Estatística Aplicada

A. Rita Gaio
Departamento de Matemática - FCUP
argaio@fc.up.pt

March 19, 2020

Índice Geral

Chapter 1

Estatística Descritiva

1.1. Amostragem

Frequentemente precisamos de obter conclusões acerca de um grande grupo de indivíduos ou objectos (chamado **população**). Por escassez de tempo, custos elevados, ou outras razões, não é possível analisar todo o grupo e decidimos escolher apenas uma pequena parte (**amostra**) dessa população. Se for razoável supor que essa amostra representa convenientemente a população, o problema reside em extrair conclusões acerca das características desconhecidas da população com base nas informações observadas na amostra. O processo através do qual se obtêm conclusões acerca de uma população com base em resultados observados numa amostra dessa população é designado por **inferência estatística**. O processo de obtenção ou extração de amostras chama-se **amostragem**. Como a inferência de uma amostra em relação à população correspondente não pode ser considerada absolutamente certa, devemos utilizar a linguagem das probabilidades na formulação de qualquer conclusão.

Exemplos:

1. Pretende-se averiguar a percentagem de comprimidos defeituosos produzidos por um laboratório durante uma semana (cinco dias úteis) examinando apenas 50 comprimidos por dia, fabricados em diferentes períodos do dia. Neste caso, a população consiste de todos os comprimidos fabricados pelo laboratório durante uma semana e a amostra consiste dos 250 comprimidos que foram sendo recolhidos ao longo da semana.

2. Pretende-se conhecer os pesos de todos os estudantes inscritos na FCUP num dado ano lectivo (população) apenas a partir de dados recolhidos em 150 estudantes (amostra) inscritos nesse ano lectivo.

3. Pretende-se averiguar a qualidade de uma moeda (percentagem de caras e coroas) jogando-a repetidamente. A população (infinita) consiste de todas os lançamentos possíveis da moeda. Uma amostra consistiria, por exemplo, dos primeiros 50 lançamentos.

A população pode ser finita ou infinita. O **tamanho da população**, usualmente designado por N, é o número de elementos da população. Analogamente, o **tamanho da amostra**, usualmente designado por n, é o número de elementos da amostra.

A amostragem pode ser feita **com reposição** ou **sem reposição**. No primeiro caso, um elemento da população pode aparecer várias vezes numa amostra enquanto que na amostragem sem reposição essa situação é impossível.
Uma população finita submetida a um processo de amostragem com reposição pode, teoricamente, ser considerada como infinita pois podemos extrair amostras de qualquer tamanho sem nunca esgotarmos a população. Em termos práticos, a amostragem de uma população finita muito grande pode ser considerada como amostragem de uma população infinita.

Um dos problemas da teoria de amostragem consiste na escolha dos indivíduos que vão fazer parte da amostra. Nem todas as amostras são representativas da população e portanto nem todas as amostras podem ser usadas para inferir acerca da população. É preciso certificarmo-nos de que cada elemento da população (suposta finita) tem a mesma probabilidade de figurar na amostra. Quando esta condição é satisfeita diz-se que a amostra é uma **amostra aleatória**. Equivalentemente, uma amostra de tamanho n de uma população de tamanho N diz-se uma amostra aleatória se qualquer outra amostra possível de tamanho n tem a mesma probabilidade de ser seleccionada.

Diz-se que uma variável aleatória é conhecida quando se conhece a sua função (densidade) de probabilidade, f. Para isso, basta conhecer os parâmetros dessa distribuição . Exemplos de parâmetros são μ e σ^2 no caso da distribuição normal, n e p no caso da binomial ou λ no caso da Poisson. Na maior parte das vezes, a função (densidade) de probabilidade não é conhecida com precisão mas, felizmente, tem-se uma ideia qualitativa do seu comportamento. Nesse caso pretende-se estimar os parâmetros da distribuição.
Suponhamos, por exemplo, que uma amostra de 150 alunos da FCUP nos faz pensar (embora não tenhamos toda a certeza) que a variável aleatória X que representa o peso dos alunos da FCUP segue uma distribuição normal. O passo seguinte consiste na **estimação da média e da variância de X (na população) a partir das informações constantes na amostra**.

