Aula 03: *Outliers* & Associação entre variáveis

Estatística e Probabilidades

André Victor Ribeiro Amaral (sala 3029)
avramaral@gmail.com

Observações atípicas (*Outliers*)

Observações atípicas (ou ***Outliers***) são valores muito altos ou muito baixos em relação ao restante do conjunto de dados.
Podem ser divido em dois tipos:

1. Não genuíno: erro de digitação, erro de medição, etc.
2. Genuíno: não são resultantes de erros e são valores importante ao estudo, podendo fornecer informações valiosas sobre a característica que está sendo estudada.

Em caso de *outliers* não genuínos, estes valores devem ser corrigidos ou excluído do banco de dados. Em contrapartida, para valores atípicos genuínos, devemos deixá-los no banco de dados, tomando cuidado com interpretações futuras.

Escore padronizado

O **escore padronizado** (ou "escore z") representa o número de desvios padrão pelo qual uma observação x_i dista da média (para mais ou para menos); e é calculado como:

$$z = \frac{x_i - \bar{x}}{s}, \text{ na amostra e}$$

$$z = \frac{x_i - \mu}{\sigma}, \text{ para população.}$$

- O escore padronizado permite distinguir entre os valores usuais e valores raros.
- São considerados valores usuais os que possuem escore z entre -2 e 2; e raros os que possuem escore z menor que -2 ou maior que 2.

Escore padronizado – Exercício

Para o exemplo da quantidade de álcool no sangue, calcule o escore z para todas as observações. Lembre-se de:

$$0.27, 0.17, 0.17, 0.16, 0.13, 0.24, 0.29, 0.24,$$
$$0.14, 0.16, 0.12, 0.16, 0.21, 0.17, 0.18.$$

E que $\bar{x} = 0.187$ ml/l e $s = 0.0512$ ml/l.

Escore padronizado – Exercício

x_i	$\frac{x_i - \bar{x}}{s}$
0.27	$\frac{(0.27-0.187)}{0.051} = 1.63$
0.17	$\frac{(0.17-0.187)}{0.051} = -0.33$
0.17	$\frac{(0.17-0.187)}{0.051} = -0.33$
0.16	$\frac{(0.16-0.187)}{0.051} = -0.53$
0.13	$\frac{(0.13-0.187)}{0.051} = -1.12$
0.24	$\frac{(0.24-0.187)}{0.051} = 1.04$
0.29	$\frac{(0.29-0.187)}{0.051} = 2.02$
0.24	$\frac{(0.24-0.187)}{0.051} = 1.04$
0.14	$\frac{(0.14-0.187)}{0.051} = -0.92$
0.16	$\frac{(0.16-0.187)}{0.051} = -0.53$
0.12	$\frac{(0.12-0.187)}{0.051} = -1.31$
0.16	$\frac{(0.16-0.187)}{0.051} = -0.53$
0.21	$\frac{(0.21-0.187)}{0.051} = 0.45$
0.17	$\frac{(0.17-0.187)}{0.051} = -0.33$
0.18	$\frac{(0.18-0.187)}{0.051} = -0.14$

Segundo a regra que estabelecemos, $x_i = 0.29$ é valor atípico.

Boxplot

Boxplot é um gráfico em forma de caixa que utiliza os quartis.

Sejam Q_1, Q_2 e Q_3 o primeiro, segundo e terceiro quartis, respectivamente. Além disso, defina $DI = Q_3 - Q_1$ como "distância interquartílica".

Note que a "distância interquartílica" também é uma medida de variabilidade do conjunto de dados.

