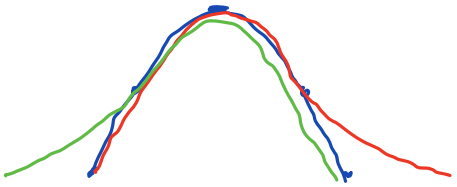
- Última aula: variáveis quantitativas –
  - indicadores amostrais ($x = (x_1, \ldots, x_n)$)
  - Medidas de localização: média ($\bar{x}$), ...
  - Medidas de dispersão: variância (s_x^2), ...
- Medidas de forma

→ Coeficiente de assimetria (skewness coefficient, sc)



simetria: $\bar{x} = me$ (= mo em dist. unimodais)

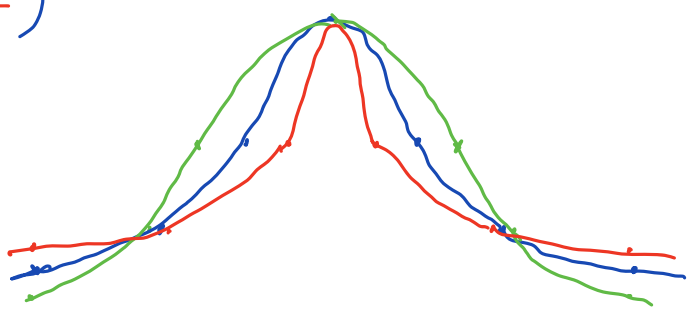
assimetria > 0: $me < \bar{x}$

assimetria < 0: $\bar{x} < me$

$$sc = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{3/2}}$$

→ Coeficiente de achatamento ou curtose (kurtosis coefficient, kc)

$$kc = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2}$$

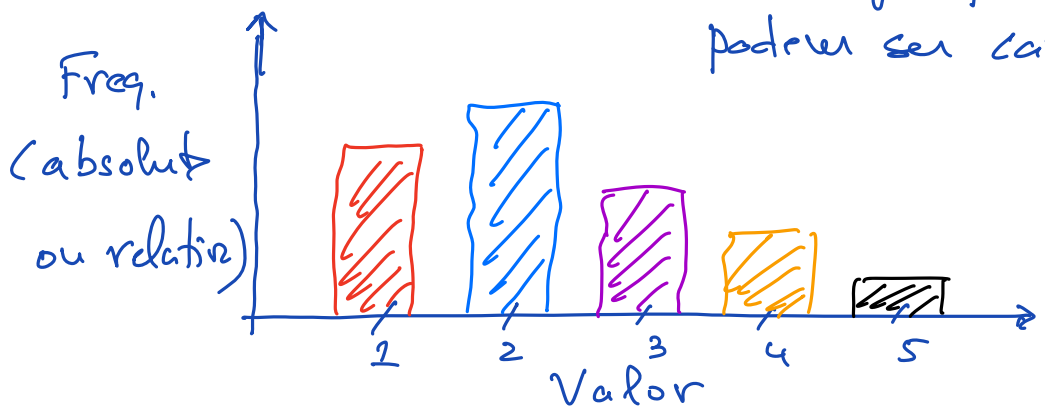


- $kc = 3$ ⇒ achatamento da distribuição normal (mesocúrtica)
- $kc > 3$ ⇒ distribuição menos achatada (leptocúrtica)
- $kc < 3$ ⇒ distribuição mais achatada (platicúrtica)

