Métodos Quantitativos (e Qualitativos)

3. Estimação

Salvatore Cosentino
D.Mat. U.Minho

15 dez 2020

Observações

Um físico tem uma teoria física, que contém um observável chamado x

(a constante de gravitação, a massa do electrão, o tempo característico do carbono C_{14}, ... a probabilidade de sair cara no lançamento de uma moeda).

Repete várias vezes uma experiência em condições que ele julga idênticas (no sentido em que controla tudo o que é controlável)

e obtém os resultados experimentais

$$x_1, x_2, \ldots, x_n$$

possivelmente com n grande.

A coisa mais honesta que ele pode dizer é que o observável está entre

$$x_{\min} \qquad \text{e} \qquad x_{\max}$$

mais ou menos.

Média aritmética

Os físicos costumam acreditar na existência do universo, e nas próprias teorias, portanto na existência do valor verdadeiro de x.

Uma aposta natural é a média (aritmética) (em inglês, (arithmetic) mean/average) dos resultados

$$\boxed{\overline{x} = \frac{1}{n}\left(x_1 + x_2 + x_3 + \cdots + x_n\right)}$$

A média aritmética $\overline{x}$ é a média mais democrática entre os valores observados.

Como tal, pode ser grandemente afetada por possíveis outliers, dados experimentais particularmente afastados da maioria dos outros.

Erros

Os físicos também sabem que não faz sentido nenhum acreditar que o valor de x seja exatamente $\overline{x}$

(as leis da física implicam que a posição de Vénus influencie a queda de uma pedra da torre de Pisa, embora não seja possível dizer qual é a sua influência!)

só acreditam em afirmações como

$$\boxed{x \simeq \overline{x} \pm \Delta x}$$

que lêm: "o verdadeiro valor do observável x está, com grande probabilidade", entre $\overline{x} - \Delta x$ e $\overline{x} + \Delta x$, mais ou menos".

O Δx é o erroestimado.

Mais significativo é o erro relativo

$$\frac{\Delta x}{\overline{x}}$$

ou seja, o número de dígitos decimais confiáveis na estimação de x.

Propagação dos erros

Se o raio de uma esfera é estimado ser $r \pm \Delta r$, então o seu volume é estimado ser

$$\frac{4}{3}\pi \ (r \pm \Delta r)^3 \simeq \frac{4}{3}\pi r^3 \pm 4\pi r^2 \, \Delta r$$

ou seja, $V \pm \Delta V$ com um erro relativo de

$$\boxed{\frac{\Delta V}{V} \simeq 3 \, \frac{\Delta r}{r}}$$

Se os dois lados de um retângulo são etimados ser $a \pm \Delta a$ e $b \pm \Delta b$, respetivamente, então a sua área é estimada ser

$$(a \pm \Delta a) \cdot (b \pm \Delta b) \simeq ab \pm (a \cdot \Delta b + b \cdot \Delta a)$$

ou seja $A \pm \Delta A$ com um erro relativo é da ordem de

$$\boxed{\frac{\Delta A}{A} \simeq \frac{\Delta a}{a} + \frac{\Delta b}{b}}$$

O que a estatística faz e não faz

Todas as medições estão afetadas por erros sistemáticos e por erros aleatórios

Os erros sistemático (ou seja, não estamos a medir exatamente o que pensamos estar a medir!) dependem da projetação das experiências, das técnicas ou dos instrumentos utilizados, e tipicamente são detetados apenas com o avance do conhecimento.

(exemplo das medições da velocidade da luz por parte de Michelson)

O que a estatística não faz é ajudar a detetar o erros sistemáticos !

Um dos problemas da estatística é estimar um valor razoável dos erros devido a fenómenos aleatórios, pequenas variações aleatórias das condições do laboratório, ou efeitos da amostragem.

Mínimos quadrados

A média aritmética $\overline{x}$ é também o valor de a que minimiza a soma

$$(x_1 - a)^2 + (x_2 - a)^2 + ... + (x_n - a)^2$$

dos quadrados dos desvios nas distintas observações.

Se acreditamos que $\overline{x}$ seja uma boa estimação do valor de x, então

$$\varepsilon_k = x_k - \overline{x}$$

pode ser interpretado como sendo o erro observado na k-ésima observação.

