

Métodos Quantitativos (e Qualitativos)

1. Descrição e visualização dos dados

Salvatore Cosentino
D.Mat. U.Minho

17 nov 2020

Observáveis/variáveis

No contexto de uma teoria física, os **observáveis** são as quantidades que conseguimos **medir** ao fazer as experiências.

Os valores possíveis, que dependem dos instrumentos usados, são tipicamente **números**, que representam múltiplos (ou, em geral, proporções) de uma unidade de medida apropriada.

(a temperatura da água em K , a distância entre a Terra e a Lua em cm , a velocidade da luz em Km/s , ...)

No contexto da teoria das probabilidades, os observáveis são chamados **variáveis (aleatórias)**.

São as funções definidas no espaço dos **acontecimentos (elementares)**, e tais que é sempre possível calcular as probabilidades associadas a conjuntos oportunos dos seus valores possíveis, chamados **eventos**.

(o número de vezes que sai cara ao lançar n moedas, a posição de um passeio aleatório no tempo t , ...)

Variáveis nas CS

No contexto típico das Ciências Sociais, onde o objeto do estudo é uma **população**, as **variáveis** são os observáveis que conseguimos medir ao observar um **indivíduo** (“indivisível”) da população, ou seja, certas suas características.

(o ordenado anual ou o sexo de um cidadão dos EUA, o número de homicídios em cada cidade portuguesa em 2019, a população de cada país da UE, o PIB per capita dos países asiáticos, ...)

As **experiências** consistem em recolher informações sobre um certo número de variáveis de uma população de N indivíduos

(censos, bases de dados, ...)

ou apenas de uma **amostra** (em inglês, **sample**), um subconjunto de n indivíduos da população, que seja suficientemente grande e aleatório para as técnicas da teoria das probabilidades fazerem sentido, i.e. $n \gg 1$, mas tipicamente muito menor da população mesma, i.e. $n \ll N$.

(inquéritos, sondagens, entrevistas, ...)

Variáveis quantitativas

Nas **ciências exatas** (em inglês, **hard sciences**), as variáveis são tipicamente **quantitativas**,

ou seja, têm **valores numéricos** que medem certos observáveis numa certa escala, fixada por uma unidade de medida.

Podem ser **discretas**, ou seja, representar o resultado de uma contagem ou de uma razão entre diferentes múltiplos de uma unidade de medida

(o número de neutrões no núcleo de um átomo, o número de pintas que tem uma joaninha, o número de cabelos na cabeça de um homem, ...)

ou **contínuas**, ou, melhor, bem aproximadas/idealizadas por quantidades que imaginamos contínuas (embora hoje sabemos que há limites físicos ditados pela mecânica quântica ...)

(a temperatura na superfície de uma estrela, a distância entre uma estrela e o Sol, a velocidade de afastamento de uma galáxia, ...)

Variáveis qualitativas

Nas CS, é também usual considerar variáveis **qualitativas**, que podem ser

em uma **escala nominal**

(as nacionalidades ou as crenças religiosas dos reclusos, ...)

ou em uma **escala ordinal**

(o grau de satisfação dos clientes de um banco, ...)

É sempre possível **associar números** aos diferentes graus de uma escala ordinal, embora pode não ser claro se todos os inquiridos concordem na tipologia de escala associada, (linear, logarítmica, ... ?)

(os inquéritos distribuídos aos alunos da UM associam os números 1, 2, 3, 4 e 5 às opiniões “Péssimo”, “Mau” , “Razoável”, “Bom” e “Excelente”, respetivamente)

Dados

Ao fazer n experiências, por exemplo ao observar uma amostra de n indivíduos escolhidos de uma população, são observados os **valores**

$$\boxed{x_1 \quad x_2 \quad x_3 \quad \dots \quad x_n}$$

da **variável** x , e possivelmente de outras variáveis $y, z, \dots$

Os **dados** são estas listas, tipicamente arranjadas em tabelas do género

	x	y	z	$\dots$
1	x_1	y_1	z_1	
2	x_2	y_2	z_2	
3	x_3	y_3	z_3	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	

Os dados recolhidos são, hoje em dia, guardados em **ficheiros**.

