Capítulo 5

Estatística básica

Quando atiramos um dardo para um alvo o resultado do lançamento tem sempre uma componente aleatória (mais ou menos imprevizível conforme o treino e o talento do atirador, a distância ao alvo, etc.). Uma das funções da estatística é utilizar a matemática para tentar descrever os resultados de uma forma que seja mais simples, concisa e intuitiva.

Por exemplo, um atirador faz dois lançamentos. Um acerta no 20 e outro no 10. Se eu dissesse que em dois lançamento a média foi 15 e o desvio entre eles foi de 10 não estaria a acrescentar informação nova. Na realidade, seria fácil chegar aos valor dos lançamentos a partir da média e do desvio entre eles.

Esta parte da estatística diz-se *descritiva* e para além de um novo ponto de vista não acrescenta nada aos dados iniciais. No entanto, este novo ponto de vista é muito útil. Se por qualquer razão o lançador tiver uma tendência para o lado direito do alvo, esse facto será facilmente visível num simples parâmetro - a pontuação média dos seu lançamentos. Olhando para esse parâmetro posso *inferir* que existe uma tendência para a direita e posso corrigir o desvio do lançador.

Se fizermos um paralelo com uma experiência laboratorial, a estatística descritiva fornece parâmetros que descrevem os dados, facilitam a sua análise e permitem planear melhor uma nova experiência.

Nesta secção vamos estudar alguns parâmetros da estatística descritiva, como fazer a sua aquisição e cálculo. Só mais tarde, quando falarmos sobre o grau de confiança (na subsecção 5.6.1) é que vamos introduzir o conceito de probabilidade.

5.1 População e amostra

Designa-se de população o conjunto de *todos* os dados que pretendemos estudar. Por exemplo, se eu quisesse uma informação completamente fiável sobre a idade dos madeirenses teria que saber a idade de todos.

Na prática nem sempre é possível ter tal informação. De facto, num laboratório como posso realizar sempre mais uma medição a população em estudo é ilimitada. Temos que nos contentar em ter apenas dados sobre um subgrupo: a amostra.

No caso da idade dos madeirenses em vez de perguntar a idade a todos os madeirenses poderia perguntar apenas a 1000. A escolha da amostra pode ser um processo complicado. Porquê? Suponhamos que realizava a amostragem num lar da terceira idade. Os resultados da medição seriam mais altos que os da população. A amostra deve ser representativa do todo. Ou seja deve haver em termos relativos tantos idosos como no todo, tantas crianças como no todo, etc.. Por isso deve ser recolhida de forma *aleatória*.

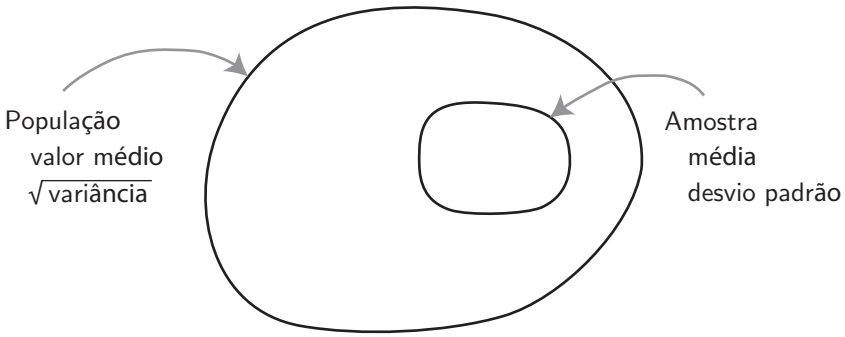

Figura 5.1: População e amostra

A média irá tender para o valor médio e o desvio padrão irá tender para a raíz quadrada da variância à medida que a amostra tender para a população. Se a amostragem for bem feita a média será uma boa estimativa do valor médio e o desvio padrão uma boa estimativa da raíz quadrada da variância mesmo para uma amostra pequena.

Tudo isto para vermos que no laboratório as medidas estatísticas que obtemos são aquelas que resultam da amostra (e.g. média, desvio padrão, etc.). Veremos em seguida como são definidas.

