FAQ sobre Como realizar un proxecto estatístico para a Incubadora de Sondaxes e Experimentos

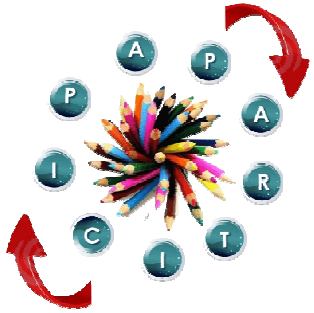
Tomás R. Cotos-Yáñez // cotos@uvigo.es

Dpto. de Estatística e I.O. – Universidade de Vigo

Indice:

1. Bases VII Incubadora de Sondaxes e Experimentos.
2. Proxecto estatítistico.
3. Redacción das preguntas.
4. Recollida de datos
    a. Fontes Públicas, Censo e Sondaxe.
    b. Mostraxe: mostra representativa.
    c. Mostraxe aleatoria Simple (mas).
        i. Selección aleatoria.
        ii. Determinar o tamaño mostral

Bases <u>VII Incubadora de Sondaxes e Experimentos</u>:



Obxectivos:

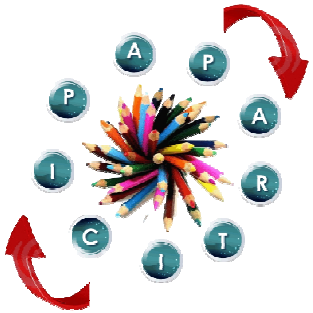
- Fomentar o ensino e aprendizaxe da Estatística nos niveis educativos non universitarios.
- Difundir a importancia e utilidade da Estatística e a Investigación de Operacións na vida real e en tódalas disciplinas académicas.
- Incentivar nos docentes o uso de ferramentas tecnolóxicas, e de novos materiais e técnicas no ensino da Estatística e da Investigación de Operacións.
- Espertar nos estudantes a curiosidade pola Estatística e a Investigación de Operacións como ferramenta fundamental na investigación en tódalas ciencias e na elaboración de proxectos.
- Elevar os coñecementos e as competencias dos estudantes en torno á Estatística e a Investigación de Operacións e a súa aplicación na vida diaria, nos estudos e na comprensión do seu contorno.
- Familiarizar aos estudantes coas fases de realización dun estudo estatístico e coas distintas linguaxes reflectidas neste proceso.

Realización dun proxecto de Estatística e/ou Investigación de Operacións:

- realizarán un proxecto de Estatística e/ou Investigación de Operacións.

Bases e +info: www.sgapeio.es

Bases VII Incubadora de Sondaxes e Experimentos:



Participantes:

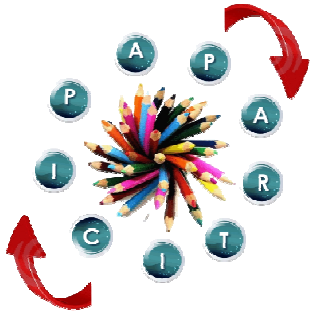
- Estudantes de ESO, FP Básica, BAC e Ciclos Formativos de Grao Medio.
  - Número: máximo de 4 estudantes. Un estudante só pode estar nun único traballo.
- Polo menos cun docente-titor do seu centro.
  - Non hai limitación de titores, nin se restrinxe a afiliación,
  - nin se limita nº máximo de traballos presentados.

Categorías:

- 3 categorías: 1º e 2º da ESO, 3º e 4º da ESO e FP Básica e Bacharelato e Ciclos Formativos de Grao Medio

Bases e +info: www.sgapeio.es

Bases <u>VII Incubadora de Sondaxes e Experimentos</u>:



Datas importantes:

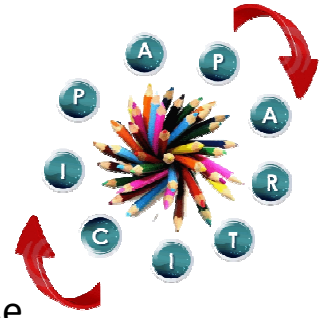
- Inscrición de traballos: dende o 1 ata o 16 de abril de 2017 via web.
- Entrega dos traballos finalizados: dende o 17 ata o 30 de abril de 2017.
- Do 29 de maio ao 4 de xuño de 2017 daranse a coñecer os traballos finalistas e a data do acto de entrega de premios

Os proxectos seleccionados poderán participar na fase Nacional*

Proposta para o 2018: 27 ao 29 de juño en Cantabria.