Para obter uma amostra de 150 alunos, começamos por escolher aleatoriamente um aluno de entre toda a população (neste caso, todos os alunos inscritos na FCUP no ano lectivo em causa) e registamos o seu peso, digamos x_1. Consideramos então a v.a.

$$X_1$$

que representa o peso do primeiro aluno seleccionado da população de todos os alunos inscritos na FCUP. O valor x_1 é apenas uma **realização** de X_1. Se tivessemos escolhido um outro indivíduo teríamos obtido possivelmente um valor diferente de x_1.
Em seguida escolhemos, aleatoriamente e de entre a população, o segundo indivíduo e voltamos a registar o seu peso, digamos x_2. Mais uma vez, o valor de x_2 não é previsível e poderia ter sido um outro qualquer de entre muitos possíveis. Seja então

$$X_2$$

a v.a. que representa o peso do segundo aluno seleccionado da população . Tal como acima, x_2 é uma realização de X_2.
Iteramos este processo até chegarmos à escolha do indivíduo número 150. O peso desse indivíduo é x_{150} e corresponde à realização da v.a.

$$X_{150}$$

que representa o peso do aluno número 150 a ser seleccionado de toda a população.

Mais geralmente, suponhamos que amostramos com reposição n elementos de uma população de tamanho N, com $n << N$ (o facto de se ter $n << N$ implica que a amostragem sem reposição daria sensivelmente os mesmos resultados).
Para $i = 1, \ldots, n$, seja

$$X_i$$

a v.a. que representa o valor da característica que está a ser avaliada no elemento i. As v.a.

$$X_1, X_2, \ldots, X_n$$

são **independentes** e **identicamente distribuídas** (*i.e.*, qualquer X_i segue a mesma distribuição). A sua função (densidade) de probabilidade conjunta é então

$$\begin{aligned} f_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) &= f_{X_1}(x_1) \ldots f_{X_n}(x_n) \\ &= f(x_1) \ldots f(x_n). \end{aligned}$$

Ao conjunto $\{X_1, X_2, \ldots, X_n\}$ chama-se **amostra aleatória**; ao conjunto dos valores observados na amostra

$$\{x_1, x_2, \ldots, x_n\}$$

chama-se uma **realização da amostra aleatória** (é o valor observado de $(X_1, \ldots, X_n)$ para aquela amostra em particular). Assim, dependendo da amostra aleatória escolhida, podemos observar diferentes realizações de $X_1, X_2, \ldots, X_n$:

$$x_1, x_2, \ldots, x_n, \quad x'_1, x'_2, \ldots, x'_n, \quad x''_1, x''_2, \ldots, x''_n, \quad \ldots$$

1.2. Medidas de localização de uma a.a.

Dados: realização de um a.a. $x_1, x_2, \ldots, x_n$

Média *Mean*	$\overline{x} = \frac{1}{n}\sum_i x_i$ [a] $= \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$
Mediana *Median*	n ímpar: observação de ordem $\frac{n+1}{2}$ n par: média das observ. de ordem $\frac{n}{2}$ e $\frac{n}{2}+1$
Moda *Mode*	observação que ocorre mais vezes

Observações:

(a) Nas conclusões de um estudo, deve sempre mencionar-se o tamanho da amostra, n.

(b) A média, a mediana e a moda exprimem-se nas unidades da v.a. correspondente. É comum apresentar estas estatísticas com uma casa decimal a mais do que os dados.

(c) Para o cálculo da mediana, as observações devem ser ordenadas por ordem crescente. A mediana corresponde ao valor para o qual metade das observações têm valores inferiores (e metade têm valores superiores).

(d) A mediana é insensível a valores extremos[a] da amostra (muito altos ou muito baixos).

(e) A média é muito sensível a valores extremos: valores muito baixos (resp. muito altos) fazem descer (resp. subir) a média, afastando-a do centro da amostra. Existindo muitos valores extremos é preferível apresentar a mediana à média.

(f) Quanto mais simétrica (em relação à mediana) for a função de frequências das observações, mais próximos serão os valores da média e da mediana. Neste caso é indiferente apresentar a média ou a mediana.

Nos casos em que a distribuição é assimétrica, é comum apresentar-se a mediana e não a média, porque a média é influenciada pelas caudas da distribuição e portanto fornece menos eficazmente informação sobre a localização das observações.