5.2 Média

Quando calculamos a média entre dois valores buscamos um valor que esteja a igual distância dos dois:

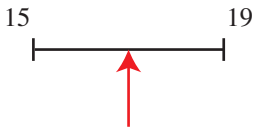

Figura 5.2: Média entre 15 e 19

Este valor é calculado somando os valores e dividindo por dois:

$$\overline{x} = \frac{15 + 19}{2} = 17 \tag{5.1}$$

O valor 17 está a igual distância de 15 e de 19 porque atribuímos *igual importância* aos valores 15 e 19. Onde é que está quantificada a importância de cada um dos valores na equação (5.1)? Para responder a esta pergunta vamos modificar a equação para:

$$\overline{x} = \frac{1}{2} \cdot 15 + \frac{1}{2} \cdot 19 \tag{5.2}$$

Os coeficientes de $\frac{1}{2}$ em cada uma das parcelas indicam que ambos os valores têm uma importância igual à qual chamaremos *peso*. Neste caso o peso de ambos os valores é de 50%.

Suponhamos que queríamos atribuír um peso maior ao valor 15 (e.g. 75%). A média passa a designar-se de média ponderada e seria calculada de acordo com:

$$\overline{x} = \frac{3}{4} \cdot 15 + \frac{1}{4} \cdot 19 = 16 \tag{5.3}$$

É de notar que o valor da média está agora mais próximo do valor com maior peso:

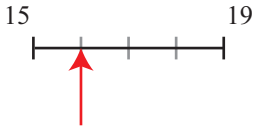

Figura 5.3: Média ponderada entre 15 e 19

De facto está a uma distância do 15 de $\frac{1}{4}$ da distância do 15 ao 19 e a uma distância do 19 de $\frac{3}{4}$ da distância do 15 ao 19. Quanto maior for o peso atribuído a um valor mais próximo estará o resultado final desse valor.

Note-se que a soma dos pesos é igual a 1. Caso contrário, seria possível que a média desse fora do intervalo entre 15 e 19!

A forma geral da equação (5.1) é:

$$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N} \tag{5.4}$$

da equação (5.2):

$$\overline{x} = \sum_{i=1}^{N} \frac{1}{N} \cdot x_i \tag{5.5}$$

e da equação (5.3):

$$\overline{x} = \sum_{i=1}^{N} p_i x_i \tag{5.6}$$

em que p_i é o peso do valor x_i e:

$$\sum_{i=1}^{N} p_i = 1 \tag{5.7}$$

Alguns exemplos:

- ao adicionar $100\,g$ de água à temperatura de $30\,°\text{C}$ com $100\,\text{g}$ de água à temperatura de $90\,°\text{C}$ a temperatura da mistura será de:

$$\frac{30\,°\text{C} + 90\,°\text{C}}{2} = 50\% \cdot 30\,°\text{C} + 50\% \cdot 90\,°\text{C} = 60\,°\text{C}$$

Ambas as temperaturas têm igual peso (50%) porque misturámos iguais massas de água.

- ao adicionar $200\,\text{g}$ de água à temperatura de $30\,°\text{C}$ com $100\,\text{g}$ de água à temperatura de $90\,°\text{C}$ a temperatura da mistura será de:

$$\frac{200\,\text{g}}{200\,\text{g} + 100\,\text{g}} \cdot 30\,°\text{C} + \frac{100\,\text{g}}{200\,\text{g} + 100\,\text{g}} \cdot 90\,°\text{C} = \frac{2}{3} \cdot 30\,°\text{C} + \frac{1}{3} \cdot 90\,°\text{C} = 50\,°\text{C}$$

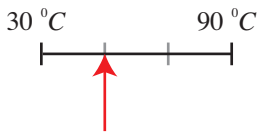

Figura 5.4: Representação gráfica do cálculo da temperatura final

A temperatura final estará mais próxima de $30\,°\text{C}$ do que de $90\,°\text{C}$ porque a massa à temperatura de $30\,°\text{C}$ é maior. O peso atribuído à massa a $30\,°\text{C}$ será igual a $\frac{2}{3}$ porque representa $\frac{2}{3}$ da massa total.