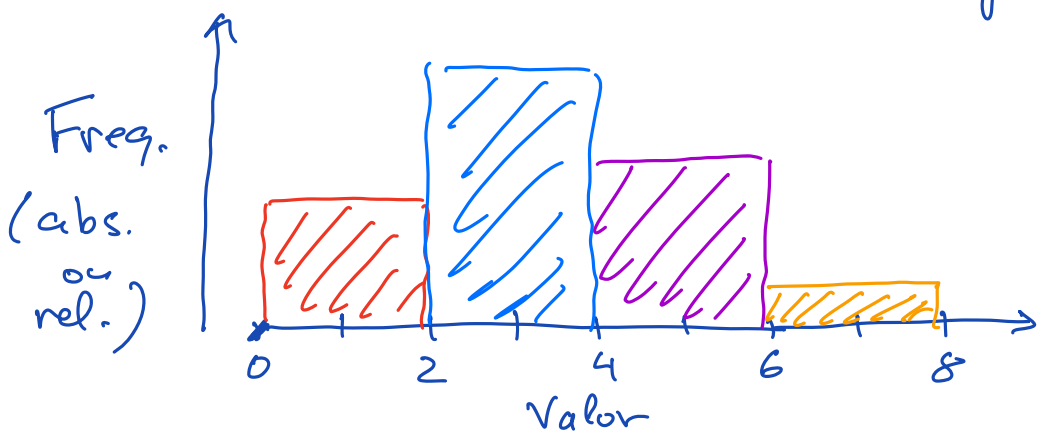
- # Algumas formas de visualização de dados

1) Diagrama de barras e Histograma
2) Caixa de bigodes (Boxplot)
3) Função distribuição empírica ($F_n(x)$)

1) **Diagrama de barras** → para dados de natureza discreta, com um nº pequeno de valores distintos (dados não agrupados) podem ser categóricos)



Histograma → para dados de natureza contínua ou com um nº elevado de valores distintos (dados agrupados)

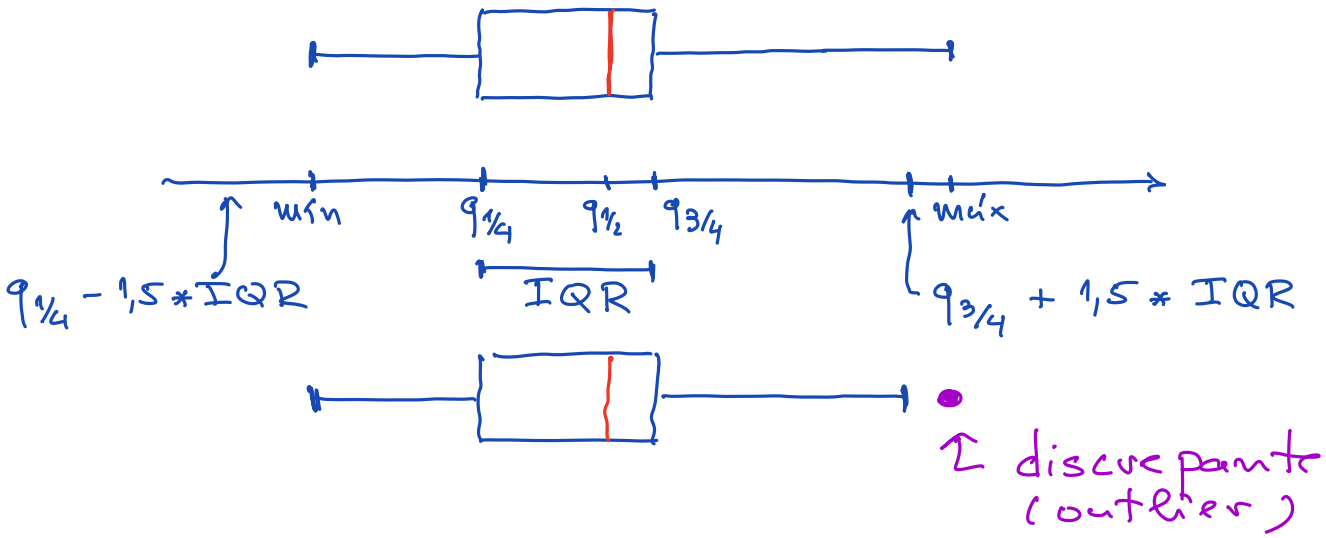


Regra de Sturges para o nº de classes (k) em termos da dim. da amostra (n): $k \approx 1 + \log_2(n)$

$= 1 + \log(n)/\log(2)$

2) Caixa de bigodes (Boxplot)

→ Fornece informações sobre posição, dispersão, assimetria, caudas e valores discrepantes (outliers)



