Técnicas de Investigación Social

Grado en Relacións Laborais e Recursos Humanos

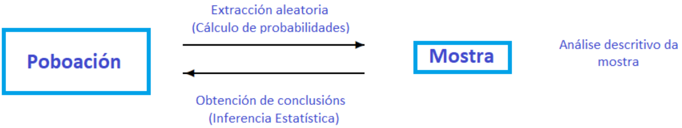
Curso 2019-2020

Tema 2. Mostraxe aleatoria simple

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

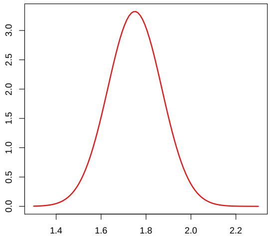
Modelo | Distribución normal

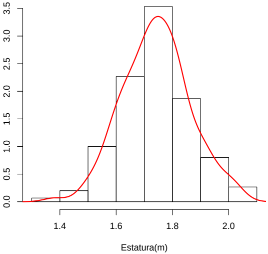
Media: $\mu = 1{,}75$; varianza: $\sigma^2 = 0{,}12^2$

Mostra estudantes RRLL, $n = 150$

$x_1 = 1.66,\ x_2 = 1.86,\ x_3 = 1.62,\ \ldots$

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





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?

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

Exemplo

Unha compañía de seguros analizou aleatoriamente 200 dos seus 10.000 expedientes para estudar as cuantías das indemnizacións que paga polos accidentes de tráfico cubertos polas súas pólizas. Da información obtida sábese que a media das indemnizacións pagadas foi de 151,25? e a cuasivarianza foi de 54,13. Cunha confianza do 95 %,

(a) Estima a cuantía total anual pagada pola compañía mediante o correspondente intervalo de confianza.

(b) Se se desexa estimar a proporción de expedientes que deron menos de dous partes de sinistro ó ano, ¿qué tamaño de mostra debe utilizarse para conseguir a estimación cun erro de mostraxe inferior ó 12 %?.

Mostraxe aleatoria simple

Mostraxe aleatoria simple

Denomínase mostraxe aleatoria simple a un método de selección dunha mostra de tamaño n, de tal modo que todas as posibles mostras teñen a mesma probabilidade de ser seleccionadas.

Existen dúas variantes das metodoloxías de mostraxe aleatoria simple que surxen ó considerar (ou non) a hipótese de reemplazo para a obtención das mostras.

- Mostraxe aleatoria simple sen reemplazo.
- Mostraxe aleatoria simple con reemplazo.

A inclusión ou non desta hipótese implica que os individuos da poboación poidan ou non ser seleccionados en máis dunha ocasión.

Estimación de θ

Supoñamos unha poboación $X_1, X_2, \ldots, X_N$ composta por N individuos, da que se extrae unha mostra aleatoria simple $x_1, x_2, \ldots, x_n$ de n observacións.

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

- **Desigualdade de TChebyshev.** IC para $\hat{\theta}$ con un nivel de confianza $1-\frac{1}{k^2}$:

$$IC_\theta = \left(\hat{\theta} - k\sqrt{V(\hat{\theta})}, \hat{\theta} + k\sqrt{V(\hat{\theta})}\right)$$

- **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

- **Desigualdade de TChebyshev.** IC para $\hat{\theta}$ con un nivel de confianza $1-\frac{1}{k^2}$:

$$IC_\theta = \left(\hat{\theta} - k\sqrt{\hat{V}(\hat{\theta})}, \hat{\theta} + k\sqrt{\hat{V}(\hat{\theta})}\right)$$

- **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 $z_{0,025} = 1{,}960$.

Mostraxe aleatoria simple sen reemplazo (MASsr)

Muestreo aleatorio simple sen reemplazo (MASsr)

Hipótese. Trátase dunha metodoloxía de mostraxe que se aplica en contextos onde o marco é perfecto, sen omisións nin duplicados de información e con tódolos elementos ben definidos.