O tamanho típico dos

$$|\varepsilon_k|$$

é uma medida da sensibilidade, ou seja, da precisão, dos instrumentos usados no laboratório.

Desvio padrão

A média aritmética dos desvios quadráticos

$$\boxed{S_x^2 = \frac{1}{n}\left((x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots (x_n - \overline{x})^2\right)}$$

é chamada variância (amostral). A sua raiz,

$$S_x = \sqrt{S_x^2}$$

é dita desvio padrão (amostral) (em inglês, standard deviation, ou standard uncertainty).

Uma apresentação honesta dos resultados das n experiências é

$$\boxed{x = \overline{x} \pm S_x}$$

que significa: "foram observadas flutuações da ordem de S_x em torno de um valor médio igual a $\overline{x}$ ".

Como apresentar os dados

Os resultados de uma experiência costumam ser apresentados numa das seguintes maneiras:

- indicando a média e o seu desvio padrão (em inglês, e.s.d., ou seja, estimated standard deviation), ou seja

$$\boxed{x = \overline{x} \pm S_{\overline{x}} \text{ (1 e.s.d. error limit)}}$$

- indicando a média e o seu desvio padrão relativo

$$\boxed{x = \overline{x} \text{ with relative standard deviation } S_{\overline{x}}/\overline{x}}$$

Dígitos significativos

O desvio padrão relativo mostra os dígitos significativos da nossa estimação.

Dizer que um observável é igual a

$$3.14159265359 \pm 0.062$$

não contém mais informação do que dizer que é igual a

$$3.14 \pm 0.06$$

(e também não há grande diferença entre 0.06 e 0.05 ou 0.07 . . .)

Por exemplo, uma tabela das constantes da física tem este valor da constante de gravitação de Newton:

$G = 6.673(10) \times 10^{-11}\mathrm{m}^3\mathrm{kg}^{-1}\mathrm{s}^{-2}$ with relative standard uncertainty 1.5×10^{-3}

Isto quer dizer que, embora a média observada seja

$$6.67310 \times 10^{-11}\mathrm{m}^3\mathrm{kg}^{-1}\mathrm{s}^{-2}$$

apenas podemos confiar nos primeiros três dígitos decimais deste valor.

Desvio padrão da média

O senso comun sugere que quanto maior for o número n das observações quanto mais próxima a média $\overline{x}$ está do verdadeiro valor de x.

Conjeturas razoáveis sobre a distribuição dos erros $x_k - x$ (sugeridas pelos histogramas dos dados experimentais) e considerações probabilísticas (o teorema do limite central) permitem quantificar esta expetativa.

Por exemplo, se n é grande e os histogramas dos dados experimentais fazem suspeitar que a distribuição dos erros é gaussiana, é possível mostrar que as flutuações da média amostral $\overline{x}$ em torno do valor verdadeiro são da ordem de

$$S_{\overline{x}} = \frac{S_x}{\sqrt{n}}$$

dito desvio padrão da média (em inglês, standard deviation of the mean).

Mais significativo é o desvio padrão relativo da média (em inglês, relative standard deviation/uncertainty),

$$\frac{S_{\overline{x}}}{\overline{x}} = \frac{S_x}{\overline{x}\sqrt{n}}$$

que diz o tamanho do erro relativo.

Fim da parte elementar

Um modelo das observações

Existe um valor verdadeiro do observável x, que chamamos m.

Cada observação é uma experiências aleatória, descrita pela variável aleatória ξ com esperança

$$\mathbf{E}\xi = m$$

(ou seja, acreditamos que os instrumentos medem o observável x, não há erros sistemáticos)

O nosso controlo das condições do laboratório não é, não pode ser, perfeito, portanto a variável ξ é mesmo variável, e tem uma certa lei (desconhecida),

e em particular variância positiva

$$\mathbf{V}\xi = \sigma^2.$$

Também é razoável assumir que as diferentes observações são independentes

(fazer física é possível precisamente na medida em que físicos que vivem em laboratórios distintos, um em Braga e outro em Guimarães, podem reproduzir e verificar as experiências dos outros: a "independência" das experiências é uma das hipóteses necessárias para a reprodutibilidade).