Boxplot

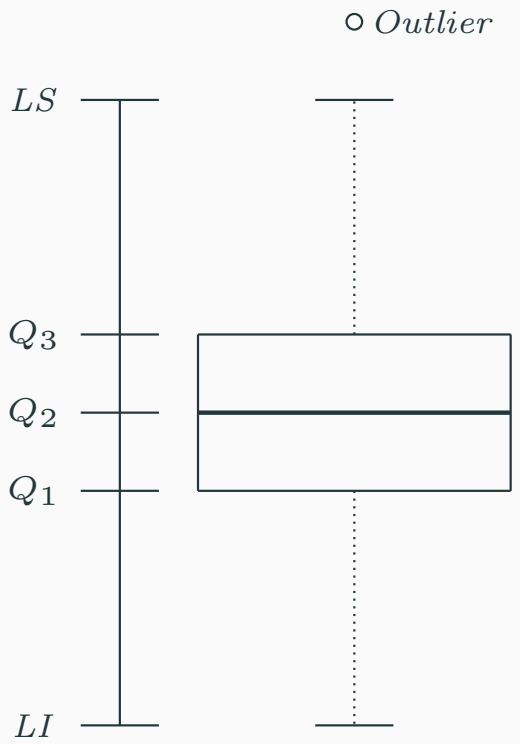


Figura 1: Boxplot.

Definimos, baseado na figura da esquerda:

- $LI = Q_1 - 1.5 \cdot DI$; e
- $LS = Q_3 + 1.5 \cdot DI$.

Assim, dados menores que LI e maiores que LS são marcados com "pontos" e são chamados de *outliers*.

Boxplot

O boxplot pode ser usado, principalmente, para:

- Detecção de valores discrepantes;
- Identificação da forma da distribuição (simetria);
- Comparação da tendência central (mediana) de dois ou mais conjuntos de dados; e
- Comparação da variabilidade de dois ou mais conjuntos de dados.

Boxplot

Exemplo (TRIOLA, Mário F. *Introdução à Estatística*)

Compare as idades (em anos completos) dos atores e atrizes na ocasião em que receberam o Oscar.

Atores:

31, 32, 32, 32, 33, 35, 36, 36, 37, 37, 38, 39, 39, 40, 40, 40, 41, 42, 42, 43, 43, 44, 45, 45, 46, 46, 47, 48, 48, 51, 53, 55, 56, 56, 60, 60, 61, 62, 76.

Atrizes:

21, 24, 25, 26, 26, 26, 26, 27, 28, 30, 30, 31, 31, 33, 33, 33, 34, 34, 34, 34, 35, 35, 35, 37, 37, 38, 39, 41, 41, 41, 42, 44, 49, 50, 60, 61, 61, 74, 80.

Boxplot

A comparação dos dois grupos pode ser realiza através do seguinte boxplot:

