Métodos Quantitativos (e Qualitativos)

4. Amostragem

Salvatore Cosentino
D.Mat. U.Minho

15 dez 2020

Populações e caraterísticas

O objetivo típico de um estudo estatístico é compreender algumas caraterísticas de uma população,

(os lobos que vivem na Serra do Gerês,
as estrelas da nossa galáxia,
os reclusos nas prisões portuguesas, . . .)

As caraterísticas são observáveis, ou seja, variáveis, quantitativas e/ou qualitativas, que podem ser associadas/medidas a cada indivíduo, ou elemento, da população

(a idade, sexo, peso . . . dos lobos,
a cor/temperatura, a luminosidade/tamanho, a velocidade . . . das estrelas,
o crimem, a duração da pena, a escolaridade, . . . dos reclusos . . .)

As variáveis são descritas por certas distribuições, e são realmente "variáveis", ou seja, há diversidades na população!

Amostragem

Pode ser praticamente impossível obter informações sobre todos os N indivíduos da população, e portanto sobre a distribuição real das variáveis em estudo.

Isto conduz à necessidade de escolher/individuar/selecionar uma amostra (em inglês, sample), ou seja, um subconjunto da população, tipicamente formada por um número de indivíduos (dimensão da amostra)

$$n \ll N$$

(algumas dezenas de lobos que conseguimos observar,
alguns milhares de estrelas nos catálogos dos astrofísicos,
algumas centenas de reclusos que aceitem responder aos inquéritos, . . .)

e observar os valores das variáveis nos indivíduos da amostra, os dados estatísticos.

na esperança de que a amostra seja representativa da população, ou seja, que permita obter uma imagem credível da distribuição das caraterísticas da população em que estamos interessados.

As diferenças entre as caraterísticas da amostra e as caraterísticas da população são chamadas genericamente erros de amostragem (em inglês, sampling errors). Uma amostra é pouco representativa de uma população quando os erros são grandes.

Sobre as generalizações

A esperança é, portanto, poder generalizar as observações feitas sobre uma amostra a toda a população

A mathematician, a physicist, and an engineer are riding a train through Scotland.

The engineer looks out the window, sees a black sheep, and exclaims, "Hey! They've got black sheeps in Scotland!"

The physicist looks out the window and corrects the engineer, "Strictly speaking, all we know is that there's at least one black sheep in Scotland."

The mathematician looks out the window and corrects the physicist, "Strictly speaking, all we know is that at least one side of one sheep is black in Scotland."

O problema das sondagens, simplificado

No lago há x peixinhos vermelhos e y peixinhos azuis, que formam uma população de

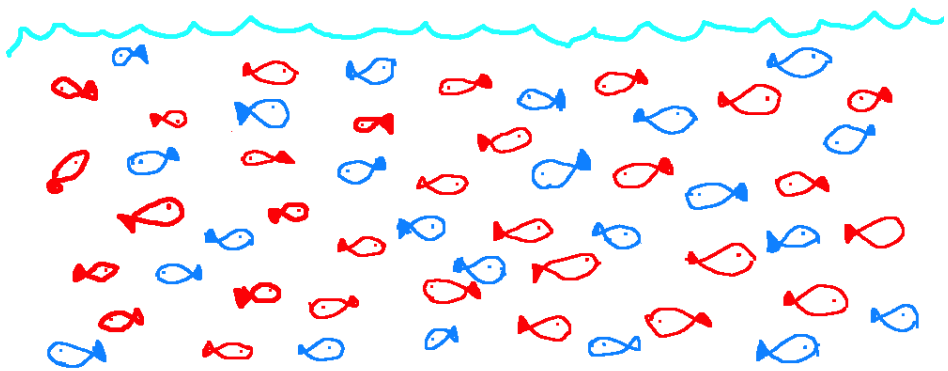
$$N = x + y$$

peixinhos.

Queremos estimar a proporção

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

ou, pelo menos, decidir se $x > y$ (quem ganha o referendum), sem ter que pescar todos os peixinhos (antes da votação).



A situação real é mais complicada, pois estes números são variáveis no tempo (decidem na hora do voto), e até pode haver peixinhos sem cor (que anulam o seu voto) ou que nunca serão pescados (que não votam) . . .

