Técnicas de Investigación Social

Grado en Relacións Laborais e Recursos Humanos

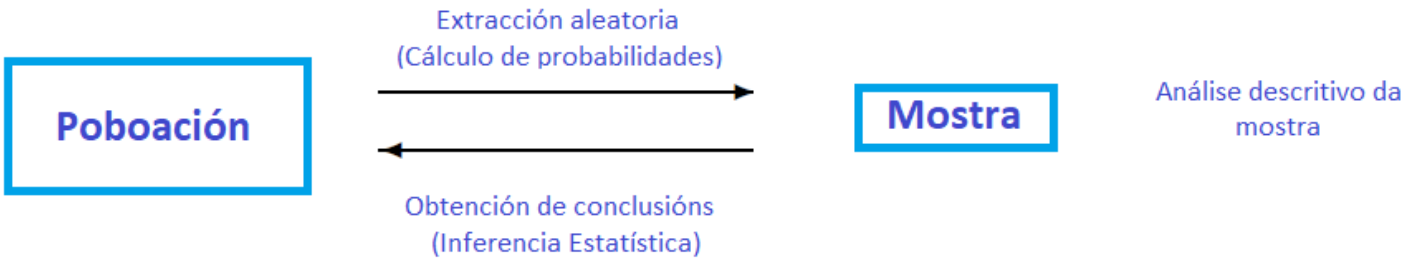
Curso 2019-2020

Tema 4. Mostraxe por conglomerados

Alejandro Saavedra Nieves

As redes sociais dos alumnos de RRLL

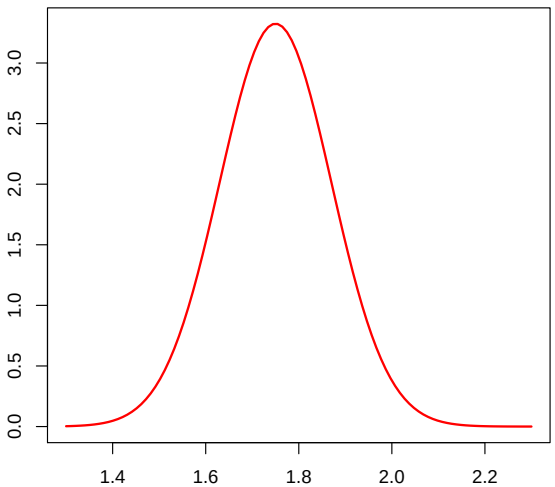
Variable: Horas empregadas en redes sociais (en media) polo estudantes de RRLL da UVigo



Estudantes Facultade RRLL

Modelo | Distribución normal

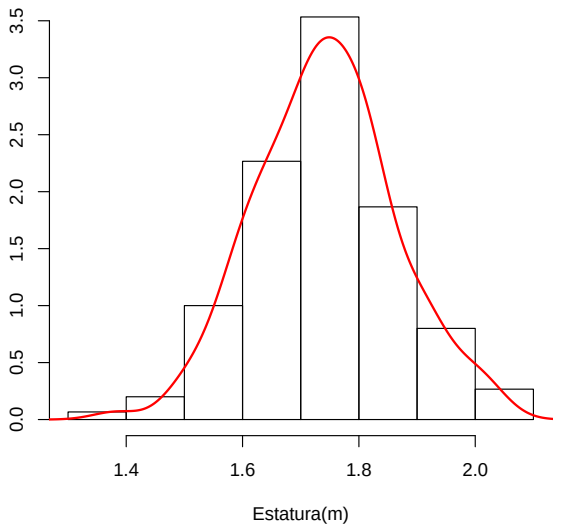
Media: $\mu = 1{,}75$; varianza: $\sigma^2 = 0{,}12^2$



Mostra estudantes RRLL, $n = 150$

$x_1 = 1.66$, $x_2 = 1.86$, $x_3 = 1.62$, ...

$\overline{X} = 1{,}74$, $s^2 = 0{,}119^2$



Algunhas cuestións naturais...

- Resulta de interés coñecer unha estimación do emprego medio diario dos alumnos de RRLL.
- En termos de productividade, poderiamos plantexarnos a necesidade de estudar o total de horas diarias que os estudantes de RRLL gastan nas redes sociais.
- Existe un problema de adicións ás redes sociais? Cal é a proporción de estudiantes que empregan máis de 2 horas diarias na súa xestión?