* Ata agora a SGAPEIO e os organizadores da fase nacional sufragaron todos os gastos dos estudantes + titores

Bases VII Incubadora de Sondaxes e Experimentos:



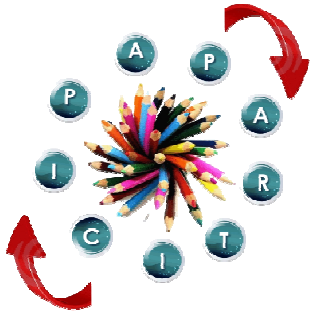
Que teño que enviar:

- Un informe especificando: Obxetivos, descrición dos datos e como se obtiveron, análise estatística e interpretación dos resultados e conclusións (máx 20 páxinas).
- Adicionalmente: arquivo cos datos, material audiovisual, ...

A quen:

- Correo-e a premio@sgapeio.es

Bases VII Incubadora de Sondaxes e Experimentos:



Valoración dos proxectos:

- A orixinalidade do tema.
- A claridade na exposición de obxectivos, fases do proxecto e resultados.
- A correcta aplicación de técnicas estatísticas e a súa adaptación ao nivel no que estudan os participantes.
- As ferramentas tecnolóxicas empregadas.
- As conclusións do traballo, de acordo aos obxectivos do mesmo.
- A análise crítica do proxecto realizado e posibles extensións do mesmo.
- O informe final (redacción, estrutura, elección apropiada de táboas e gráficos, etc.)

Xurado (6 membros):

- Representando a Sociedade/Consellería de Educación/Universidades/IGE/Profesorado

Plantexamento do problema:

2. Proxecto estatístico:

Dividir un problema xeral en pequenos subproblemas que se converteran en **obxectivos** concretos.

Estes **obxectivos** concretos traduciranse no argot estatístico fundamentalmente en **parámetros** a determinar mediante a investigación empírica.

Natureza dos obxectivos:

- observación de cantidades cualitativas.
- observación de cantidades cuantitativas.

Cantidades cualitativas:

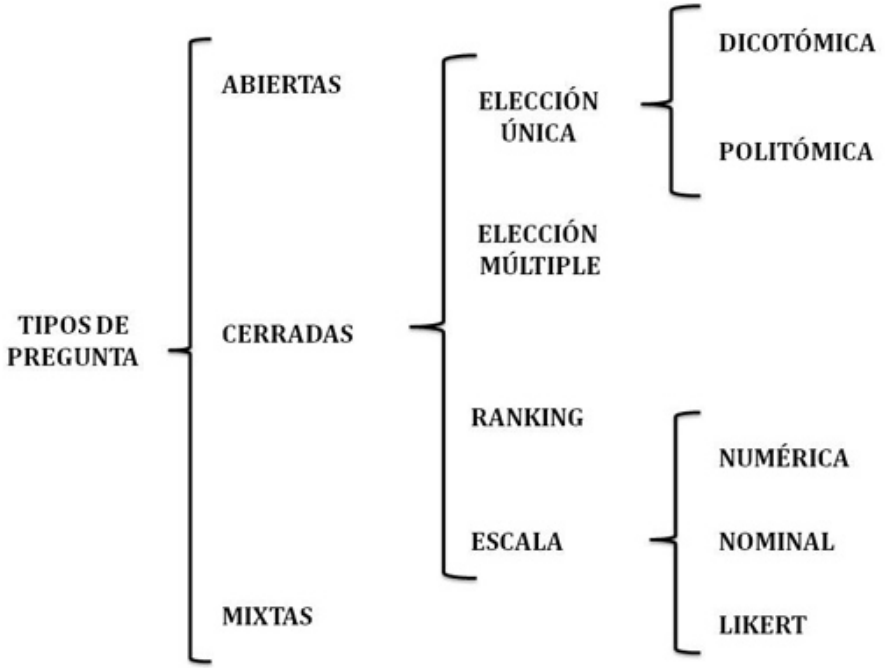
- Respostas dicotómicas: presentas dúas opción, p.e. Si-Non
- Respostas politómicas: presentan varias alternativas
- Respostas en escala de Likert (cuantificación cualitativa en diferentes graos):
  - Escala de Likert de 5 niveis:
    - Totalmente en desacordo 1
    - En desacordo 2
    - Nin de acordo nin en desacordo 3
    - De acordo 4
    - Totalmente de acordo 5

Cantidades cuantitativas:

- Respostas numéricas: discretas ou continuas
- Respostas en intervalos: 0, [1,5], … (ou en versión agrupada: Nunca (0), Poucas veces (0,5], A veces (5,10] e Moitas veces >10)

3. Redactar as Preguntas

- Elección múltiple: as opcións non son excluíntes entre si.
- Ranking: pídese ordenar as respostas



Parámetro:

- Resumo dunha variable estatística.