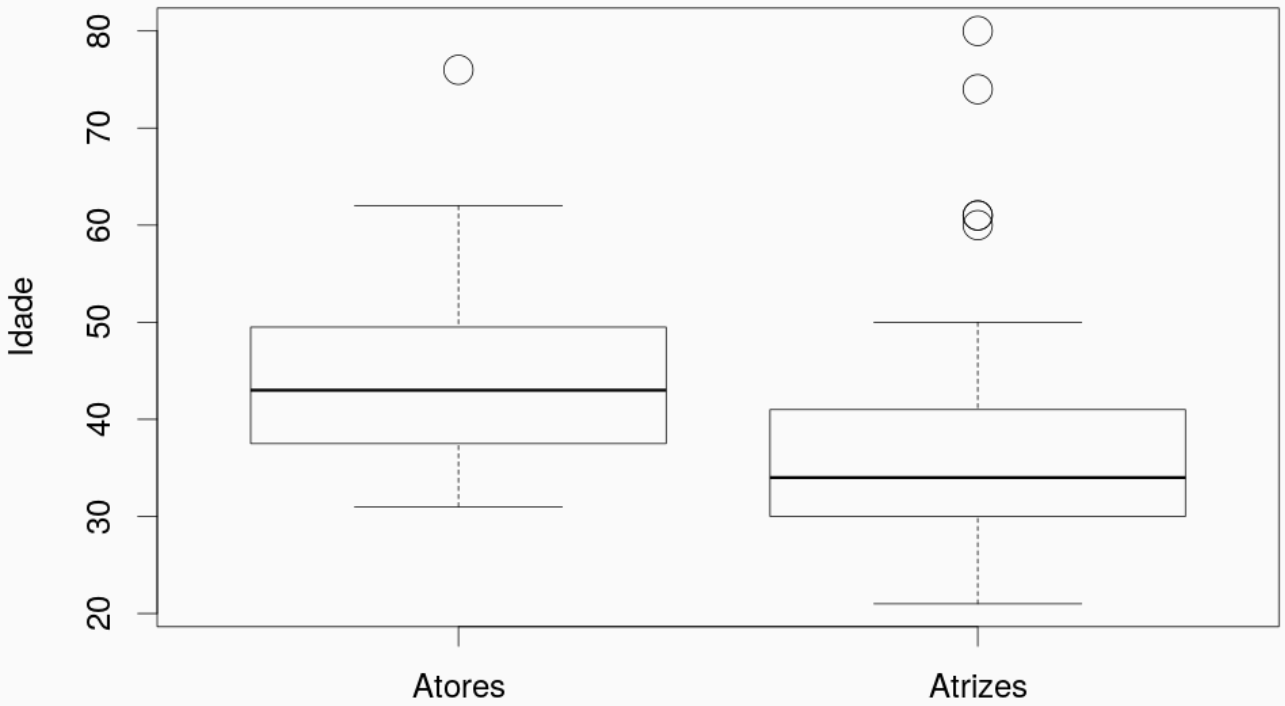


Figura 2: Boxplots para comparação das idades dos Atores e Atrizes.

Associação entre variáveis

Um dos principais objetivos de uma distribuição conjunta é descrever a associação que existe entre as variáveis.

Nesse caso, conhecer o grau de dependência entre as variáveis nos ajuda a predizer melhor o resultado de uma delas quando conhecemos a realização da outra.

Exemplos: a relação que existe ao compararmos indivíduos com respeito às seguintes características

- Sexo e peso;
- Idade e peso;
- País de moradia e peso (?), etc.

Associação entre variáveis categóricas

Exemplo: verifique se existe dependência entre sexo e carreira escolhida (com opções em "Economia" e "Administração").

	Sexo		
Carreira	Masculino	Feminino	Total
Economia	85 (61%)	35 (58%)	120 (60%)
Administração	55 (39%)	25 (32%)	80 (40%)
Total	140 (100%)	60 (100%)	200 (100%)

Tabela 1: Distribuição conjunta de "sexo" e "carreira escolhida" (1).

Visualmente, não parece existir diferença entre as proporção de indivíduos que escolhem "Economia" ou "Administração" com respeito à característica "sexo".

Associação entre variáveis categóricas

Exemplo: verifique se existe dependência entre sexo e carreira escolhida (com opções em "Física" e "Ciências Sociais").

	Sexo		
Carreira	Masculino	Feminino	Total
Física	100 (71%)	20 (33%)	120 (60%)
Ciências Sociais	40 (29%)	40 (67%)	80 (40%)
Total	140 (100%)	60 (100%)	200 (100%)

Tabela 2: Distribuição conjunta de "sexo" e "carreira escolhida" (2).

Visualmente, parece existir diferença entre as proporção de indivíduos que escolhem "Física" ou "Ciências Sociais" com respeito à característica "sexo".

Coeficiente de contingência

Exemplo: verifique se existe dependência entre uso de capacete e o tipo de lesão sofrida na cabeça.

	Uso de capacete		
Tipo de lesão	Sim	Não	Total
Grave	15 (40.5%)	22 (59.5%)	37 (100%)
Leve	45 (71.4%)	18 (28.6%)	63 (100%)
Total	60 (60%)	40 (40%)	100 (100%)

Tabela 3: Distribuição conjunta de "uso de capac." e "tipo de lesão".