A massa à temperatura de $90\,°\text{C}$ é menor e representa $\frac{1}{3}$ da massa total logo o peso atribuído a esta massa será de $\frac{1}{3}$. Tal como previsto pela equação (5.7): $\frac{2}{3} + \frac{1}{3} = 1$.

5.3 Desvio padrão

O desvio padrão é uma medida estatística da dispersão em relação à média dos valores medidos.

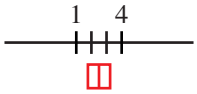

Figura 5.5: Dados menos dispersos

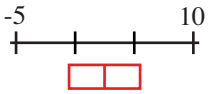

Figura 5.6: Dados mais dispersos

Esta dispersão é avaliada a partir da distância entre cada um dos valores x_i e a média $\overline{x}$ e do número de dados N segundo a equação:

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \overline{x})^2}{N-1}} \tag{5.8}$$

À partida temos um sistema de N equações $(x_i = valor)$. Para o cálculo da média reduzimos uma equação no sistema. Isto quer dizer que para o cálculo do desvio padrão dispomos de $N-1$ *graus de liberdade*. É esta a razão para termos $\sqrt{N-1}$ em vez de $\sqrt{N}$ no denominador da equação (5.8).

Se x tem uma distribuição normal (ver 5.6), à medida que o número de medições da amostra aumenta o valor de σ_x tende para um valor constante positivo. Ou seja, para grandes amostras o desvio padrão é independente de N (a largura da distribuição não se altera com N).

5.4 Erro padrão

É razoável aceitar que quanto maior for uma amostra mais confiança terei na inferência de que a média obtida é uma boa estimativa do valor médio. De facto é possível demonstrar que quanto maior for o número de medições (N), mais próxima estará a média do valor médio [4] (desde que x tenha uma distribuição normal - ver 5.6):

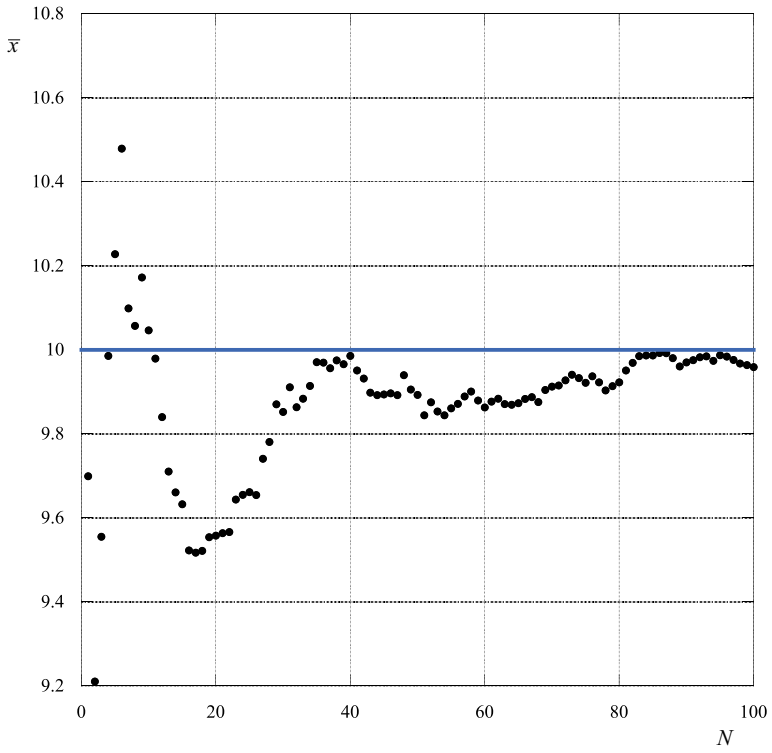

Figura 5.7: Exemplo da variação da média com o número de medições

Na figura 5.7 vemos que a média converge para o valor médio (10) à medida que o número de medições aumenta. A distância entre os dois valores diminui com $1/\sqrt{N}$.

Podemos então definir como *erro padrão* ou *erro da média* (μ_x) a seguinte quantidade:

$$\mu_x = \frac{\sigma_x}{\sqrt{N}} \tag{5.9}$$

Este erro será então uma medida do erro estatístico cometido quando afirmamos que a média de N medições da grandeza x coincide com o valor real (x_R) dessa grandeza.