Tipos:

- de posición: Media aritmética/Proporción, Mediana, Cuantis,
- de dispersión: desviación típica, rango intercuartílico,
- de forma: asimetría,
- …

O obxectivo do traballo (estatisticamente falando) é a consecución numérica dos

parámetros

¿Cantos parámetros?, 1,2, …, tantos como preguntas?

4. Recollida de datos:

Fontes públicas	Datos nosos
• IGE e equivalentes • INE • CIS • Eurostat	• Experimento • Sondaxe - Censo

4.a <u>Fontes públicas</u>:

- Arquivo en formato xls, ods, csv, coas variables indicadas.

<u>Experimentos</u>:

- Recoller os datos da experimentación nun soporte informático.
- Se é posible e ten sentido, repetir os experimentos nas mesmas condicións para eliminar o erro de medición, p.e. *tempo de ebulición dun líquido*.

Censo

- Examínanse todas as unidades da poboación obxectivo.
  - Principal vantaxe: Ausencia de erro. Non hai estimacións, os valores dos parámetros obtidos son os reais.
  - Principal desvantaxe: consume moitos recursos. Xeralmente non é viable por falta de tempo, medios económicos, excesivo persoal, …

Sondaxe - Enquisa:

- examinar unhas poucas unidades da poboación que son representativas: *mostra*.
  - Principal vantaxe: Razoable consumo de recursos.
    - A poboación é inabarcable para o investigador.
    - A poboación é suficientemente homoxénea.
    - O proceso de medida o investigación é destrutivo. . . .
  - Principal desvantaxe: Información non exacta → Medidas con Erro.

Censo ou Sondaxe

- Desexase obter información sobre a porcentaxe de alumnos nun centro de ensino que teñen móbil*.
  - Censo: datos de todos os alumnos!!!
  - Sondaxe: seleccionar unha mostra representativa.

* que significa ter móbil: con tarxeta/sen tarxeta? levalo ao centro? nº de horas mínimo de uso?...

4.b Mostra representativa

- **NON EXISTEN**
- Depende de para qué? e unha vez determinado, do erro máximo permitido na obtención da/s cantidade/s descoñecida/s
- Exemplo:
  - Estimar a proporción de alumnos que usan o móbil nun centro ou a homoxeneidade nas notas dos alumnos por clase (a través do Coef. de Variación)
  - Un erro máximo de +-0.1 ou de +-0.5 nunha proporción.

Unha mostra representativa será aquela que me permita estimar os parámetros descoñecidos cun erro máximo e para un nivel de confianza determinado.

Cómo se fai a selección e a cantos (n) 

Como selecciono unha mostra

¿Selección aleatoria ou intencional?

- Selección intencional: un experto establece os elementos.
- Selección aleatoria: usando modelos probabilísticos.

Sondaxe de hábitos de estudo (dúas opcións):

1. A equipa directiva determina os 100 estudantes pertencen ao estudo.
2. Selecciónanse ao azar:
   a. Os 100 primeiros estudantes que entren despois do recreo, os voluntarios, ...
   b. Enuméranse os estudantes e selecciónanse ao azar 100 números.

4.c Mostraxe Aleatoria Simple (mas) para poboacións finitas

- (H1) Todos os elementos da poboación teñen idéntica probabilidade de ser seleccionados.
- (H2) A selección realizase de forma independente.

Con reemplazamento ou sen reemplazamento (en poboacións finitas)

Baixo condicións xerais: $n_s < n_r = n_\infty$

Polo tanto, o mecanismo de recollida da mostra debe asegurar estas dúas premisas. O non cumprimento delas leva a resultados non fiables e xeralmente non corrixibles

Exemplo mas:

100 Estudantes nun centro

- N: total de estudantes
- Enumerar os estudantes de 1 a N
- Elexir aleatoriamente 100 números entre 1 e N

Pódense establecer cuotas:

proporcionalidade por sexo

por curso, …

Recursos:

- Excel, Calc: aleatorio.entre(linf;lsup)
- http://www.numeroalazar.com.ar/

100 Hab. nun concello

- Idem … pero como precísanse datos censais… dificiles de conseguir
- Alternativa: rutas aleatorias. Obxectivo cubrir xeográficamente a superficie do concello
- Non é 100% aleatorio

Proporcionalidade hab urbanos/rural

- http://pinetools.com/es/generador-numeros-aleatorios
- http://rextester.com/l/r_online_compiler

  cbind(sort(sample(1:1000, size=100, replace=FALSE)))

Cantos – ¿n?

n tamaño da mostra para estimar un parámetro θ verificando:

- Tipo de mostraxe (mas)
- Nivel de confianza 1-α. Xeralmente 95%
- Erro máximo permitido – Establécese subxectivamente!!

