ESTIMACIÓN

puntual y por intervalo

•   (   )

V.E.Rohen

¿Podemos conocer el comportamiento del ser humano?

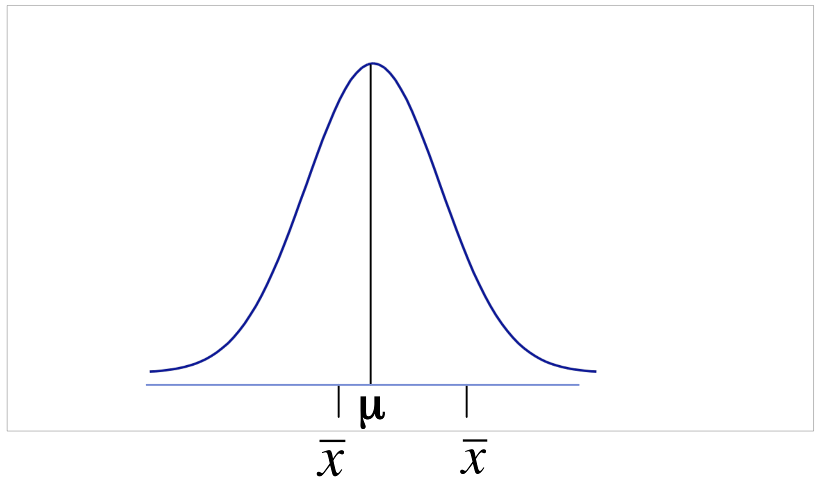
V.E.Rohen

Podemos usar la información contenida en la muestra para tratar de "adivinar" algún aspecto de la población bajo estudio y sustituirla en lo que sería nuestra "verdad desconocida"

Esto, por supuesto, implica que la información que obtenemos de nuestras observaciones debe ser representativa del particular aspecto de la población.

V.E.Rohen

Es importante notar que no siempre coincide la información que hemos observado con la información real de la población.



Sin embargo, es una buena aproximación y la podemos utilizar para la estimación de las características propias de dicha población.

V.E.Rohen

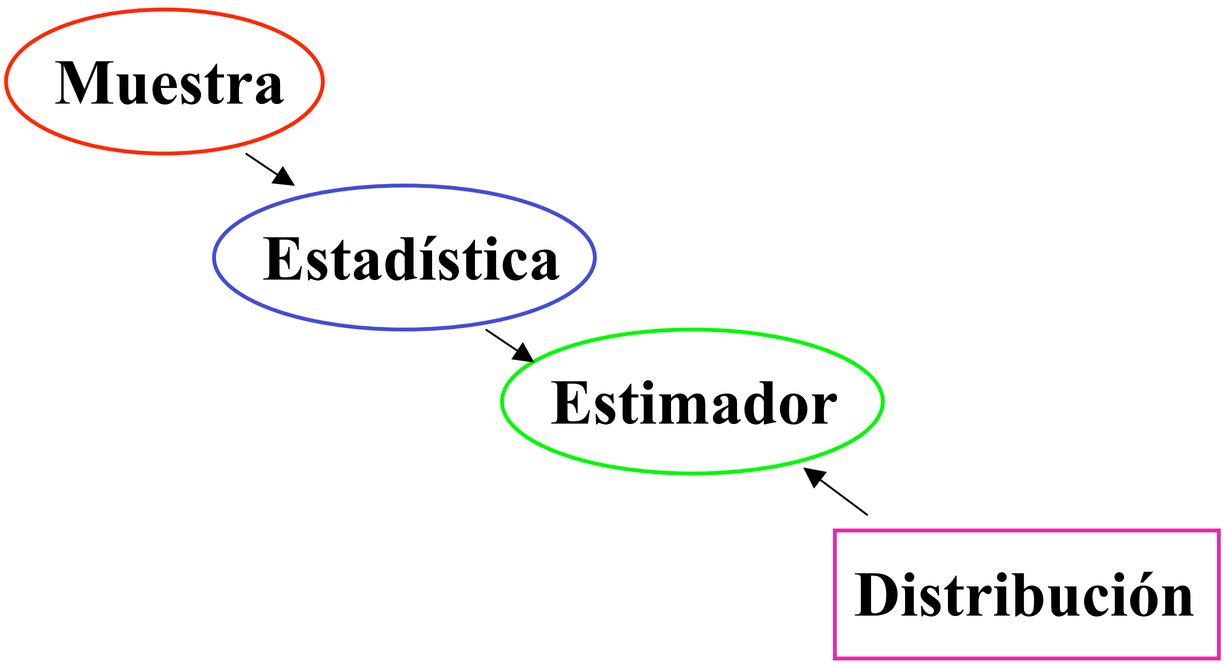
Podemos entonces dar una medida de dicha incertidumbre:

$$\boxed{\varepsilon = \left|\theta - \hat{\theta}\right|}$$

(Esta medida nos ayudará a crear estimadores por intervalo para medias y proporciones muestrales)

V.E.Rohen

La distribución de la muestra y de las "estadísticas" juega un papel crítico en la inferencia estadística porque la bondad de los estimadores se mide en base a la media y varianza de éstas.