Sondagens ideais

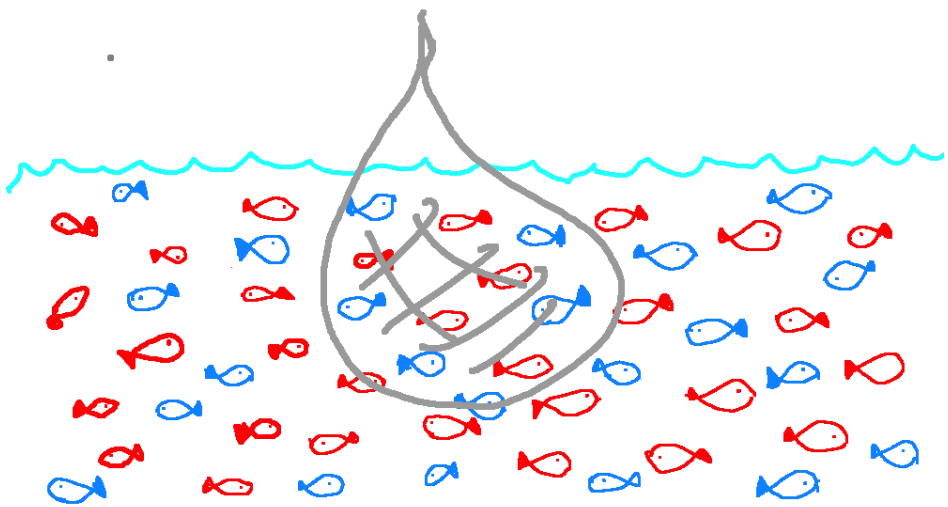
A ideia das sondagens é pescar, sem ver, n peixinhos e depois observar as cores: por exemplo, x_n peixinhos vermelhos e $y_n = n - x_n$ peixinhos azuis

(mas pode haver peixinhos que não querem dizer a cor!)

A melhor aposta para o verdadeiro valor de p é

$$p \simeq f_n = \frac{x_n}{n}$$

a proporção dos peixinhos vermelhos pescados.



Precisão das sondagens ideais

Se $n \ll N$, escolher uma amostra de n peixinhos não difere muito de pescar n vezes um peixinhos (e voltar a repor o peixinho no lago), logo a lei da variável X_n que conta o número de peixinhos vermelhos dentro da amostra é bem aproximada por uma binomial

$$X_n \sim B(p, n)$$

com valor médio $\mathbf{E}X_n = np$ e variância $\mathbf{V}X_n = n\sigma^2 = np(1-p)$.

Se $n \gg 1$, o teorema limite central diz que a variável

$$\xi_n = \frac{(f_n - p)}{\sigma/\sqrt{n}} \sim N(0,1)$$

e portanto $\mathbf{P}(|\xi_n| \leq 2) \simeq 95\%$ e $\mathbf{P}(|\xi_n| \leq 2.6) \simeq 99\%$.

Sendo o desvio padrão limitado a priori por

$$\sigma = \sqrt{p(1-p)} \leq \sqrt{1/4}$$

podemos afirmar que

$$\boxed{p \simeq f_n \pm 2\,\frac{1}{\sqrt{4n}}} \qquad \text{ou} \qquad \boxed{p \simeq f_n \pm 2.6\,\frac{1}{\sqrt{4n}}}$$

com probabilidade não inferior a 95% ou 99%, respetivamente.

Margens de erro

Por exemplo, usando um nível de confiança 95%,

uma amostra de $n \simeq 100$ peixinhos garante uma margem de erro de

$$\boxed{\pm 13\%}$$

uma amostra de $n \simeq 1000$ peixinhos garante uma margem de erro de

$$\boxed{\pm 3.1\%}$$

(uma precisão razoável)

uma amostra de $n \simeq 10000$ peixinhos garante uma margem de erro de

$$\boxed{\pm 1.3\%}$$

(uma precisão mais que razoável!)

Estimação de outros parâmetros

Outros parâmetros,como a média ou a variâcia, ou a própria distribuição, de uma variável x podem também ser estimados usando os dados estatísticos observados numa amostra.

Uma ideia da distribuição de x na população é o próprio istograma dos dados estatísticos.

Por exemplo, um intervalo de confiança 95% para a média $\mathbf{E}x = m$ é

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

onde $\overline{x}$ é a média amostral e S_x é o desvio padrão amostral.

Assim, a margem de erro, ou o sampling error, é

$$\boxed{\Delta x \simeq 2\,\frac{S_x}{\sqrt{n}}}$$

Margens de erro (sampling error) e dimensões

A margem de erro, logo a representatividade de uma amostra, depende da dimensão da amostra e da variabilidade da população.

