

BIOESTATÍSTICA

Parte 1 - Estatística descritiva e análise exploratória dos dados

Aulas Teóricas de 17/02/2011 a 03/03/2011

1.1. População, amostra e dados estatísticos. Dados qualitativos e quantitativos

Dados Qualitativos:

- Escala nominal – não existe relação entre categorias. Ex: solteiro, casado, viúvo (...)
- Escala ordinal – existe relação entre categorias. Ex: forte, médio, fraco

1.2. Organização e representação gráfica de dados

Box-Plot (caixa com Bigodes)

O conjunto dos valores da amostra compreendidos entre o 1º e o 3º quartis é representado por um retângulo com a mediana indicada por uma barra. Desenham-se então as linhas que unem o meio [...]

$$\text{Barreira inferior} \Rightarrow BI = Q_{\frac{1}{4}} - 1,5 (Q_{\frac{3}{4}} - Q_{\frac{1}{4}})$$

$$\text{Barreira superior} \Rightarrow BS = Q_{\frac{3}{4}} + 1,5 (Q_{\frac{3}{4}} - Q_{\frac{1}{4}})$$

Dizemos que uma observação é um outlier quando não está compreendida no intervalo [BI, BS]. Na representação box-plot, os outliers assinalam-se com *.

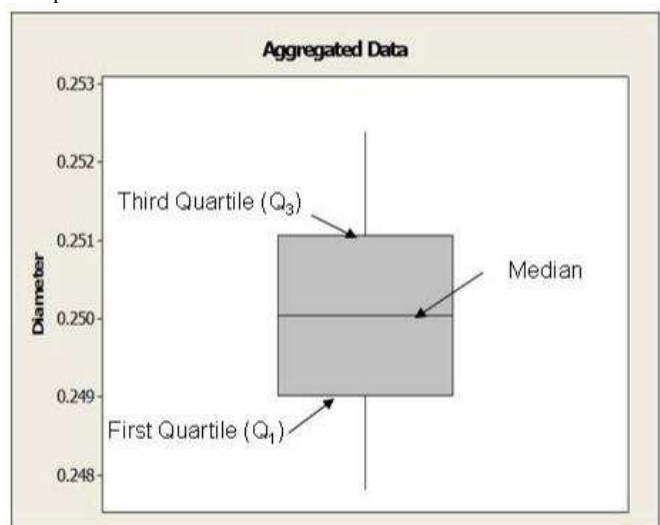
$$\text{Barreira (exterior) inferior} \Rightarrow Q_{\frac{1}{4}} - 3 (Q_{\frac{3}{4}} - Q_{\frac{1}{4}})$$

$$\text{Barreira (exterior) superior} \Rightarrow Q_{\frac{3}{4}} + 3 (Q_{\frac{3}{4}} - Q_{\frac{1}{4}})$$

As observações com valores inferiores ou superiores às barreiras externas são outliers severos e são representados em box-plot com o símbolo ☒.

Quando não há outliers a box-plot coincide com o diagrama de extremos e quartis.

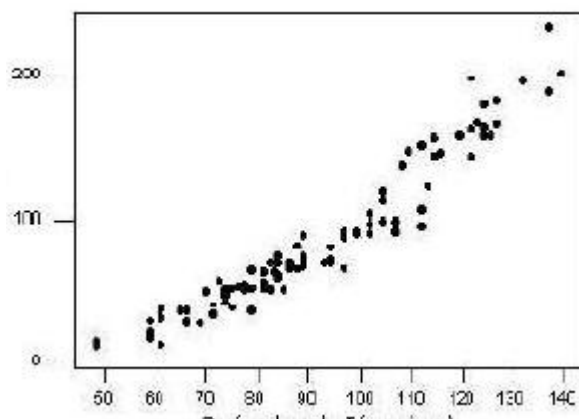
A box-plot realça:



- Centro da amostra
- Variabilidade
- Simetria
- Existência de outliers
- Útil na comparação de amostras

Diagrama de Dispersão

É uma representação gráfica adequada para dados bivariados quantitativos, em que cada par de valores (x_i, y_i) é representado por um ponto de coordenadas (x_i, y_i) num sistema de eixos coordenados. É muito útil, pois permite realçar algumas características importantes dos dados, nomeadamente a existência de alguns tipos de associação entre variáveis.