V.E.Rohen

Las muestras son tomadas para Estimar *parámetros* y para Probar Hipótesis acerca de los *parámetros*

Un parámetro es una medida numérica de algún aspecto de la población

Cuando no tenemos la información sobre toda la población es necesario estimar el valor del parámetro en base a la información de la muestra sobre dicho aspecto de interés y tenemos lo que se llama "estadística"

**Supongamos que tomamos una muestra de una población y obtenemos la media muestral.
Si tomamos otra muestra obtendremos otro valor de la media muestral, y así sucesivamente.**

Todas estas medias serán variables aleatorias que tienen asociada una función de densidad.

Lo mismo sucede con las varianzas muestrales que cambian su valor de muestra a muestra y con las proporciones muestrales.

Pero el <u>promedio de todas las medias muestrales posibles</u> con o sin reemplazo (cada una del mismo tamaño n) es igual a la media poblacional μ.

La fluctuación en el número que representa a estas medias muestrales se ve en un histograma de todos los posibles valores de éstas. Estas fluctuaciones son menores que las fluctuaciones de los valores en la población.

Estas variaciones entre las medias muestrales se conoce como error estándar de la media y se obtiene como

$$\boxed{\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}}$$

Se puede observar que si el tamaño de la muestra aumenta, el error estándar disminuye.

¿Qué distribución sigue la media muestral?

Teorema Central del Límite

Consideremos muestras aleatorias de una población con media μ y varianza σ^2, conforme el tamaño de la muestra crece, la distribución de las medias muestrales es aproximadamente NORMAL, sin importar la forma de la distribución de la población.

DISTRIBUCIÓN DE LA MEDIA MUESTRAL

$$\bar{X}$$

Recordemos que la media muestral $\bar{X}$ obtenida de una muestra aleatoria de tamaño n de una población con media μ y varianza σ^2, tiene una distribución normal con media μ y varianza σ^2/n

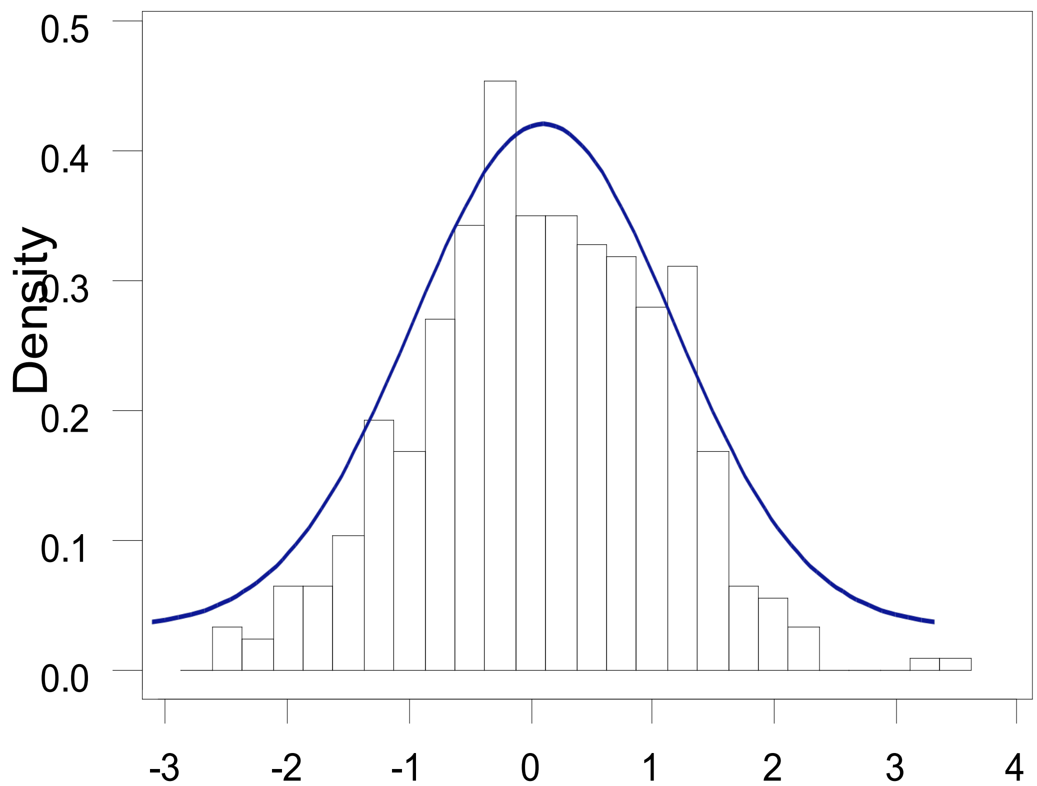
V.E.Rohen

Vamos a poder medir qué tanto se desvía la media muestral de la media poblacional a través del valor *Z*, de la siguiente manera

$$Z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\left(\overline{X} - \mu\right)\sqrt{n}}{\sigma}$$

Es fácil ver que la *Z*, que es una estadarización de la media muestral, sigue una distribución *N(0,1)*

V.E.Rohen



V.E.Rohen

Con frecuencia estamos interesados en determinar si la media de una población es diferente de la media de otra poblacion.

