Unoeste

“R” Com Data Science

Gabrielle Gomes dos Santos Ribeiro

2024

Medidas-Resumo

As medidas-resumo descrevem e sintetizam as principais características observadas em um conjunto de dados, permitindo ao pesquisador melhor compreensão do comportamento dos dados. Essas medidas fazem parte da Análise Descritiva dos dados, a etapa inicial de um estudo estatístico. Elas podem ser subdivididas da seguinte maneira:

Medidas de Posição e de Dispersão

POSIÇÃO

- DÁ UMA IDÉIA DE ONDE SE LOCALIZA O CENTRO DE UM CONJUNTO DE DADOS

DISPERSÃO

- MODO COMO OS DADOS SE POSICIONAM AO REDOR DE UM PONTO CENTRAL

Medidas de Posição

MÉDIA

- É definida pela soma de todas as observações do conjunto de dados dividida pelo número de observações;

- Em certos casos ela pode não ser o parâmetro mais adequado para descrever um conjunto de dados. Isto pode ocorrer, entre outros casos, quando existem dados aberrantes, extremos ou discrepantes. Pois todos os valores entram para o cálculo da média, então os valores extremos afetam no valor calculado e em alguns casos pode haver uma grande distorção, tornando, neste caso, a média indesejável como medida de tendência central.

MÉDIA

Exemplo: Uma amostra de salário de 10 funcionários da empresa X (em mil reais):

1,2 – 1,2 – 1,3 – 1,5 – 1,7 – 1,8 – 1,9 – 2,1 – 2,3 – 55,0

Note que provavelmente um dos salários deve ser de um dos diretores da empresa. Sua inclusão vai alterar sensivelmente o salário médio dos funcionários. O salário médio dos funcionários sem o maior salário é (1,2 + 1,2 + 1,3 + 1,5 + 1,7 + 1,8 + 1,9 + 2,1 + 2,3) = 15/9 = 1,67 **(R$ 1670,00)** e considerando o maior salário a média é (15 + 55) /10 = 7,0 **(R$ 7000,00)**, mostrando uma situação totalmente enganosa.

Média no R

dados<-read.csv("amostra.csv", header=TRUE, sep=";")

mean(dados$Idade)

- Para calcular as médias das linhas ou das colunas de uma tabela você pode usar as funções **rowMeans()** e **colMeans()**, respectivamente.

colMeans(dados[,3:6]) #Selecionei as colunas numéricas

colMeans(dados[dados$EstCivil=="Divorciado" ,3:6])
#Média das variáveis numéricas, mas somente dos Divorciados

Medidas de Posição

MEDIANA

- A mediana (Md) é o valor central da variável quando os valores estão dispostos **em ordem crescente ou decrescente** de magnitude. É o valor que divide o conjunto de dados em dois subconjuntos com o mesmo número de elementos.

- Se o número de elementos “n” for ímpar, a Md será o elemento central da sequência de dados; se for par, a Md será a média entre os dois elementos centrais da sequência de dados.

- **Exemplo:** O número de empresas falidas no mês de Janeiro nos últimos 8 anos são:

52 – 41 – 37 – 58 - 82 – 24 – 63 – 68

Ordenando estes valores por ordem crescente, temos:

24 – 37 – 41 – 52 - 58 – 63 – 68 – 82.

Md=55

Medidas de Posição

MEDIANA no R

median(dados$Idade)

Medidas de Posição

MODA

A moda (Mo) é o valor que ocorre com mais frequência. A moda é também conhecida como tipo dominante, valor popular e valor de densidade máxima de um conjunto de dados.

Apesar de seu significado ser bem simples, a moda nem sempre existe e nem sempre é única.

- Quando não há valores repetidos, a série é **amodal.**
- Quando tem apenas uma moda, a série é **unimodal.**
- Quando tem duas modas, a série é **bimodal.**
- Quando tem várias modas, a série é **multimodal.**

Por exemplo:
X: 2, 3, 4, 4, 4, 5, 7, 7, 8, 10 → Moda = 4

Y: 1, 1, 2, 5, 5, 7, 9, 11, 11 → Modas = 1, 5, 11

Medidas de Posição

MODA

No R não há uma função para moda. A função *mode()* retorna o formato da variável (e.g.: numérico).

Para obter a moda, a alternativa é obter a frequência de cada valor e a partir daí utilizar a função ***max***. Veja o exemplo:

obtém a tabela com frequência das variáveis

freq=table(dados$Filhos)

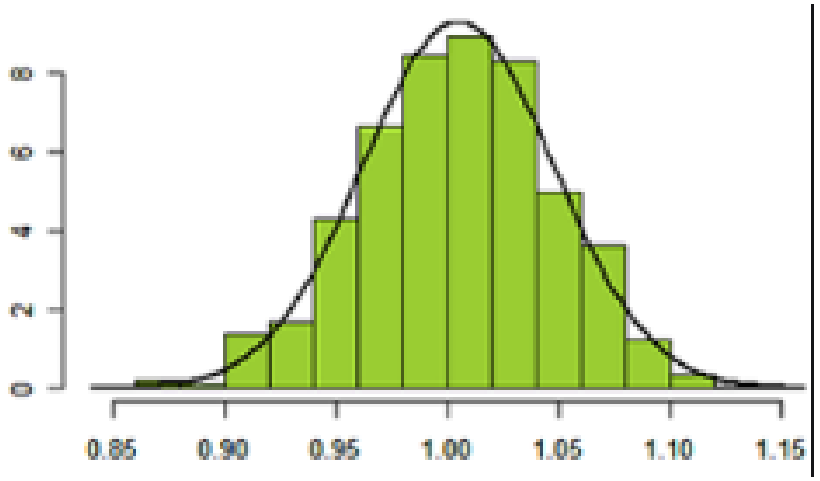
max(freq)

