Estatística
X

M. Cristina Miranda e Anabela Rocha

Introdução à estimação.

Técnicas de amostragem

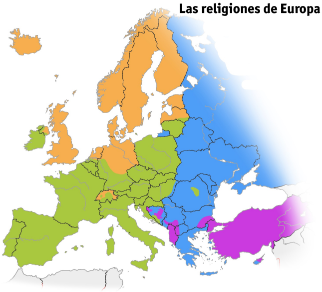
- Amostragem estratificada
- Amostragem por grupos
- Amostragem sistemática
- Amostragem aleatória simples sem reposição

(*Revisão dos métodos de amostragem (artigo) | Khan Academy* 2021; R. Antunes 2011)

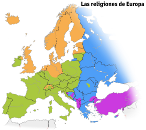
Amostragem estratificada

A população admite estratos e faz-se amostragem aleatória simples em cada estrato.

População



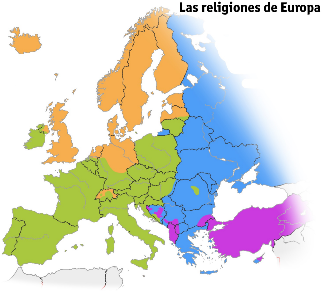
Amostra



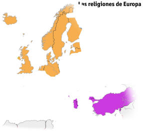
Amostragem por grupos

É feita uma amostragem aleatória simples de grandes grupos e selecionados todos os elementos dentro de cada grupo seleccionado.

População



Amostra



Amostragem sistemática

Escolhe-se o primeiro elemento de forma aleatória e os restantes a intervalos regulares.

População

Amostra

Amostragem aleatória (a.a.) simples sem reposição

- selecção aleatória;
- elementos independentes;
- cada elemento tem igual probabilidade de ser selecionado.

$(X_1, X_2, ..., X_n)$ é uma amostra aleatória (a.a.),

significa que as v.a.s X_i são independentes e identicamente distribuídas (*i.i.d.*).

Perante uma a.a. $(X_1, X_2, ..., X_n)$ temos interesse em obter a informação contida na amostra. Para esse efeito podem usar-se estatísticas.

Estatística

Estatística é uma função que depende apenas da amostra, *i.e.*, T é uma estatística

$$\Leftrightarrow T = f(X_1, X_2, ..., X_n).$$

Estimação pontual. Propriedades.

Estimação pontual

ESTIMADOR:

Sendo θ um parâmetro desconhecido da distribuição de uma dada população, pode estimar-se o seu valor com base numa estatística T. A essa estatística damos o nome de **Estimador** $T = \hat{\theta} = f(X_1, X_2, ..., X_n)\}$

Estimativa

O valor assumido por T para cada amostra observada $(x_1, x_2, ..., x_n)$ fornece uma estimativa: $t = \hat{\theta}$, ou seja, fornece um valor aproximado para θ.

Estimação pontual

A **média da população**, μ, é usualmente desconhecida. A partir de uma amostra representativa dessa população, pode calcular-se a respetiva **média amostral**, $\bar{x}$, e usar esse valor como estimativa da média populacional.

Assim, um **estimador** da média populacional é a **média amostral** e tem-se

$$E[\bar{X}] = \mu.$$

De modo análogo, a **variância da população**, σ^2, é usualmente desconhecida.

Um **estimador** da variância da população é a **variância amostral corrigida**,

$$S_c^2,$$

a qual permite obter um valor aproximado para σ^2, a partir de uma amostra representativa da população. Verifica-se ainda que

$$E[S_c^2] = \sigma^2.$$

Propriedades dos estimadores

Estimador Centrado

Seja T um estimador para um parâmetro θ de uma população. Dizemos que o estimador T é um estimador **Centrado** para θ, se

$$E[T] = \theta.$$

Estimador Eficiente

Entre dois estimadores centrados T_1 e T_2 para o parâmetro θ, é mais **eficiente** aquele que tiver menor variância, *i.e.*, T_1 é mais eficiente do que T_2, se $V[T_1] < V[T_2]$. (Pestana e Velosa 2002; Murteira e M. Antunes 2012)

1) Exemplo

Considere a estatística T, definida para uma a.a., de tamanho $n = 10$, proveniente de uma população $N(\mu, \sigma^2)$,

$$T = \frac{5X_1 + 5X_{10}}{10}.$$

1. Deduza a distribuição de T;
2. Para estimar a média da população, μ, qual dos estimadores é preferível, T ou $\bar{X}$? Justifique.