Si la Población 1 tiene una media μ_1 y una desviación estándar σ_1 y la Población 2 tiene una media μ_2 y una desviación estándar σ_2 , nos gustaría determinar si $\mu_1 = \mu_2$ o si una es mayor que la otra ($\mu_1 > \mu_2$ ó $\mu_1 < \mu_2$)

para lo cual nos basamos en la evidencia que tenemos al considerar dos muestras aleatorias, una de cada una de las poblaciones y observar la diferencia de las medias muestrales $\overline{X}_1 - \overline{X}_2$ **.**

Como $\overline{X}_1$ y $\overline{X}_2$ **son variables aleatorias normalmente distribuidas, entonces**

$\boxed{\overline{X}_1 - \overline{X}_2}$ **es una variable aleatoria distribuida normalmente con media**

$\boxed{\mu_1 - \mu_2}$ **y con varianza** $\boxed{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$ **.**

En muchas ocasiones no conocemos la probabilidad de éxito en un experimento binomial y tiene que ser estimado de la muestra. Como *p* es la probabilidad de éxitos en cualquier prueba, en una población finita, *p* mide la proporción de éxitos en esa población.

Así, si en una muestra de tamaño *n* de una población, *X* es el número de éxitos, estimamos la proporción de éxitos en esta muestra: $\frac{X}{n}$

Entonces $\hat{p} = \frac{X}{n}$ **tiene una distribución**

normal con media *p* y varianza *p(1-p)/n*

siempre y cuando $np(1-p)>5$ ***(Rosner)***

Muchos problemas están enfocados en determinar si la proporción de gente o cosas en una población que posee cierta característica es la misma que la proporción que posee dicha característica en otra población: $p_1 = p_2$, ó si es mayor: $p_1 > p_2$ ó menor: $p_1 < p_2$.

Cuando desconocemos estas proporciones es necesario tomar una muestra de cada población y estimar dichas proporciones

Tomemos dos muestras de tamaño n_1 y n_2 de las dos poblaciones bajo estudio.

Encontremos el número (X_1) de individuos en la muestra de la Población 1 que posee la característica de interés y el número (X_2) de individuos en la muestra de la Población 2 que poseen la misma característica, entonces las proporciones muestrales

$$\hat{p}_1 = \frac{X_1}{n_1} \quad \text{y} \quad \hat{p}_2 = \frac{X_2}{n_2}$$

serán los estimadores de p_1 y p_2 respectivamente

La distribución de la variable aleatoria $\boxed{\hat{p}_1 - \hat{p}_2}$

es aproximadamente normal con media $\boxed{p_1 - p_2}$

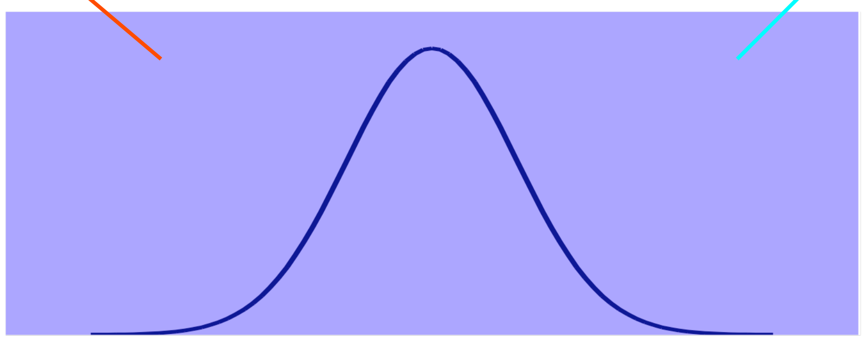
y varianza

$$\boxed{\sigma^2_{\hat{p}_1 - \hat{p}_2} = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

siempre y cuando $n_1 p_1(1- p_1) > 5$, $n_2 p_2(1- p_2) > 5$

(Rosner)

$$\bar{x} \qquad \sum z_i^2 \qquad \frac{z}{\sqrt{\chi^2/\nu}} \qquad \frac{w_1/\nu_1}{w_2/\nu_2}$$



$$x_1, x_2, \cdots, x_n$$

Distribuciones de Muestreo

V.E.Rohen

Algunas distribuciones que se derivan de la distribución normal

Si $\boxed{Z \sim N(0,1)}$ **entonces** $\boxed{Z^2 \sim \chi_1^2}$

Si $\boxed{Z_i \sim N(0,1)}$ **para** *i=1,...,n,* **entonces** $\boxed{\sum_{i=1}^{n} Z_i^2 \sim \chi_n^2}$

V.E.Rohen

$\boxed{Z \sim N(0,1)}$ $\boxed{W \sim \chi_n^2}$

$$\boxed{\frac{Z}{\sqrt{\frac{W}{n}}} \sim t_n}$$

Si $\boxed{W_1 \sim \chi_n^2}$ **y** $\boxed{W_2 \sim \chi_m^2}$ **y** W_1 **y** W_2 **son independientes, entonces**

$$\boxed{\frac{W_1 / n}{W_2 / m} \sim F_{n,m}}$$

V.E.Rohen

Si nuestro interés es sobre la medida de variación, tendremos que hacer uso de la expresión

$$\boxed{\frac{(n-1)S^2}{\sigma^2}}$$

donde S^2 es la varianza muestral.