para obter o nome da categoria da variável

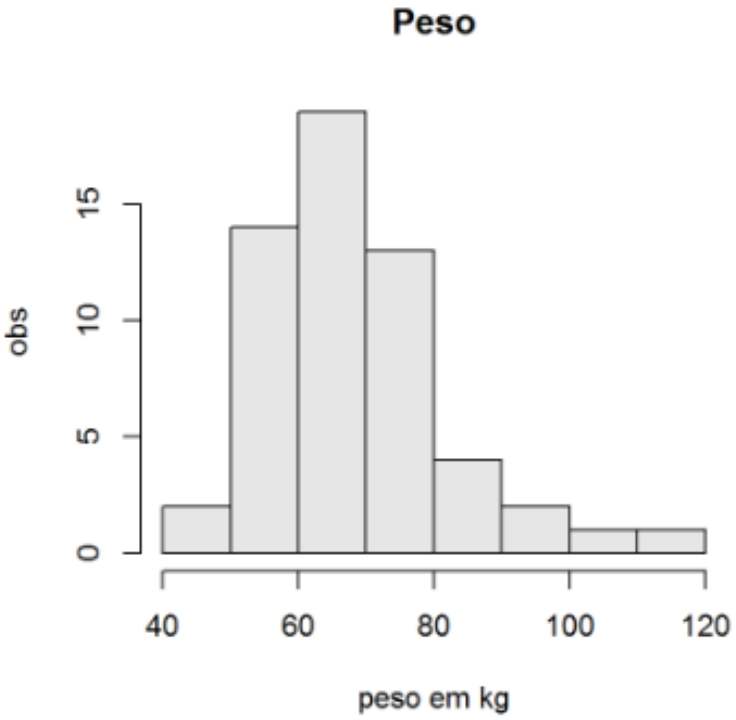
names(table(dados$Filhos))[table(dados$Filhos) == max(table(dados$Filhos))]

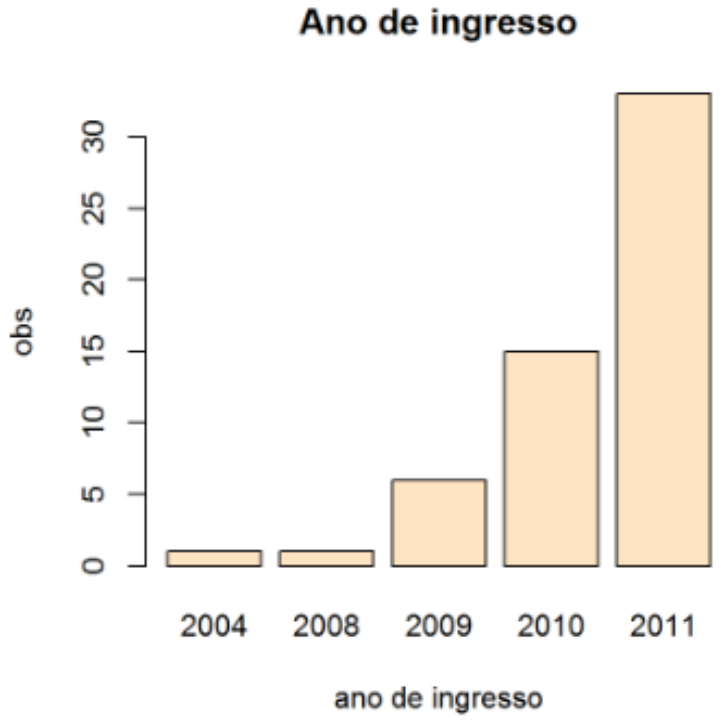
Relação entre média, moda e mediana

- Na maioria dos casos, optamos por escolher entre a média e a mediana.
- Essa escolha depende da simetria (ou assimetria) da distribuição dos dados.
- Em distribuições assimétricas, optamos por utilizar a mediana, pois ela não é influenciada por valores extremos.



ASSIMETRIA


Peso
obs
0
5
10
15
40
60
80
100
120
peso em kg


Ano de ingresso
obs
0
5
10
15
20
25
30
2004
2008
2009
2010
2011
ano de ingresso

Medidas Separatrizes

Para entender bem uma distribuição, pode-se conhecer valores acima ou abaixo dos quais se encontra uma determinada porcentagem dos dados através das medidas separatrizes.

As separatrizes são números reais que dividem os dados ordenados em partes que contêm a mesma quantidade de elementos da série.

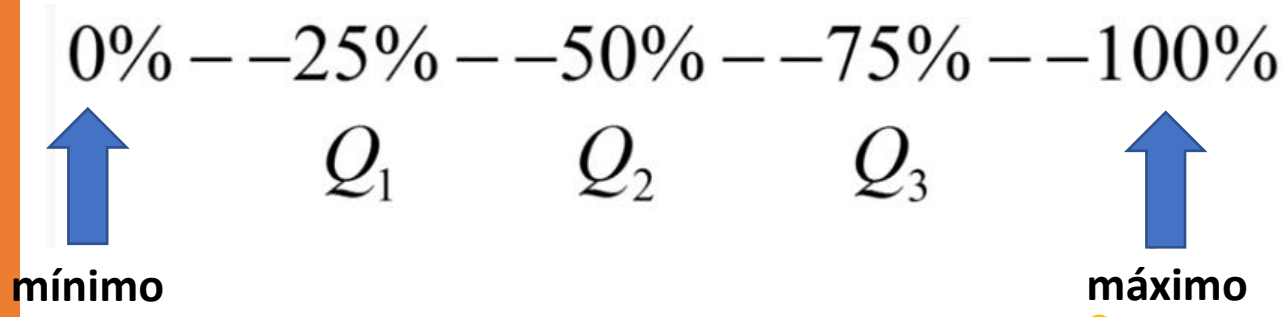
Desta forma, a mediana que divide a sequência ordenada em dois grupos, cada um deles contendo 50% dos valores, também é uma medida separatriz.

Medidas de Posição

QUARTIS

São medidas que dividem os dados ordenados em quatro partes iguais: primeiro quartil (Q1), segundo quartil (Q2) e terceiro quartil (Q3).

Pode-se dizer que 25% dos valores estão abaixo de Q1 e 75% dos valores estão acima de Q1. A diferença entre Q3 e Q1 é chamada de amplitude interquartílica . O segundo quartil é exatamente igual à mediana.