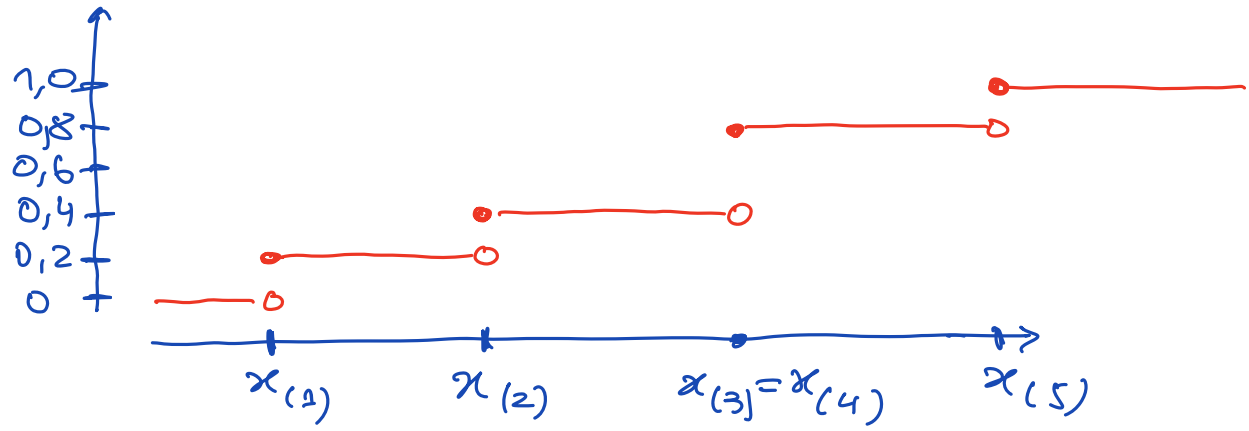
→ Método de Tukey: a observação x é classificada como discrepante (outlier) se

$$x < q_{1/4} - k \times IQR \quad \text{ou} \quad x > q_{3/4} + k \times IQR$$

em que os valores mais usados de k são 1.5 e 3.

3) Função distribuição empírica ($F_n(x)$)

$$F_n : \mathbb{R} \to \mathbb{R}, \quad F_n(x) = \frac{\#\text{observações} \leq x}{n}$$



- **Amostra bivariada**: duas variáveis observadas em cada unidade estatística $(x,y)=((x_1,y_1),\dots,(x_n,y_n))$

Objetivo: por em evidência uma potencial associação linear entre x e y.

→ **Covariância** de x e y (s_{xy}):

$$s_{xy}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}=\frac{\left(\sum_{i=1}^{n}x_iy_i\right)-n\bar{x}\bar{y}}{n-1}$$

- $s_{xy}=0 \Rightarrow$ x e y não associados linearmente
- $s_{xy}>0 \Rightarrow$ associação linear positiva
- $s_{xy}<0 \Rightarrow$ " " negativa

→ **Coeficiente de correlação linear** de x e y (r_{xy}):

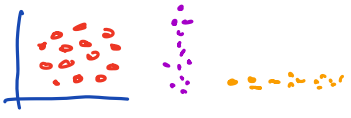
$$r_{xy}=\frac{s_{xy}}{s_x\cdot s_y} \qquad (-1\le r_{xy}\le 1)$$

Adimensional. Se $y=mx+b$, e portanto $y_i=mx_i+b$, $\forall i$, então $s_{xy}=m(s_x)^2$ e $s_y=|m|s_x$ pelo que $r_{xy}=\begin{cases}1, & \text{se } m>0\\ -1, & \text{se } m<0\end{cases}$

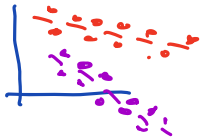
Assim:

- $r\approx 1 \Rightarrow$ pontos próximos de reta c/ declive >0

- $r\approx 0 \Rightarrow$ nuvem de pontos arredondada ou alongada segundo um eixo



- $r\approx -1 \Rightarrow$ pontos próximos de reta c/ declive <0



- R/RStudio: algum matrial adicional disponivel em www.math.tecnico.ulisboa.pt/rmabreu/PE