E se a poboacións se divide en subpoboacións homoxéneas entre si e semellantes á poboación?

Introdución

No que segue supoñemos a existencia dunha poboación finita de N individuos, sobre a que pode resultar de interese unha certa característica:

- a media dos seus valores,
- o total, en termos da suma dos valores,
- a proporción de individuos con certos valores da característica.

As metodoloxías de mostraxe permiten a estimación destes parámetros poboacionais, analizando o problema desde unha perspectiva estatística.

Notación.

- Os valores que esta variable toma para cada un dos individuos da poboación denotaranse por $X_1, \cdots, X_N$.
- Desa poboación extráese unha mostra con n observacións, $x_1, x_2, \ldots, x_n$.

Mostraxe por conglomerados

Mostraxe por conglomerados

Denomínase mostraxe por conglomerados a un método de selección dunha mostra de tamaño n, ó supoñer que:

- a poboación está dividida en grupos internamente heteroxéneos e
- de cada un deles, extráese un subconxunto de observacións.

Chámanse conglomerados a estas unidades amplas onde se agrupan os elementos da poboación. Os conglomerados refírense a formas de agrupación física das unidades no espazo ou no tempo.

É unha metodoloxía de mostraxe menos costosa, xa que non é necesario o coñecemento de información sobre todos os individuos.

Uso da mostraxe por conglomerados

Enquisa sobre os estudantes universitarios dun pais

- O uso de mostraxe aleatoria simple requiría do uso dun censo, o que é difícil en termnos de costes.
- Sen embargo, sabemos que os estudantes están clasificados en universidades.
- A mostraxe por conglomerados garante que nunha primeira etapa se seleccionan algunhas universidades; logo, unhas certas facultades; dentro das facultades, certas clases e dentro das clases, estudantes mediante mostraxe aleatoria simple.