Exemplo

- **Exemplo: amostra de salários (em salário mínimo) de 160 professores de uma escola.**
- **Q1 = 4 salários mínimos**

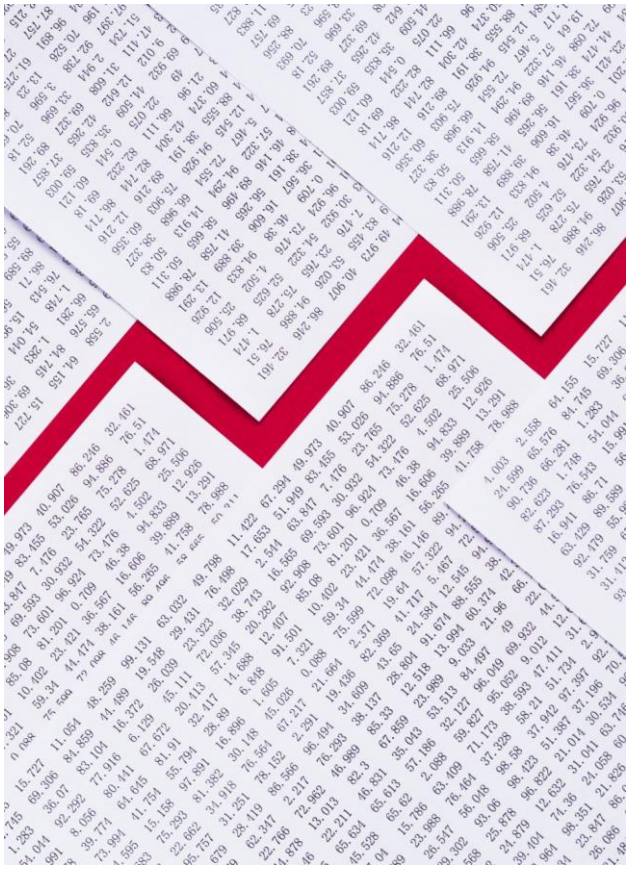
Interpretação: 25% dos professores da escola ganham até 4 salários mínimos ou 75% dos professores ganham mais de 4 salários mínimos

- **D4 = 5,13 salários mínimos**

Interpretação: 40% dos professores da escola ganham até 5,13 salários mínimos ou 60% dos professores ganham mais de 5,13 salários mínimos.

- **C85 = 8,07 salários mínimos**

Interpretação: 85% dos professores da escola ganham até 8,07 salários mínimos ou 15% dos professores ganham mais de 8,07 salários mínimos.



Medidas de Posição

QUARTIS no R

quantile(dados$AnosServico)

Medidas de Dispersão

MEDIDAS DE DISPERSÃO

Considere a seguinte situação: tenho notas de provas de 3 turmas de alunos:

Tabela 1 – Notas das turmas A, B e C.

Turma	Notas	Média
A	2 3 4 6 6 8 9 10	6
B	4 5 5 6 6 7 7 8	6
C	6 6 6 6 6 6 6 6	6

Calculei a média de cada turma e observei que todas deram iguais a 6 ($\overline{x}_A = \overline{x}_B = \overline{x}_C = 6$). Então, posso concluir que as turmas mostraram ter adquirido o mesmo conhecimento?

MEDIDAS DE DISPERSÃO

Considere a seguinte situação: tenho notas de provas de 3 turmas de alunos:

Tabela 1 – Notas das turmas A, B e C.

Turma	Notas	Média
A	2 3 4 6 6 8 9 10	6
B	4 5 5 6 6 7 7 8	6
C	6 6 6 6 6 6 6 6	6

Evidente que NÃO. Mesmo as médias sendo iguais, a variância entre as notas foi diferente em cada turma. Por exemplo, na turma A as notas são muito diferentes entre si, repetindo apenas o valor 6, ou seja, há uma grande variação entre os dados. Já na turma C todas as notas são iguais, portanto a variância é zero. Por isso a importância de saber a variância de um conjunto de dados.

MEDIDAS DE DISPERSÃO

O resumo de variável observada apenas por uma medida de posição, ignora a informação sobre a sua variabilidade.

Não é seguro analisar um conjunto de dados somente pelo emprego de medidas de tendência central.

Por isso, precisamos de medidas que caracterizem a dispersão ou variabilidade dos dados em relação a um valor central.

AMPLITUDE

A primeira medida de dispersão que vamos comentar é a amplitude total. Ela é definida pela diferença entre o maior valor e o menor valor do seu conjunto de dados:

$$A = x_{máx} - x_{mín}$$

Desvantagens da amplitude:

- Considera somente os dois valores extremos, por isso é apenas uma indicação aproximada da dispersão.

- Apresenta muita variação de uma amostra para outra, mesmo que ambas sejam extraídas da mesma população.

- Portanto, você deve trabalhar com uma medida que leve em consideração todas as observações, ou seja, a variância e o D.P.

AMPLITUDE no R

No R ela pode ser calculada da seguinte maneira:

range(dados$Idade)
diff(range(dados$Idade))

VARIÂNCIA

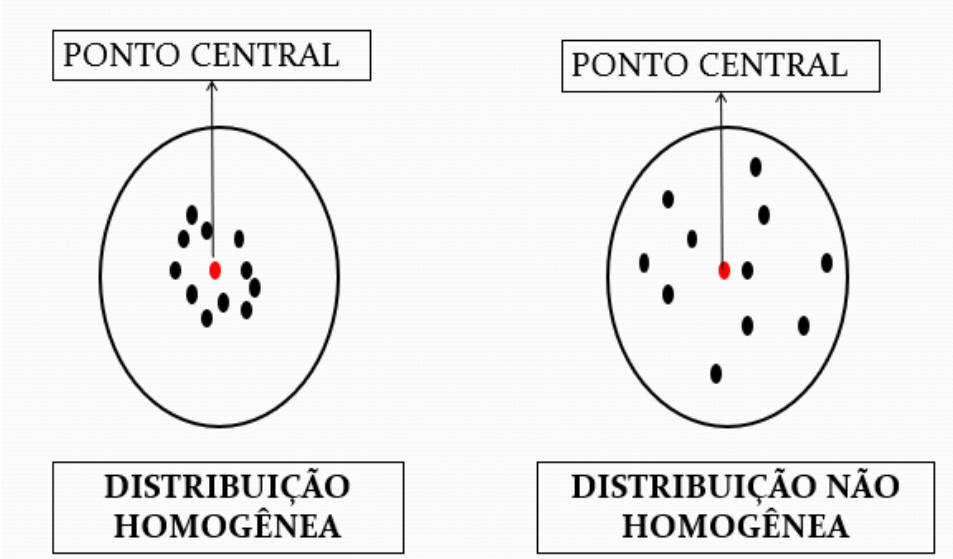
- A variância de uma amostra $\{x_{1,}x_2, \ldots\ldots, x_n\}$ de *n* elementos é definida como a soma ao quadrado dos desvios dos elementos em relação à sua média $\bar{x}$ dividido por *(n-1)*. Ou seja, a variância amostral é dada por:

$$\boldsymbol{s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}}$$

DESVIO-PADRÃO

Medida que expressa o grau de dispersão dos valores do conjunto de dados em torno da média;

- **É a raiz quadrada da variância.**



DESVIO-PADRÃO

- O motivo de calcular o desvio padrão é que, como elevamos os desvios ao quadrado para calcular a variância, sua unidade de medida é o quadrado da unidade de medida dos dados, o que muitas vezes não faz muito sentido.
- Por exemplo, suponha que estejamos trabalhando com dados de idade e chegamos numa variância igual a 25. A unidade de medida deste número é idade ao quadrado, o que não faz sentido. Mas, se extrairmos sua raiz quadrada, obtemos um desvio padrão de 5 anos, o que significa que os dados estão concentrados cerca de 5 anos em torno de sua média.

VARIÂNCIA E DESVIO-PADRÃO no R

Funções:

var(dados$Idade)

sd(dados$Idade)

DESVIO-PADRÃO

O desvio-padrão, quando analisado isoladamente, não dá margem a muitas conclusões. Por exemplo, para uma distribuição cuja média é 300, um desvio-padrão de 2 unidades é pequeno, mas para uma distribuição cuja média é 20, ele já não é tão pequeno. Por isso ele é mais recomendável para comparar 2 ou mais grupos (TAVARES, 2007).

Importante!

Condições para se usar o desvio-padrão ou variância para comparar a variabilidade entre grupos:

- mesmo número de observações;
- mesma unidade;
- mesma média

Medidas-Resumo

COEFICIENTE DE VARIAÇÃO

$$CV = \frac{s}{\bar{x}} .100$$

- Interpretado como a variabilidade dos dados em relação à média. Quanto menor o CV mais homogêneo é o conjunto de dados;
- Expresso em porcentagem;

Interpretação:

Baixa dispersão: CV≤15% → **dados homogêneos**

Média dispersão: 15%<CV<30%

Alta dispersão: CV≥30%

CV=(sd(dados$Idade)/mean(dados$Idade))*100

Função *summary*

- Função que faz um resumo dos dados apresentando seis medidas de posição que descrevem os dados (os valores mínimo e máximo, a média e a mediana, o primeiro e o terceiro quartis.

summary(dados)

ANÁLISE DE CORRELAÇÃO LINEAR

ANÁLISE DE CORRELAÇÃO LINEAR

- Área da estatística que analisa o **comportamento conjunto** de duas variáveis quantitativas e verifica se existe algum tipo de **relação** entre elas.

- Neste caso estamos interessados apenas nas relações do tipo lineares entre as variáveis – **Análise de Correlação Linear Simples.** A palavra "simples" indica que a análise será apenas entre 2 variáveis.

- Primeiro, iremos verificar visualmente a existência de associação entre as variáveis a partir do **diagrama de dispersão**. Ele representa, em um sistema coordenado cartesiano ortogonal, os pares ordenados (x_i, y_i) das variáveis X e Y, obtendo uma nuvem de pontos.

ANÁLISE DE CORRELAÇÃO LINEAR

- Depois de detectada visualmente a correlação entre as variáveis, vamos verificar se existe correlação significativa entre o par de variáveis através do **teste de Pearson.** Se for encontrada correlação significativa, vamos medir o grau dessa associação (ou dependência) por meio de um único número. Para isso usaremos uma medida chamada **coeficiente de correlação linear de Pearson**.

- Uma vez caracterizada a relação, procuramos descrevê-la por meio de uma função matemática. A **regressão** é o instrumento adequado para a determinação dos parâmetros dessa função.

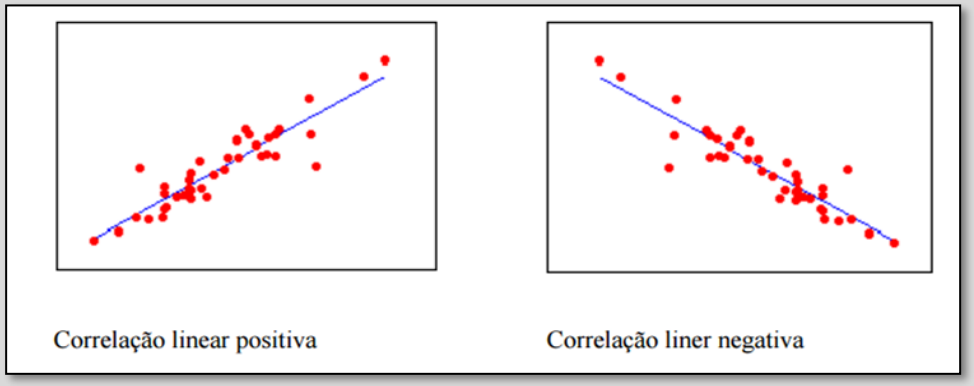
1. Diagrama de Dispersão

- Usado para verificar visualmente a existência de associação entre as variáveis a partir do **diagrama de dispersão**.

- Ele representa, em um sistema coordenado cartesiano ortogonal, os pares ordenados (x_i, y_i) das variáveis X e Y, obtendo uma nuvem de pontos.

- Afirma-se que existe uma relação linear entre as variáveis se os dados se aproximarem de uma linha reta.