Esta estadística tiene una distribución $\boxed{\chi^2_{n-1}}$

con $n-1$ grados de libertad

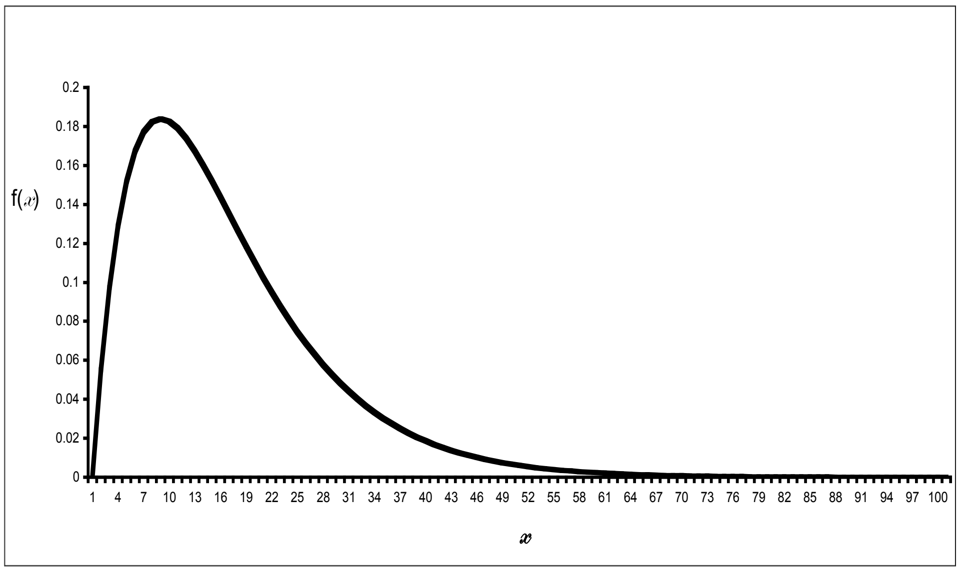
Distribución χ^2

Sesgo derecho

Un solo parámetro (grados de libertad)

Modela entre otras cosas a espacios continuos entre eventos discretos

Modela la distribución de la varianza muestral



V.E.Rohen

Cuando desconocemos la varianza poblacional, es preciso estimarla.

La expresión

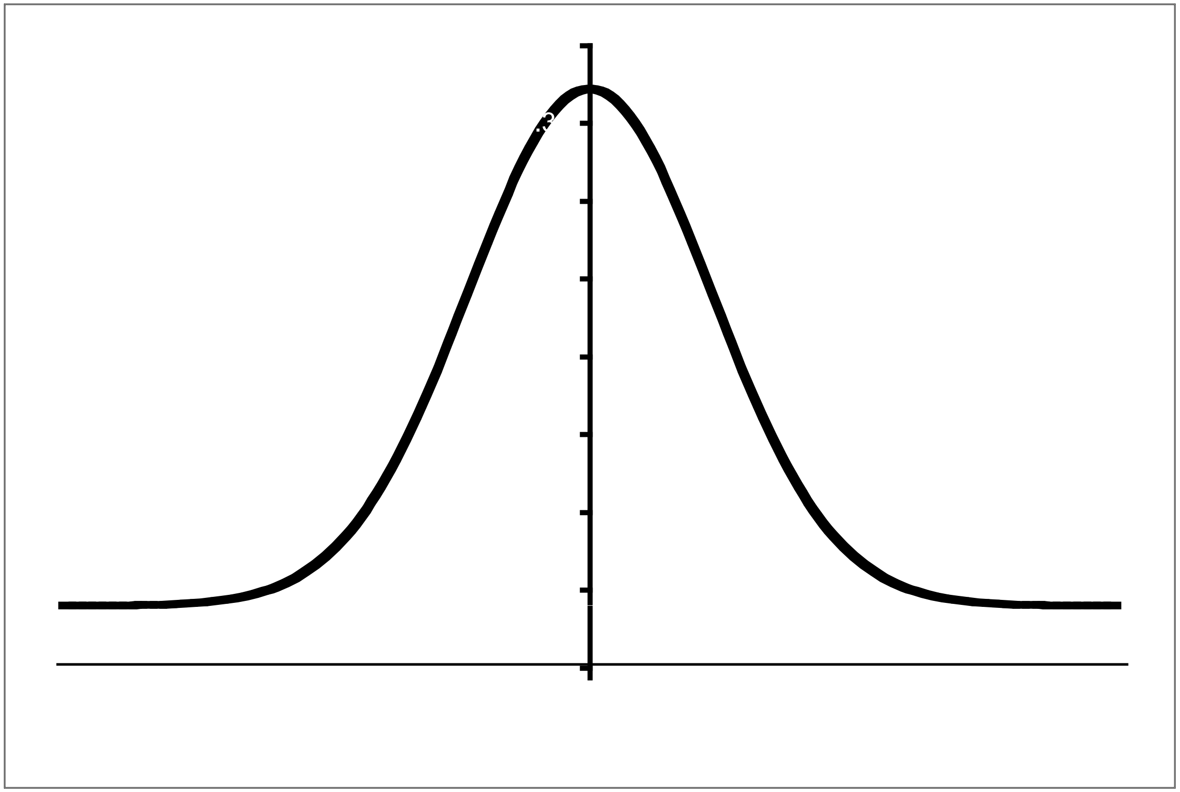
$$\boxed{Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}$$