A margem de erro é inversamente proporcional a $\sqrt{n}$, ou seja, diminue quando a dimensão da amostra cresce.

A margem de erro é proporcional ao desvio padrão S_x, ou seja, cresce com a variabilidade da caraterística da população.

A margem de erro não depende da razão

$$\frac{n}{N}$$

entre a dimensão n da amostra e a dimensão N da população, desde que $n \ll N$ (assim que N pode ser arbitrariamente grande!)

Por exemplo, uma amostra de

$$1000 \qquad \text{ou} \qquad 1500$$

indivíduos é suficiente para garantir margens de erros razoáveis, da ordem de 2% ou 3%, na estimação de uma probabilidade.

Amostragens aleatórias

É realmente muito difícil, para não dizer impossível, em condições reais, realizar uma amostragem aleatória.

Em teoria, basta ter uma lista (em inglês, sampling frame) de toda a população

(por exemplo, os NIF dos residentes portugueses,
os nomes de todos os reclusos em Portugal, . . .),

e usar um computador para selecionar n elementos, ou, melhor, n vezes um elemento, com probabilidade uniforme

(é o que faz o sorteio da e-fatura, imagino, mas usando as faturas registadas).

Qualquer linguagem de programação moderna dispoe de geradores de sequências numéricas que simulam muito bem sequências aleatórias, distribuidas de acordo com as principais leis probabilísticas (uniforme, binomial, gaussiana, . . .).

Amostragens aleatórias sistemáticas

Também existem técnicas para reduzir tempos e custos necessários para obter amostragem satisfatórios.

Por exemplo, a partir de uma lista, logo de uma ordenação dos N induviduos da população,

$$x_1 \qquad x_2 \qquad x_3 \qquad \dots \qquad x_N$$

é possível escolher "aleatoriamente" apenas um primeiro elemento, por exemplo o k-ésimo,

$$x_k$$

e depois seleccionar rapidamente os n elementos da amostra pretendida

$$x_k \qquad x_{k+d} \qquad x_{k+2d} \qquad \dots \qquad x_{k+(n-1)d}$$

sendo d o intervalo de amostragem, um inteiro da ordem de

$$d \simeq \frac{N}{n}$$

Amostragens aleatórias estratificadas

Os N indivíduos da população são catalogados de acordo com uma certa caraterística, logo divididos em estrados

$$a_1 \quad a_2 \quad \ldots \quad a_{N_a} \quad b_1 \quad b_2 \quad \ldots \quad b_{N_b} \quad \ldots$$

sendo $N = N_a + N_b + \ldots$ a dimensão da população.

(por exemplo, a crença religiosa ou a etnia dos reclusos . . .)

Uma amostra aleatória é seleccionada dentro de cada estrado, formando uma amostra de tamanho

$$n = n_a + n_b + \ldots$$

respeitando a proporção que cada estrado representa dentro da população total,

$$\frac{n_a}{N_a} = \frac{n_b}{N_b} = \ldots$$

(se 60% dos reclusos são católicos e 40% dos reclusos são protestantes, uma amostra de tamanho n será produzida escolhendo $0.4 \times n$ católicos e $0.6 \times n$ protestantes)

Este método garante que cada estrado da população é representado na amostra com o seu peso real.

Amostragens aleatórias em "cluster"

Os N individuos da população podem ser naturalmente divididos em (sub)grupos (em inglês, clusters), cada um representativo da população inteira,

(os alunos de Braga estão divididos entre as diferentes escolas do conselho,
os reclusos portugueses estão divididos entre as diferentes prisões do pais . . .)

Por razões práticas, por exemplo de distribuição dos inquéritos, pode não ser viável escolher uma amostra com individuos que pertencem a todos os grupos.

A estratégia é então selecionar aleatóriamente alguns cluster, e finalmente uma amostra aleatória dentro de cada cluster selecionado.

(escolher 3 escolas de Braga, e uma amostra de $n/3$ alunos de cada uma destas escolas . . .)

. . . ou multi-etapas

Cada grupo pode estar dividido em subgrupos, cada subgrupo em mais subgrupos . . .

(os alunos de Portugal estão divididos entre as diferentes regiões, depois entre os diferentes conselhos, depois entre as diferentes escolas, . . .
os reclusos dos EUA estão divididos entre os diferentes estados, e em cada estado entre as diferentes prisões daquele estado, . . .)