Agora, ao invés de analisarmos *visualmente* a relação que existe entre as variáveis, vamos quantificar essa dependência através **coeficiente de contingência**.

Coeficiente de contingência

Antes de continuarmos, vamos tentar determinar as proporções esperadas no caso de **não** haver dependência entre as variáveis (notação: vamos determinar e_{ij}, $\forall i, j$[1]), onde $e_{ij} = \frac{n_i \cdot n_j}{n} = n_i \cdot f_j$. Assim:

$$e_{11} = \frac{37 \cdot 60}{100} = 22.2 \quad e_{12} = \frac{37 \cdot 40}{100} = 14.8$$
$$e_{21} = \frac{63 \cdot 60}{100} = 37.8 \quad e_{12} = \frac{63 \cdot 40}{100} = 25.2$$

	Uso de capacete		
Tipo de lesão	Sim	Não	Total
Grave	$e_{11} = 22.2$	$e_{12} = 14.8$	37
Leve	$e_{21} = 37.8$	$e_{11} = 25.2$	63
Total	60	40	100

[1] i linha e j coluna.

Coeficiente de contingência

Agora, vamos definir uma medida de afastamento das quantidades observadas em relação às quantidades esperadas:

$$\chi^2 = \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}},$$

onde o_{ij} diz respeito aos valores observados.

No nosso exemplo,

$$\begin{aligned}\chi^2 &= \frac{(15 - 22.2)^2}{22.2} + \frac{(22 - 14.8)^2}{14.8} + \frac{(45 - 37.8)^2}{37.8} + \frac{(18 - 25.2)^2}{25.2} \\ &= 2.33 + 3.50 + 1.37 + 2.06 = 9.26.\end{aligned}$$

O problema é que o número χ^2 não diz muita coisa por si só (Pelo menos nesse momento!).

Coeficiente de contingência

Para uma tabela 2×2 (como a que acabamos de analisar), o cálculo de χ^2 pode ser simplificado para:

$$\chi^2 = \frac{n \cdot (o_{11} \cdot o_{22} - o_{12} \cdot o_{21})^2}{(o_{11} + o_{12}) \cdot (o_{21} + o_{22}) \cdot (o_{11} + o_{21}) \cdot (o_{12} + o_{22})}.$$

Relembre a tabela:

$o_{11} = 15$	$o_{12} = 22$
$o_{21} = 45$	$o_{22} = 18$

Tabela 4: Tabela simplificada.

Então, temos que:

$$\chi^2 = \frac{100 \cdot (15 \cdot 18 - 22 \cdot 45)^2}{37 \cdot 63 \cdot 60 \cdot 40} \approx 9.26.$$

Coeficiente de contingência

Ao invés de olharmos para χ^2, vamos analisar o coeficiente de contingência C:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}},$$

onde, como já vimos, n é o tamanho da amostra.

Para o nosso exemplo, $C = \sqrt{\frac{9.26}{9.26+100}} = 0.29$.

Agora, C é um número entre 0 e 1. Se $C = 0$, então as variáveis **não** estão associadas; ao passo que, se C cresce, então a associação entre as variáveis começa a ficar mais forte.

Coeficiente de contingência

Não há consenso sobre a interpretabilidade de C; porém, de acordo com esse[2] artigo podemos utilizar a seguinte classificação:

1. $C \leq 0.1 \implies$ associação fraca;
2. $0.1 \leq C \leq 0.3 \implies$ associação moderada; e
3. $C > 0.3 \implies$ associação forte.

Ponto extra: valendo 2 pontos extras na 1ª prova (limitado a 100% do valor da prova), faça a atividade "Coeficiente de Contingência", disponibilizada na aba "Exercícios" do site, e entregue na próxima aula.

[2] Clique na palavra "esse" para acessar o link.