Trátase da metodoloxía de mostraxe máis básica e simple que se poden realizan. É o método máis importante en poboacións finitas.

- Non se presupón información a priori e serve para comparar con outros tipos de mostraxe máis complexos.
- Se outras metodoloxías de mostraxe non son mellores, desde o punto de vista téorico con costes iguais, adoitan rexeitarse pola súa maior complexidade.

Notación.

- O factor $f = \frac{n}{N}$ é coñecido como fracción da mostraxe. Representa a proporción de unidades da poboación contidas na mosta.
- $f = 0$ cando non hai representación da poboación na mostra e vale 1 cando toda a poboación está na mostra.
- O factor $1 - f = \frac{N-n}{N}$ é o coeficiente de corrección por finitude.

Proceso para realizar un MASsr

Para realizar un MASsr débense ordear tódalas posibles mostras de n unidades diferentes da poboación (feito pouco práctico se N é o suficientemente grande), e seleccionar unha delas.

Este proceso é excesivamente laborioso, polo que se aplica o seguinte esquema alternativo:

1. Partimos dunha poboación con N individuos e quérese extraer unha mostra de tamaño n.
2. Extraemos sucesiva e independentemente as unidades da poboación con igual probabilidade en cada extracción. É igual a $\frac{1}{N-t}$, para todo $t = 0, 1, \ldots, n-1$.
3. Unha vez seleccionada unha unidade da poboación, esta é excluída para que todas sexan diferentes na mosta.

A estimación da media baixo MASsr

A estimación da media poboacional

Neste caso o parámetro descoñecido é a media poboacional:

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i = \frac{X_1 + X_2 + \cdots + X_N}{N}.$$

O estimador máis utilizado na estimación da media poboacional é a media mostral:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

- Trátase dun estimador insesgado da media poboacional.

A varianza do estimador da media

Tras certas operacións, a varianza do estimador para a media baixo MASsr é:

$$V(\bar{x}) = \frac{N-n}{N-1}\frac{\sigma^2}{n},$$

onde

- N é o tamaño da poboación,
- n o tamaño da mostra, e
- σ^2 denota a varianza poboacional.

Dado que σ^2 é descoñecida, utilizamos como aproximación a cuasivarianza muestral

$$S_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Desta maneira, un estimador insesgado da varianza do estimador é

$$\hat{V}(\bar{x}) = \frac{(1-f)}{n}S_c^2.$$

Intervalos de confianza para a media

- **Desigualdade de TChebyshev.** IC para $\bar{X}$ con un nivel de confianza $1 - \frac{1}{k^2}$:

$$IC_{\bar{X}} = \left(\bar{x} - k\sqrt{\frac{(1-f)}{n} S_c^2}, \bar{X} + k\sqrt{\frac{(1-f)}{n} S_c^2} \right)$$

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

$$IC_{\bar{X}} = \left(\bar{x} - z_{\alpha/2}\sqrt{\frac{(1-f)}{n} S_c^2}, \bar{X} + z_{\alpha/2}\sqrt{\frac{(1-f)}{n} S_c^2} \right)$$

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

A estimación do total baixo MASsr

A estimación do total poboacional

Neste caso o parámetro descoñecido é o total poboacional:

$$X = \sum_{i=1}^{N} X_i = X_1 + X_2 + \cdots + X_N.$$

O estimador máis utilizado na estimación do total poboacional é:

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

- Trátase dun estimador insesgado do total poboacional.

A varianza do estimador do total

Tras certas operacións, a varianza do estimador para o total baixo MASsr é:

$$V(\hat{X}) = N^2 V(\bar{x}) = N^2 \frac{(1-f)}{n} S^2,$$

onde

- N é o tamaño da poboación,
- n o tamaño da mostra, e
- S^2 denota a cuasivarianza poboacional.