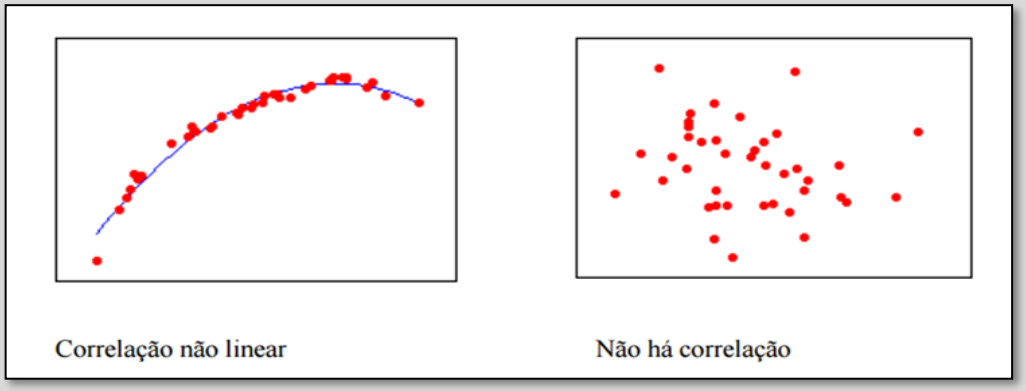
Diagrama de Dispersão



- Se, quando uma das variáveis "cresce", a outra, em média, também "cresce", dizemos que entre as duas variáveis existem **correlação linear positiva**;

- Se, quando uma das variáveis "cresce", a outra, em média, "decresce", dizemos que entre as duas variáveis existem **correlação linear negativa**;

Diagrama de Dispersão



- Se os pontos estiverem dispersos, sem definição de direção, dizemos que a correlação é nula. As variáveis nesse caso são ditas não correlacionadas.
- Se os pontos estiverem dispostos em outra forma geométrica que não seja uma reta, dizemos apenas que não existe relação linear.

Obs.: O comportamento de Y em relação a X pode se apresentar de diversas maneiras: linear, quadrático, cúbico, exponencial, logarítmico, etc...

EXEMPLO

Para ilustrar, vamos trabalhar com o conjunto de dados abaixo (Bussab e Morettin, 2010):

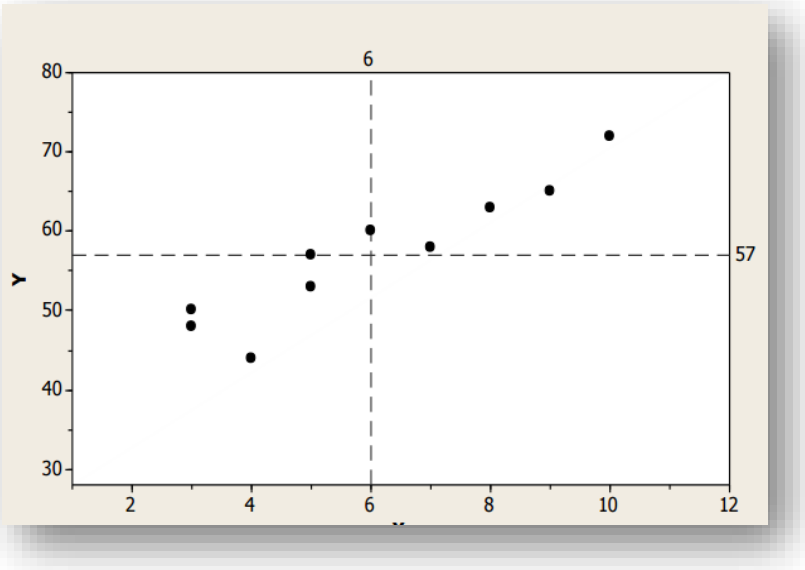
Exemplo 1- Amostra de 10 agentes de uma companhia de seguros

X: anos de serviços e Y: número de clientes.

Agente	A	B	C	D	E	F	G	H	I	J
X	3	3	5	5	4	6	8	7	9	10
Y	48	50	57	53	44	60	63	58	65	72

Diagrama de Dispersão

Segue abaixo o diagrama de dispersão do Exemplo 1:



Correlação Linear Positiva

Teste e Coeficiente de Correlação linear de Pearson

- **Teste de Pearson**

Hipóteses do teste:

- H_0: As variáveis são independentes.
- H_a: As variáveis são dependentes.

- **Cálculo do Coeficiente de Correlação de Pearson:**

```
cor(dados$Renda, dados$Idade)
```

TESTES ESTATÍSTICOS – P-valor

Os testes estatísticos, por meio de fórmulas específicas resultam em um valor de probabilidade - **p-valor -** que varia de 0 a 1 .

Compara-se o p-valor com o nível de significância adotado para a pesquisa, para saber se o resultado do teste foi significativo ou não (rejeita ou não a hipótese nula).

O nível de significância (α) é a probabilidade de rejeitar a hipótese nula quando ela é verdadeira, ou seja, é a probabilidade de erro.

$\alpha = 1 -$ nível de confiança

TESTES ESTATÍSTICOS – P-valor

O QUE É NÍVEL DE CONFIANÇA??

- Também conhecida popularmente como “confiabilidade”, o nível de confiança é o grau de certeza de que o valor obtido ao pesquisar a amostra representa o valor que seria obtido ao pesquisar toda a população.

- O valor mais comum utilizado para o nível de confiança é de 95%, mas também são usuais os valores de 90% e 99%.