1.3. Características amostrais

Medidas de Localização: localizam o centro da amostra

- Média
- Mediana
- Moda
- Quantis

Medidas de Dispersão: medem a variabilidade dos dados

- Desvio Padrão
- [...]

Medidas de Localização

Média (\bar{x})

[Média aparada – retiram-se uma determinada % dos extremos]

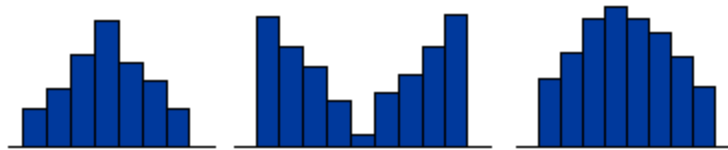
É uma medida pouco resistente, pois é muito sensível à existência de valores muito grandes ou pequenos.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

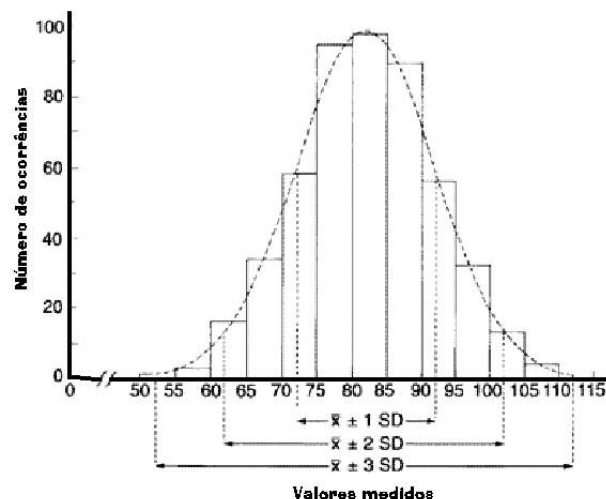
No caso de dados em intervalos não se pode obter um valor exacto, apenas um valor aproximado:

[multiplica-se a frequência absoluta da classe pela sua mediana e divide-se pelo número de classes]

A média é uma boa medida de localização do centro da amostra quando a distribuição dos dados for simétrica.



É a melhor medida de localização quando a distribuição dos dados é normal.



Mediana (M)

Ordenados os elementos da amostra, a mediana é o valor (pertencente ou não à amostra) que a divide ao meio, ou seja 50% dos elementos estão acima da mediana e 50% abaixo.

- Se n é *ímpar*, a mediana é a observação central da amostra

$$M = x_{\frac{n+1}{2}:n}$$

- Se n é *par*, a mediana é a média aritmética das 2 observações centrais

$$M = \frac{\frac{x_{\frac{n}{2}:n} + x_{\frac{n+1}{2}:n}}{2}}{2}$$

Ex: 10, 10, 10, 11, 11, 11, 11, 11, 12 $M = 11$ $\bar{x} = 10,75$

Ex: 10, 10, 11, 11, 11, 11, 11, 12, 18 $M=11$ $\bar{x} = 11,75$

A mediana é uma medida mais resistente que a média, pois não é tão sensível quanto à existência de valores extremos.

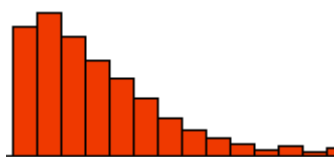
Centro da amostra:

A média é o centro de gravidade da distribuição dos dados (a média reflecte o valor de todas as observações)

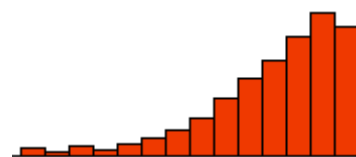
A mediana é o centro posicional da distribuição dos dados.