tiene que ser sustituida por

$$\boxed{T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}}$$

Esta estadística tiene una distribución *t* con *n-1* grados de libertad

Distribución *t*- Student



Simétrica con respecto al cero

Un solo parámetro (grados de libertad)

Tiene las colas más pesadas que la normal

Cuando los grados de libertad aumentan converge a una normal estánadar

V.E.Rohen

La comparación de dos varianzas poblacionales se realiza por medio del cociente de las mismas.

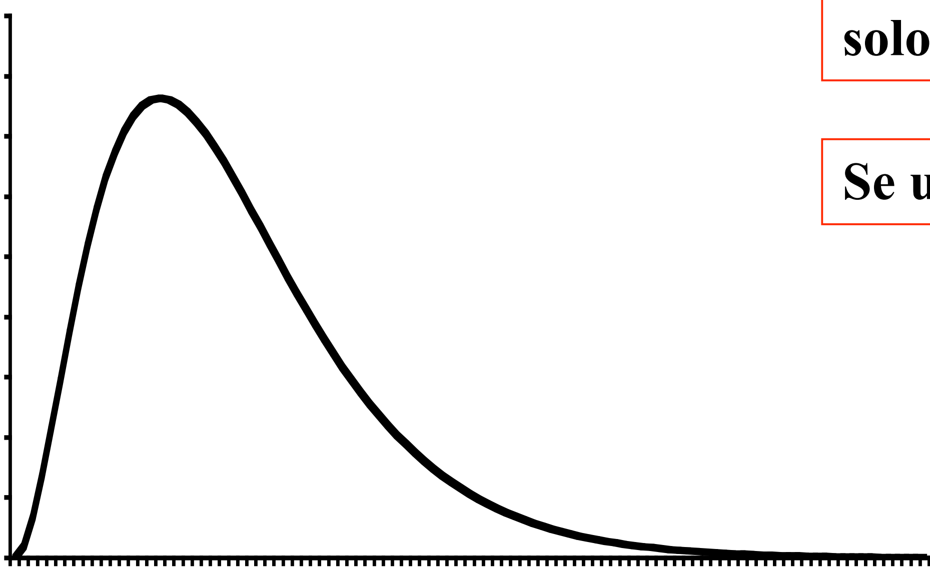
La estadística de prueba que involucra este cociente incluye las varianzas muestrales de la siguiente manera:

$$F = \frac{\left[\dfrac{(n_1 - 1)S_1^2}{\sigma_1^2}\right] \Big/ (n_1 - 1)}{\left[\dfrac{(n_2 - 1)S_2^2}{\sigma_2^2}\right] \Big/ (n_2 - 1)}$$

que tiene una distribución F con (n_1-1) y (n_2-1) grados de libertad

V.E.Rohen

Distribución *F*



Tiene una pareja de grados de libertad

Tiene sesgo derecho y toma solo valores positivos

Se usa para contrastar varianzas

V.E.Rohen

Existen dos tipos principales de estimadores:

Estimadores puntuales que consisten en un sólo valor o estadística muestral que se usa para estimar el verdadero valor del parámetro poblacional

$$\overline{X} = \frac{1}{n}\sum X_i \longrightarrow \mu$$

$$S^2 = \frac{\sum\left(X - \overline{X}\right)^2}{n-1} \longrightarrow \sigma^2$$

$$\frac{X}{n} \longrightarrow p$$

Estimadores por Intervalo que consiste en dos valores entre los cuales esperamos que se encuente el verdadero valor del parámetro

$$\boxed{\hat{\theta}_1 < \theta < \hat{\theta}_2}$$

donde $\hat{\theta}_1$ y $\hat{\theta}_2$ son función del estimador puntual de θ

Algunas propiedades deseables de los estimadores son las siguientes:

Que en promedio los estimadores sean igual al parámetro poblacional que estiman. Es decir, que el estimador sea Insesgado

Que tenga varianza mas pequeña que otros estimadores. A esta propiedad se le llama eficiencia.

