteses são representadas da seguinte forma:

Ho: $\mu = 30$ minutos

H_1: $\mu \neq 30$ minutos

Além da definição acerca das hipóteses, o nível de significância também deve ser escolhido pelo analista.

Nível de significância: Consiste na probabilidade de rejeitar a hipótese nula, dado que ela é verdadeira.

Geralmente é representado pela letra grega alfa (α). O nível de significância também é conhecido como erro tipo I.

Exemplos:

Ho = a proporção de homens fumantes é igual à proporção de mulheres fumantes, na população de estudo.

H_1 = a proporção de homens fumantes é diferente da proporção de mulheres fumantes, na população de estudo.

Exemplos:

Ho = Em média, as vendas não aumentam com a introdução da propaganda.

H_1 = Em média, as vendas aumentam com a introdução da propaganda.

Qual seria o significado da expressão "... rejeitar a hipótese nula, dado que ela é verdadeira"?

- Nesse caso, pode ser obtida uma amostra muito ou pouco parecida com a população. Tanto no primeiro como no segundo caso existem probabilidades associadas.

- Existem chances de coletar uma amostra que dê evidências de que a hipótese seja rejeitada, mesmo quando, na verdade, a hipótese seja verdadeira.

O analista sempre corre o risco de tomar uma decisão equivocada no que se refere à rejeição ou não da hipótese nula, cabendo a ele escolher quanto risco aceita correr.

Esse risco é conhecido como nível de significância e geralmente é estipulado em 10%, 5% ou 1%.

Dessa forma, ao efetuar um teste de hipóteses com 5% de significância, podemos afirmar que exista 5% de probabilidade de rejeitar a hipótese nula, quando na verdade ela é verdadeira, ou seja, 5% de chance de cometer o erro tipo I.

É comum Ho ser apresentada em termos de igualdade de parâmetros populacionais, enquanto H_1 em forma de desigualdade (maior, menor ou diferente) → testes unilaterais e bilaterais

p-value (ou valor p)

Atualmente, em vez de fixar o nível de significância de um teste, usa-se o valor-p.

Compara-se o valor-p obtido para a amostra com o alfa fixado.

Rejeita-se Ho quando o valor-p for menor ou igual a alfa.

- Assim, usando o valor-p, o procedimento para o teste seria:

A) Formular Ho e H_1 (e definir alfa, se for de interesse);

B) Especificar a estatística do teste;

C) Determinar o valor da estatística do teste e o valor-p correspondente baseado na amostra;

D) Comparar o valor-p com alfa:

Se o valor-p for ≤ α, rejeita Ho

Se o valor-p for ≥ α, não rejeita Ho

Exemplo de testes de hipóteses:

- A distribuição da estatística de teste tende para o formato de uma distribuição normal quando o tamanho da amostra é relativamente grande (geralmente maior ou igual a 30).