Um formato relativamente standard é o **.csv** (ou seja, **comma separated values**)

Diferentes software de análise de dados ou de computação trabalham com diferentes outros formatos das bases de dados, por exemplo:

IBM SPSS software usa ficheiros **.sav**

R (Project for Statistical Computing) usa ficheiros **.Rdata**

Excel usa ficheiros **.xlsx**

Dados qualitativos

Se a variável qualitativa x assume valores

blue red orange green ...

tudo o que é possível fazer é calcular frequências

$$n_{\text{blue}} = \#\{\text{observações } x_i = \text{blue}\} \quad n_{\text{red}} \quad \dots$$

e portanto frequências relativas

$$f_{\text{blue}} = \frac{n_{\text{blue}}}{n} \quad f_{\text{red}} \quad \dots$$

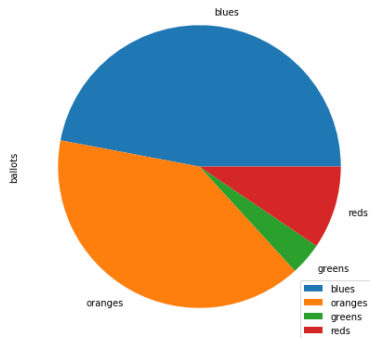
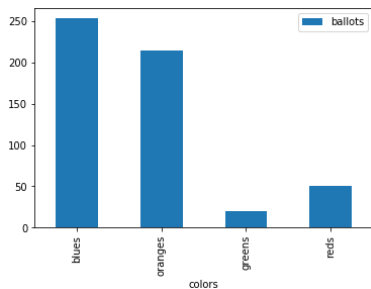
que são estimativas de probabilidades.

Outras manipulações algébricas não fazem muito sentido, sendo as escalas subjetivas/arbitrárias.

Diagramas de barras e circulares

As frequências podem ser representadas em **diagramas de barras**

ou em **diagramas circulares** (em inglês, **pie charts**)



Dados quantitativos

Ao fazer n medições, por exemplo de amostra de n indivíduos escolhidos de uma população, são observados os valores

$$x_1 \quad x_2 \quad x_3 \quad \dots \quad x_n$$

da variável quantitativa x .

O **intervalo de variação** (em inglês, **range**) das observações, é

$$[x_{\min}, x_{\max}]$$

onde $x_{\min}$ e $x_{\max}$ são o menor e o maior dos x_k 's, respetivamente.

Se os valores são muitos, como desejável, uma simples lista não permite ver grande coisa.

Mais conveniente é sintetizar os dados por meio de um número pequeno de parâmetros, medidas de **localização** e de **dispersão**,

e visualizar uma imagem desfocada da distribuição dos mesmos, tipicamente com o uso de **histogramas**.

Outliers

Pode acontecer que um (ou alguns) dos valores observados, por exemplo

$$x_{13}$$

esteja(m) muito afastado(s) dos outros. Tais valores são chamados **outliers** (em português, **discrepâncias**, ou **valores atípicos**).

Os outliers podem ser indícios de experiências mal feitas, de perturbações das condições do laboratório, mas também indícios de um fenómeno ainda não considerado ...

Aceitar ou rejeitar eventuais outliers é uma questão debatida e **não consensual**.

Alguns estatísticos desenvolveram “testes” para decidir quando aceitar ou rejeitar um outlier x_{13} , o **teste Q**, o **critério de Chauvenet**, ... , baseados em conjecturas acerca da distribuição dos erros e em considerações estatísticas.