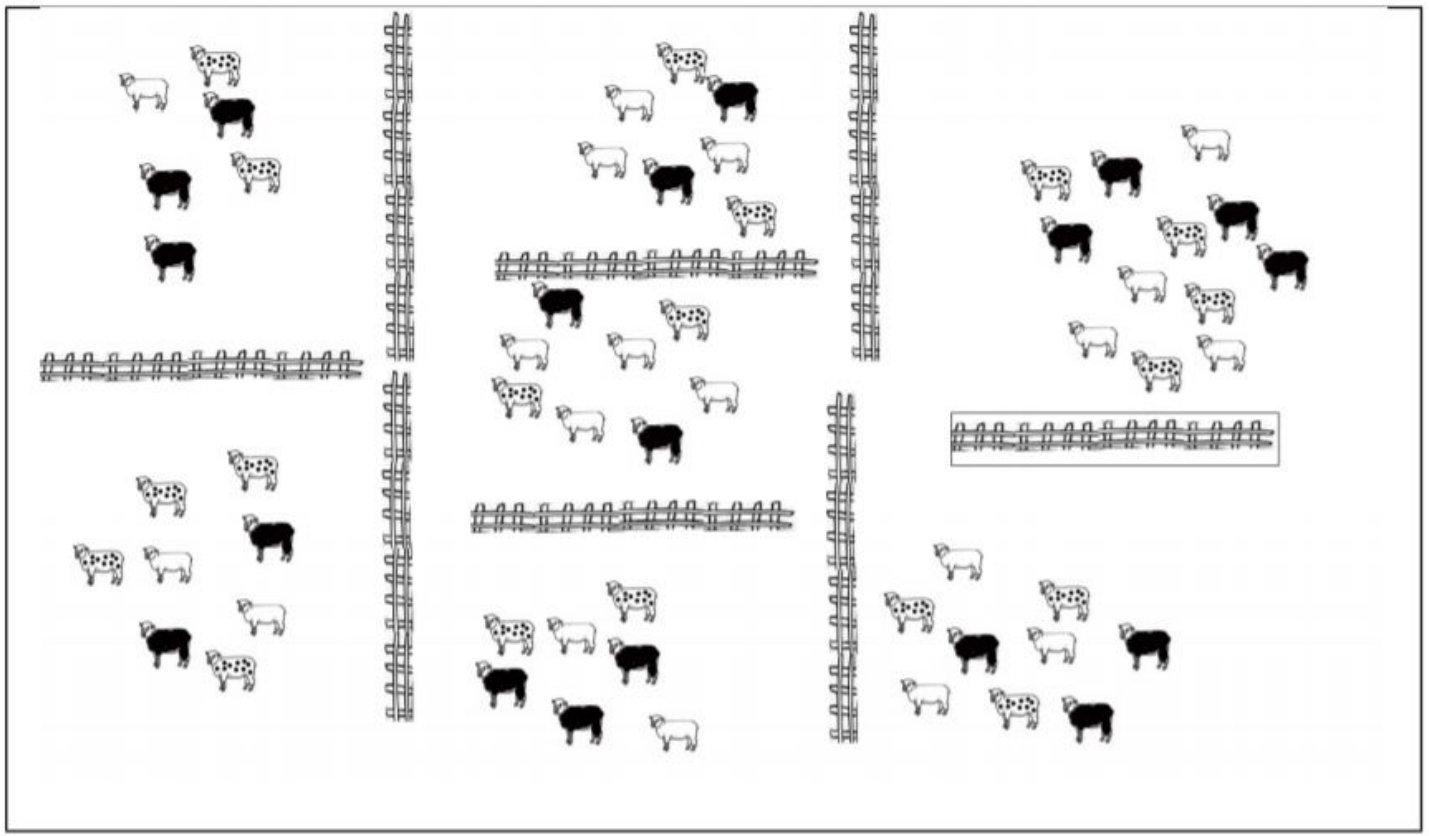
Vantaxes da mostraxe por conglomerados

- Adoita ser menos costoso en tempo e recursos que outros tipos de mostraxe, xa que a disposición das unidades en grupo facilita as tarefas administrativas de localización, desplazamento e toma de datos.
- A miúdo non se dispón de información exhaustiva de tódalas unidades da poboación ou é costoso conseguila.
- Neste tipo de mostraxe só se precisa recabar información sobre os conglomerados seleccionados, diminuíndo os costes.
- Frecuentemente, os conglomerados xa existen como unidades administrativas, facilitando a tarefa de particionar a poboación.

Mostraxe estratificada vs. mostraxe por conglomerados

- Debemos ter en conta as diferenzas entre estrato e conglomerado.
- Os estratos teñen en conta a homoxeneidade das unidades ou subpoboacións dentro da poboación a investigar.
- Interésanos que os estratos sean internamente tan homoxéneos como sexa posible e tan diferentes entre si como se poida.
- Os conglomerados deben ser tan homoxéneos internamente como a poboación e tan homoxéneos entre si como sexa posible.

Mostraxe por conglomerados



Hipótese. Á vista da figura os conglomerados son parecidos entre sí, e a variabilidade interna de cada un deles representa en certo modo a variabilidade da poboación.

Organización do método

Un conglomerado pode, á súa vez, ser dividido internamente en subgrupos que conformen unha partición interna de unidades.

De existir esta xerarquía, temos:

- Unidades de primeira etapa, conformada polos primeiros conglomerados que dividen a poboación.
- Unidades de segunda etapa, conformada polas segundas unidades que dividen os conglomerados.
- ...

O método de mostraxe en varias etapas consiste en seleccionar por mostraxe algúns dos primeiros grupos e, a continuación, e dentro deles, seleccionar tamén por mostraxe subgrupos. Se se repite sucesivamente este procedemento, chégase ós individuso (unidades elementais).

Elementos da mostraxe por conglomerados

Dispoñemos dunha poboación $X_1, X_2, \ldots, X_N$, composta por N individuos e dividida en L subgrupos disxuntos de tamaño $N_1, N_2, \ldots, N_L$.

- O tamaño medio dos conglomerados é $\bar{N} = \frac{1}{L}\sum_{i=1}^{L} N_i = \frac{N}{L}$.
- A media do conglomerado i é $\bar{X}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} X_{ij}$.
- O total do conglomerado i é $X_i = N_i\bar{X}_i = \sum_{j=1}^{N_i} X_{ij}$.

.

Elementos da mostraxe por conglomerados

Dispoñemos dunha poboación $X_1, X_2, \ldots, X_N$, composta por N individuos e dividida en L subgrupos disxuntos de tamaño $N_1, N_2, \ldots, N_L$.