O máis habitual é estimar como parámetro unha media ou unha porcentaxe:

1. Recursos online.
2. Folla Excel "Tamaño-mostral.xls".
   https://www.dropbox.com/s/uhh0u07ykpqkxaf/Tama%C3%B1o_mostral.xls?dl=0

Estimación por Intervalos/ Intervalos de Confianza

Definición:

La estimación confidencial o estimación por intervalos consiste en determinar un posible rango de valores o intervalo, en los que pueda precisarse con una determinada probabilidad que el valor de un parámetro se encuentra dentro de esos límites.

Definición:

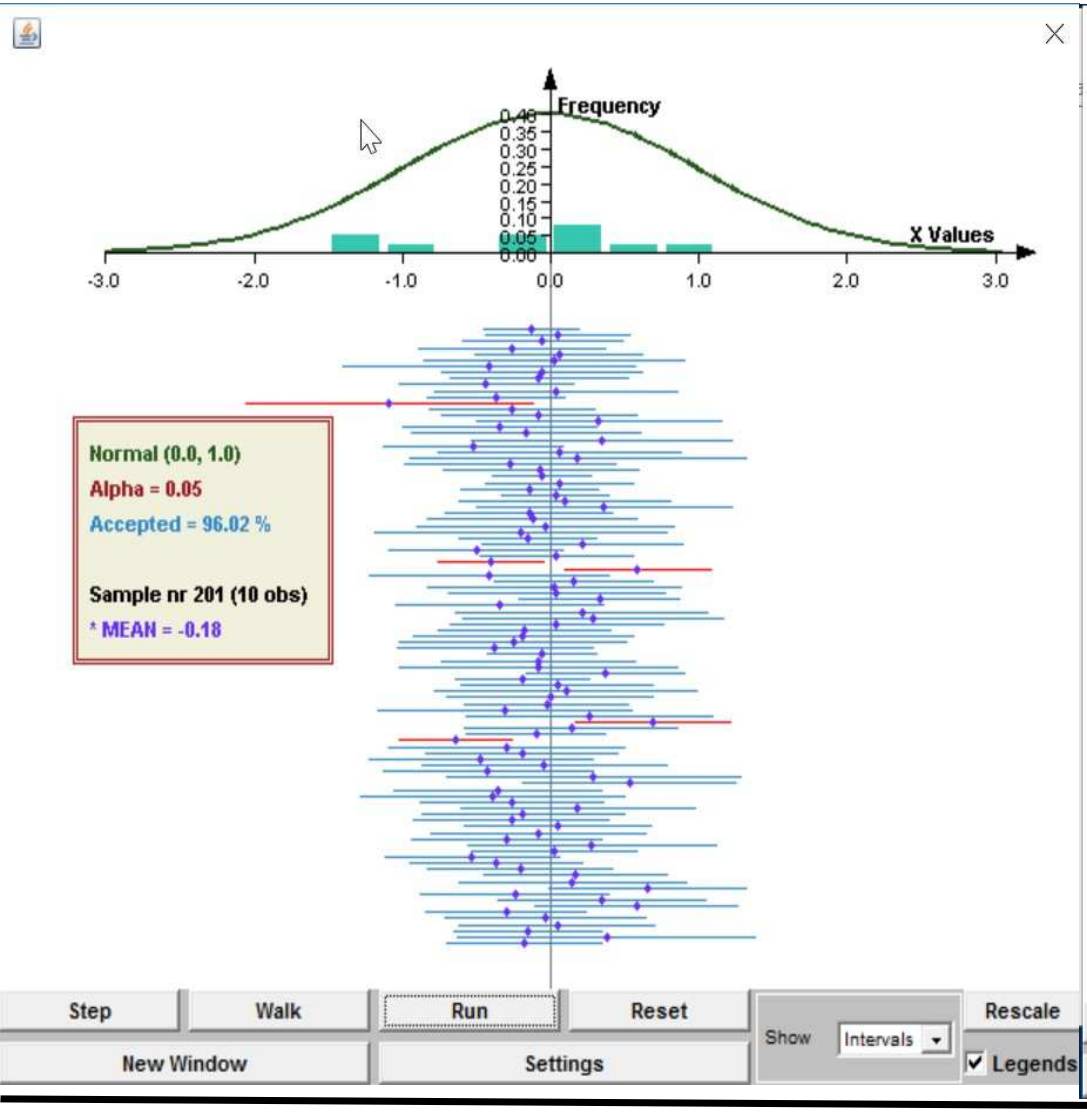
Dada una muestra aleatoria $X_1, \ldots, X_n$, se denomina Intervalo de Confianza para el parámetro θ con nivel de confianza $1-\alpha$, a un intervalo aleatorio $\left(\widehat{\Theta}_{inf}, \widehat{\Theta}_{sup}\right)$ (cuyos límites dependen de la muestra) de manera que verifica:

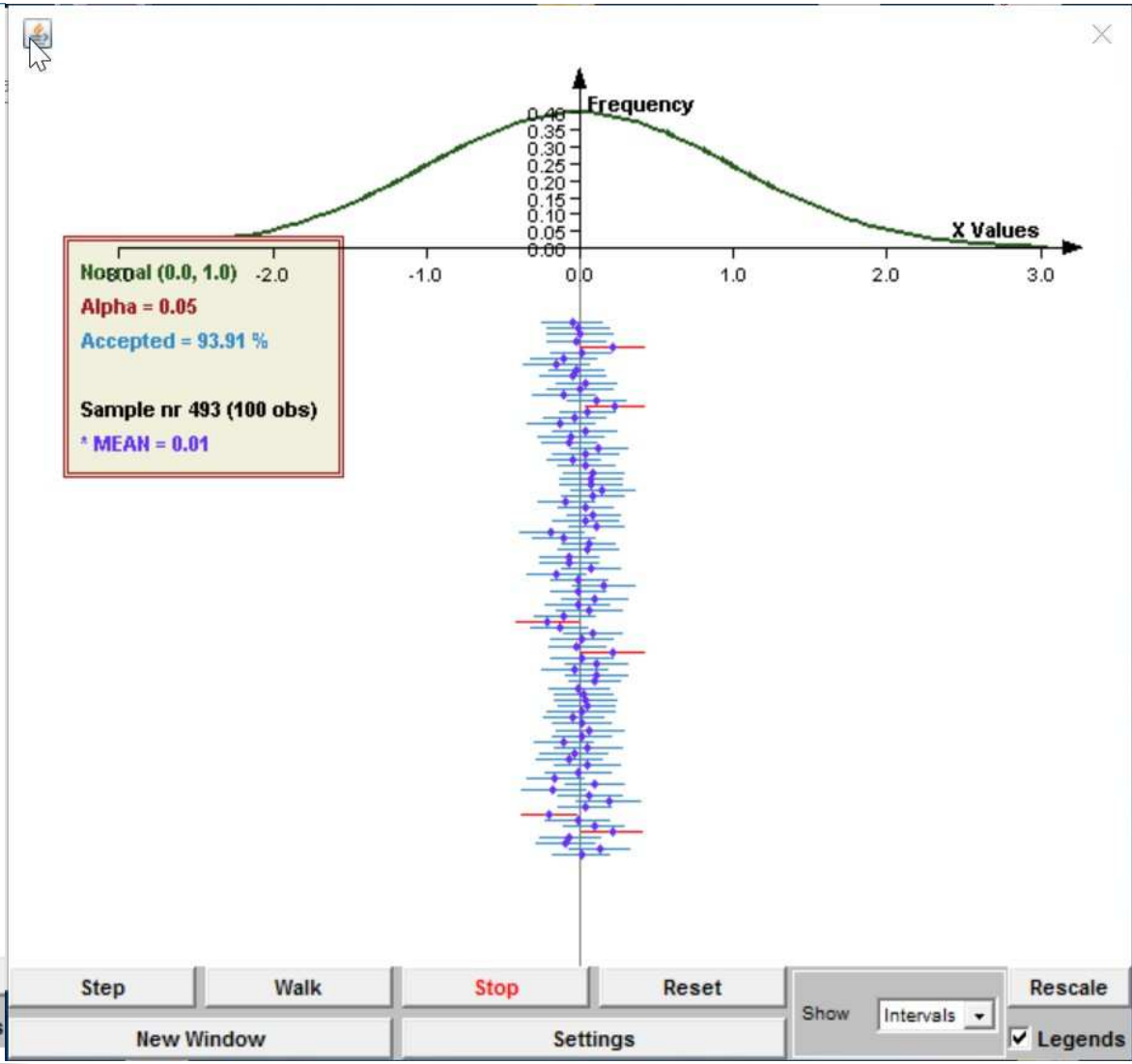
$$P\left(\widehat{\Theta}_{inf}(X_1, \ldots, X_n) < \theta < \widehat{\Theta}_{sup}(X_1, \ldots, X_n)\right) = 1-\alpha \text{ para cada } \theta \in \Theta$$

