

Amostragem e Distribuições por Amostragem

Departamento de Matemática
Escola Superior de Tecnologia de Viseu

População, amostra e inferência estatística

Exemplo 1

O João é licenciado em Gestão pela ESTV e está a trabalhar para a fábrica VisaTexteis. A fábrica recebeu recentemente uma encomenda de 500 rolos de linhas de várias cores. O João tem de verificar se os rolos recebidos estão de acordo com as especificações feitas. Entre as especificações de qualidade, exigia-se que cada rolo tivesse pelo menos 500 metros de comprimento. Como deve o João proceder? Obviamente não é possível observar um a um, todos os 500 rolos que compõem esta população.

Exemplo 2

Para elaborar uma notícia, um determinado jornal semanal pretende saber qual a opinião dos portugueses relativamente a um dado projecto governamental. Obviamente, o jornal não poderá inquirir todos os Portugueses.

População, amostra e inferência estatística

Em situações como as ilustradas nos exemplos anteriores, o estudo é feito com base numa parte representativa da população, à qual se dá o nome de **amostra**.

A informação obtida a partir da observação dos elementos de uma amostra, conduz-nos a certas conclusões que depois **inferimos** para toda a população. Estamos a fazer **inferência estatística**.

Seja

X : característica numérica em estudo numa dada população

Na prática, em geral a distribuição de X não é conhecida ou sabe-se qual a forma geral da distribuição de X , mas não são conhecidos os parâmetros dessa distribuição, ditos **parâmetros populacionais**.

População, amostra e inferência estatística

Como estimar o valor de uma média populacional μ_X , de uma variância populacional σ_X^2 ou de uma proporção populacional p ?

- ▶ Estimamos a média da população, μ_X , através da média da amostra, \bar{x} .
- ▶ Usamos o desvio padrão da amostra, s , para aproximar ou estimar o desvio padrão na população, σ_X .
- ▶ Usamos a proporção encontrada na amostra, \hat{p} , para estimar a proporção populacional p .

Amostra aleatória

Para fazer inferência estatística, a amostra deve ser recolhida obedecendo a certos critérios, caso contrário, as conclusões decorrentes do estudo da amostra poderão não ser válidas para toda a população.

Exemplo 2

Suponhamos que o director do jornal, por razões de comodidade, decide recolher opiniões numa amostra de pessoas residentes na cidade onde o jornal é editado.

A amostra retirada naquela cidade não é representativa da população portuguesa e, portanto, as conclusões retiradas no estudo não se podem estender a toda a população portuguesa.

Para obter uma amostra representativa da população portuguesa, parece razoável que se exija que **cada português tenha a mesma probabilidade de vir a figurar na amostra**. Temos então que **escolher os portugueses para a nossa amostra perfeitamente ao acaso, i.e., aleatoriamente**.

Amostra aleatória

Exemplo 1

O João escolheu ao acaso seis rolos do lote recebido e observou o comprimento de cada um deles, tendo obtido os seguintes valores:

$459m$ $455m$ $502m$ $501.5m$ $500.5m$ $456m$

Estes valores constituem uma amostra concreta de tamanho seis.

Antes de seleccionar os rolos para a amostra, o João não é capaz de prever qual irá ser o comprimento do primeiro rolo, do segundo rolo, etc.

O comprimento do i -ésimo rolo seleccionado para a amostra é uma variável aleatória que representamos por X_i .

Amostra aleatória

Tem-se:

- ▶ cada variável aleatória X_i tem a mesma distribuição de X ;
- ▶ as variáveis aleatórias X_1, X_2, \dots, X_n são independentes (selecção aleatória dos rolos - a selecção de um rolo não tem influência na selecção de qualquer outro - o valor que uma das variáveis assume não tem qualquer efeito sobre o valor que outra qualquer assume)

Definição

Seja X uma variável aleatória que representa uma característica numérica em estudo numa determinada população. Chama-se **amostra aleatória de tamanho n** ao conjunto das variáveis aleatórias X_1, X_2, \dots, X_n independentes e com a mesma distribuição de X .

Os valores observados das variáveis aleatórias X_1, X_2, \dots, X_n numa amostra concreta são representados por letras minúsculas:

$$x_1, x_2, \dots, x_n$$

Amostra aleatória

Quando a **população é finita** obtém-se uma amostra aleatória se a selecção dos elementos para a amostra é feita **com reposição**, pois neste caso as sucessivas extracções são independentes. Isto assegura que as variáveis aleatórias identicamente distribuídas X_1, X_2, \dots, X_n sejam independentes.