A estimativa de x (x_{est}) será dada por:

$$x_{est} = \overline{x} \pm \mu_x \tag{5.10}$$

Se todas as medições de x são independentes e igualmente distribuídas então as médias obtidas para amostras de vários tamanhos têm uma distribuição normal em torno do valor médio com um desvio padrão igual a $\sigma_x/\sqrt{N}$. Logo podemos dizer que temos um grau de confiança de 68% de que x_R estará entre $\overline{x} - \mu_x$ e $\overline{x} + \mu_x$.

5.5 Histograma

Um histograma é basicamente uma representação gráfica da frequência de uma grandeza num conjunto discreto de intervalos. Cada intervalo denomina-se de classe. Para uma série de medições sucessivas faz-se a contagem do número de eventos que ocorrem dentro de cada classe.

O histograma pode ser visto como uma tentativa rudimentar de estimar a forma da densidade de probabilidade a partir de uma amostra. Utilizando critérios de construção (ver 5.5.1) podemos até vislumbrar a forma da densidade de probabilidade para amostras pequenas.

5.5.1 Construção de um histograma

Existem muitas formas de optimizar um histograma. Ou seja, ser capaz de vizualizar o todo (população) a partir da parte (amostra). Um parâmetro decisivo neste propósito é o número de classes (n_C) em que a amplitude dos dados (Δ) será dividida. A amplitude de dados é definida como:

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

em que x_{max} é o valor máximo medido e x_{min} o valor mínimo.

Se o número de classes for muito grande para um número pequeno de dados, o histograma terá uma quantidade visível de classes sem eventos (será muito irregular). No entanto, se o número de classes for demasiado pequeno o histograma pode reduzir-se a um só bloco que pouco ou nada tem a ver com a forma da densidade de probabilidade.

A decisão do número de classes é portanto um problema de optimização em que a variável mais importante é o número de elementos da amostra. Como é habitual em problemas de optimização há muitas soluções possíveis. Vejamos algumas:

- Critério de Sturges
$$n_C = \lceil \log_2 N \rceil + 1 \tag{5.11}$$
- Critério de Scott
$$n_C = \frac{\Delta}{3.5\sigma} \sqrt[3]{N} \tag{5.12}$$
- Critério da raíz quadrada
$$n_C = \sqrt{N} \tag{5.13}$$
- Critério de Freedman Diaconis
$$n_C = \frac{\Delta}{2IQR} \sqrt[3]{N} \tag{5.14}$$

Uma vez decidido o número de classes podemos definir as classes C_i em que $i \in \{1, \ldots, n_C\}$:

$$C_i = \left[x_{min} + (i-1)\frac{\Delta}{n_C}, x_{min} + i\frac{\Delta}{n_C} \right[\, , \forall i < n_C$$

$$C_{n_C} = \left[x_{max} - \frac{\Delta}{n_C}, x_{max} \right]$$

5.5.2 Análise de histogramas: individual, do grupo e da turma.

Com base no histograma posso estimar a probabilidade de medir dentro de um certo intervalo. Por exemplo, se um aluno acertou 25% dos lançamentos entre 5.0 e 9.0 então posso inferir que um novo lançamento terá uma probabilidade de 25% de acertar dentro desse intervalo.

5.6 A distribuição normal

A densidade de probabilidade para uma distribuição normal é dada por:

$$\phi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\overline{x})^2}{2\sigma^2}} \tag{5.15}$$

Podemos observar que a função depende apenas de dois parâmetros: a média ($\overline{x}$) e o desvio padrão (σ). Isto implica que se uma grandeza está distribuída normalmente podemos facilmente estimar qual é a sua distribuição a partir dos valores da média e do desvio padrão de uma amostra. A forma da função é a apresentada na figura 5.9. Podemos observar que:

- a média é o valor mais provável (pico da distribuição) por isso quando repetirmos uma medição tomaremos como *estimativa do valor real* dessa grandeza a média desses valores.
- o desvio padrão é uma medida possível da largura da distribuição. Ou seja, a largura da distribuição está relacionada com a dispersão dos dados.