(g) Se adicionarmos uma constante a cada uma das observações, ou multiplicarmos cada observação por uma constante, quer a média quer a mediana são alteradas dessa constante.

(h) A moda é uma estatística usada essencialmente em dados discretos com muitas observações. Quando os dados são contínuos, podem existir várias modas ou a moda pode mesmo não existir, tornando-se uma estatística irrelevante.

[a] Equivalentemente, $\overline{x} = \sum_i x_i \, \widehat{f}(x_i)$ onde $\widehat{f}$ representa a frequência empírica relativa.

[a] *outliers*

(i) A média de variáveis ordinais ou nominais tem de ser usada com muito cuidado. Podem ser feitas várias observações a este respeito:

- a média não será necessariamente um dos valores da variável (numa variável contínua, a média toma sempre um dos valores da variável)
- considere-se uma variável ordinal cujas k categorias foram ordenadas por ordem crescente segundo 1, 2, 3, ..., k. Nesta situação a média só pode ser considerada se a passagem de uma categoria para a seguinte tiver sempre o mesmo significado.
- no caso de uma variável binária, ordinal ou nominal, (sem perda de generalidade, tomando 0's e 1's), a média corresponde à proporção de 1's na amostra.

(j) Em Estatística, a média é muito mais usada do que a mediana, essencialmente porque a soma de variáveis aleatórias é muito recorrente e, enquanto que a média da soma de variáveis aleatórias é a soma das médias, para a mediana essa propriedade não se verifica.

1.3. Medidas de escala de uma a.a.

Dados: a.a. $x_1, x_2, \ldots, x_n$ de uma v.a. X.

Variância *Variance*	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ $= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)$
Desvio Padrão *Standard Deviation*	$s = \sqrt{\text{variância}}$
p-Quantil/Percentil $100p$ *p-Quantile/Percentile* $100p$	observação de ordem $p(n+1)$ ou média pesada das observações vizinhas.
Amplitude *Range*	(maior observação)-(menor observação)
Amplitude Interquartil *Interquartile Range*	AIQ=(percentil 75)-(percentil 25)

Observações:

(a) É adequado apresentar o máximo e o mínimo da a.a.

(b) É adequado apresentar o desvio padrão com duas casas decimais a mais do que os dados.

(c) A mediana corresponde ao percentil 50.

(d) As unidades da variância são (unidades da v.a.)^2; as unidades do desvio padrão são as unidades da v.a.

(e) Para uma amostra de tamanho n, resulta da definição de média que

$$\sum_{i=1}^{n}(x_i - \overline{x}) = 0.$$

Daí a necessidade de considerar os quadrados na fórmula de s^2.

(f) Há estudos em que a dispersão é medida usando $\sum |x_i - \overline{x}|$ mas esta fórmula traz complicações computacionais adicionais.

(g) Se adicionarmos uma constante a cada uma das observações, o valor do desvio padrão permanece inalterado. Se multiplicarmos cada observação por uma constante, o desvio padrão é multiplicado também por essa constante.

(h) Se for razoável assumir que a amostra provém de uma população que segue uma distribuição simétrica, devemos apresentar a média e o desvio-padrão [a]. Caso contrário, devemos apresentar a mediana e alguns percentis superiores e inferiores (por exemplo, os percentis 5, 25, 75 e 95) de forma a fornecer informações sobre a distribuição.

[a] Apresentamos o desvio-padrão e não a variância, apenas por ser mais fácil de visualizar.

(i) Veremos mais tarde que, na fórmula da variância, a divisão por $n-1$ e não por n, garante que a variância amostral seja um estimador não enviesado da variância populacional. Pode-se ainda argumentar usando os graus de liberdade: as n observações da amostra não estão todas livres porque já se calculou a média, tendo-se imposto portanto uma condição.