A assimetria da amostra pode estudar-se considerando a posição relativa das medidas de localização média e mediana, visto que quando a distribuição dos dados:

- É simétrica \Rightarrow média = mediana
- É assimétrica positiva \Rightarrow média > mediana
- É assimétrica negativa \Rightarrow média < mediana



2. Assimetria Positiva



1. Assimetria Negativa

Quantil de ordem p (quantil empírico [ou amostral])

Ordenada a amostra, o quantil de ordem p ($0 < p < 1$) é o valor Q_p tal que $100\%p$ dos elementos da amostra são inferior ou iguais a Q_p e os restantes $100(1-p)\%$ elementos da amostra são superiores ou iguais a Q_p

$$\text{Se } np \text{ é inteiro} \Rightarrow Qp = \frac{x_{np+1:n} + x_{np:n}}{2}$$

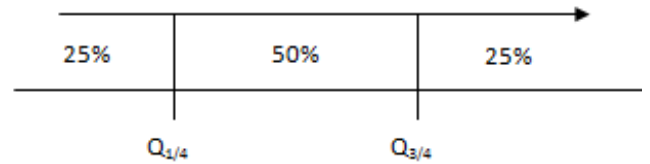
$$\text{Se } np \text{ não é inteiro} \Rightarrow Qp = x_{[np]+1:n}$$

Onde $[k]$ representa a ordem inteira de k

$Q_{1/2}$ – mediana

$Q_{1/4}$ – 1º Quartil

$Q_{3/4}$ – 3º Quartil



Medidas de Dispersão

$$\left. \begin{array}{l} \text{Ex1: 15, 15, 15, 15, 15} \\ \text{Ex2: 10, 13, 15, 17, 20} \\ \text{Ex3: 0, 7, 15, 23, 30} \end{array} \right\} \bar{x} = 15 \quad [\text{mas a medida de dispersão é diferente}]$$

Variância e Desvio Padrão (amostrais)

São medidas de dispersão de dados, relativamente ao centro da amostra que é a média.

$$\text{Variância: } s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Para obter uma medida de dispersão que esteja expressa na mesma unidade de dispersão que os dados, consideramos o desvio padrão:

$$\text{Desvio Padrão: } s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\text{Ex1: } s = 0$$

$$\text{Ex2: } s = 3,8$$

$$\text{Ex3: } 12,0$$

Quanto maior o desvio padrão (s), maior será a dispersão dos dados? Não necessariamente. Um valor grande de desvio padrão pode ser devido à grande dispersão dos dados ou à existência de outliers. O desvio padrão é muito sensível à existência de outliers, é portanto uma medida de dispersão pouco resistente.

$$\text{Amplitude (amostral): } R = x_{n:n} - x_{1:n}$$

$$[R = \text{máximo} - \text{mínimo}]$$

É a medida de variabilidade mais simples, mas é muito sensível à existência de uma observação muito grande ou muito pequena.

Amplitude interquartil: $AIQ = Q_{\frac{3}{4}} - Q_{\frac{1}{4}}$

Esta é a amplitude que contém 50% das observações no centro da amostra

Ex: 1,2,3,3,3,3,3,4,5

n=10

$$Q_{\frac{1}{4}} = x_{3:10} = 3$$

$$Q_{\frac{3}{4}} = x_{8:10} = 3$$

s=1,5

$$AIQ = 0$$

A amplitude inter-quartil é uma medida mais resistente que o desvio padrão. No entanto, o desvio padrão reflecte o conjunto dos dados e a amplitude inter-quartil não.

[s e AIQ expressam informações diferentes e devem por isso ser utilizados em conjunto]

Dispersão Relativa

Para acompanharmos vários conjuntos de dados convém utilizar uma medida que quantifique a variabilidade dos dados relativamente à localização.

Uma medida de dispersão relativamente muito usada para dados positivos é o coeficiente de variação [expresso em %].

$$CV = \frac{s}{\bar{x}} \times 100\%$$

[É um desvio padrão “normalizado”, útil para comparações]

Profundidade

A profundidade é um indicador de quão interior à amostra é um certo valor. O máximo e o mínimo da amostra (extremos) têm profundidade 1.