1. T é uma combinação linear de v.a.s i.i.d. com distribuição Normal, logo tem distribuição Normal, com parâmetros:

$$E(T) = E\left(\frac{5X_1 + 5X_{10}}{10}\right) =$$

$$= \frac{5E[X_1] + 5E[X_{10}]}{10} = \frac{5\mu + 5\mu}{10} = \mu;$$

$$V(T) = V\left(\frac{5X_1 + 5X_{10}}{10}\right) =$$

$$= V\left(\frac{5}{10}(X_1 + X_2)\right) = \frac{25}{100}(\sigma^2 + \sigma^2) = \frac{\sigma^2}{2};$$

$$\therefore T \sim N(\mu, \frac{\sigma^2}{2}).$$

2. De modo análogo se mostra que $\bar{X}$ tem distribuição Normal com parâmetros $E(\bar{X}) = \mu$ e $V(\bar{X}) = \dfrac{\sigma^2}{10}$;

Portanto são os dois estimadores centrados! Será preferível aquele que tiver aquele que tiver menor variância, *i.e.*, o que for mais eficiente. Como

$$V(\bar{X}) = \frac{\sigma^2}{10} < V(T) = \frac{\sigma^2}{2},$$

concluímos que $\hat{\mu} = \bar{X}$ é o estimador mais eficiente para μ.

Estimação pontual

A **estimação pontual** não permite medir a qualidade (confiança) de cada estimativa, logo não sabemos **qual a melhor estimativa**.

Em alternativa, a **estimação intervalar** (intervalo de confiança) permite:

- medir a qualidade da estimativa (através da confiança);
- obter um intervalo para o parâmetro desconhecido, para uma determinada confiança, tendo esta um valor elevado. Os valores mais comuns para a confiança de um intervalo são 90%, 95% ou 99%.

Intervalos de confiança.

Estimação intervalar

Interessa obter um intervalo no qual seja plausível encontrar o verdadeiro valor do parâmetro - estimação intervalar que se traduz numa estimativa intervalar ou,

Intervalo de Confiança (IC).

Intervalos de confiança para o valor médio em populações Normais.

Intervalos de confiança (IC) para o valor médio

1. Populações Normais:

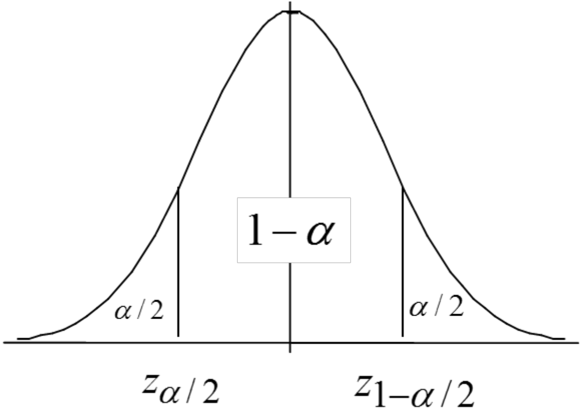
Suponhamos uma a.a. $(X_1, X_2, ..., X_n)$ e $X_i \sim N(\mu, \sigma^2)$ 1.1 Caso da variância conhecida (σ^2) - Estatística:

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

Podemos calcular quantis de tal modo que

$$P\left(\left|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right| \leq z\right) = 1 - \alpha.$$

Intervalos de confiança (IC) para o valor médio



$$\left[\bar{X} - z_{1-\alpha/2}\sigma/\sqrt{n}; \bar{X} + z_{1-\alpha/2}\sigma/\sqrt{n}\right] \qquad (1)$$

é um I. C. com confiança $100(1-\alpha)\%$ para μ, onde $z_{1-\alpha/2}$ é o quantil de probabilidade $1-\alpha/2$ para a N(0,1).