Muitas vezes, no entanto, a selecção dos elementos para uma amostra é feita **sem reposição**. Neste caso as variáveis aleatórias X_1, X_2, \dots, X_n não serão independentes pois os valores que os primeiros elementos da amostra tomam condicionam os seguintes.

Amostra aleatória

Porém, se a amostra é pequena relativamente à população, a diferença entre reposição e não reposição é atenuada, já que a retirada de alguns elementos não altera drasticamente a composição da população e por isso a não reposição do item examinado terá efeito desprezível.

Na prática, quando é feita amostragem sem reposição, é usual assumir a independência entre as variáveis aleatórias X_1, X_2, \dots, X_n , se a amostra não exceder 5% do tamanho da população. Assim, se, ao contrário, a amostra exceder 5% do tamanho da população, deve-se fazer amostragem com reposição.

Do que foi dito podemos concluir que quando a população é infinita é indiferente fazer ou não reposição; a amostra recolhida será sempre aleatória.

Números aleatórios

Exemplo 1

O João decidiu recolher uma amostra aleatória de tamanho 10. Para que a escolha dos rolos fosse feita ao acaso, aleatoriamente, o João procedeu da seguinte maneira:

- ▶ numerou os rolos;
- ▶ colocou dentro de uma caixa quinhentos pedaços de papel numerados de 1 a 500;
- ▶ depois de os mexer bem, retirou um papel e anotou o número (1º rolo da amostra);
- ▶ continuou o processo até obter os números de identificação dos dez rolos que irão constituir a amostra.

Recorrendo a uma tabela de números aleatórios, podemos recolher uma amostra sem ter de passar por aquele esforço.

Tabela de números aleatórios

Uma [tabela de números aleatórios](#) é uma tabela com números obtidos por um processo aleatório.

Extracto de uma tabela de números aleatórias:

77921	06907	11008
42751	99562	72905
56420	69994	96301
91977	05463	07972

Estas tabelas são construídas de forma a garantir que numa dada sequência de algarismos, a probabilidade de aparecimento de um algarismo num determinado ponto da sequência é igual para todos os algarismos de 0 até 9 (0.1).

Números pseudo-aleatórios

Hoje em dia existem programas de computador que geram números aleatórios.

No entanto, é de salientar que estes números não são verdadeiramente aleatórios, pois não estão sujeitos à chance. De facto, os números são produzidos por um algoritmo e portanto são resultado de um processo determinístico.

Por isso estes números são, por vezes, designados [pseudo-aleatórios](#).

Exemplo 1

Como recolher uma amostra de 10 rolos do lote de 500 recebido pela fábrica?

- ▶ Cada rolo é identificado com um número de três algarismos (000–499).
- ▶ Todos os números de identificação devem ter o mesmo número de algarismos e devem ser sucessivos.
- ▶ Obtemos uma sequência de algarismos aleatoriamente (usando um computador ou uma tabela de números aleatórias).
- ▶ Com esta sequência formamos números com três algarismos e desprezamos os números que não correspondem a nenhum rolo.

Note que, se estivermos a fazer amostragem sem reposição, os números repetidos que eventualmente apareçam não são considerados.

Estatísticas, Estimadores e Estimativas

Definição

Sejam X_1, X_2, \dots, X_n variáveis aleatórias que constituem uma amostra aleatória. Chama-se **estatística** a uma função das variáveis aleatórias X_1, X_2, \dots, X_n que não contenha parâmetros desconhecidos. Uma estatística é, assim, uma nova variável aleatória (o valor assumido por ela é variável de amostra para amostra) e terá uma distribuição de probabilidade que é designada por **distribuição por amostragem**.

Definição

Chama-se **estimador** a qualquer estatística usada para estimar um parâmetro da população ou uma função desse parâmetro. Designa-se por **estimativa** o valor que um estimador assume para uma dada amostra concreta.

Estatísticas, Estimadores e Estimativas

Exemplos

1. $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad \hookrightarrow$ média amostral;
2. $S_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad \hookrightarrow$ variância amostral;
3. $T_1 = \max(X_1, X_2, \dots, X_n)$ ou $T'_1 = \min(X_1, X_2, \dots, X_n)$;
4. $T_2 = X_1 + X_2 + \dots + X_n$;
5. $T_3 = X_1 \times X_2 \times \dots \times X_n$;
6. $T_4 = n \sum_{i=1}^n X_i$.

A função $T_5 = (\sum_{i=1}^n X_i)^{ny}$ não é uma estatística porque y não é conhecido, i.e., não é observável na amostra.

\bar{X} e S_X^2 , a média e a variância amostral, são exemplos de **estimadores**, usados para estimar, respectivamente, a média populacional μ_X e a variância populacional σ_X^2 .