Lei dos grandes números

Então a média aritmética das n observações, que os estatísticos chamam média amostral, tem boa probabilidade de estar próxima da esperança m, o valor verdadeiro de x, ou seja

$$\boxed{\overline{x} = \frac{1}{n}\left(x_1 + x_2 + ... + x_n\right) \simeq m}$$

quando o número das observações n é grande.

Este é o conteúdo da lei dos grandes números, uma coleção de teoremas que os matemáticos sabem provar desde os tempos de Bernoulli, dependendo de certas condições razoáveis sobre a variável.

Teorema limte central e hipótese gaussiana

Os histogramas dos resultados das experiências reais podem ser muito parecidos com o gráfico de uma distrubuição normal.

Uma hipótese de trabalho pode ser que ξ tem lei normal/gaussiana

$$\xi \sim N(m, \sigma^2)$$

Então a média aritmética $\overline{x}$ tem lei normal com média $\mathbf{E}\overline{x} = m$ e variância $\mathbf{V}\overline{x} = \sigma^2/n$, ou seja

$$\overline{x} \sim N(m, \sigma^2/n)$$

Em particular, a varãncia de $\overline{x}$ é inversamente proporcional ao número de observações (uma manifestação do princípio da raiz quadrada).

O modelo faz previsões quantitativas: diz que a variável normalizada

$$\frac{\overline{x} - m}{\sigma/\sqrt{n}}$$

tem lei normal reduzida, ou seja,

$$\boxed{\frac{\overline{x} - m}{\sigma/\sqrt{n}} \sim N(0, 1)}$$

Somas de pequenos erros aleatórios

Por outro lado, mesmo se ξ não fôr normal, o teorema limite central de De Moivre e Laplace diz que, quando n é muito grande, a lei de $\overline{x}$ é bem aproximada pela lei normal.

Se n não é muito grande, também existe uma "justificação" para a hipótese gaussiana.

É razoável pensar que a variável ξ seja igual ao vaor verdadeiro m mais uma soma de pequenos erros aleatórios

$$\xi = m + \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \dots$$

devidos a pequenas perturbações incontroláveis das condições de laboratório

(uma borboleta que posou no aparelho, uma eclipse de lua, . . .).

Mais uma vez, se os erros são muitos, e se "em média" são nulos, o teorema limite central sugere que a lei de ξ é bem aproximada por uma lei normal com média m.

Variância amostral

Como estimar a variância $\mathbf{V}\xi = \sigma^2$?

Os estatísticos chamam variância amostral

$$\boxed{S_x^2 = \frac{1}{n-1}\left((x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots (x_n - \overline{x})^2\right)}$$

Dentro do nosso modelo das observações esta é uma variável aleatória, porque cada x_k é uma variável. É possível mosrar que a sua esperança é

$$\mathbf{E}S_x^2 \simeq \sigma^2$$

Também útil é saber que a variância de S_x^2 é

$$\mathbf{V}S_x^2 \simeq \frac{2\sigma^2}{n-1}$$

e portanto se n é grande a variância amostral tem boa probabilidade de estar próxima da variância de ξ (lei dos grandes números).

Teorema de Cochran

O modelo diz que a variável

$$\frac{n-1}{\sigma^2} \cdot S_x^2$$

tem lei qui quadrado χ^2_{n-1}, mas esta variável ainda contém a variância desconhecida σ^2.

Mais útil, enfim, é saber que o modelo diz que a variável

$$\boxed{t = \frac{\overline{x} - m}{S_x/\sqrt{n}}}$$

onde só aparece o parâmetro m, o observável que queremos estimar, tem lei de Student com $n-1$ graus de liberdade T_{n-1}.

Em conclusão

Se n é grande, os quantis da lei de Student não diferem significativamente dos quantis da lei normal reduzida $N(0,1)$.

Uma variável normal reduzida está

entre ± 2 com probabilidade superior a 95%,

ou entre ± 3 com probabilidade superior a 99%.

Em conclusão, um físico pode razoavelmente dizer que o verdadeiro valor do observável x está no intervalo de confiança

$$\boxed{x = \overline{x} \pm 3 \cdot \frac{S_x}{\sqrt{n}}}$$

com probabilidade superior a 99%.

Intervalos de confiança

Os resultados de uma experiência podem ser apresentados da seguinte forma: o valor do observável x está no intervalo

$$\boxed{a \leq x \leq b}$$

dito intervalo de confiança, com probabilidade $\geq 1 - \alpha$, dita nível (do intervalo de confiança).