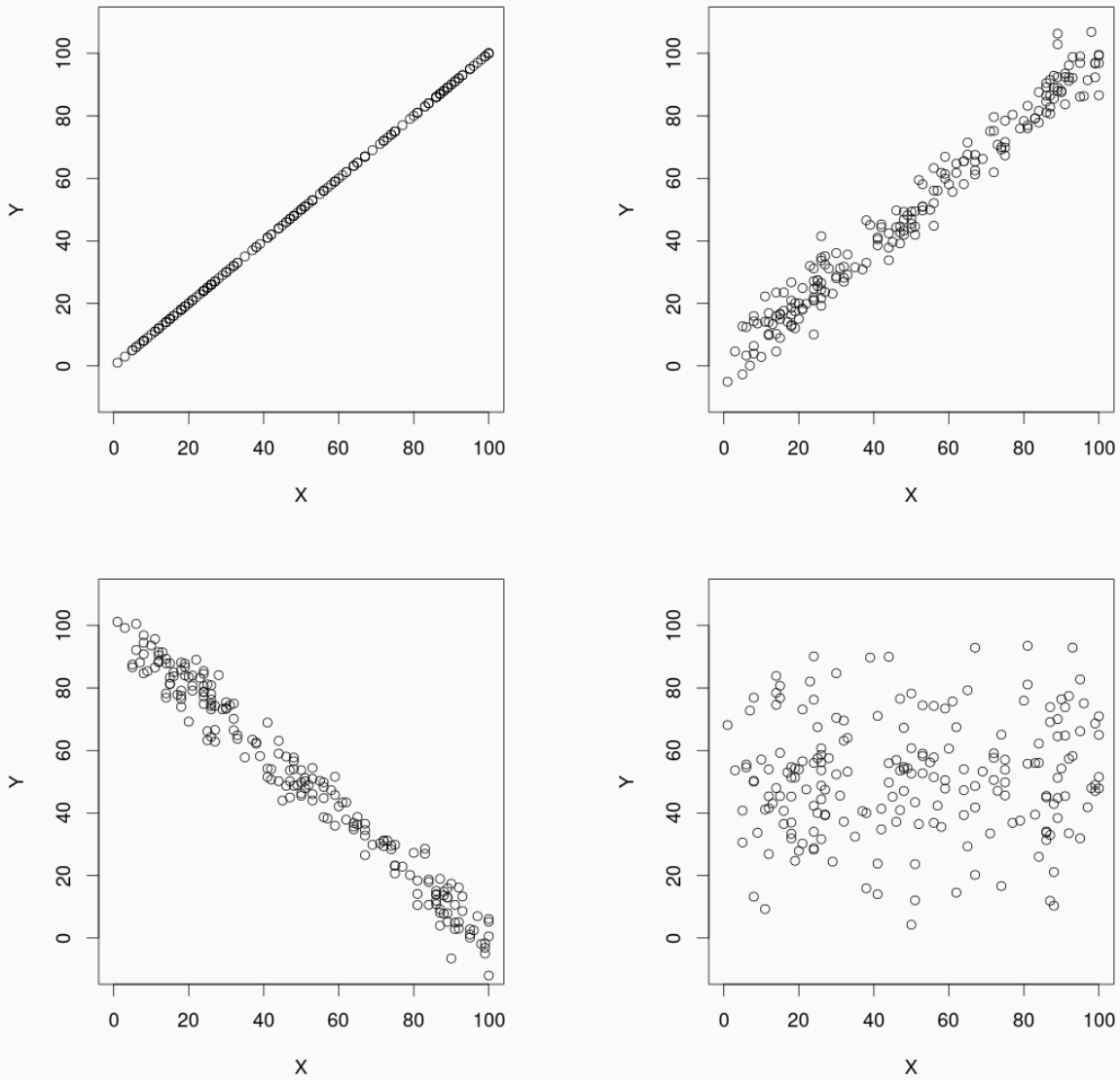
Associação entre variáveis quantitativa

Para analisar a associação que existe entre variáveis quantitativas, a nossa primeira ferramenta será o gráfico de dispersão.