Dado que S^2 é descoñecida, úsase como aproximación a cuasivarianza muestral

$$S_c^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Desta maneira, un estimador insesgado da varianza do estimador é

$$\hat{V}(\hat{x}) = \frac{N^2(1-f)}{n} S_c^2.$$

- **Desigualdade de TChebyshev.** IC para X con un nivel de confianza $1-\frac{1}{k^2}$:

$$IC_X = \left(\hat{X} - k\sqrt{\frac{N^2(1-f)}{n}S_c^2}, \hat{X} + k\sqrt{\frac{N^2(1-f)}{n}S_c^2}\right)$$

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

$$IC_X = \left(\hat{X} - z_{\alpha/2}\sqrt{\frac{N^2(1-f)}{n}S_c^2}, \hat{X} + z_{\alpha/2}\sqrt{\frac{N^2(1-f)}{n}S_c^2}\right)$$

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

A estimación da proporción baixo MASsr

A estimación da proporción poboacional

Neste caso o parámetro descoñecido é a proporción poboacional:

$$P = \frac{1}{N}\sum_{i=1}^{N} A_i = \frac{A_1 + A_2 + \cdots + A_N}{N},$$

onde $A_i = 1$ se o individuo i presenta a característica e $A_i = 0$, noutro caso.

Denótase por $Q = 1 - P$ a proporción de individuos que non presenta a característica.

O estimador máis utilizado na estimación do total poboacional é:

$$p = \frac{1}{n}\sum_{i=1}^{n} A_i.$$

- Trátase dun estimador insesgado da proporción poboacional.

A varianza do estimador da proporción

Tras certas operacións, a varianza do estimador para o total baixo MASsr é:

$$V(p) = \frac{N-n}{N-1}\frac{PQ}{n},$$

onde

- N é o tamaño da poboación,
- n o tamaño da mostra, e
- P e Q son as proporcións da presenza (e non presenza) dunha característica na poboación.

Dado que P e Q son descoñecidos, un estimador insesgado da varianza do estimador é

$$\hat{V}(p) = \frac{N-n}{N-1}\frac{pq}{n-1},$$

sendo $q = 1 - p$.

Intervalos de confianza para a proporción

- **Desigualdade de TChebyshev.** IC para P con un nivel de confianza $1-\frac{1}{k^2}$:

$$IC_P = \left(p - k\sqrt{\frac{N-n}{N-1}\frac{pq}{n-1}}, p + k\sqrt{\frac{N-n}{N-1}\frac{pq}{n-1}} \right)$$

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

$$IC_P = \left(p - z_{\alpha/2}\sqrt{\frac{N-n}{N-1}\frac{pq}{n-1}}, p + z_{\alpha/2}\sqrt{\frac{N-n}{N-1}\frac{pq}{n-1}} \right)$$

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

Exemplo: mostraxe sen reemplazo

Unha compañía de seguros analizou aleatoriamente 200 dos seus 10.000 expedientes para estudar as cuantías das indemnizacións que paga polos accidentes de tráfico cubertos polas súas pólizas. Da información obtida sábese que a media das indemnizacións pagadas foi de 151,25? e a cuasivarianza foi de 54,13. Cunha confianza do 95 %,

(a) Estima a cuantía total anual pagada pola compañía mediante o correspondente intervalo de confianza.

Datos.

- Tamaño da poboación: $N = 10000$ expedientes.
- Tamaño da mostra: $n = 200$.
- Media das indemnizacións: $\bar{x} = 151{,}25$ euros.
- Cuasivarianza das indemnizacións: $S_c^2 = 54{,}13$.
- Nivel de confianza: $1 - \alpha = 0{,}95$ (polo tanto, $\alpha = 0{,}05$).

Exemplo: mostraxe sen reemplazo

Obxectivo. Determinar un IC para o total (X) cunha confianza do 95 %.

Teorema Central do Límite. IC para X con un nivel de confianza $1-\alpha$:

$$IC_X = \left(\hat{X} - z_{\alpha/2}\sqrt{\frac{N^2(1-f)}{n}S_c^2}, \hat{X} + z_{\alpha/2}\sqrt{\frac{N^2(1-f)}{n}S_c^2}\right)$$

e sabendo que a estimación do total $\hat{X}$ se obtén como

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

Cal é o intervalo de confianza para o total X?