Consistencia cuando la diferencia entre el estimador y el parámetro se hace mas pequeña conforme el tamaño de muestra crece.

V.E.Rohen

Cuando tratamos de evaluar la bondad de un estimador, tratamos de poner alguna cota en el error de estimación que pudiera ocurrir. Este error de estimación es $\left|\hat{\theta}-\theta\right|$**, y debe ser menor a** $k\left(\sigma_{\hat{\theta}}\right)$

donde k **es un factor que especifica los límites de confianza en la distribución de** $\hat{\theta}$ **(porcentiles de la Normal o de la t-Student:** $z_{\alpha/2}$ **ó** $t_{\alpha/2}$ **)**

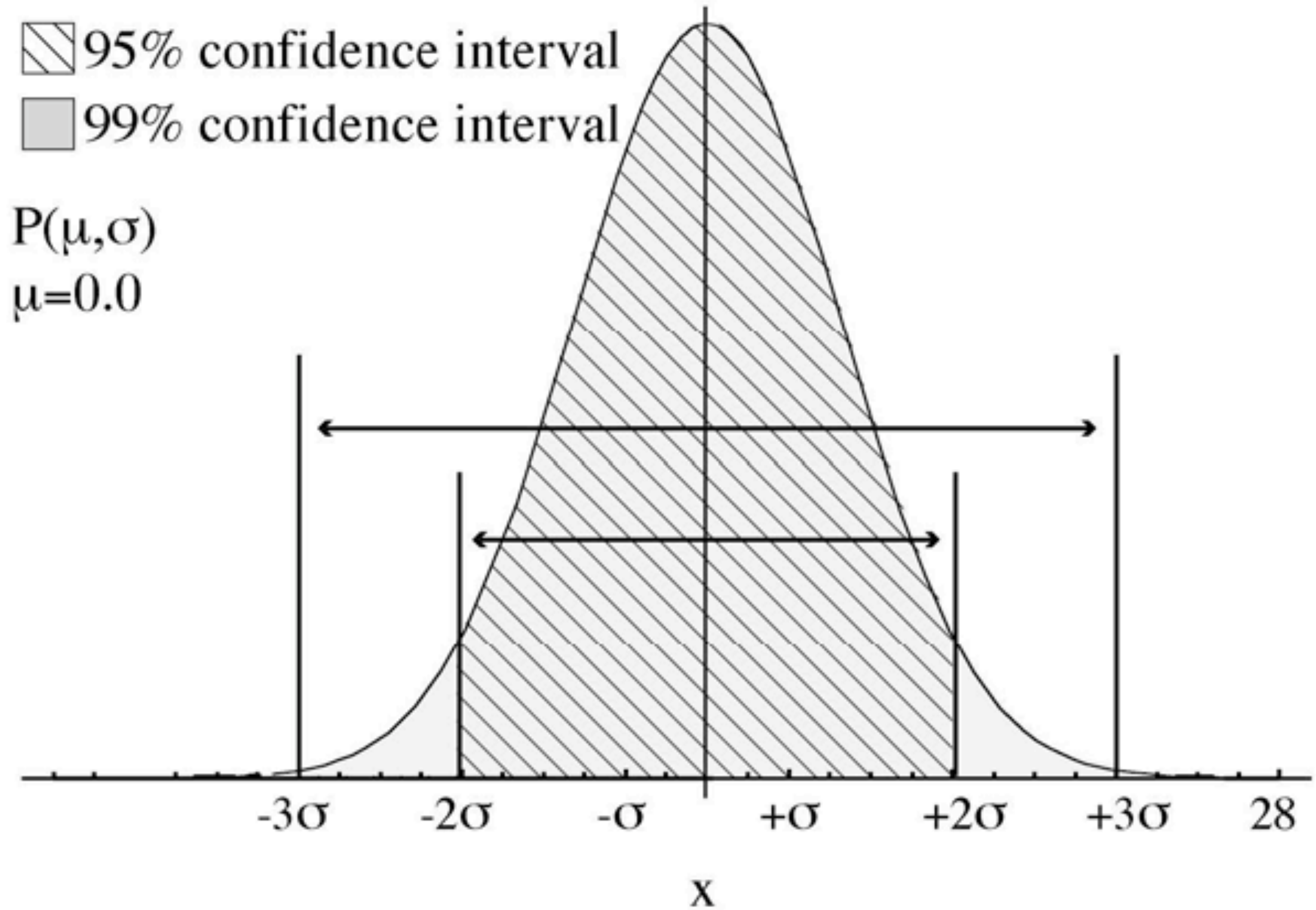
Si $\hat{\theta}$ **tiene una distribución Normal con media** θ **y varianza** $\sigma^2_{\hat{\theta}}$**, entonces** k **toma el valor 1.96 para un nivel de confianza** $(1-\alpha)$ **de 0.95 (ó 95%)**

La amplitud de un intervalo de confianza para la media poblacional depende de tres factores:

- el nivel de confianza

- la desviación estándar poblacional

- el tamaño de muestra.