Dada uma certa amostra x_i, \dots, x_n , a ordem de uma observação pode ser definida como:

- Ordem ascendente – posição da observação na amostra, a partir do menor valor
- Ordem descendente – posição da observação na amostra, a partir do maior valor

Profundidade de uma observação é o mínimo das suas 2 ordens ascendente e descendente.

n = 23

8+8+7 = 23

{

3	5	027
8	6	14589
(8)	7	11236799
7	8	0145
3	9	367

[Profundidade máxima das observações na linha]

}

[Na linha da mediana escreve-se o número de folhas da linha

Ex:

elemento 98: profundidade 1

elemento 61: profundidade 4

elemento 93: profundidade 3

A profundidade máxima é dada por $\left\lceil \frac{n+1}{2} \right\rceil$

1.4. Associação entre variáveis. Correlação. Relação: regra dos mínimos quadrados

Dados Bivariados

Há muitas situações em que se pretende estudar 2 variáveis em simultâneo; interessa portanto conhecer a sua variabilidade conjunta.

Estudar a relação entre:

- Idade e pressão arterial;
- Concentração de uma droga injectável e a frequência cardíaca;
- Peso e nível de glicose no sangue em adultos saudáveis.

A população em estudo surge então sob a forma de pares de valores, i.e., cada indivíduo ou resultado experimental contribui com um conjunto de 2 valores. Uma amostra de dimensão n será então $(x_i, y_i), \dots, (x_n, y_n)$.

- **Correlação** – [2 variáveis em pé de igualdade. Procura-se saber se à alguma relação entre as variáveis em estudo]
- **Regressão** – [uma das variáveis está sob o controlo do investigador – variável controlada, independente – (ex. concentração de droga administrada) e uma das variáveis é a resposta – variável resposta ou dependente – (ex. frequência cardíaca)]

Para organizar a informação correspondente a dados bivariados utiliza-se uma tabela de contingências. É adequada para dados qualitativos ou quantitativos, sendo necessário este caso agrupar os dados em classes.

Exemplo:

n = 1000 habitantes

	Não vacinados	1 dose	2 doses
Gripe	24	9	13
Sem gripe	289	100	565

[representação gráfica \Rightarrow diagrama de dispersão]

Dada uma amostra bivariada de dimensões $(x_i, y_i), \dots, (x_n, y_n)$, define-se uma medida de variabilidade conjunta existente entre as variáveis em estudo que é a **covariância amostral**.

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

- Se houver associação linear positiva [variáveis variam no mesmo sentido; x elevado corresponde a y elevado] entre as variáveis, predominam parcelas tais que $(x_i - \bar{x})(y_i - \bar{y}) \geq 0 \Rightarrow$ covariância positiva
- Se houver associação linear negativa [variáveis variam em sentido oposto] entre as variáveis, predominam parcelas tais que $(x_i - \bar{x})(y_i - \bar{y}) \leq 0 \Rightarrow$ covariância negativa

Grande inconveniente: depende da unidade de medida usada [Kg/m e Kg/cm – a covariância vem com valores diferentes]

[só medem associação linear]

Coeficiente de Correlação Amostral ou Coeficiente de Correlação de Pearson

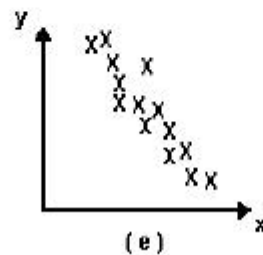
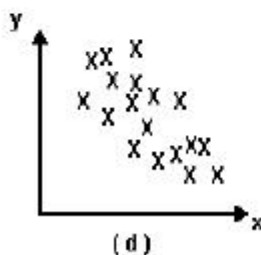
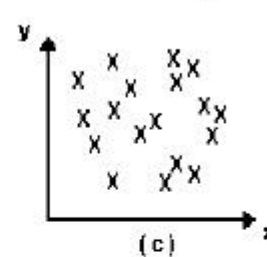
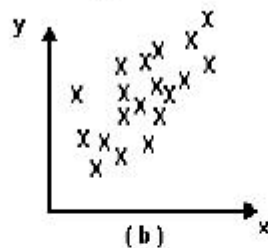
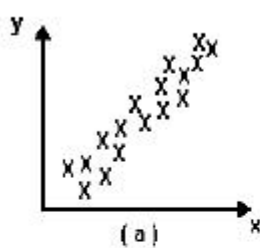
A medida mais usada para medir o grau de associação linear entre 2 variáveis quantitativas é o coeficiente de correlação dado por:

$$r = \frac{\text{cov}(x, y)}{s_x \times s_y}$$

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

- O Coeficiente de correlação não depende de medidas das variáveis;
- r toma valores no intervalo [-1, 1];
- O valor do coeficiente de correlação mede a intensidade da associação linear entre as variáveis;
- Quando $r = 0$ não existe associação linear entre as variáveis, mas elas podem estar associadas não linearmente;
- O coeficiente de correlação não é uma medida resistente [está dependente da média].