- Por exemplo,

se o **Nível de confiança for de 95% → α = 5% = 0,05**

TESTES ESTATÍSTICOS

TESTE SIGNIFICATIVO

- P<0,05 – Rejeita-se a hipótese H_0

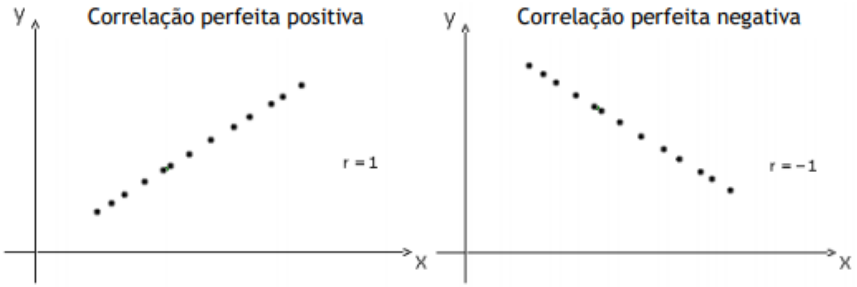
TESTE NÃO SIGNIFICATIVO

- P>0,05 – Não Rejeita a hipótese H_0

Coeficiente de Correlação linear de Pearson (r)

O valor de *r* está sempre entre -1 e +1. Quanto maior o valor de *r* (positivo ou negativo), mais forte a associação. Assim:

- a) Se a correlação entre duas variáveis é perfeita (todos os pontos no gráfico de dispersão caem exatamente numa linha reta) e positiva, então r = 1
- b) Se a correlação é perfeita e negativa, então r = -1.
- c) Se r = 0, ou não há correlação entre as variáveis, ou a relação que possa existir não é linear.



Coeficiente de Correlação linear de Pearson

INTERPRETAÇÃO DE r:

- Se 0,9 ≤ $|r|$< 1,0→ há uma correlação muito forte
- Se 0,7 ≤ $|r|$ < 0,9→ há uma correlação forte
- Se 0,4 ≤ $|r|$ < 0,7 → há uma correlação moderada;
- Se 0,2 ≤ $|r|$ < 0,4 → há uma correlação fraca;
- Se 0 <$|r|$ < 0,2 → a correlação é muito fraca

E o sinal (positivo ou negativo) indica o sentido da correlação. Quando o sinal é positivo, a correlação é positiva, ou seja, conforme uma variável cresce a outra também cresce. E quando é negativo, a correlação é negativa, ou seja, conforme uma variável cresce a outra diminui.

3. ANÁLISE DE REGRESSÃO $Y = a + b X$

Tem por objetivo descrever, por meio de um modelo matemático, a relação entre duas variáveis, partindo de *n* observações das mesmas.

- A variável sobre a qual desejamos fazer uma estimativa recebe o nome de variável resposta (de interesse ou dependente) e a outra recebe o nome de variável explicativa (auxiliar ou independente). Assim, supondo X a variável explicativa e Y a variável resposta, vamos procurar determinar o ajustamento de uma reta à relação entre essas variáveis, ou seja, vamos obter uma função definida por: **Y = a + b X**, sendo a e b parâmetros.

- A partir dessa equação, podemos fazer estimativas de Y a partir de valores de X.

OBSERVAÇÕES:

- Para cálculo das medidas de associação, é necessário que as duas variáveis sejam medidas sobre os mesmos elementos (indivíduos) da amostra (medidas pareadas).

- A Análise de Correlação Linear só é possível para variáveis quantitativas.

Análise de Correlação Linear Simples no R

```
dados<-read.table("clipboard", header=TRUE, dec=',')
Ou
dados<-read.csv("DadosPIB.csv", header=TRUE, sep=";", dec=",")

plot(dados$PIBAGRO, dados$CULTIVO, main="Diagrama de Dispersão",
ylab="Cultivo", xlab="PIB Agro")
```

Análise de Correlação Linear Simples no R

```
cor.test(dados$PIBAGRO, dados$CULTIVO,method='pearson')$p.value
```

```
cor(dados$PIBAGRO, dados$CULTIVO)
```

Análise de Correlação Linear Simples no R

modelo = lm(formula = CULTIVO~PIBAGRO, data = dados)

Note que função lm() é chamada com o formato lm(y ~ x), ou seja, a variável resposta é y e a explicativa é x, sempre nessa ordem.

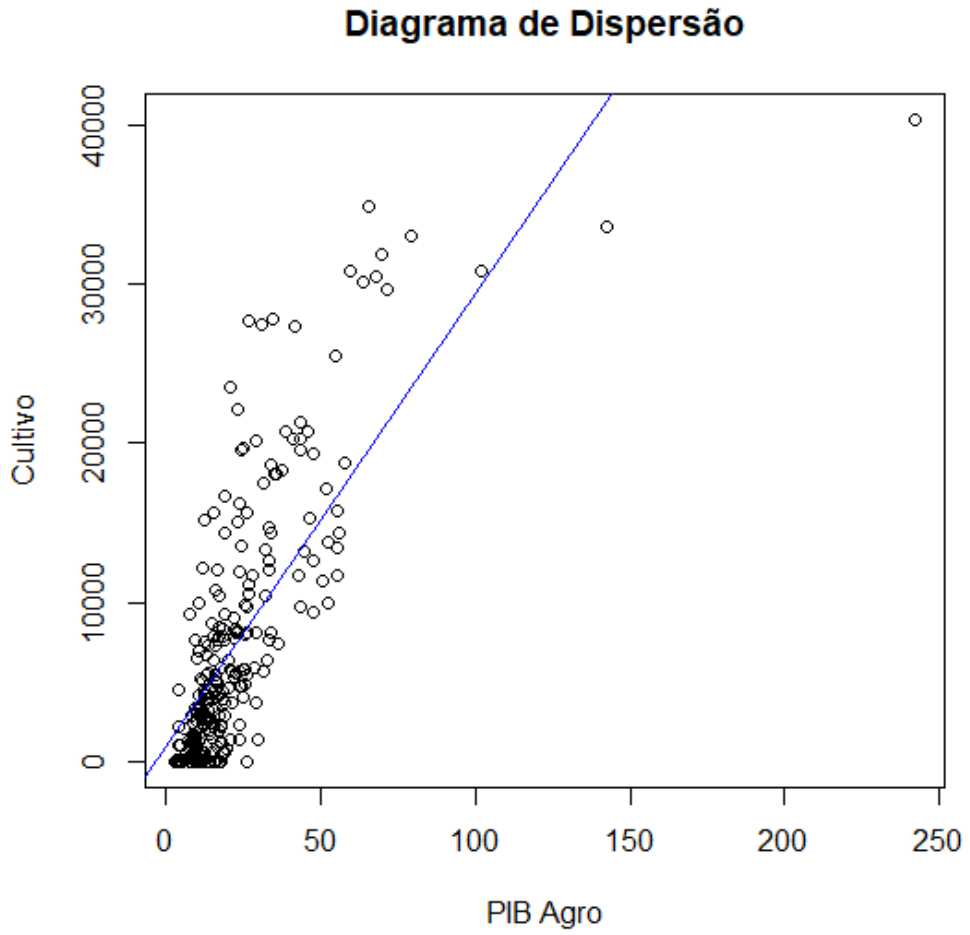
```
Call:
lm(formula = CULTIVO ~ PIB_AGRO, data = tabela)

Coefficients:
(Intercept)      PIB_AGRO
      985.1         284.6
```