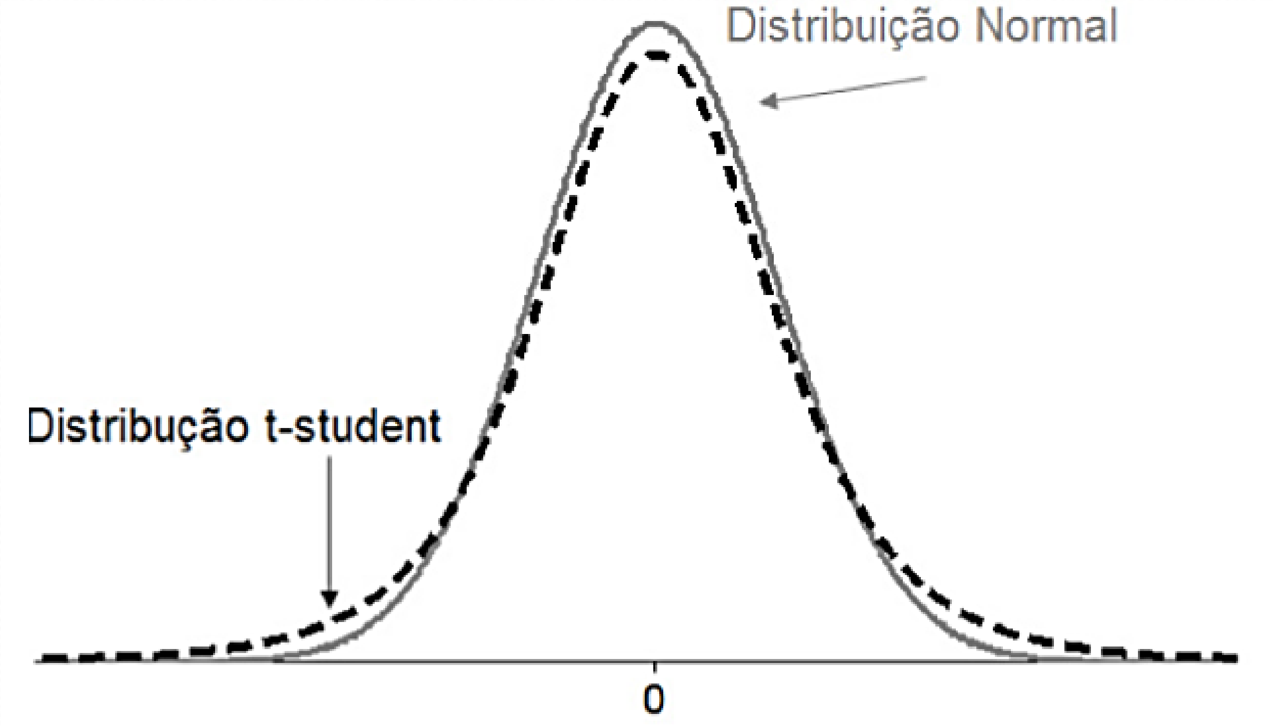
- Se o tamanho da amostra for pequeno (menor do que 30) e o desvio padrão for desconhecido, a distribuição da estatística de teste apresenta formato mais próximo da distribuição t de Student.

- Essa informação é importante porque definirá até que valor da estatística de teste a hipótese deve ser rejeitada.

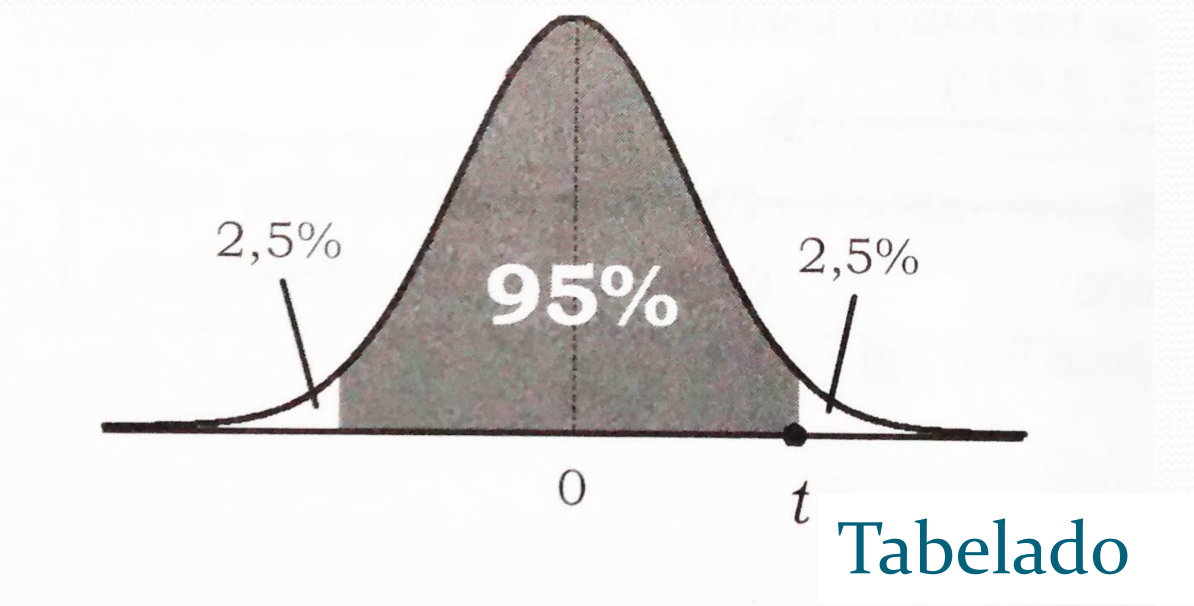
2. Teste t-Student

Essa distribuição é muito semelhante à distribuição normal: tem forma de sino, é simétrica e tem média zero.