A estratégia é então selecionar aleatóriamente alguns grupos, depois algum sub-grupos dentro dos grupos selecionados, . . .

. . . e finalmente uma amostra aleatória dentro de cada sub-sub-. . . -grupo selecionado.

(para estimar o número de erros num livro, podemos selecionar uma amostra de páginas, e dentro de cada página uma amostra de linhas, . . .)

Também é possível misturar a amostragem por cluster e a amostragem estratificada . . .

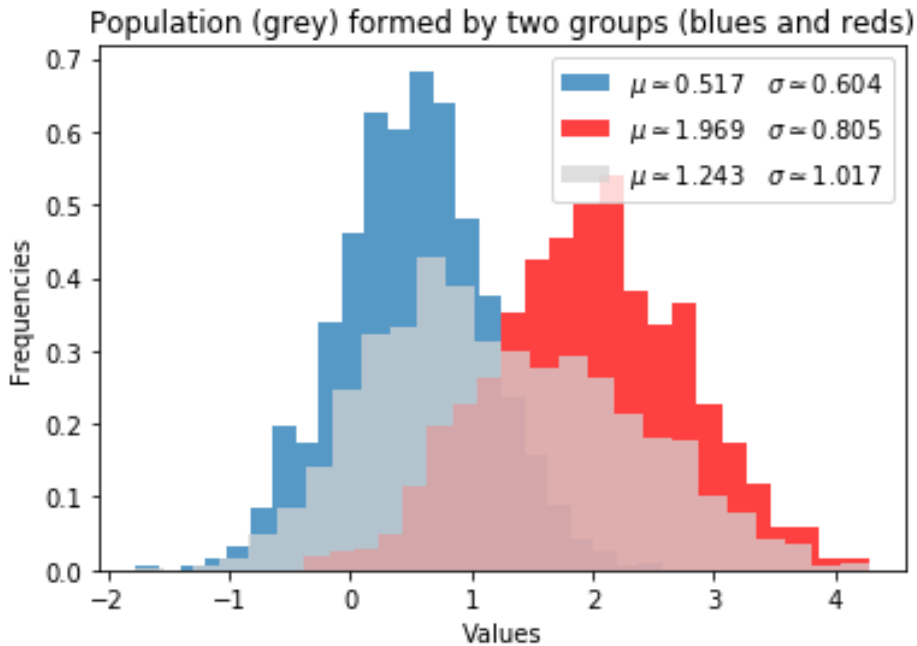
Populações reais

As populações reais não são homogéneas

costumam ser divididas em grupos com caraterísticas diferentes

(as comunidades rurais ou urbanas, as diferentes famílias mafiosas, . . .),

o que coloca dúvidas sobre a "existência" ou "relevância" de um único parâmetro que descreva uma caraterística da população.



Amostragens reais

Os sampling frame das populações reais são raramente corretos, ou podem até não existir

(imigração e emigração mudam as populações,
os sem abrigos são dificilmente observados, . . .)

Outro problema típico da investigação com inquéritos é que os individuos podem não responder,

(porque não sabem ler, ou por outras razões . . .)

e estes indivíduos tipicamente representam grupos da população com caraterísticas distintas dos indivíduos que respondem!

Amostragens empíricas (não aleatórias)

Naturalmente, amostragem não aleatórias não são representivas de uma população !

No entanto, podem ser úteis, ou até ser a única opção, quando não existe um sampling frame, ou em estudos preliminares, ou quando a população é pequena ou identificável com dificuldade.

Amostragem por conveniência
(inquéritos de rua, chamadas dos telespetadores, formulários por e-mail, . . .)

Amostragem por quotas
(amostra que respeita os pesos que diferentes categorias de indivíduos têm dentro da população)

Purposive or judgement sampling
(é selecionado um grupo particular, de interesse do pesquisador)

Snowball sampling
(um contacto sugere/indica os contactos seguintes . . .)

. . . que podem ler no Chapter 5: Sampling de
R.D. Bachman and R.K. Schutt, The practice of Research in Criminology and Criminal Justice, SAGE, 2018.

R.A.

Amostras e populações.

Representatividade de uma amostra e erros de amostragem.

Amostragens aleatórias.

Margem de erro e sua dependência das dimensões da amostra e da variabilidade da população.

Amostragens empíricas.