Também há cientistas que simplesmente decidem não fazer nada,

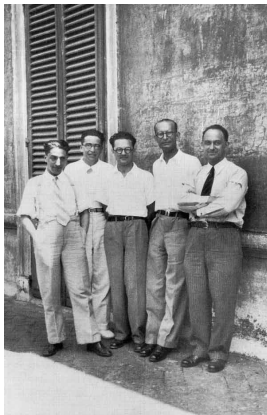
ou, podendo, decidem **repetir muitas mais vezes** as observações até que os outliers deixem de serem importantes nas estimações.

Para ser honestos, é importante decidir uma estratégia antes de obter os dados.

a sora Cesarina

Famosa é a história do **balde de água** escondido de noite pela **sora Cesarina** por baixo da mesa do laboratório de física de **Fermi, Pontecorvo, ...** em via Panisperna, em Roma.

Este episódio conduziu a equipa a uma descoberta científica devastante, e, dez anos depois, à explosão da primeira **bomba nuclear** em Alamogordo, New Mexico.



Classificação dos dados ...

É conveniente, por exemplo quando n é grande, **classificar** os dados em grupos,

ou seja, dividir o intervalo de variação $[x_{\min}, x_{\max}]$ em um número razoável (tipicamente da ordem de ~ 20) de **subintervalos** de comprimentos iguais

$$I_1 = [x_{\min}, a_2] \quad I_2 = [a_2, a_3] \quad \dots \quad I_m = [a_m, x_{\max}]$$

e calcular a **frequência** de observações em cada intervalo

$$n_k = \#\{\text{observações } x_i \in I_k\}$$

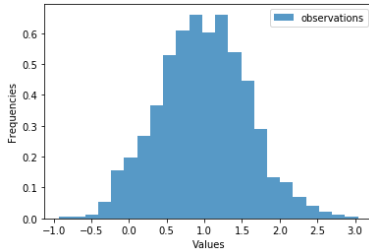
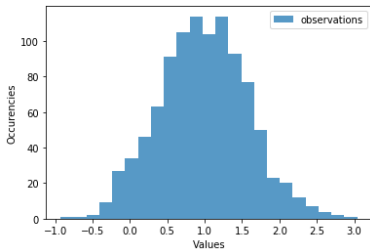
e portanto a **frequência relativa**

$$f_k = \frac{n_k}{n}$$

... e histogramas

Um gráfico das frequências ou das frequências relativas dos dados classificados é chamado **histograma**.

Proporciona uma imagem das **importâncias relativas** dos diferentes valores observados, uma imagem de uma hipotética **distribuição de probabilidade** subjacente.



(amostra de 1000 variáveis i.i.d. com lei normal $N(1, 0.6)$)

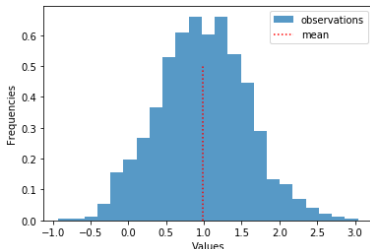
Média

A **média (aritmética)** (em inglês, **mean** ou **average**) é

$$\bar{x} := \frac{x_1 + x_2 + \cdots + x_n}{n}$$

A média aritmética é a medida de localização mais **democrática**, pois depende linearmente de todas as observações, e cada observação é pesada com o mesmo peso.

Sendo assim, depende significativamente das observações extremas, e portanto também dos outliers.



Uma descrição numérica essencialmente completa dos dados é a seguinte.

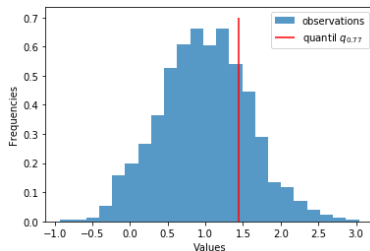
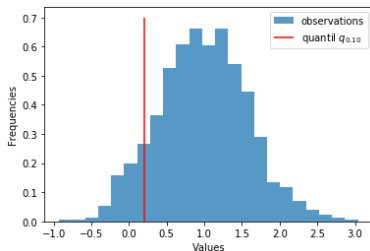
Fixada uma probabilidade α , entre 0 e 1,