95% confidence interval
99% confidence interval
P(μ,σ)
μ=0.0
-3σ
-2σ
-σ
+σ
+2σ
+3σ
28
x

V.E.Rohen

Propiedades que satisface un intervalo de confianza.

1. Mientras mayor sea el nivel de confianza $(1-\alpha)$, mayor será el valor de $Z_{\alpha/2}$ y más amplio será el intervalo de confianza, manteniendo constantes la varianza y el tamaño de muestra.

2. Mientras mas pequeña sea la desviación estándar, el intervalo será mas angosto.

3. Conforme el tamaño de muestra se incrementa, la amplitud del intervalo de confianza será menor.

V.E.Rohen

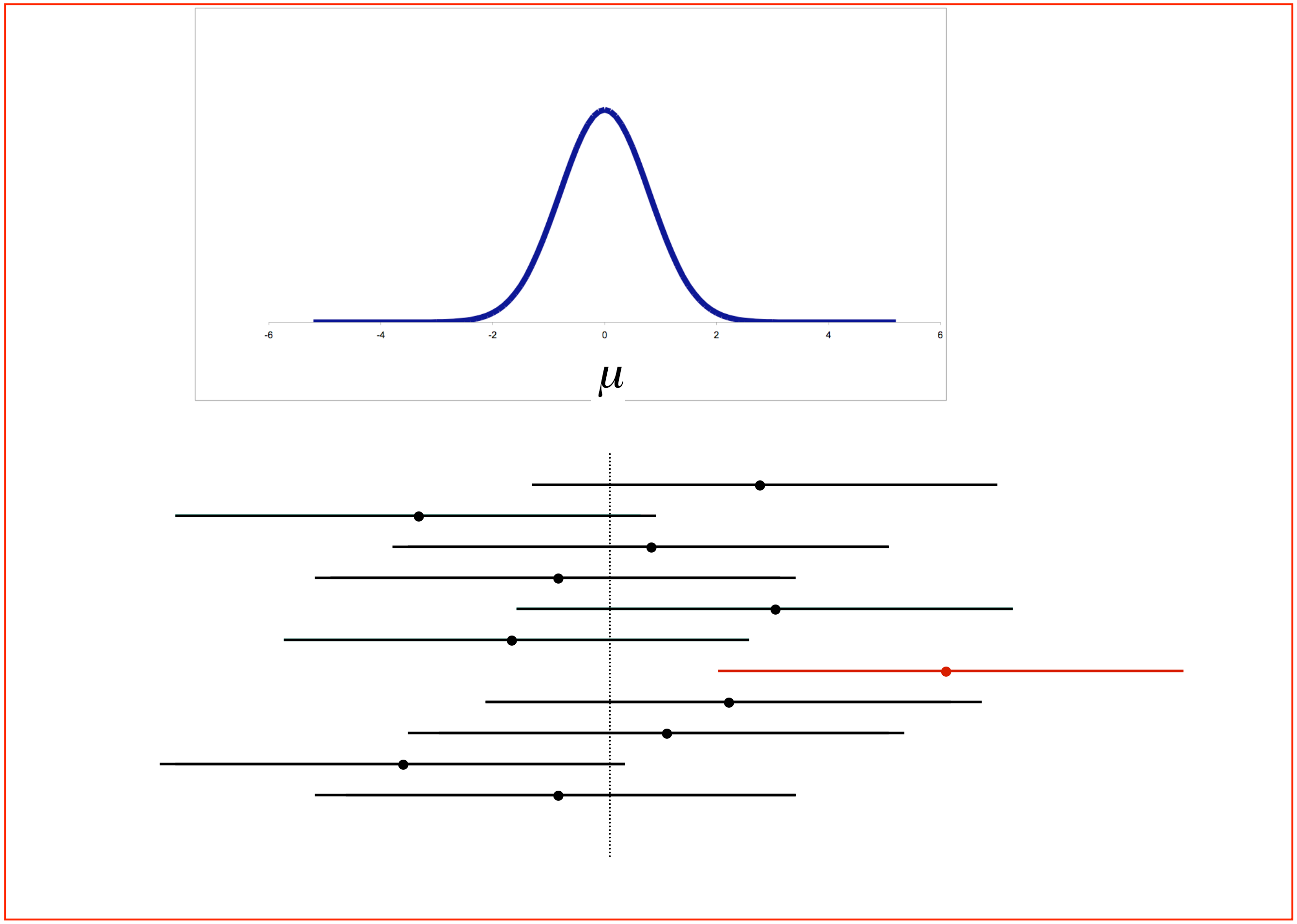
El valor α indica la proporción de veces que supondremos incorrectamente que el intervalo contiene el parámetro poblacional.

La interpretación del intervalo de confianza para μ es como sigue:

de una gran cantidad de intervalos que se construyan para el parámetro poblacional μ , 100(1–α)% contendrán a μ dentro de los límites encontrados.