Possíveis Padrões para Diagramas de Dispersão.



Legenda

- (a) - Elevada correlação positiva
- (b) - Moderada correlação positiva
- (c) - Ausência de correlação
- (d) - Moderada correlação negativa
- (e) - Elevada correlação negativa

O coeficiente de correlação pode ser calculado como:

$$r = \frac{\sum_{i=1}^n x_i \times y_i - n\bar{x} \times \bar{y}}{(n-1)s_x \times s_y}$$

Correlação elevada indica apenas associação estatística. É um erro grave confundir associação estatística com causalidade.

[Exemplo: O consumo de gelados e os fogos florestais apresentam uma relação de associação estatística, não de causalidade]

[Podem existir variáveis de confundimento]

Regressão

Uma das variáveis está sob o controlo do investigador e a outra variável pode ser considerada uma “resposta”.

$x \Rightarrow$ variável controlada, independente ou explicativa

$y \Rightarrow$ variável resposta ou dependente

Quando o diagrama de dispersão realça a existência de associação linear entre 2 variáveis quantitativas x e y é possível resumir através de uma recta a forma como uma variável resposta y é influenciada por uma variável explicativa x .

Recta de Regressão

Pretende-se exprimir a variável resposta como função da variável independente através de um modelo de regressão \Rightarrow regressão linear simples

Os valores observados podem ser escritos como:

$$y_i = \hat{y}_i + \varepsilon_i = ax_i + b + \varepsilon_i$$

[Resíduos:

Idealmente $\varepsilon_i = 0$, quando os pontos pertencem todos à recta;
são quantidades aleatórias]

Recta dos mínimos quadrados

Para que um conjunto de dados $(x_i, y_i), \dots, (x_n, y_n)$ que seguem um padrão linear vamos ajustar uma recta na forma $\hat{y} = ax + b$. Um dos métodos mais conhecidos para ajustar uma recta aos dados é usar o critério dos mínimos quadrados.

Determinar os coeficientes da recta de modo que a soma dos quadrados dos desvios entre os valores observados y_i e os valores \hat{y}_i obtidos a partir da recta que se pretende ajustar seja mínima, ou seja, minimizar a soma dos quadrados dos resíduos.

$$\sum_{i=1}^n e_1^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Determinar os coeficientes a e b tais que $Q(a,b) = \sum_{i=1}^n (y_i - ax_i - b)^2$ seja mínimo.

O declive da recta da regressão é

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

E a ordenada na origem é $b = \bar{y} - a\bar{x}$.

A recta dos mínimos quadrados é uma medida pouco resistente.

A recta de regressão é frequentemente utilizada para fazer predição. Para um novo valor da variável x que não foi usada na construção da recta, pretende-se estimar o valor de y que lhe corresponderá.

[predição \Rightarrow considerar um valor de x que não é amostra e calcular o y esperado através da recta dos mínimos quadrados]

Uma forma de verificar o modelo ajustado é adequado é através dos resíduos.

Se o modelo ajustado for adequado:

- Os resíduos devem dispor-se aleatoriamente, sem um padrão definido, em torno de zero.
- É de esperar que o resíduo quadrado médio seja moderado.

No contexto da regressão, os outliers são valores a que correspondem grandes resíduos.

Note-se que, teoricamente, a soma dos resíduos quadrados é zero. Na prática, devido aos arredondamentos, a soma pode não ser nula, mas terá sempre um valor muito próximo de zero.

O valor de r^2 representa a proporção da variabilidade de y que é explicada pela recta de regressão; é uma medida de precisão da recta.

r^2 designa-se por coeficiente de determinação.

$$[-1 \leq r \leq 1 \Rightarrow 0 \leq r^2 \leq 1]$$

$$\text{Resíduos} = y_{\text{observado}} - y_{\text{predicto}}$$