- A diferença é que a distribuição t-student é mais achatada (tem caudas mais pesadas). Com isso, as estimativas obtidas a partir dessa distribuição serão menos precisas.



Para construção de intervalos de confiança, devemos olhar os valores para área em duas caudas e a área deve se referir ao valor de α/2 + α/2.



α = significância (5% no caso acima)

Os graus de liberdade são dados por n - 1, ou seja, o tamanho da amostra menos 1.

Exemplo: Encontrar o valor de t para um intervalo de 95% de confiança para uma amostra de 13 elementos.

α = 5% (0,05)

n = 13 elementos

Consultar tabela t

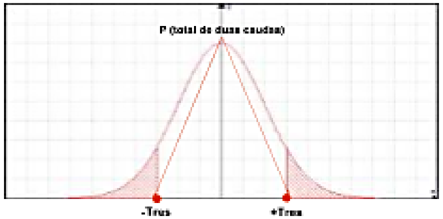
g.l. = n – 1

g.l. = 13 – 1

g.l. = 12

Tabela T: Distribuição de t-Student segundo os graus de liberdade e uma dada probabilidade num teste bicaudal (primeira linha)

Para um teste monocaudal, considere metade do valor de probabilidade apontado



Nº de graus de liberdade	Probabilidade para um teste bicaudal													
	0,95	0,90	0,80	0,70	0,60	0,50	0,40	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,0787	0,1584	0,3249	0,5095	0,7265	1,0000	1,3764	1,9626	3,0777	6,3138	12,7062	31,8205	63,657	636,619
2	0,0708	0,1421	0,2887	0,4447	0,6172	0,8165	1,0607	1,3862	1,8856	2,9200	4,3027	6,9646	9,9248	31,5991
3	0,0681	0,1366	0,2767	0,4242	0,5844	0,7649	0,9785	1,2498	1,6377	2,3534	3,1824	4,5407	5,8409	12,9240
4	0,0667	0,1338	0,2707	0,4142	0,5686	0,7407	0,9410	1,1896	1,5332	2,1318	2,7764	3,7469	4,6041	8,6103
5	0,0659	0,1322	0,2672	0,4082	0,5594	0,7267	0,9195	1,1558	1,4759	2,0150	2,5706	3,3649	4,0321	6,8688
6	0,0654	0,1311	0,2648	0,4043	0,5534	0,7176	0,9057	1,1342	1,4398	1,9432	2,4469	3,1427	3,7074	5,9588
7	0,0650	0,1303	0,2632	0,4015	0,5491	0,7111	0,8960	1,1192	1,4149	1,8946	2,3646	2,9980	3,4995	5,4079
8	0,0647	0,1297	0,2619	0,3995	0,5459	0,7064	0,8889	1,1081	1,3968	1,8595	2,3060	2,8965	3,3554	5,0413
9	0,0645	0,1293	0,2610	0,3979	0,5435	0,7027	0,8834	1,0997	1,3830	1,8331	2,2622	2,8214	3,2498	4,7809
10	0,0643	0,1289	0,2602	0,3966	0,5415	0,6998	0,8791	1,0931	1,3722	1,8125	2,2281	2,7638	3,1693	4,5869
11	0,0642	0,1286	0,2596	0,3956	0,5399	0,6974	0,8755	1,0877	1,3634	1,7959	2,2010	2,7181	3,1058	4,4370
12	0,0640	0,1283	0,2590	0,3947	0,5386	0,6955	0,8726	1,0832	1,3562	1,7823	$\boxed{2,1788}$	2,6810	3,0545	4,3178
13	0,0639	0,1281	0,2586	0,3940	0,5375	0,6938	0,8702	1,0795	1,3502	1,7709	2,1604	2,6503	3,0123	4,2208
14	0,0638	0,1280	0,2582	0,3933	0,5366	0,6924	0,8681	1,0763	1,3450	1,7613	2,1448	2,6245	2,9768	4,1405