Observação acerca dos quantis:

O **R** permite escolher a definição de quantil a usar, de entre 9 possíveis. A definição apresentada na tabela atrás é a mais comum. É simples de calcular e generaliza o método de determinação da mediana.
Represente-se a função característica de um número real x por

$$[x],$$

correspondente ao menor inteiro que não excede x.
O quantil de ordem p, tal como o definimos atrás, corresponde ao valor obtido por interpolação linear entre

$$([p(n+1)], x_{[p(n+1)]}) \text{ e } ([p(n+1)]+1, x_{[p(n+1)]+1});$$

isto é, corresponde a

$$x_{[p(n+1)]} + (p(n+1) - [p(n+1)])(x_{[p(n+1)]+1} - x_{[p(n+1)]})$$

depois de termos ordenado a amostra.

Outras definições de p-quantil associado a uma amostra de tamanho n: é o número q (não necessariamente na amostra) tal que

- $\widehat{F}(q) = p$.

  Esta definição é a análoga para o caso amostral da definição de p-quantil de uma v.a. contínua.

- $\frac{q-1}{n-1} = p$.

  Definição usada, por defeito, no **S** e no **R**

- $\frac{q}{n+1} = p$.

  Definição usada no **SPSS**;

A diferença entre os vários métodos é pequena e irrelevante para os nossos propósitos.

Quantis na calculadora:

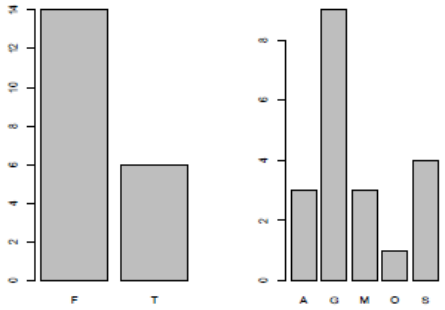
Não existem algoritmos de cálculo para determinação de quantis amostrais numa calculadora. Tudo o que se consegue são os quantis 0.25 e 0.75, por causa da construção de boxplots.
O que a calculadora faz é encontrar a mediana amostral M pela definição usual e depois encontrar a mediana de todas as observações que são estritamente superiores a M - que será 0.75-quantil - e a mediana de todas as observações que são estritamente inferiores a M - que será o 0.25-quantil.
Este é, aliás, o procedimento seguido no **R**, por omissão, para a determinação dos quantis 0.25 e 0.75 a usar no boxplot.

1.4. Medidas sumárias descritivas de uma v.a. categórica

As medidas sumárias descritivas mais comuns para v.a. categóricas consistem apenas:

- do tamanho da amostra
- da designação das categorias
- do número de observações em cada uma das categorias, e correspondente percentagem relativamente ao total de elementos na amostra (frequências absoluta e relativa de cada categoria, respectivamente).

O gráfico correspondente será simplesmente um gráfico de barras usando frequências absolutas ou relativas, conforme se achar mais adequado.



1.5. Erro padrão (amostral) da média

Dada uma a.a. $x_1, x_2, \ldots, x_n$ de tamanho n, o erro padrão amostral da média (sample standard error of the mean) é representado por $s.e.$ [a] e definido por

$$s.e. = \frac{s}{\sqrt{n}}.$$

Trata-se de uma estatística que estima a precisão com que a média da a.a. estima a média da população.
É sempre inferior ao desvio-padrão amostral.
Por vezes é mencionado nalguns estudos porque fá-los parecer melhores!...em geral, os dados devem ser sumariados usando o desvio-padrão, que quantifica a variabilidade na população, e não o desvio padrão da média.

[a] *S.E.*, *s.e.m.*, *sem* ou *SEM*

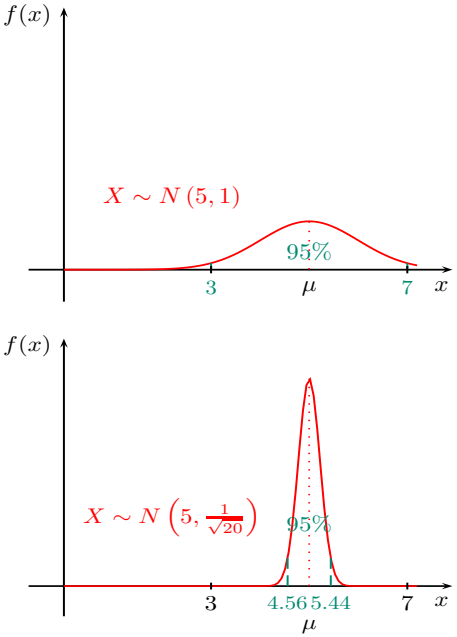
Considere-se uma a.a. de tamanho $n = 20$ com

$$\overline{x} = 5 \qquad \text{e} \qquad s = 1.$$

Tem-se

$$s.e. = \frac{1}{\sqrt{20}} \approx 0.22.$$

Confundir o desvio padrão da média com o desvio padrão, numa população que segue uma distribuição normal com os parâmetros amostrais, confere uma ideia errada acerca da dispersão da distribuição:



1.6. Coeficiente de Variação

Dada uma realização de uma a.a. $x_1, x_2, \ldots, x_n$ de tamanho n em que todas as observações são positivas[a], o coeficiente de variação (coefficient of variation), representado por CV é definido por

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

Trata-se de uma medida de dispersão estandardizada; uma quantidade adimensional particularmente útil para comparar a dispersão relativa de variáveis aleatórias com:

- médias bastante diferentes, ou
- unidades de medida diferentes.

De facto, o desvio padrão dos dados só pode ser interpretado tendo em conta a média amostral;

$$s = 10 \text{ e } \overline{x} = 20$$

tem consequências diferentes de

$$s = 10 \text{ e } \overline{x} = 100.$$

Na primeira situação tem-se $CV = 50\%$ enquando que na segunda $CV = 10\%$ portanto a dispersão relativa é menor no segundo caso.

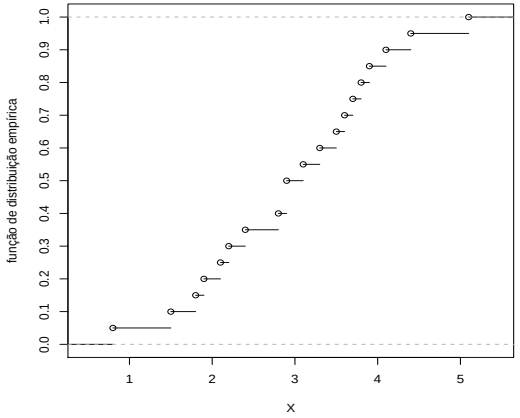
[a] esta condição garante que a média seja positiva

1.7. Gráficos associados a uma amostra de uma variável contínua

Dados: Uma amostra aleatória de tamanho n de uma v.a. contínua, $X_1, \ldots, X_n$.

A **função de distribuição empírica** $\widehat{F}$ é

$$\widehat{F} : \mathbb{R} \longrightarrow \mathbb{R}, \qquad x \mapsto \left(\sum_{x_i \leq x} x_i \right) / n$$



Trata-se sempre de uma função em escada crescente, cujo primeiro valor é 0 e o último é 1.
A função de distribuição empírica é a função de distribuição associada à função de probabilidade dada pela frequência com que cada observação aparece na amostra.

Para o cálculo de $\widehat{F}$ pode construir-se a seguinte **tabela de frequências**, em que as observações x_1, $\ldots$, x_n são dispostas por ordem crescente.

valores	frequência **absoluta**	frequência **relativa**	frequência **relativa acumulada**
x_i	n_i	n_i/n	$\widehat{F}(x_i) = \sum_{x_j \leq x_i} n_j/n$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
$\ldots$	$\ldots$	$\ldots$	1

A soma de todas as frequências absolutas é igual ao tamanho da amostra (n) e o último valor da frequência relativa acumulada é sempre 1.

Diagrama de extremos e quartis e **Diagrama de caixa e bigodes** (***Boxplot***)
A construção de qualquer um destes gráficos é feita essencialmente a partir dos quartis (0.25, 0.5 e 0.75-quantil) do conjunto de dados.
Desenhe-se uma caixa vertical com ordenada da base igual ao 0.25-quantil e ordenada do topo igual ao 0.75-quantil. A largura da caixa é arbitrária. No interior dessa caixa construa-se, com uma espessura maior do que a usada anteriormente, um segmento de recta com ordenada igual à mediana.

Pode-se agora proceder de 2 formas:

- a mais simples consiste em marcar traços horizontais correspondentes aos valores máximo e mínimo da amostra (valores extremos) e ligar esses traços horizontais à caixa inicial através de segmentos de recta verticais - **diagrama de extremos e quartis**.