Mostraxe aleatoria simple con reemplazo (MAScr)

Muestreo aleatorio simple con reemplazo (MAScr)

Hipótese. Trátase dunha metodoloxía de mostraxe que se aplica en contextos onde o marco é perfecto, sen omisións nin duplicados de información e con tódolos elementos ben definidos.

Trátase da metodoloxía de mostraxe básica e sinxela que se poden realizan. É aplicable en poboacións finitas e infinitas.

Proceso para realizar un MAScr

Para realizar un MAScr débense ordear tódalas posibles mostras de n unidades diferentes da poboación (feito pouco práctico se N é o suficientemente grande), e seleccionar unha delas.

Este proceso é excesivamente laborioso, polo que se aplica o seguinte esquema alternativo:

1. Partimos dunha poboación con N individuos e quérese extraer unha mostra de tamaño n.
2. Extraemos sucesiva e independentemente as unidades da poboación con igual probabilidade en cada extracción. É igual a $\frac{1}{N}$.
3. Unha vez seleccionada unha unidade da poboación, esta é devolta á poboación.

A estimación da media baixo MAScr

A estimación da media poboacional

Neste caso o parámetro descoñecido é a media poboacional:

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i = \frac{X_1 + X_2 + \cdots + X_N}{N}.$$

O estimador máis utilizado na estimación da media poboacional é a media mostral:

$$\hat{\bar{X}} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

- Trátase dun estimador insesgado da media poboacional.

A varianza do estimador da media

Tras certas operacións, a varianza do estimador para a media baixo MAScr é:

$$V(\bar{x}) = \frac{\sigma^2}{n},$$

onde

- n o tamaño da mostra, e
- σ^2 denota a varianza poboacional.

Dado que σ^2 é descoñecida, utilizamos como aproximación a cuasivarianza muestral

$$S_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Desta maneira, un estimador insesgado da varianza do estimador é

$$\hat{V}(\bar{x}) = \frac{S_c^2}{n}.$$

Intervalos de confianza para a media

- **Desigualdade de TChebyshev.** IC para $\bar{X}$ con un nivel de confianza $1 - \frac{1}{k^2}$:

$$IC_{\bar{X}} = \left(\bar{x} - k\sqrt{\frac{S_c^2}{n}}, \bar{X} + k\sqrt{\frac{S_c^2}{n}} \right)$$

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

$$IC_{\bar{X}} = \left(\bar{x} - z_{\alpha/2}\sqrt{\frac{S_c^2}{n}}, \bar{X} + z_{\alpha/2}\sqrt{\frac{S_c^2}{n}} \right)$$

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

A estimación do total baixo MASsr

A estimación do total poboacional

Neste caso o parámetro descoñecido é o total poboacional:

$$X = \sum_{i=1}^{N} X_i = X_1 + X_2 + \cdots + X_N.$$

O estimador máis utilizado na estimación do total poboacional é:

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

- Trátase dun estimador insesgado do total poboacional.

A varianza do estimador do total

Tras certas operacións, a varianza do estimador para o total baixo MAScr é:

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

onde

- N é o tamaño da poboación,
- n o tamaño da mostra, e
- σ^2 denota a varianza poboacional.

Dado que S^2 é descoñecida, úsase como aproximación a cuasivarianza muestral

$$S_c^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Desta maneira, un estimador insesgado da varianza do estimador é

$$\hat{V}(\hat{x}) = \frac{N^2 S_c^2}{n}.$$

Intervalos de confianza para o total

- **Desigualdade de TChebyshev.** IC para X con un nivel de confianza $1 - \frac{1}{k^2}$:

$$IC_X = \left(\hat{X} - k\sqrt{\frac{N^2 S_c^2}{n}}, \hat{X} + k\sqrt{\frac{N^2 S_c^2}{n}} \right)$$

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

$$IC_X = \left(\hat{X} - z_{\alpha/2}\sqrt{\frac{N^2 S_c^2}{n}}, \hat{X} + z_{\alpha/2}\sqrt{\frac{N^2 S_c^2}{n}} \right)$$

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

A estimación da proporción baixo MAScr

A estimación da proporción poboacional

Neste caso o parámetro descoñecido é a proporción poboacional:

$$P = \sum_{i=1}^{N} A_i = A_1 + A_2 + \cdots + A_N,$$

onde $A_i = 1$ se o individuo i presenta a característica e $A_i = 0$, noutro caso.