α = 5% (0,05)

2.1 Intervalo de confiança para média populacional

Erro padrão para distribuição de médias usando a estatística t:

a) Para populações finitas:

$$\sigma_\mu = \frac{\sigma_S}{\sqrt{n-1}} \sqrt{\frac{N-n}{N-1}}$$

σ_S = erro padrão amostral

N = número total de elementos da população

n = número de elementos da amostra

Erro padrão para distribuição de médias usando a estatística t:

b) Para populações infinitas ou muito grandes:

$$\sigma_\mu = \frac{\sigma_S}{\sqrt{n-1}}$$

Erro admitido pelo pesquisador:

$$e = t * \sigma_\mu$$

Os valores padronizados de t são dados pela fórmula:

$$t = \frac{\mu_s - \mu}{\sigma_\mu}$$

μ_s = média amostral

μ = média populacional

O intervalo de confiança para verdadeira média populacional será dado por:

$$\mu_s - e < \mu < \mu_s + e$$

Ou

$$\mu_s \pm e$$

Exemplo: A altura de 17 crianças de um jardim de infância, escolhidos aleatoriamente, apresentou média igual a 107cm com desvio padrão de 10cm. Estabelecer o intervalo de 98% de confiança para a verdadeira altura média da população de crianças desse jardim.

$$\sigma_\mu = \frac{\sigma_S}{\sqrt{n-1}}$$

$$\sigma_\mu = \frac{10}{\sqrt{17-1}}$$

$$\sigma_\mu = 2{,}5$$

Dados:

n = 17 crianças

g.l.= n – 1 → 16

μ_S = 107cm

σ_S = 10cm

IC 98% → t = 2,58

$$e = t * \sigma_\mu$$

$$e = 2{,}58 * 2{,}5$$

$$e = 6{,}45$$

O intervalo de confiança para verdadeira média populacional será dado por:

$$\mu_s - e < \mu < \mu_s + e$$

$$107 - 6{,}45 < \mu < 107 + 6{,}45$$

$$100{,}55 < \mu < 113{,}45$$

A verdadeira altura média da população de crianças desse jardim é 107cm com um erro de 6,45cm para mais ou para menos, num intervalo de 98% de confiança.

2.2 Intervalo de confiança para a diferença de médias de duas populações

- Erro padrão:

$$\sigma_{\mu sA-\mu sB} = \sqrt{\frac{1}{n_A}+\frac{1}{n_B}} * \sqrt{\frac{n_A\sigma^2{}_{sA}+n_B\sigma^2{}_{sB}}{n_A+n_B-2}}$$

- Erro admitido pelo pesquisador:

$$e = t * \sigma_{\mu sA - \mu sB}$$

- Intervalo de confiança:

$$(\mu sA - \mu sB) - e < (\mu A - \mu B) < (\mu sA - \mu sB) + e$$

- Valores de t padronizados:

$$t = \frac{(\mu_{sA} - \mu_{sB})}{\sigma_{\mu sA - \mu sB}}$$

- Graus de liberdade:

$$g.l. = nA + nB - 2$$

Exemplo: A altura média de 16 crianças de um determinado jardim de infância apresenta média 107cm, com desvio padrão de 10cm, enquanto que a altura média de 14 crianças de um outro jardim de infância apresenta altura média 112cm, com desvio padrão de 8cm.

Estabelecer o intervalo de 95% de confiança para a verdadeira diferença entre as alturas média dos dois grupos.