- a segunda forma tem em consideração a existência de observações no conjunto de dados que estão *muito distantes* da maior parte das outras (***outliers***). Determine-se a amplitude interquartil

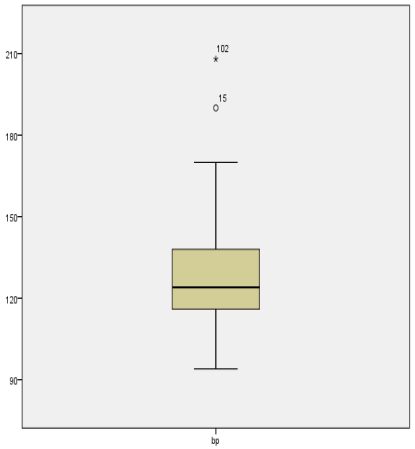
$$AIQ = q_{0.75} - q_{0.25}$$

e as barreiras internas superior (BIS) e inferior (BII), e as barreiras externas superior (BES) e inferior (BEI):

$$BII = q_{0.25} - 1.5AIQ, \quad BIS = q_{0.75} + 1.5AIQ$$
$$BEI = q_{0.25} - 3AIQ, \quad BES = q_{0.75} + 3AIQ$$

Marquem-se agora dois traços horizontais: um correspondente ao maior valor da amostra que não é superior à BIS, e um outro correspondente ao menor valor da amostra que não é inferior à BII.

Liguem-se esses traços horizontais à caixa inicial através de segmentos de recta verticais (os bigodes).



Todas as observações que se localizem entre as barreiras BIS e BES, ou BII e BEI, são designadas por **outliers moderados** e representadas na figura por pequenos círculos abertos, ∘ ; todas as observações que tomem valores para além das barreiras externas são designadas por **outliers severos** e representadas por asteriscos,∗.
O gráfico assim obtido é designado por **gráfico de caixa e bigodes**, do inglês *boxplot*, abreviatura de *box-and-whiskers plot*.
Alguns softwares não fazem distinção entre os dois tipos de outliers. Outros apresentam o número da observação em causa associado a cada outlier.

O boxplot é um gráfico que traduz informação acerca da zona de concentração, dispersão e simetria do conjunto de dados observados.

- 50% das observações estão contidas na caixa $(q_{0.75} - q_{0.25})$, e essas são precisamente as observações centrais.
- Para avaliar a dispersão pode-se usar:
  - a amplitude do intervalo de variação: $H = \max - \min$
  - a amplitude interquartil: $AIQ = q_{0.75} - q_{0.25}$.

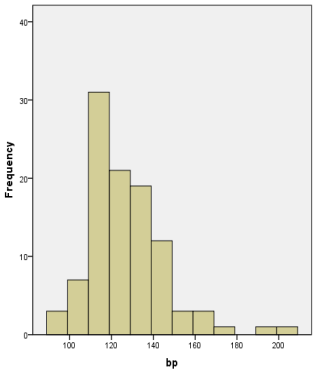
  Enquanto que a amplitude do intervalo de variação é sensível à presença de outliers, a AIQ é uma medida de dispersão resistente.
- Quanto ao estudo da simetria da distribuição:
  - se $q_{0.75} - q_{0.5} \approx q_{0.5} - q_{0.25}$ então a distribuição é aproximadamente simétrica
  - se $q_{0.75} - q_{0.5} >> q_{0.5} - q_{0.25}$ então a distribuição é assimétrica à direita
  - se $q_{0.75} - q_{0.5} << q_{0.5} - q_{0.25}$ então a distribuição é assimétrica à esquerda.

Histograma

Divide-se a recta real em k intervalos disjuntos consecutivos $(-\infty < a_1 < a_2 < \cdots < a_k < +\infty)$

$$I_1 =]-\infty, a_1], \qquad I_2 =]a_1, a_2], \qquad \ldots$$
$$\ldots, \qquad I_{k-1} =]a_{k-2}, a_{k-1}], \qquad I_k =]a_{k-1}, +\infty[.$$

O histograma é o gráfico de barras contíguas que atribui a cada intervalo o número de observações na amostra que tomam valores nesse intervalo.
Só está definido para variáveis contínuas; aliás, as barras são contíguas para reflectirem esse carácter contínuo da variável subjacente, por oposição ao gráfico de barras das variáveis categóricas.
Também pode ser definido usando frequências relativas, em lugar das absolutas. Nesse caso, será uma representação discreta da função densidade de probabilidade.