O objetivo desse tipo de gráfico é tentar descobrir se existe relação entre duas variáveis quantitativas (por exemplo, X e Y).

No gráfico de dispersão, cada indivíduo da amostra é representado por um ponto de tal forma que o eixo horizontal represente seu valor para a variável X e o eixo vertical represente o seu valor para a variável Y.

Gráfico de dispersão



Coeficiente de Correlação de Pearson

Mais uma vez, ao invés de nos prendermos, somente, à interpretação visual dos gráficos, vamos quantificar a medida correlação entre duas variáveis X e Y. Para isso, considere o "Coeficiente de Correlação de Pearson".

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_{i=1}^{n}(x_i - \bar{x})^2)(\sum_{i=1}^{n}(y_i - \bar{y})^2)}},$$

onde r é o Coeficiente de Correlação de Pearson **amostral**.

Similarmente, poderíamos definir o Coeficiente de Correlação de Pearson **populacional** como:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{(\sum_{i=1}^{n}(x_i - \mu_x)^2)(\sum_{i=1}^{n}(y_i - \mu_y)^2)}} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\cdot\mathrm{Var}(Y)}}^{\star}.$$

Coeficiente de Correlação de Pearson

Observações importantes

1. $-1 \leq r \leq 1$, de modo que:
   - se $r \approx +1 \rightarrow$ forte correlação **linear** positiva entre X e Y;
   - se $r \approx -1 \rightarrow$ forte correlação **linear** negativa entre X e Y; e
   - se $r \approx \ 0 \rightarrow$ não existe correlação **linear** entre X e Y.
2. Uma possível classificação mais refinada para essa medida é dada por:
   - se $0 \leq |r| < 0.4$, então existe correlação fraca;
   - se $0.4 \leq |r| < 0.7$, então existe correlação moderada;
   - se $0.7 \leq |r| < 1$, então existe correlação forte; e
   - se $|r| = 1$, então existe correlação perfeita.

Coef. de Correlação de Pearson – Exercício

Considere o conjunto de dados abaixo, que apresenta o autonomia (em mi/gal) de $n = 10$ automóveis e o peso (em centenas de libras) de cada um desses veículos.

X	Peso	29	35	28	44	25	34	30	33	28	24
Y	Autonomia	31	27	29	25	31	29	28	28	28	33

Calcule o Coeficiente de Correlação de Pearson.

Coef. de Correlação de Pearson – Exercício

Considere o conjunto de dados abaixo, que apresenta o autonomia (em mi/gal) de $n = 10$ automóveis e o peso (em centenas de libras) de cada um desses veículos.

X	Peso	29	35	28	44	25	34	30	33	28	24
Y	Autonomia	31	27	29	25	31	29	28	28	28	33

Calcule o Coeficiente de Correlação de Pearson.

Aqui, $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = -32$, $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 306$ e $\sum_{i=1}^{n}(y_i - \bar{y})^2 = 46.9$.

Assim, $r = \frac{-102}{\sqrt{306 \cdot 46.9}} = -0.851$.

Coef. de Correlação de Pearson – Exercício

Figura 3: Gráfico de dispersão para "Peso" e "Autonomia".

Correlação e causalidade

Alto coeficiente de correlação **não** se traduz, necessariamente, em causalidade.

Correlação indireta: existe correlação entre duas variáveis, mas isso é justificado por uma terceira variável (que pode ser conhecida ou não).

Exemplo: Número de palavras que uma criança conhece com a altura desse(a) menino(a).

Correlação espúria: $|r|$ é alto, mas não existe relação alguma entre as variáveis X e Y.

Exemplo: Em certa região da Europa registrou-se aumento de avistamento de cegonhas e aumento da taxa de natalidade.