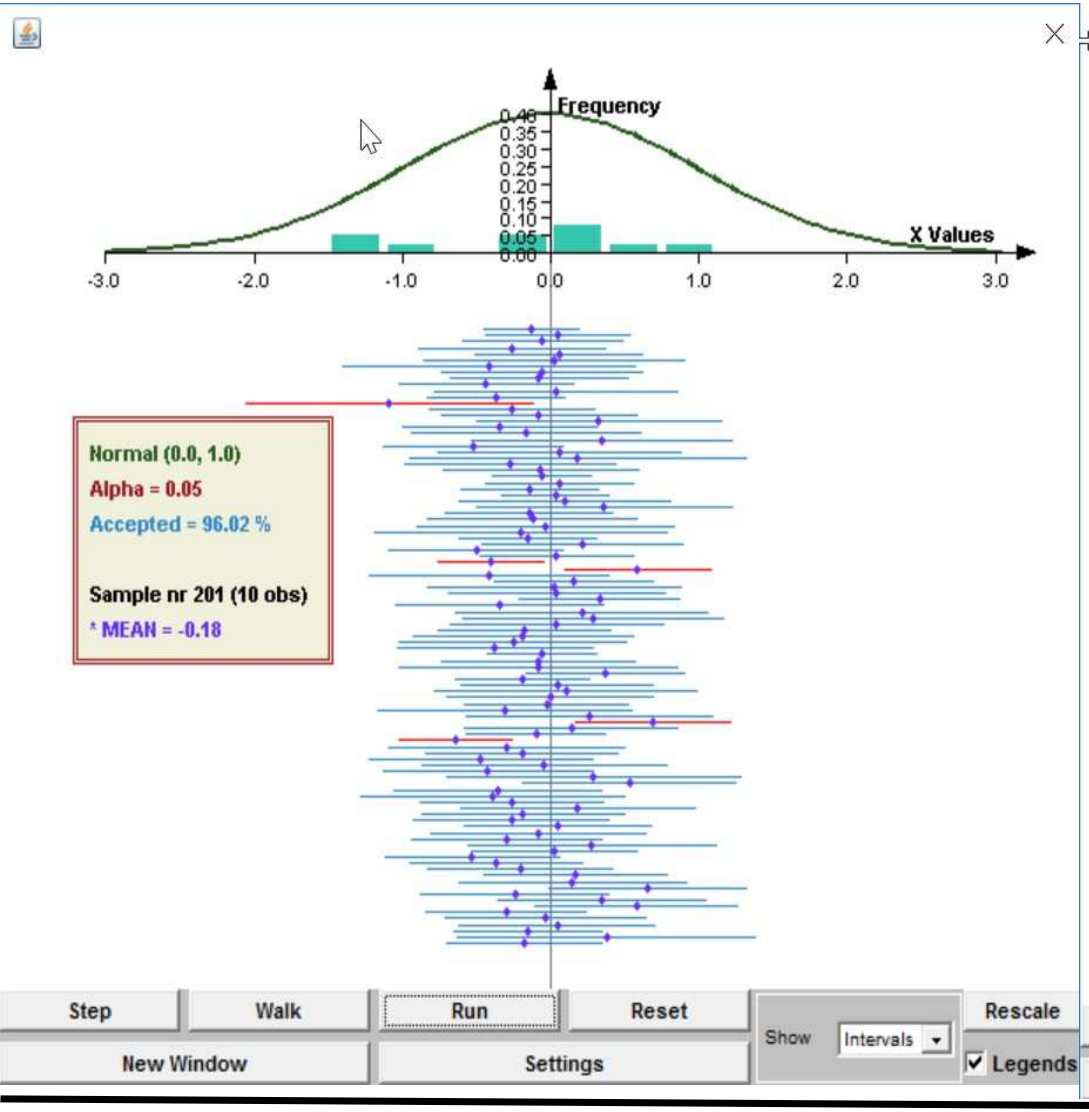
Interpretación:

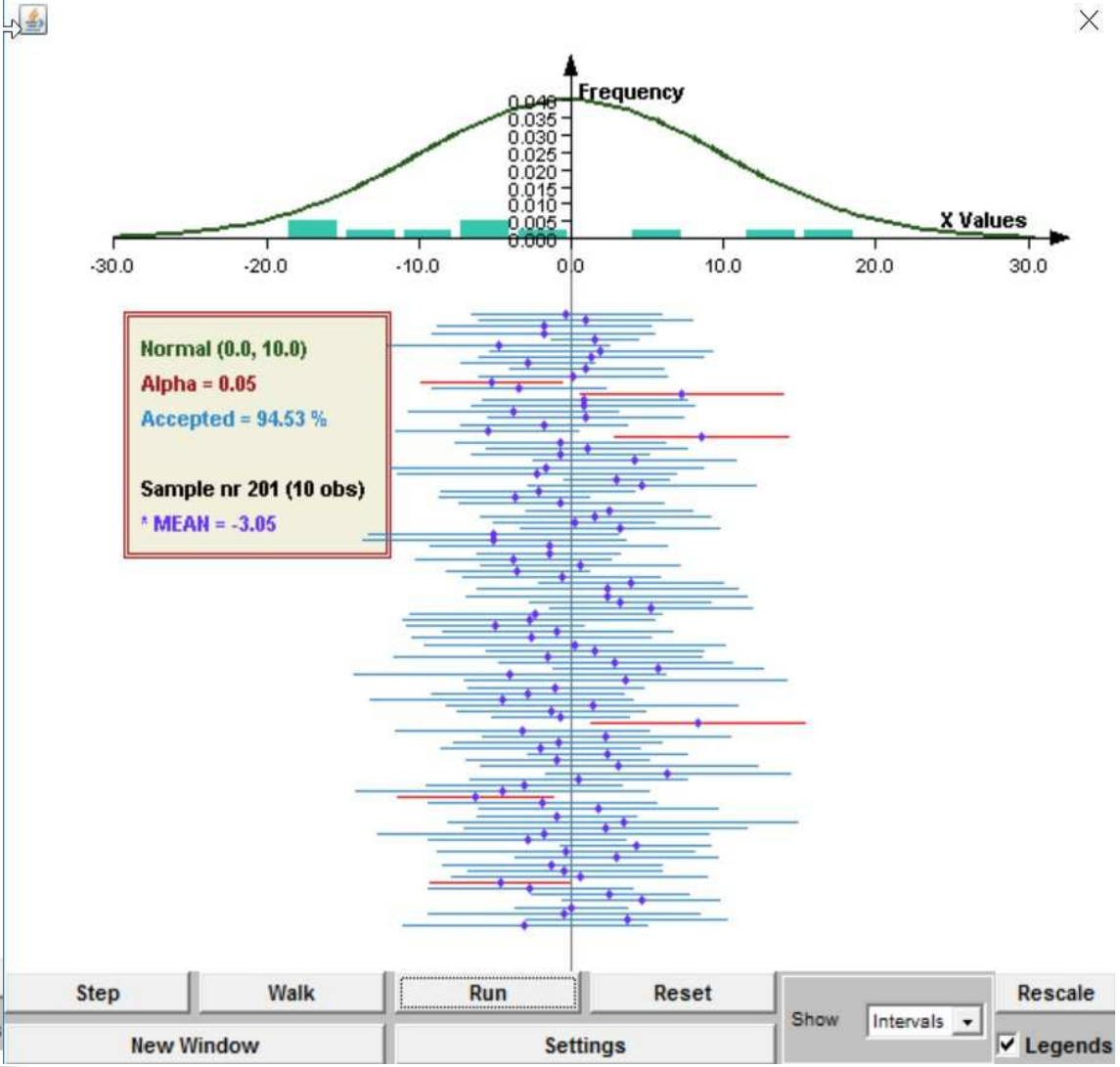
Para una muestra $x_1, \ldots, x_n$, una vez calculados los límites inferior $\widehat{\Theta}_{inf}(x_1, \ldots, x_n)$ y superior $\widehat{\Theta}_{sup}(x_1, \ldots, x_n)$, se dice entonces que el intervalo que forman $\left(\widehat{\theta}_{inf}, \widehat{\theta}_{sup}\right)$ constituye una **estimación por intervalo de confianza** de θ con **nivel de confianza** $1 - \alpha$ en el sentido de que si se tomasen infinitas muestras $\{x_1, \ldots, x_n\}$ y construyésemos los correspondientes intervalos de confianza, el $100(1 - \alpha)\,\%$ de ellos contendrían el verdadero valor del parámetro.