Dados:

n_A = 16 crianças

μ_{sA} = 107cm

σ_{SA} = 10cm

n_B = 14 crianças

μ_{sB} = 112cm

σ_{SB} = 8cm

g.l.= nA + nB −2 → 28

IC 95% → t = 2,05

$$\sigma_{\mu sA-\mu sB} = \sqrt{\frac{1}{n_A}+\frac{1}{n_B}} * \sqrt{\frac{n_A\sigma^2{}_{sA}+n_B\sigma^2{}_{sB}}{n_A+n_B-2}}$$

$$\sigma_{\mu sA-\mu sB} = \sqrt{\frac{1}{16}+\frac{1}{14}} * \sqrt{\frac{16*10^2+14*8^2}{16+14-2}}$$

$\sigma_{\mu sA-\mu sB}$ = 3,455

e = t * $\sigma_{\mu sA - \mu sB}$

e = 2,05 * 3,455

e = 7,08

- Intervalo de confiança:

$$(\mu sB - \mu sA) - e < (\mu B - \mu A) < (\mu sB - \mu sA) + e$$

$$(112 - 107) - 7{,}08 < (\mu A - \mu B) < (112 - 107) + 7{,}08$$

$$-2{,}08 < (\mu A - \mu B) < 12{,}08$$

A diferença entre as alturas médias dos dois grupos é 5cm com um erro de 7,08cm para mais ou para menos, num intervalo de 95% de confiança.

- Resolvendo com o R:

Teste t para uma média:

- Vamos testar se x tem média estatisticamente igual a 35 ou maior:

$$H_o: \mu_x = 35$$
$$H_1: \mu_x > 35$$

```
x<-c(30.5,35.3,33.2,40.8,42.3,
     41.5,36.3,43.2,34.6,38.5)
t.test(x,  #amostra a ser testada
        mu=35,  #hipótese nula
        alternative = "greater") #unilateral à direita
```

```
        One Sample t-test

data:  x
t = 1.9323, df = 9, p-value = 0.04268
alternative hypothesis: true mean is greater than 35
95 percent confidence interval:
 35.13453      Inf
sample estimates:
mean of x
    37.62
```

- Um geólogo afirmou que a resistência média à compressão de uma rocha é de 285 Mpa. Desconfiado dessa afirmação, um estudante resolveu fazer um teste de resistência usando amostras provenientes da mesma região e encontrou os seguintes valores:

254.29, 165, 189.02, 277.46, 235.56, 198.32

Se o estudante realizou um teste bilateral (ele não sabe se a resistência é maior ou menor), para um nível de significância de 1%, a qual conclusão chegou?

$$H_o: \mu_x = 285\ MPa$$
$$H_1: \mu_x \neq 285\ MPa$$

```
amostra<-c(254.29, 165, 189.02, 277.46,
           235.56, 198.32)
t.test(amostra,     #teste t para "amostra"
       mu=285,      #valor paramétrico de referência
       alternative = "two.sided",   #teste bilateral
       conf.level = 0.99)           #significância de 1%
```

t tabelado = 4,0321

```
        One Sample t-test

data:  amostra
t = -3.7203, df = 5, p-value = 0.01371
alternative hypothesis: true mean is not equal to 285
99 percent confidence interval:
 149.4298 290.4536
sample estimates:
mean of x
 219.9417
```

> que 0,01

- Resolvendo com o R:

Teste t para médias de duas amostras independentes:

- Vamos testar se x e y possuem médias estatisticamente iguais, a 1% de significância, oriundas de distribuição normal:

$$H_o: \mu_x = \mu_y$$
$$H_1: \mu_x \neq \mu_y$$

```
x<-c(30.5,35.3,33.2,40.8,42.3,
     41.5,36.3,43.2,34.6,38.5)
y<-c(28.2,35.1,33.2,35.6,40.2,
     37.4,34.2,42.1,30.5,38.4)
t.test(x,y,  #amostras a serem testadas
       conf.level = 0.99)
```

```
        Welch Two Sample t-test

data:  x and y
t = 1.1148, df = 17.999, p-value = 0.2796
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 -3.369829  7.629829
sample estimates:
mean of x mean of y
    37.62     35.49
```

> que 0,01

- Suponha que duas amostras de lâmpadas incandescentes de dois fabricantes A e B foram testadas quanto a duração do filamento de tungstênio.