Y = a + b.X

Portanto, a = 985,1 e b = 284,6

Análise de Correlação Linear Simples no R

abline(modelo, col="blue")



Análise de Correlação Linear Simples no R

Com a função *summary*, diversas medidas descritivas úteis para a análise do ajuste podem ser obtidas:

summary(modelo)

```
Call:
lm(formula = CULTIVO ~ PIB_AGRO, data = tabela)

Residuals:
   Min     1Q Median     3Q    Max
-29565  -2972  -1770   1813  19239

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   985.08     473.82   2.079   0.0386 *
PIB_AGRO      284.61      15.21  18.707   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5360 on 258 degrees of freedom
Multiple R-squared:  0.5756,    Adjusted R-squared:  0.574
F-statistic:   350 on 1 and 258 DF,  p-value: < 2.2e-16
```

Hipóteses

Hipóteses do teste t:

- H_0: A variável não é significativa para o modelo.
- H_a: A variável é significativa para o modelo.

Hipóteses do teste F:

- H_0: O modelo não é significativo
- H_a: O modelo é significativo (o modelo é eficiente para explicar a variável Y).

ANÁLISE BIVARIADA

Na análise bivariada, procuramos identificar relações entre duas variáveis.

O tipo de resumo estatístico informativo vai depender dos tipos das variáveis envolvidas.

QUANTITATIVA x QUANTITATIVA

ANÁLISE DE CORRELAÇÃO:

- Construção do gráfico de dispersão
- Teste de Pearson
- Cálculo do Coeficiente de Correlação
- Definição da Equação de Regressão

QUALITATIVA x QUALITATIVA

- Forma de representar o cruzamento dos dados – **Tabela de Contingência**
- Vamos considerar as variáveis **EstCivil** (estado civil) e **Escolaridade** (grau de instrução) do conjunto amostra.csv

```
amostra<-read.csv("amostra.csv", header=T, sep=";")
table(amostra$EstCivil,amostra$Escolaridade)
prop.table(table(amostra$EstCivil,amostra$Escolaridade))*100
```

	Fundamental completo	Médio completo	Superior completo
Casado	17.5	38.5	13.0
Divorciado	3.0	8.5	3.0
Solteiro	4.5	7.5	4.5

Teste Qui-Quadrado (χ^2)

- Avalia a existência de associação entre duas variáveis qualitativas (nominais ou ordinais)

- Testa as hipóteses:

H_0: As variáveis são independentes.
H_a: As variáveis são dependentes.

- Decisão do teste:

P≤0,05 – Rejeita-se a hipótese H_0 → existe associação entre as variáveis

P>0,05 – Não Rejeita a hipótese H_0 → **NÃO** existe associação entre as variáveis

Teste Qui-Quadrado (χ^2)

CONDIÇÕES:

- Exclusivamente para variáveis nominais e ordinais;
- Os grupos devem ser independentes;
- Os itens de cada grupo são selecionados aleatoriamente;
- As observações devem ser frequências ou contagens;
- Cada observação pertence a uma e somente uma categoria
- A amostra deve ser relativamente grande (pelo menos 5 observações em cada célula e, no caso de poucos grupos - pelo menos 10).

Teste Qui-Quadrado (χ^2)

Qui-quadrado no R

Função → *chisq.test*

Exemplo:
chisq.test(table(amostra$EstCivil,amostra$Escolaridade))

Teste Qui-Quadrado (χ^2)

Exemplo:

chisq.test(table(amostra$EstCivil,amostra$Escolaridade))

Pearson's Chi-squared test

data: table(amostra$EstCivil, amostra$Escolaridade)
X-squared = 1.8436, df = 4, p-value = 0.7645

Teste Qui-Quadrado (χ^2)

Outra opção: Entrando com a Tabela de contingência pronta:

Ex.:

Sexo \ Opção	C.S.	C.H.	C.E.	TOTAL
Masculino	80	55	70	205
Feminino	65	85	45	195
TOTAL	145	140	115	400

H_0: A área do curso independe do sexo.

H_a: A área do curso dependente do sexo.

Teste Qui-Quadrado (χ^2)