Um intervalo de confiança simétrico, i.e. do tipo

$$a - \varepsilon \leq x \leq a + \varepsilon$$

costuma ser apresentado pela expressão

$$\boxed{x = a \pm \varepsilon}$$

Intervalos para a média

Num modelo em que temos que estimar a variância amostral (ou seja, sempre!), um intervalo de confiança de nível $1-\alpha$ é

$$\boxed{m = \overline{x} \pm t_{1-\alpha/2} \cdot \frac{S_x}{\sqrt{n}}}$$

onde $t_{1-\alpha/2}$ é o quantil da lei de Student T_{n-1}.

Se n é grande, os quantis da lei de student T_{n-1} n ao diferem significativamente dos quntis da lei normal reduzida. Assim, um intervalo de confiança de nível $1-\alpha$ para a média é

$$\boxed{m = \overline{x} \pm \phi_{1-\alpha/2} \cdot \frac{S_x}{\sqrt{n}}}$$

onde $\phi_{1-\alpha/2}$ é o quantil da lei normal.

Valores típicos são $\phi_{0.975} \simeq 1.96$ se o nível é $1-\alpha = 95\%$, ou $\phi_{0.995} \simeq 2.6$ se o nível é $1-\alpha = 99\%$.

Intervalos para uma probabilidade/proporção

Um caso particular é uma experiência em que queremos estimar uma probabilidade p, ou seja os resultados possíveis são $x_k = 0$ ou 1 e o resultado das experiências é a frequência observada

$$f = \overline{x} = \frac{\text{“número de sucessos em } n \text{ provas”}}{n}$$

Se n é suficientemente grande, a lei da variável

$$\frac{\overline{x} - p}{\sqrt{p(1-p)}/\sqrt{n}} \sim N(0,1)$$

é bem aproximada pela lei normal reduzida.

Uma boa ideia é estimar a variância $p(1-p)$ com o seu máximo

$$p(1-p) \leq \frac{1}{4}$$

Um intervalo generoso de nível $\geq 1-\alpha$ para a probabilidade p é portanto

$$\boxed{p = \overline{x} \pm \phi_{1-\alpha/2} \cdot \frac{1}{2\sqrt{n}}}$$

Exemplo: sondagens

N americanos podem escolher entre os candidatos B ou T.

Dentro de uma amostra de $n < N$ eleitores, b' eleitores afirmam estar intencionados em votar o senhor B e os outros $t' = n - b'$ afirmam estar intencionados em votar o senhor T.

O problema é estimar o número b de eleitores, dentro da população total, que estão intencionados em votar B, e portanto a percentagem

$$p = \frac{b}{N}$$

A variável b' tem lei hipergeométrica,

que, se $n \ll N$, é bem aproximada pela lei binomial $B(n, p)$

(escolher uma amostra de n eleitores não difere muito de escolher n vezes um eleitor).

Portanto, um intervalo de confiança 95% para a percentagem $p = b/N$ é

$$\boxed{p = \frac{b'}{n} \pm \frac{1}{\sqrt{n}}}$$

O $\pm 1/\sqrt{n}$ é o que os técnicos chamam margem de erro da sondagem.

Exemplo: a agulha de Buffon

O assoalho é feito de tábuas de largura ℓ. Lanço n vezes uma agulha de comprimento ℓ e registo a frequência f_n das vezes que a agulha toca uma das junções.

Estas são provas de Bernoulli onde a probabilidade de sucesso (que os matemático sabem calcular) é

$$p = \frac{2}{\pi}$$

Um intervalo de confiança de níivel 95% para a probabilidade p é

$$p = f_n \pm \Delta f_n$$

onde $\Delta f_n \simeq 1/\sqrt{n}$.

O observável π, a área di um disco de raio um, é igual a $2/p$. Portanto, usando a lei de propagação dos erros, um físico que quer estimar π escreve

$$\pi = \frac{2}{f_n} \pm \frac{2}{f_n^2} \Delta f_n$$

Exercício: estimar π

Em 1901, o senhor Lazzarini afirmou ter lançado $n = 3408$ agulhas obtendo

$$\frac{2}{f_n} \simeq 3.1415929$$

Quantos dígitos desta estimação são confiáveis?