- O experimento visou identificar o tempo em horas que se iniciava no momento em que a lâmpada era acesa e terminava com o rompimento do filamento. Os dados coletados no experimento encontram-se a seguir:

```
A<-c(147.84,153.54,101.52,111.53,
     181.50,148.73,157.55,113.51)
B<-c(137.04,124.10,157.84,160.83,
     136.50,59.60,114.50,123.89,137.36,
     172.78)
```

- Verifique se as lâmpadas produzidas pelo fabricante A têm duração maior que as produzidas pelo outro fabricante, usando 5% de significância.

$$H_o: \mu_A = \mu_B$$
$$H_1: \mu_A > \mu_B$$

```
t.test(A,B,
       alternative = "greater",  #unilateral à direita
       var.equal = T)    #variâncias homogêneas
```

```
        Two Sample t-test

data:  A and B
t = 0.49593, df = 16, p-value = 0.3133
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -17.69573       Inf
sample estimates:
mean of x mean of y
  139.465   132.444
```

p > 0,05, não rejeito Ho

3. Teste F

- Esse é um teste usado para verificar se as variâncias de amostras oriundas de distribuições normais são idênticas.

- Nesse caso, dizemos que as amostras são oriundas de populações com variâncias homogêneas.

Resolvendo no R:

- Verificar se duas máquinas produzem peças com a mesma homogeneidade de resistência à tensão.

- Para isso, foram sorteadas amostras independentes que consistiam de seis peças de cada máquina. Foram obtidas as seguintes resistências:

```
maqA<-c(145,127,136,142,141,137)
maqB<-c(143,128,132,138,142,132)
```

- O que se pode concluir fazendo um teste de hipótese adequado para um nível de significância de 5%?

Segundo o teste *F*, podemos montar as seguintes hipóteses:

$$H_o: \sigma_A^2 = \sigma_B^2$$
$$H_1: \sigma_A^2 > \sigma_B^2$$

```
> maqA<-c(145,127,136,142,141,137)
> maqB<-c(143,128,132,138,142,132)
>
> var.test(maqA, maqB,
+          alternative = "greater")

        F test to compare two variances

data:  maqA and maqB
F = 1.0821, num df = 5, denom df = 5, p-value = 0.4666
alternative hypothesis: true ratio of variances is greater than 1
95 percent confidence interval:
 0.2142545       Inf
sample estimates:
ratio of variances
          1.082056
```

p > 0,05, não rejeito Ho

4. Teste para a Normalidade

- Por vezes temos necessidade de identificar, com certa confiança, se urna amostra ou conjunto de dados segue a distribuição normal.

- Podemos usar o teste de Shapiro-Wilk, que pode ser realizado no R com o uso do comando `shapiro.test( )`.

- Resolvendo no R:

- Selecionam-se 50 pessoas ao acaso e mensuram-se suas massas em quilogramas (kg). Queremos saber se esse conjunto de dados segue a distribuição normal.