Denótase por $Q = 1 - P$ a proporción de individuos que non presenta a característica.

O estimador máis utilizado na estimación do total poboacional é:

$$p = \frac{1}{n}\sum_{i=1}^{n} A_i.$$

- Trátase dun estimador insesgado da proporción poboacional.

A varianza do estimador da proporción

Tras certas operacións, a varianza do estimador para o total baixo MAScr é:

$$V(p) = \frac{PQ}{n},$$

onde

- n é o tamaño da mostra, e
- P e Q son as proporcións da presenza (e non presenza) dunha característica na poboación.

Dado que P e Q son descoñecidos, un estimador insesgado da varianza do estimador é

$$\hat{V}(p) = \frac{pq}{n-1},$$

sendo $q = 1 - p$.

- **Desigualdade de TChebyshev.** IC para P con un nivel de confianza $1 - \frac{1}{k^2}$:

$$IC_P = \left(p - k\sqrt{\frac{pq}{n-1}}, p + k\sqrt{\frac{pq}{n-1}}\right)$$

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

$$IC_P = \left(p - z_{\alpha/2}\sqrt{\frac{pq}{n-1}}, p + z_{\alpha/2}\sqrt{\frac{pq}{n-1}}\right)$$

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

Exemplo: mostraxe aleatoria simple

Unha compañía de seguros analizou aleatoriamente 200 dos seus 10.000 expedientes para estudar as cuantías das indemnizacións que paga polos accidentes de tráfico cubertos polas súas pólizas. Da información obtida sábese que a media das indemnizacións pagadas foi de 151,25? e a cuasivarianza foi de 54,13. Cunha confianza do 95 %,

(a) Estima a cuantía total anual pagada pola compañía mediante o correspondente intervalo de confianza.

Datos.

- Tamaño da poboación: $N = 10000$ expedientes.
- Tamaño da mostra: $n = 200$.
- Media das indemnizacións: $\bar{x} = 151{,}25$ euros.
- Cuasivarianza das indemnizacións: $S_c^2 = 54{,}13$.
- Nivel de confianza: $1 - \alpha = 0{,}95$ (polo tanto, $\alpha = 0{,}05$).

Obxectivo. Determinar un IC para o total (X) cunha confianza do 95 %.

Teorema Central do Límite. IC para X con un nivel de confianza $1-\alpha$:

$$IC_X = \left(\hat{X} - z_{\alpha/2}\sqrt{\frac{N^2 S_c^2}{n}}, \hat{X} + z_{\alpha/2}\sqrt{\frac{N^2 S_c^2}{n}} \right)$$

e sabendo que a estimación do total $\hat{X}$ se obtén como

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

Cal é o intervalo de confianza para o total X?

Mostraxe aleatoria simple: determinación do tamaño da mostra

Tamaño da mostra para a media

Fixado $\varepsilon > 0$, o tamaño de mostra requerido para que $|\bar{x} - \bar{X}| < \varepsilon$ é

MAS sen reemplazo	MAS con reemplazo
$n = \dfrac{n_0}{\frac{n_0}{N} + 1}$	$n = n_0$

- **Desigualdade de TChebyshev.**

$$n_0 = \frac{k^2\sigma^2}{\varepsilon^2}$$

- **Teorema Central do Límite.**

$$n_0 = \frac{z_{\alpha/2}^2\sigma^2}{\varepsilon^2}$$

sendo z_α o punto que deixa unha probabilidade á súa dereita igual a α e σ^2 é a varianza da poboación. O seu valor, usualmente descoñecido, é aproximado por

$$S_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \text{ (cuasivarianza da mostra)}$$