- A media poboacional é $\bar{X} = \frac{1}{N}\sum_{i=1}^{L}\sum_{j=1}^{N_i} X_{ij}$.
- O total da poboación é $X = N\bar{X} = \sum_{i=1}^{L}\sum_{j=1}^{N_i} X_{ij}$.
- A proporción poboacional é $\bar{X} = \frac{1}{N}\sum_{i=1}^{L}\sum_{j=1}^{N_i} A_{ij}$, con $A_{ij} = 1$ se o individuo j no conglomerado i ten a característica e, $A_{ij} = 0$, en caso contrario.

.

Estimación de θ

Supoñamos unha poboación $X_1, X_2, \ldots, X_N$ composta por N individuos organizados en L conglomerados, da que se extrae unha mostra aleatoria simple de n conglomerados.

O obxectivo pasa por analizar algunha característica da poboación, θ, que é descoñecida e da que resulta de interese a súa estimación.

- Usaremos unha función obtida da mostra, o estimador $\hat{\theta}$, como aproximación de θ.
- Como é construido a partir da mostra, terá unha certa distribución.
- Debe ser fácil obter $E(\hat{\theta})$ e $V(\hat{\theta})$, que será aproximado por $\hat{V}(\hat{\theta})$.

Estimación de θ

Desde o punto de vista estatítico, cómpre controlar o erro na estimación, isto é,

$$P(|\theta - \hat{\theta}| < \varepsilon) \geq 1 - \alpha,$$

para:

- unha certa cota do erro cometido, ε, e
- un nivel de confianza (probabilidade) $1 - \alpha$.

Nesta dupla xogará un papel fundamental o tamaño da mostra seleccionado.

Equivalentemente, este problema tradúcese en determinar valores L_1 e L_2 (dependentes da mostra) tal que se satisfagan

$$P(\theta \in [L_1, L_2]) \geq 1 - \alpha.$$

Intervalos de confianza!

Intervalos de confianza para θ

- **Teorema Central do Límite.** IC para $\hat{\theta}$ con un nivel de confianza $1-\alpha$ (con $n > 30$):

$$IC_\theta = \left(\hat{\theta} - z_{\alpha/2}\sqrt{V(\hat{\theta})}, \hat{\theta} + z_{\alpha/2}\sqrt{V(\hat{\theta})}\right)$$

  sendo z_α o punto que deixa unha probabilidade á súa dereita igual a α.

Si $Z \sim N(0,1)$, $P(Z \geq z_\alpha) = \alpha$. Entonces, $z_{0,1} = 1{,}281$, $z_{0,05} = 1{,}649$ y $z_{0,025} = 1{,}960$.

Polo xeral, $V(\hat{\theta})$ é descoñecido.

Intervalos de confianza para θ

- **Teorema Central do Límite.** IC para $\hat{\theta}$ con un nivel de confianza $1-\alpha$ (con $n > 30$):

$$IC_\theta = \left(\hat{\theta} - z_{\alpha/2}\sqrt{\hat{V}(\hat{\theta})}, \hat{\theta} + z_{\alpha/2}\sqrt{\hat{V}(\hat{\theta})}\right)$$

  sendo z_α o punto que deixa unha probabilidade á súa dereita igual a α.

Si $Z \sim N(0,1)$, $P(Z \geq z_\alpha) = \alpha$. Entonces, $z_{0,1} = 1{,}281$, $z_{0,05} = 1{,}649$ y $z_{0,025} = 1{,}960$.

A estimación da media baixo mostraxe por conglomerados

A estimación da media poboacional

Neste caso o parámetro descoñecido é a media poboacional:

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{L}\sum_{j=1}^{N_i} X_{ij} = \frac{1}{N}\sum_{i=1}^{L} X_i.$$

O estimador máis utilizado na estimación da media poboacional é:

$$\bar{x}_c = \frac{\sum_{h=1}^{n} X_h}{\sum_{h=1}^{n} N_h},$$

con X_h a suma da característica no conglomerado h.