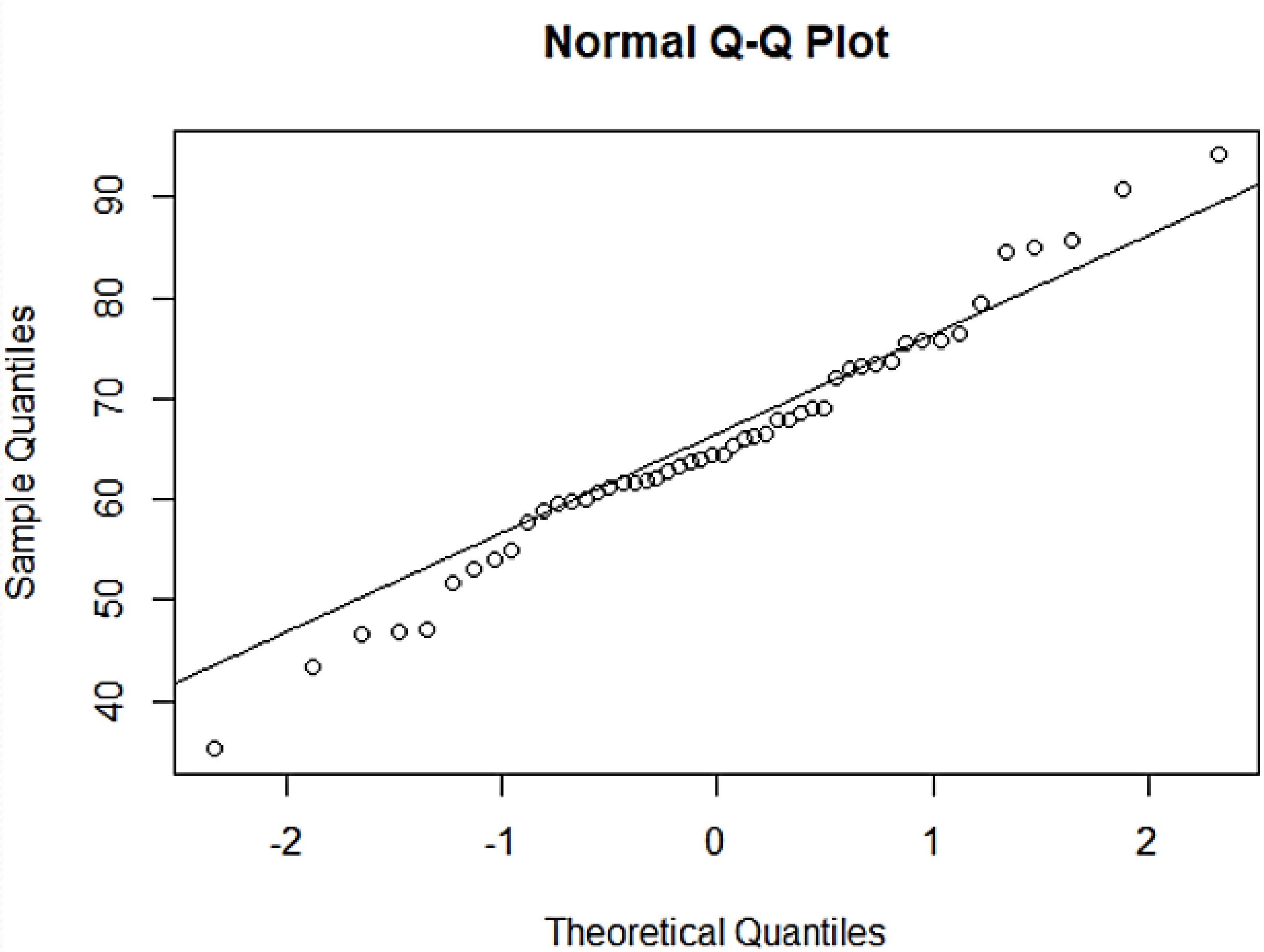
```
massas<-c(46.88,47.17,64.46,67.84,85.76,65.41,60.10,
          75.84,61.21,61.65,63.87,53.95,63.66,69.06,
          76.41,75.56,69.04,35.18,66.42,58.78,73.02,
          51.69,90.88,53.01,64.31,61.91,79.42,57.78,
          62.73,60.63,63.29,46.53,84.64,61.76,85.08,
          59.66,54.89,94.18,59.89,68.56,75.66,72.06,
          62,43.43,73.38,73.31,66.37,73.72,66.15,67.79)
```

```
> shapiro.test(massas)

        Shapiro-Wilk normality test

data:  massas
W = 0.98355, p-value = 0.7078
```

A hipótese nula do teste de Shapiro-Wilk é que a população **possui distribuição normal**. Portanto, um valor de p < 0,05 indica que você rejeitou a hipótese nula, ou seja, seus dados não possuem distribuição normal.

- Podemos plotar o teste:

```
> qqnorm(massas)  #obtendo o normal probability plot
> qqline(massas)  #colocando uma linha auxiliar
```



- O comando qqnorm( ) nos fornece diretamente um gráfico da distribuição de percentagens acumuladas, chamado de gráfico de probabilidade normal.

- Se os pontos desse gráfico seguem um padrão aproximado de uma reta, este fato evidencia que a variável aleatória em questão tem a distribuição aproximadamente normal.

5. Exercícios