o **quantil** q_α é o valor tal que uma fração α das observações satisfazem

$$x_k \leq q_\alpha$$

ou seja,

$$\frac{\#\{\text{observações } x_k \leq q_\alpha\}}{n} \simeq \alpha$$

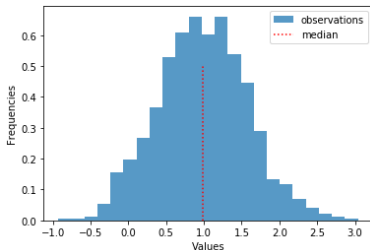


Mediana

A **mediana** é o quantil

$$m := q_{0.5}$$

que separa o 50% das observações menores do 50% das observações maiores.



Os valores extremos, pequenos ou grandes, podem variar sem alterar a mediana.

Em particular, a mediana não depende dos outliers.

Percentis

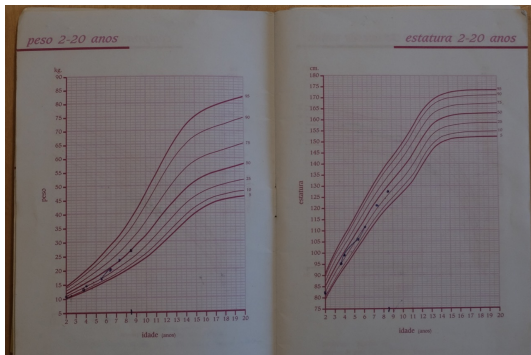
Uma das divisões preferidas do intervalo de variação é em **100 partes**, contendo uns números aproximadamente iguais de observações.

Os **percentis** são os quantis

$$P_k := q_k\%$$

com $k = 1, 2, 3, \dots, 99$.

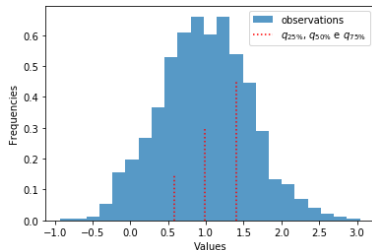
(por exemplo, dizer que a altura da minha filha aos 4 anos de idade está no percentil P_{25} significa que apenas 25 em cada 100 meninas de 4 anos são mais baixas do que a minha filha)



Os **quartis** são os quantis

$$Q_1 := q_{25\%} \quad Q_2 := q_{50\%} \quad \text{e} \quad Q_3 := q_{75\%}$$

que dividem o intervalo de variação em 4 intervalos com um número igual de observações.



Os quartis não dependem sensivelmente dos valores extremos das observações, por exemplo, dos outliers.

Diagrama caixa-com-bigodes

O **intervalo interquartil** (em inglês **IQR**, ou seja, **inter-quartile range**)

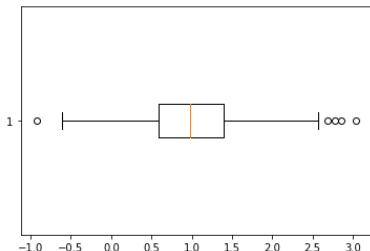
$$[Q_1, Q_3] = [q_{25\%}, q_{75\%}]$$

é o intervalo central onde estão concentradas o 50 % das observações.

Juntamente com o intervalo,

$$[q_{5\%}, q_{95\%}]$$

(ou entre $q_{10\%}$ e $q_{90\%}$), forma o **diagrama caixa-com-bigodes** (em inglês, **box plot**)



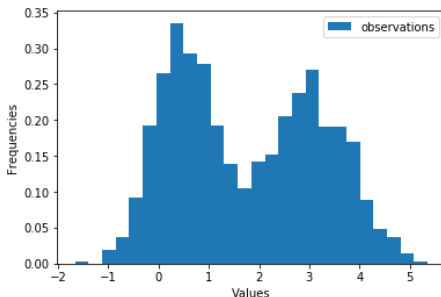
Ficam fora os valores extremos, em particular outliers.