A varianza do estimador da media

Tras certas operacións, a varianza do estimador para a media baixo mostraxe por conglomerados dunha etapa é:

$$\hat{V}(\bar{x}_c) = \frac{L(L-n)}{N^2 n}\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x}N_i)^2.$$

A estimación do total baixo mostraxe por conglomerados

A estimación do total poboacional

Neste caso o parámetro descoñecido é o total poboacional:

$$X = N\bar{X} = \sum_{i=1}^{L}\sum_{j=1}^{N_i} X_{ij}.$$

O estimador máis utilizado na estimación do total poboacional é:

$$\hat{X}_c = N\bar{x}_c.$$

A varianza do estimador do total

Tras certas operacións, a varianza do estimador para o total baixo mostraxe por conglomerados é:

$$\hat{V}(\hat{X}_c) = N^2\,\hat{V}(\bar{x}_c)$$

A estimación da proporción baixo mostraxe por conglomerados

A estimación da proporción poboacional

Neste caso o parámetro descoñecido é a proporción poboacional:

$$P = \frac{\sum_{h=1}^{L} A_h}{\sum_{h=1}^{L} N_h},$$

con $A_h = \sum_{j=1}^{N_h} A_{ij}$ e $A_{ij} = 1$ se o individuo ten a característica e $A_{ij} = 0$, noutro caso.

O estimador máis utilizado na estimación da media poboacional é:

$$p_c = \frac{\sum_{h=1}^{n} A_h}{\sum_{h=1}^{n} N_h}.$$

A varianza do estimador da proporción

A varianza do estimador para a proporción baixo mostraxe por conglomerados dunha etapa é:

$$\hat{V}(p_c) = \frac{L(L-n)}{N^2 n} \frac{1}{n-1} \sum_{i=1}^{n} (A_i - \hat{P} N_i)^2.$$

Estimación dos intervalos de confianza

IC con nivel de confianza $1-\alpha$

- **Media.**

$$IC_{\bar{X}} = \left(\bar{x}_c - z_{\alpha/2}\sqrt{\hat{V}(\bar{x}_c)}, \bar{x}_c + z_{\alpha/2}\sqrt{\hat{V}(\bar{x}_c)} \right)$$

- **Total.**

$$IC_X = \left(\hat{X}_c - z_{\alpha/2}\sqrt{\hat{V}(\hat{X}_c)}, \hat{X}_c + z_{\alpha/2}\sqrt{\hat{V}(\hat{X}_c)} \right)$$

- **Proporción.**

$$IC_P = \left(p_c - z_{\alpha/2}\sqrt{\hat{V}(p_c)}, p_c + z_{\alpha/2}\sqrt{\hat{V}(p_c)} \right)$$

sendo z_α o punto que deixa unha probabilidade á súa dereita igual a α.

Coeficiente de correlación intraconglomerados, δ

É o coeficiente de correlación lineal entre todos os pares de valores da variable en unidades dos conglomerados e estendido a tódolos conglomerados. En definitiva, é unha medida da homoxeneidade no interior dos conglomerados.

- É fundamental usar conglomerados heteroxéneos.
- Canto menor sexa δ, maior eficiencia aporta a mostraxe por conglomerados.
- Se $\delta = 0$, a mostraxe por conglomerados e a mostraxe aleatoria simple son equivalentes.

Determinación do tamaño de mostra

O tamaño de mostra necesario ó empregar mostraxe por conglomerados para lograr a mesma precisión da mostraxe aleatoria simple é:

$$n_c = n_{m.a.s.}(1 + (L - 1)\delta),$$

con L o número de conglomerados e n_c e $n_{m.a.s.}$ os tamaños de mostra na mostraxe por conglomerados e m.a.s., respectivamente.

O factor $(1+(L-1)\delta)$ é a variación do tamaño da mostra que precisamos debido ó uso de conglomerados. É o que se coñece como efecto de deseño.

Mostraxe por conglomerados en dúas (ou máis) etapas

- Se en cada un dos conglomerados seleccionados na primeira etapa se realiza un proceso de mostraxe en lugar dun estudo completo o censo teremos mostraxe por conglomerados en dúas etapas.

  Exemplo. Realizamos nunha cidade mostraxe por conglomerados en dúas etapas. Na primeira etapa escóllense 5 seccións censais mediante o método de Madow. Na segunda etapa realízase en cada sección censal escollida mostraxe estratificada con afixación uniforme. Temos mostraxe por conglomerados en dúas etapas.

- Sen embargo, esto dificulta a obtención de resultados teóricos.