Se n é grande, podemos aproximar

$$f_n \sim p \sim 2/\pi \sim 0.637$$

A magem de erro de um intervalo de confiança de nível 95% é portanto

$$\Delta\pi \sim \frac{2}{f_n^2}\,\frac{1}{\sqrt{n}} \sim 0.08$$

Assim,

$$\pi = 3.14 \pm 0.08$$

ou seja, apenas o segundo dígito decimal é confiável, embora os primeiros 6 sejam corretos!

O mais provável é que o senhor Lazzarini fez batota . . .

Intervalos para a variância

Às vezes é importante estimar a variância σ^2 dos resultados das experiências

(é uma medida da reproducibilidade da experiência, ou da sensibilidade dos instrumentos do laboratório).

No modelo, a variável

$$\frac{n-1}{\sigma^2} \cdot S_x^2 \sim \chi^2(n-1)$$

tem lei qui-quadrado com $n-1$ graus de liberdade.

Fixado um nível $1-\alpha$, dois intervalos de confiança (asimétrico ou simétrico) são

$$\boxed{0 \leq \sigma^2 \leq \frac{n-1}{q_\alpha} \cdot S_x^2}$$

e

$$\boxed{\frac{n-1}{q_{1-\alpha/2}} \cdot S_x^2 \leq \sigma^2 \leq \frac{n-1}{q_{\alpha/2}} \cdot S_x^2}$$

onde q_α, $q_{1-\alpha/2}$ e $q_{\alpha/2}$ são os quantis da lei $\chi^2(n-1)$.

Na prática, . . .

Se não temos tabelas no laboratório, basta lembrar que intervalos de confiança "generosos" com nível de confiança $\geq 95\%$ ou $\geq 99\%$ ou ≥ 99.7 são da ordem de

$$\boxed{\overline{x} \pm 2\,\frac{S_x}{\sqrt{n}}}$$

ou

$$\boxed{\overline{x} \pm 2.6\,\frac{S_x}{\sqrt{n}}}$$

ou

$$\boxed{\overline{x} \pm 3\,\frac{S_x}{\sqrt{n}}}$$

respetivamente

(os quantis da lei de Student não dão erros relativos significativamente diferentes dos quantis da lei normal, se n é grande).

Como apresentar os dados

Se o número n de observações é grande e/ou depois de ter visto os histogramas julgamos que a distribuição dos erros é normal, os resultados de uma experiência costumam ser apresentados numa das seguintes maneiras:

- indicando a média e o seu desvio padrão estimado (e.s.d., ou seja "estimated standard deviation" of the mean), ou seja

$$\boxed{x = \overline{x} \pm S_{\overline{x}} \text{ (1 e.s.d. error limit)}}$$

- indicando a média e o seu desvio padrão relativo estimado

$$\boxed{x = \overline{x} \text{ with relative standard deviation } S_{\overline{x}}/\overline{x}}$$

- indicando um intervalo de confiança, por exemplo

$$\boxed{x = \overline{x} \pm t_{0.975} \cdot S_m \text{ (95\% confidence, } \nu = n - 1)}$$

onde lembramos ao leitor que o quantil $t_{0.975}$ da lei de Student foi obtido com $\nu = n - 1$ graus de liberdade (e portanto que foram feitas n observações).

Quantas observações fazer

Vale a pena observar que o erro (e o erro relativo) na estimação de um observável é inversamente proporcional a raiz quadrada

$$\sqrt{n}$$

do número n de observações.

Portanto, para reduzir o erro de um factor 10, um físico tem que multiplicar por 100 as observações.

Por exemplo, se para estimar uma probabilidade p com um erro da ordem de 3% é preciso fazer uma sondagem com uma amostra de

$$n \sim 1000$$

pessoas, para reduzir o erro a 0.3% é preciso escolher uma amostra de

$$n \sim 100000$$

pessoas!

R.A.

Erros aleatórios e erros sistemáticos.

Apresentação dos resultados: média e desvio padrão amostrais.

Desvio padrão da média, desvio padrão relativo.

Intervalos de confiança, nível do intervalo.

Intervalos de confiança para uma média.

Intervalos de confiança para uma variância.

Intervalos de confiança para uma probabilidade.

Margem de erro de uma sondagem e sua dependência da dimensão da amostra.