Intervalos de confiança (IC) para o valor médio

1. Populações Normais 1.2 Caso da Variância desconhecida $\sigma^2 =?$

É necessário estimá-la recorrendo à variância amostral corrigida. - Estatística:

$$\frac{\bar{X}-\mu}{s_c/\sqrt{n}} \sim t_{n-1} \text{ e}$$

$$\left[\bar{X} - t_{n-1;1-\alpha/2}\frac{Sc}{\sqrt{n}}; \bar{X} + t_{n-1;1-\alpha/2}\frac{Sc}{\sqrt{n}}\right] \qquad (2)$$

é um I. C. com confiança $100(1-\alpha)\%$ para μ, onde $t_{n-1;1-\alpha/2}$ é o quantil de probabilidade $1-\alpha/2$ para a distribuição t_{n-1}.

Intervalos de confiança (IC) para o valor médio

Quando a dimensão da amostra é grande, a distribuição t-Student pode ser aproximada à distribuição Normal padrão e os quantis desta Normal podem ser utilizados na construçao do IC para o valor médio de uma população Normal com variância desconhecida.

Como regra prática é usual considerar n elevado se $n \geq 30$.

2) Exemplo

Num estudo feito em áreas de serviço, questionaram-se 20 indivíduos sobre as quantias gastas em outros bens, diferentes de combustíveis, e verificou-se que tinham gasto uma média de 20.57€, com um desvio corrigido de 6.6€ Admita que as quantias gastas seguem uma distribuição Normal.

1. Determine um intervalo com 95% confiança para o valor médio dessa distribuição.
2. Supondo que a distribuição Normal tem variância igual a 9, determine um intervalo de confiança para o valor médio, com 99% de confiança.

1. X representa a quantia gasta noutros produtos e $X \sim N(\mu, \sigma^2)$,

um IC com 95% de confiança para μ pode ser obtido com base em (2) - população Normal, com variância desconhecida.

Assim, para a amostra observada vem:

$$\left[20.57 - t_{19;0.975}\frac{6.6}{\sqrt{20}}; 20.57 + t_{19;0.975}\frac{6.6}{\sqrt{20}}\right] = [17.48; 23.66].$$

2. Agora $X \sim N(\mu, \sigma^2)$, com variância conhecida, $\sigma^2 = 9$, e queremos um IC com 99% de confiança para μ.

É o caso de uma população Normal, com variância conhecida, e o IC pode ser obtido com base na forma (1) e na amostra dada:

$$\left[20.57 - z_{0.995}\frac{3}{\sqrt{20}}; 20.57 + z_{0.995}\frac{3}{\sqrt{20}}\right] = [18.84; 22.30].$$

Intervalos de confiança para o valor médio em grandes amostras.

Intervalos de confiança (IC) para o valor médio em grandes amostras

2. Populações não Normais (grandes amostras - $n \geq 30$)

Se $n \geq 30$ podemos aplicar o Teorema do Limite Central e obter o IC para o valor médio da população, μ, com confiança $100(1-\alpha)\%$:

$$\left[\bar{X} - z_{1-\alpha/2}\frac{Sc}{\sqrt{n}}; \bar{X} + z_{1-\alpha/2}\frac{Sc}{\sqrt{n}}\right] \qquad (3)$$

onde $z_{1-\alpha/2}$ é o quantil de probabilidade $1-\alpha/2$ para a N(0,1).

Interpretação do intervalo de confiança (IC)

Interpretação de um IC

Um IC para o valor médio μ, com 95% de confiança, traduz que: Se recolhermos 100 amostras aleatórias de dimensão n e construirmos os 100 IC correspondentes, então espera-se que 95 desses intervalos contenham o verdadeiro valor de μ e apenas 5 intervalos não contenham esse valor.
A interpretação deve ser feita de modo análogo para um IC de qualquer outro parâmetro e para qualquer valor da confiança.