A determinação da amplitude e do número de intervalos envolve um compromisso entre precisão e generalização. A forma do histograma obtido pode variar consideravelmente conforme as escolhas consideradas. Existem vários algoritmos que permitem determinar o número de intervalos a usar. Sturges (1926) sugeriu que o número de intervalos seja dado por

$$1 + \log_2(n)$$

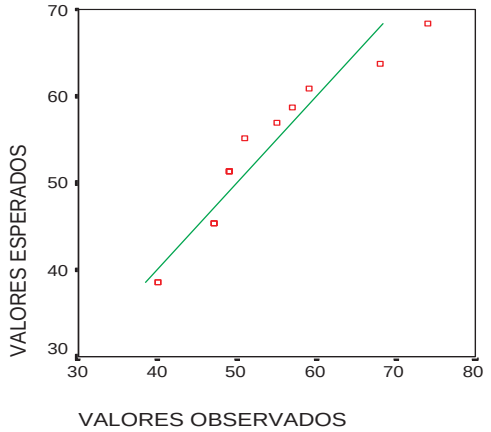
onde n é o tamanho amostral. Pequenos ajustes a este número são bem tolerados.

O histograma é uma representação discreta da função densidade de probabilidade. Em particular, permite avaliar a forma da distribuição da v.a. em causa.

Q-Q plot: gráfico dos quantis

Trata-se do gráfico dos quantis de uma distribuição de probabilidade teórica previamente especificada de uma variável aleatória Z (usualmente a distribuição normal) contra os quantis amostrais da distribuição empírica $\widehat{F}$ associada a uma amostra $y_1, \ldots, y_n$. Permite detectar de forma "visual" se a distribuição empírica está ou não próxima da distribuição teórica especificada. Constrói-se do modo a seguir descrito.

Seja y_i uma observação da amostra. Tem-se $\widehat{F}(y_i) = p_i$ para algum $p_i \in [0,1]$ e, por definição, y_i é o p_i-quantil amostral.
O p_i-quantil (teórico) de Z é o número z_i tal que $P(Z \leq z_i) = p_i$.
O Q-Q plot é o gráfico que contém os pares ordenados (z_i, y_i), para todos os valores $y_1, \ldots, y_n$ da amostra.

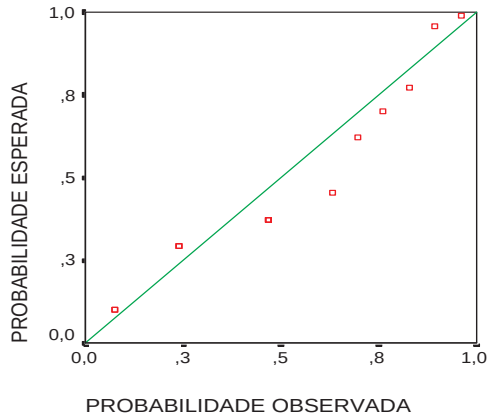


A linha recta no gráfico representa a correspondência perfeita entre os quantis amostrais e os quantis teóricos.

P-P plot: gráfico de probabilidades

Trata-se do gráfico de uma distribuição de probabilidade G de uma variável aleatória Z contra a distribuição empírica $\widehat{F}$ associada a uma amostra $y_1, \ldots, y_n$. Tal como o Q-Q plot, permite também detectar de forma "visual" se $F = G$ ou $F \neq G$. Constrói-se do modo a seguir descrito.

Seja $y \in \mathbb{R}$. A probabilidade empírica de termos $Y \leq y$ é $\widehat{F}(y)$. A probabilidade (teórica) de termos $Z \leq y$ é $G(y)$.
O P-P plot é o gráfico que contém os pares ordenados $(G(y_i), \widehat{F}(y_i))$, para todos os valores $y_1, \ldots, y_n$ da amostra.



A linha recta no gráfico representa a correspondência perfeita entre a distribuição amostral e a distribuição teórica.

Tal como no caso do Q-Q plot, a distribuição teórica usada mais frequentemente é a distribuição normal.