Teorema Limite Central

O teorema limite central é considerado um dos teoremas mais importantes na Estatística. Este pode ser enunciado da seguinte maneira:

Teorema Limite Central

Sejam X_1, X_2, \dots, X_n variáveis aleatórias **independentes** e **identicamente distribuídas** com média μ e variância σ^2 (finita).

Seja

$$S_n = X_1 + X_2 + \dots + X_n.$$

Então,

$$\frac{S_n - E(S_n)}{\sqrt{\text{Var}(S_n)}} = \frac{S_n - n\mu}{\sqrt{n\sigma}}$$

é uma v.a. cuja distribuição se aproxima da distribuição normal reduzida - $N(0, 1)$, quando n tende para infinito.

Teorema Limite Central

Isto é,

$$\frac{S_n - E(S_n)}{\sqrt{\text{Var}(S_n)}} = \frac{S_n - n\mu}{\sqrt{n}\sigma} \sim N(0, 1) \quad \text{ou} \quad S_n \sim N(n\mu, n\sigma^2)$$

Este teorema afirma que uma soma de variáveis aleatórias independentes e com a mesma distribuição (qualquer que ela seja) tem distribuição aproximadamente normal, desde que o número de parcelas, n , seja suficientemente grande. Na prática considera-se válida esta aproximação desde que n seja superior a 30.

Distribuição da média amostral

Suponha que se extrai de uma população uma amostra aleatória, de tamanho n , com vista ao estudo de uma sua característica aleatória X , de média μ e variância σ^2 .

Já sabemos que a amostra aleatória é constituída por n variáveis aleatórias, X_1, X_2, \dots, X_n , independentes e com a mesma distribuição de X .

Qual é o valor esperado e a variância da média amostral?

Valor esperado da média amostral

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{1}{n}(X_1 + X_2 + \dots + X_n)\right) \\ &= \frac{1}{n}(E(X_1) + E(X_2) + \dots + E(X_n)) = \frac{1}{n}(\mu + \mu + \dots + \mu) = \end{aligned}$$

Distribuição da média amostral

Variância da média amostral

$$\begin{aligned} \text{Var}(\bar{X}) &= \text{Var}\left(\frac{1}{n}(X_1 + X_2 + \dots + X_n)\right) \\ &= \frac{1}{n^2} (\text{Var}(X_1) + \text{Var}(X_2) + \dots + \text{Var}(X_n)) \\ &= \frac{1}{n^2} (\sigma^2 + \sigma^2 + \dots + \sigma^2) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n} \end{aligned}$$

Então,

$$\mu_{\bar{X}} = \mu \quad \text{e} \quad \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$$

Distribuição da média amostral

E quanto à distribuição de \bar{X} ?

- ▶ Se $X \sim N(\mu, \sigma^2)$ então, pelo Teorema da Aditividade da Distribuição Normal,

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim N(\mu, \sigma^2/n) \quad \Leftrightarrow \quad Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1),$$

pois X_1, \dots, X_n são independentes e têm distribuição normal.

- ▶ Se o tamanho da amostra n é suficientemente grande ($n > 30$), então pelo Teorema Limite Central

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim N(\mu, \sigma^2/n) \quad \Leftrightarrow \quad Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1),$$

pois X_1, \dots, X_n são independentes e têm a mesma distribuição.

Distribuição da média amostral

Vimos que, se $X \sim N(\mu, \sigma^2)$, então

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Suponhamos que a variância populacional σ^2 é desconhecida. Qual será a distribuição de Z se substituirmos a variância populacional σ^2 pelo seu estimador S^2 ?

Se $X \sim N(\mu, \sigma^2)$, então

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Vamos usar esta estatística na construção de intervalos de confiança e nos testes de hipóteses, relativos à média μ de uma população normal.

Se X_1, X_2, \dots, X_n é uma amostra aleatória de uma população normal com média μ e variância σ^2 , então

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2.$$

Iremos usar esta v. a. para construir intervalos de confiança para a variância e para testar hipóteses relativas a este parâmetro, quando a distribuição populacional é normal.

Suponhamos que temos duas populações com distribuição normal e com variâncias σ_1^2 e σ_2^2 respectivamente. Sejam S_1^2 e S_2^2 as variâncias amostrais baseadas em amostras independentes de tamanhos n_1 e n_2 , respectivamente, dessas populações. Então,

$$\frac{S_1^2}{S_2^2} \frac{\sigma_2^2}{\sigma_1^2} \sim F_{n_1-1, n_2-1}.$$

Esta variável aleatória vai ser usada, mais adiante, na construção de intervalos de confiança e nos testes de hipóteses, relativos à comparação de variâncias de populações normais.