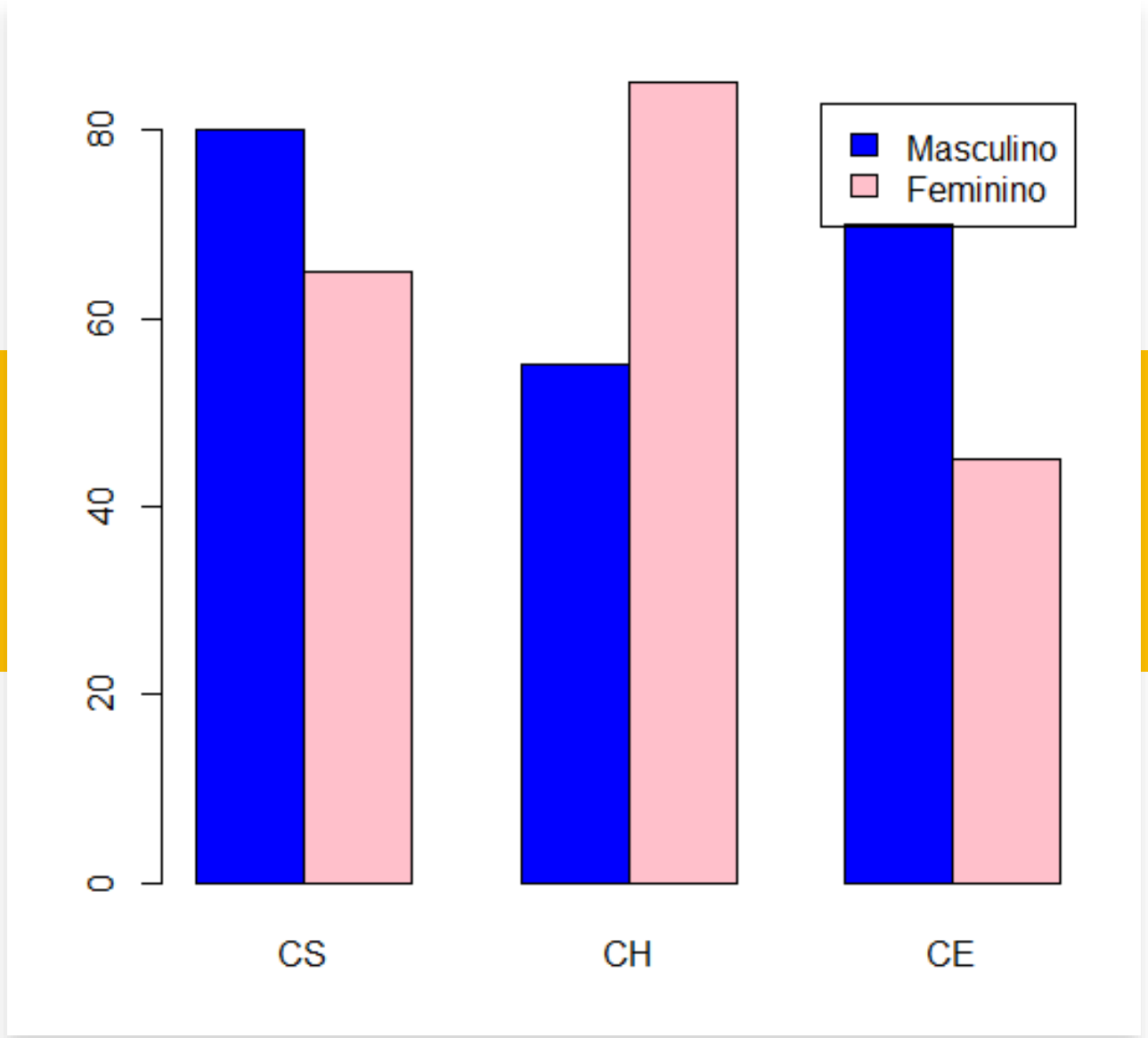
Ex.:

Sexo \ Opção	C.S.	C.H.	C.E.	TOTAL
Masculino	80	55	70	205
Feminino	65	85	45	195
TOTAL	145	140	115	400

```
tabela<-matrix(c(80,55,70,65,85,45), nrow=2, ncol=3, byrow=T)
rownames(tabela)<-c("Masculino","Feminino")
colnames(tabela)<-c("CS", "CH", "CE")
chisq.test(tabela)
```

Para visualizar a distribuição da tabela:

```
barplot(tabela, beside=T, legend=T, col=c("blue", "pink"))
```



Exemplo

Teste exato de Fisher

- Em amostras pequenas o erro do valor de Qui-quadrado é alto e, portanto, o teste não é recomendável.
- Assim, em amostras pequenas deve-se executar esse teste, pois produz erro menor que o teste de Qui-Quadrado.
- De modo geral usa-se o Teste exato de Fisher quando:

1. **o valor de n < 20 ou**
2. **20 < n < 40 e a menor frequência for menor que 5.**

- A análise do teste de Fisher é feita como a de $\chi 2$

Teste exato de Fisher

- **Teste Exato de Fisher no R**

Função → ***fisher.test***

Exemplo: Sexo x Opinião sobre a pena de morte

	Favor	**Contra**
Homem	10	2
Mulher	6	12

Teste exato de Fisher

Exemplo: Sexo x Opinião sobre a pena de morte

opiniao<-matrix(c(10,2,6,12), nrow=2, ncol=2, byrow=T)

fisher.test(opiniao)

	Favor	Contra
Homem	10	2
Mulher	6	12

Fisher's Exact Test for Count Data

data: opiniao

p-value = 0.01061

QUALITATIVA VS QUANTITATIVA

- Para se obter uma tabela de contingência é necessário agrupar a variável quantitativa em classes.
- Para exemplificar este caso vamos considerar as variáveis Escolaridade e Renda
- No exemplo a seguir, vamos agrupar a variável Renda em 4 classes, definidas pelos quartis, usando ***cut().***
- Após agrupar esta variável, obtemos a(s) tabela(s) de cruzamento como mostrado anteriormente.

QUALITATIVA VS QUANTITATIVA

Exemplo:

quantile(amostra$Renda)

Renda.cl <- cut(amostra$Renda, quantile(amostra$Renda))
#transforma os dados em categóricos

tabela <- table(amostra$Escolaridade, Renda.cl)

	Renda.cl			
	(1.05,5.31]	(5.31,9.55]	(9.55,16]	(16,59.1]
Fundamental completo	11	10	16	12
Médio completo	27	31	24	26
Superior completo	10	9	10	12

QUALITATIVA VS QUANTITATIVA

Para as medidas estatísticas, o usual é obter um resumo da variável quantitativa para cada nível do fator qualitativo.

Função → ***tapply***

Exemplo: Resumos da variável Renda, para cada nível de instrução.

tapply(amostra$Renda, amostra$Escolaridade, mean)

```
Fundamental completo       Medio completo    Superior completo
            13.07900             11.54862             12.07439
```

tapply(amostra$Renda, amostra$Escolaridade, sd)

```
Fundamental completo       Medio completo    Superior completo
            11.895652             8.605886             8.503721
```

QUALITATIVA VS QUANTITATIVA

chisq.test(table(amostra$Escolaridade, Renda.cl))

QUALITATIVA VS QUANTITATIVA

chisq.test(table(amostra$Escolaridade, Renda.cl))

```
        Pearson's Chi-squared test

data:  table(amostra$Escolaridade, Renda.cl)
X-squared = 3.0327, df = 6, p-value = 0.8047
```