Tamaño da mostra para o total

Fixado $\varepsilon > 0$, o tamaño de mostra requerido para que $|\hat{X} - X| < \varepsilon$ é

MAS sen reemplazo	**MAS con reemplazo**
$n = \dfrac{n_0}{\frac{n_0}{N} + 1}$	$n = n_0$

- **Desigualdade de TChebyshev.**

$$n_0 = \frac{N^2 k^2 \sigma^2}{\varepsilon^2}$$

- **Teorema Central do Límite.**

$$n_0 = \frac{N^2 z_{\alpha/2}^2 \sigma^2}{\varepsilon^2}$$

sendo z_α o punto que deixa unha probabilidade á súa dereita igual a α e σ^2 é a varianza da poboación. O seu valor, usualmente descoñecido, é aproximado por

$$S_c^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \text{ (cuasivarianza da mostra)}$$

Tamaño da mostra para a proporción

Fixado $\varepsilon > 0$, o tamaño de mostra requerido para que $|P - p| < \varepsilon$ é

MAS sen reemplazo	**MAS con reemplazo**
$n = \dfrac{n_0}{\frac{n_0}{N} + 1}$	$n = n_0$

- **Desigualdade de TChebyshev.**

$$n_0 = \frac{k^2 PQ}{\varepsilon^2}$$

- **Teorema Central do Límite.**

$$n_0 = \frac{z_{\alpha/2}^2 PQ}{\varepsilon^2}$$

sendo z_α o punto que deixa unha probabilidade á súa dereita igual a α, e sendo P a proporción de individuos coa característica na poboación e $Q = 1 - P$.

Exemplo: mostraxe aleatoria simple

Unha compañía de seguros analizou aleatoriamente 200 dos seus 10.000 expedientes para estudar as cuantías das indemnizacións que paga polos accidentes de tráfico cubertos polas súas pólizas. Da información obtida sábese que a media das indemnizacións pagadas foi de 151,25? e a cuasivarianza foi de 54,13. Cunha confianza do 95 %,

(b) Se se desexa estimar a proporción de expedientes que deron menos de dous partes de sinistro ó ano, ¿qué tamaño de mostra debe utilizarse para conseguir a estimación cun erro de mostraxe inferior ó 12 %?.

Datos.

- Tamaño da poboación: $N = 10000$ expedientes.
- Tamaño da mostra: $n = 200$.
- Media das indemnizacións: $\bar{x} = 151{,}25$ euros.
- Cuasivarianza das indemnizacións: $S_c^2 = 54{,}13$.
- Nivel de confianza: $1 - \alpha = 0{,}95$ (polo tanto, $\alpha = 0{,}05$).

Exemplo: mostraxe aleatoria simple

Neste caso, o erro debe ser inferior ó 12 % do valor do total, isto é:

$$\varepsilon = 0{,}12\hat{X}.$$

MAS sen reemplazo	**MAS con reemplazo**
$n = \dfrac{n_0}{\frac{n_0}{N} + 1}$	$n = n_0$

- Teorema Central do Límite.

$$n_0 = \frac{N^2 z_{\alpha/2}^2 \sigma^2}{\varepsilon^2}$$

Cal é o tamaño de mostra requerido?

Algúns comentarios no cálculo dos tamaños da mostra

- Se $n \leq n_0$, o MASsr asegura tamaños de mostra menores que os MAScr para un mesmo valor de ε.
- Se n aumenta, ε diminúe e polo tanto, o nivel de confianza $1-\alpha$.
- Se S^2 aumenta, fixado ε, n tamén aumenta.
- Debe tomarse como tamaño de mostra n o valor enteiro máis próximo por exceso ó que proporciona a fórmula.
- Se σ^2 (e polo tanto S^2) é descoñecido, débense tomar aproximacións ou cotas (a cuasivarianza da mostra, por exemplo).
- Se P é descoñecido, débese usar o valor 0,5.