Intervalos de confiança para uma proporção.

Intervalos de confiança para uma proporção

Para grandes amostras ($n \geq 30$) é possível aplicar o Teorema do Limite Central e obter um IC para uma proporção p.

Consideremos uma a.a. $(X_1, X_2, ..., X_n)$, com $X_i \sim B(1, p), i.e.,$

$$X_i = \begin{cases} 1 & \text{se houver sucesso, com probabilidade p} \\ 0 & \text{se não houver sucesso, com probabilidade (1-p),} \end{cases}$$

- p - proporção de "sucessos" (casos com $X_i = 1$) na população,
- $\bar{X}$ - proporção de "sucessos" na amostra.

Utilizando a média amostral como estimador do parâmetro p, um I.C. para p com uma confiança $100(1-\alpha)\%$ é dado por

$$\left[\bar{X} - z_{1-\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}; \bar{X} + z_{1-\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}\right], \qquad (4)$$

onde $z_{1-\alpha/2}$ é o quantil de probabilidade $1-\alpha/2$ da N(0,1).

Como escolher a dimensão da amostra n que permita obter um determinado coeficiente de confiança $100(1-\alpha)\%$ e assegurar que amplitude do IC não exceda d?

A amplitude do IC é igual a:

$$\bar{X} + z_{1-\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}} - \left(\bar{X} - z_{1-\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}\right) =$$

$$= 2z_{1-\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}.$$

Vamos impor que a amplitude do IC não exceda d:

$$2z_{1-\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \leq d \Rightarrow n \geq \left(\frac{2}{d}z_{1-\alpha/2}\right)^2 \bar{X}(1-\bar{X}).$$

Pode provar-se que $\bar{X}(1-\bar{X})$ atinge o máximo em $\bar{X} = 0.5$, logo obtemos a relação

$$n \geq \left(\frac{z_{1-\alpha/2}}{d}\right)^2.$$

Este cálculo pode ser adaptado para qualquer tipo de IC referido anteriormente.

3) Exemplo

Numa epidemia de gripe pretende-se estimar a proporção p de doentes que apresentam certa complicação sintomatológica. Numa amostra de 200 doentes, verificou-se que 140 não apresentavam a referida complicação.

Determine um IC para p, com uma confiança de 95%.

$$X_i = \begin{cases} 1 & \text{se o doente i apresenta a complicação;} \\ 0 & \text{se não apresenta.} \end{cases}$$

X =nº de doentes com a complicação entre os 200 e $X \sim B(200, p)$.

O IC é obtido com base na forma (4) e na amostra dada,

$$\left[\frac{60}{200} - z_{1-0.05/2}\sqrt{\frac{\frac{60}{200}(1-\frac{60}{200})}{200}}; \frac{60}{200} + z_{1-0.05/2}\sqrt{\frac{\frac{60}{200}(1-\frac{60}{200})}{200}}\right]$$

ou seja,

$$[0.236; 0.364].$$

Calculando IC com base em 100 amostras de igual dimensão, esperamos encontrar o valor da proporção de doentes que com complicação em aproximadamente 95 dos intervalos obtidos - interpretação da confiança de 95%.

- Antunes, R. (2011). *Amostragem aleatória estratificada.* URL: https://sondagenseestudosdeopiniao.wordpress.com/amostragem/amostras-probabilisticas-e-nao-probabilisticas/amostragem-aleatoria-estratificada/.
- Murteira, B. e M. Antunes (2012). *Probabilidades e Estatística*. Vol. 1. Escolar Editora. ISBN: 9789725923559.
- Pestana, Dinis e Sílvio Filipe Velosa (2002). *Introdução à Probabilidade e à Estatística*. Lisboa: Fundação Calouste Gulbenkian.
- *Revisão dos métodos de amostragem (artigo) | Khan Academy* (2021). URL: https://pt.khanacademy.org/math/statistics-probability/designing-studies/sampling-methods-stats/a/sampling-methods-review/ (acedido em 05/03/2021).