Outras medidas de localização

Outra medida de localização é a **moda**, a classe com a frequência relativa maior.

A moda, assim como a mediana, não depende dos outliers.

Podem existir mais classes que maximizam localmente a frequência relativa. Estas distribuições são chamadas **multimodais**, por exemplo, **bimodais** ...



Variância e desvio padrão

A **variabilidade** das observações em torno do valor médio é medida pela **variância**

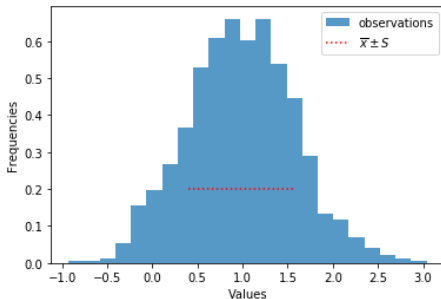
$$S^2 := \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2$$

e portanto pelo **desvio padrão** (em inglês, **standard deviation**)

$$S := \sqrt{S^2}$$

(que tem as dimensões de x).

Uma medida adimensional dos desvios típicos observados é o **desvio padrão relativo** (ou **coeficiente de variação**), a razão $S/|\bar{x}|$.



Outras medidas de variabilidade

Existem outras medidas de variabilidade, todas com o defeito de ser pouco úteis para ser manipuladas matematicamente.

A amplitude do intervalo de variação, ou **range** das observações, é

$$x_{\max} - x_{\min}$$

O **desvio absoluto médio**

$$\frac{1}{n} \sum_{k=1}^n |x_k - \bar{x}|$$

é o valor médio dos desvios absoluto.

A **amplitude interquartil**

$$Q_3 - Q_1$$

é a amplitude do intervalo onde estão concentradas a metade central das observações, usado nos diagramas caixa-com-bigodes. .

Assimetria

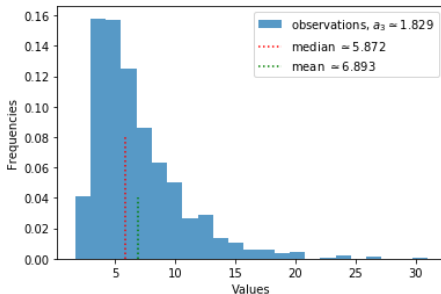
A **assimetria** das observações em torno do valor médio é medida, por exemplo, pelo primeiro momento centrado ímpar não nulo, que é o momento centrado de grau 3

$$m_3 := \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^3$$

e portanto pelo **coeficiente de assimetria**

$$a_3 := \frac{m_3}{S^3}$$

que é positivo ou negativo dependendo se os maiores desvios $x_k - \bar{x}$ são positivos ou negativos.



Curtose

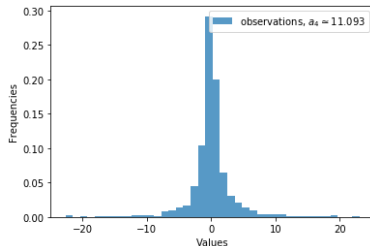
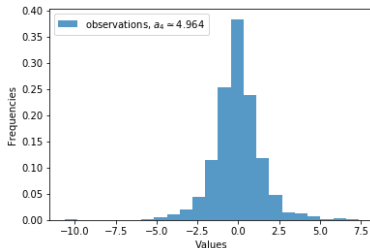
O **achatamento** das observações, ou seja, o valor relativo das caudas dos histogramas, é medido, por exemplo, pelo momento centrado de grau 4

$$m_4 := \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^4$$

e portanto pelo **coeficiente de curtose**

$$a_4 := \frac{m_4}{S^4}$$

que é grande ou pequeno dependendo se as caudas são importantes ou desprezáveis.



Observáveis qualitativos e quantitativos.

Diagramas de barras e circulares para frequências.

Histogramas.

Medidas de localização (média, mediana e moda), e suas dependências dos possíveis outliers.

Medidas de dispersão, ou escala (desvio padrão, intervalo inter-quartil, . . .)