Ver Java-Applet on line *Test*: https://lstat.kuleuven.be/newjava/vestac


Frequency
X Values
Normal (0.0, 1.0)
Alpha = 0.05
Accepted = 96.02 %
Sample nr 201 (10 obs)
* MEAN = -0.18
Step
Walk
Run
Reset
Show
Intervals
Rescale
New Window
Settings
Legends


Frequency
X Values
Normal (0.0, 1.0)
Alpha = 0.05
Accepted = 93.91 %
Sample nr 493 (100 obs)
* MEAN = 0.01
Step
Walk
Stop
Reset
Show
Intervals
Rescale
New Window
Settings
Legends


Frequency
X Values
Normal (0.0, 1.0)
Alpha = 0.05
Accepted = 96.02 %
Sample nr 201 (10 obs)
* MEAN = -0.18
Step
Walk
Run
Reset
Show
Intervals
Rescale
New Window
Settings
Legends


Frequency
X Values
Normal (0.0, 10.0)
Alpha = 0.05
Accepted = 94.53 %
Sample nr 201 (10 obs)
* MEAN = -3.05
Step
Walk
Run
Reset
Show
Intervals
Rescale
New Window
Settings
Legends

Cantos – ¿n?

n tamaño da mostra para estimar un parámetro θ verificando:

- Tipo de mostraxe (mas)
- Nivel de confianza 1-α. Xeralmente 95%
- Metade da lonxitude do I.C.:
  - Erro máximo absoluto: $e = \hat{\theta} - \theta$ ou
  - Erro máximo relativo (%): $er = 100 * \frac{\hat{\theta}-\theta}{\theta}$

Fórmulas para o tamaño da mostra (2 BAC) :

1. Para μ baixo hipótese de Normalidade e poboación infinita:
2. Para μ baixo hipótese de Normalidade e poboación finita:
3. Para p para mostras grandes:

- Para la media y el total

Exacto	Aproximado
$n \geq \frac{z^2_{1-\frac{\alpha}{2}} N S^2}{z^2_{1-\frac{\alpha}{2}} S^2 + (N-1)E^2}$ Si $1-\frac{\alpha}{2}=0{,}955$ $\approx \frac{4NS^2}{4S^2+(N-1)E^2}$	$n \geq \frac{z^2_{1-\frac{\alpha}{2}} S^2}{E^2}$ Si $1-\frac{\alpha}{2}=0{,}955$ $\approx \frac{4S^2}{E^2}$

- Para la proporción y el total de clase

Exacto	Aproximado
$n \geq \frac{z^2_{1-\frac{\alpha}{2}} P(1-P)N}{E^2(N-1)+z^2_{1-\frac{\alpha}{2}} P(1-P)}$	$n \geq \frac{z^2_{1-\frac{\alpha}{2}} P(1-P)}{E^2}$

¿Cantos parámetros?, 1,2, …, tantos como preguntas?

Para θ_1 precísase unha mostra de tamaño n_1 cunha confianza de nivel 1-α e error máximo relativo e_1
Para θ_2 precísase unha mostra de tamaño n_2 cunha confianza de nivel 1-α e error máximo relativo e_2
Para θ_3 precísase unha mostra de tamaño n_3 cunha confianza de nivel 1-α e error máximo relativo e_3

Para (θ_1, θ_2) o nivel de confianza é agora (supoñendo independencia) $(1-\propto)^2$ que no caso do 95% teríamos unha confianza de 90.25%
Para $(\theta_1, \theta_2, \theta_3)$ o nivel de confianza é $(1-\propto)^3$, polo tanto no caso do 95% teríamos unha confianza de 85.73%
Para $(\theta_1, \ldots, \theta_4)$ o nivel de confianza é $(1-\propto)^4$, polo tanto no caso do 95% teríamos unha confianza de 81.45%
…
Para $(\theta_1, \ldots, \theta_{10})$ o nivel de confianza é $(1-\propto)^{10}$, e ao 95% teríamos unha confianza de 59.87%

Datos missing:

Son aqueles que non constan debido a calquera acontecemento que imposibilita o seu análise.

Múltiples causas: erros na transcrición , non resposta, perda das enquisas, …

O seu tratamento depende da tipoloxía da perda:

- Perda completamente aleatoria: o efecto é soamente no tamaño da mostra, que loxicamente será menor. Poderíase substituír por outro elemento da poboación
- Perda aleatoria (MAR) e non aleatoria (NMAR): nestes casos o valor da variable que se perde está relacionada co motivo da perda. Ademais do efecto anterior, sesgo nas estimacións!!! Non se pode substituír!!!

← FIN →