5.6.1 Intervalo de confiança e grau de confiança

Ao olhar para o histograma de um conjunto de N medições experimentais podemos avaliar a percentagem de medições que ocorreram (passado) dentro de um certo intervalo de valores. Podemos com esta informação saltar do passado para o futuro. Para tal podemos assumir que a frequência de ocorrência de valores dentro de um certo intervalo é uma estimativa da probabilidade da próxima medição $(N+1)$ acontecer com um valor dentro desse intervalo. O intervalo denomina-se de *intervalo de confiança* porque temos uma estimativa do *grau de confiança* (dado pela probabilidade anterior) da próxima medição aí estar incluída.

Por exemplo, suponhamos que uma medição foi executada um número suficiente de vezes para que possamos assumir que o histograma obtido denuncia tratar-se de um grandeza que segue uma distribuição normal. Sendo assim, sabemos pela equação (5.15) que toda a distribuição pode ser descrita a partir de apenas dois parâmetros (a média e o desvio padrão). Sendo assim consideremos o seguinte intervalo: $[\overline{x}-\sigma, \overline{x}+\sigma]$. É possível provar que a área da densidade de probabilidade neste intervalo é de 0.68. Ou seja, temos um grau de confiança de 68% de que uma medição suplementar ocorrerá dentro do intervalo que vai desde $\overline{x}-\sigma$ até $\overline{x}+\sigma$.

Este intervalo de confiança para uma distribuição normal será utilizado recorrentemente ao longo destas aulas e será representado por um rectângulo seccionado por um segmento de recta vertical:

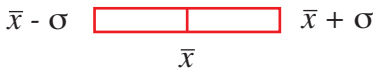

Figura 5.8: Representação gráfica de um intervalo de confiança

O segmento de recta vertical representa a posição da média $\overline{x}$. O lado esquerdo do rectângulo assinala a posição de $\overline{x}-\sigma$ e o direito a posição de $\overline{x}+\sigma$. Isto quer dizer que quando virmos um resultado experimental representado desta forma temos um grau de confiança de 68% de que uma medição posterior ocorrerá dentro deste intervalo de confiança (desde que a sua distribuição seja normal).

Se quizermos ter um grau de confiança superior podemos aumentar o intervalo de confiança:

- o intervalo de confiança $[\overline{x}-2\sigma, \overline{x}+2\sigma]$ tem um grau de confiança de 95%
- o intervalo de confiança $[\overline{x}-3\sigma, \overline{x}+3\sigma]$ tem um grau de confiança maior que 99%

5.6.2 Propriedades

Outra forma de quantificar a dispersão dos valores numa distribuição normal é através da largura da distribuição a meia altura (Full width half maximum - FWHM). É possível demonstrar (ver Apêndice A) que esta quantidade está relacionada com o desvio padrão de acordo com:

$$FWHM = 2\sigma\sqrt{2\ln 2} \simeq 2.355\sigma$$

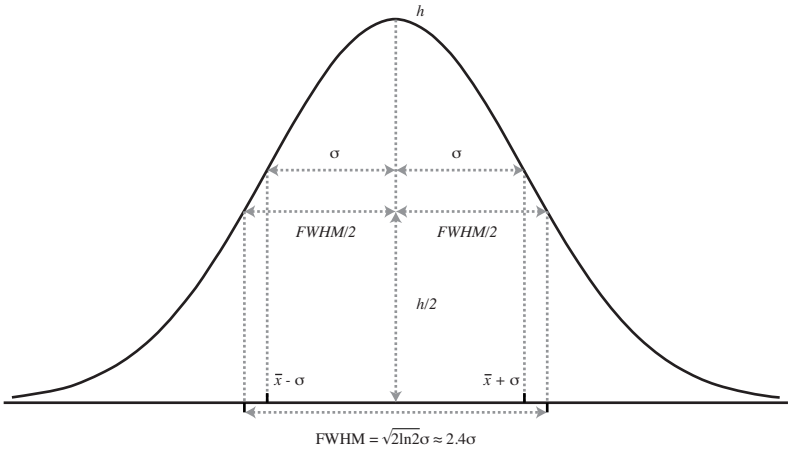

Figura 5.9: Largura a meia altura de uma distribuição normal