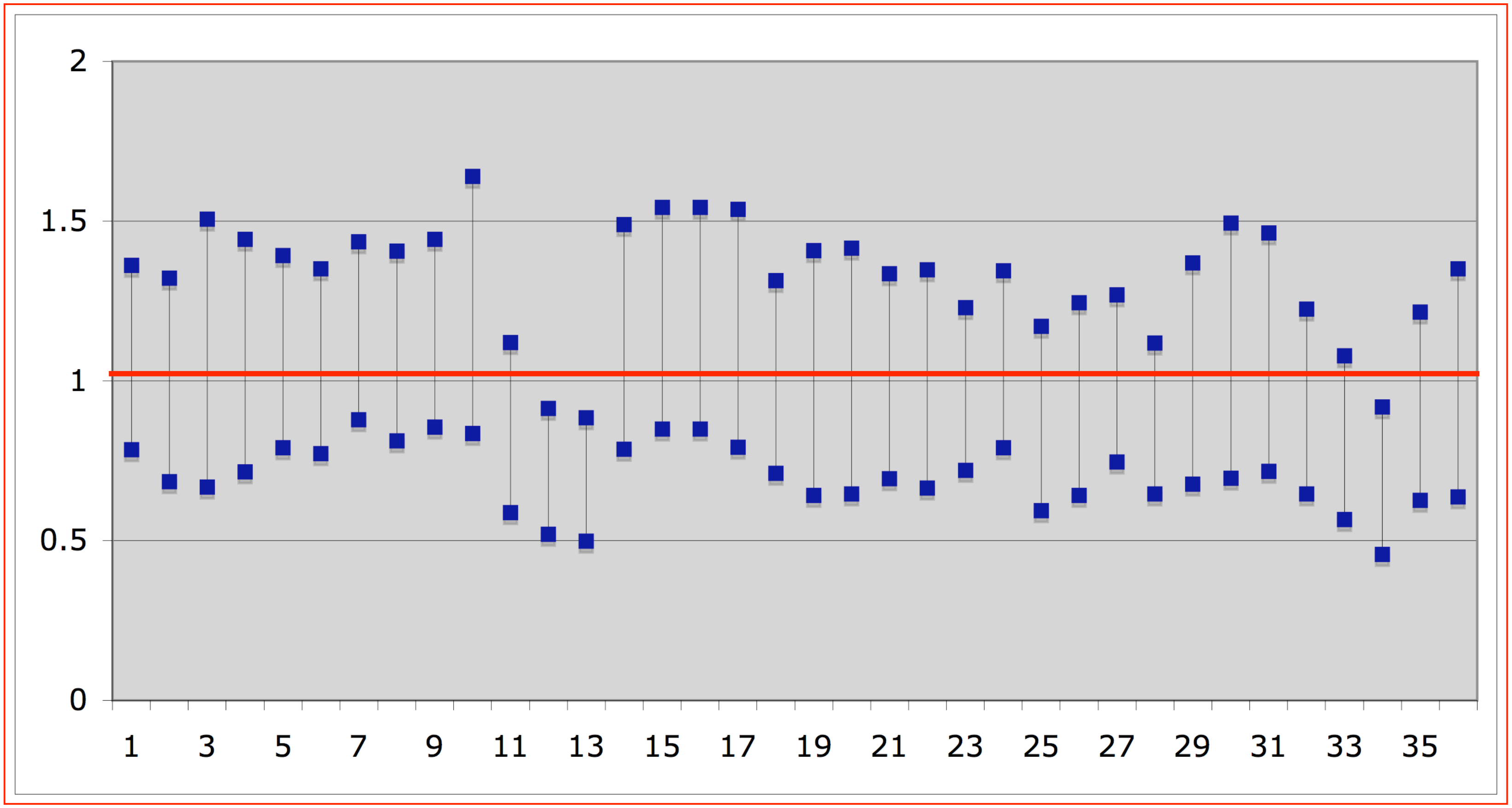
V.E.Rohen

(Intervalos de Confianza)



V.E.Rohen

Intervalos de confianza del 95% para el parámetro de una exponencial con media $\beta=1$



V.E.Rohen

Aclaremos que aunque no conozcamos el valor real de μ, éste es una cantidad fija y constante.

Puede suceder que μ se encuentre entre $\hat{\mu}_1$ y $\hat{\mu}_2$ pero también puede suceder que NO se encuentre entre esos dos valores, y sería incorrecto asignar una probabilidad a cualquiera de estas posibilidades, aún cuando μ permanezca desconocida

Así, un intervalo de confianza para μ del 100($1-\alpha$)% está dado por

$$\left(\overline{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \overline{X} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

cuando σ es conocida, pero si ésta es desconocida (casi siempre), se sustituye por su estimador puntual y el intervalo queda de la forma

$$\left(\overline{X} - t_{(\alpha/2),n-1} \frac{s}{\sqrt{n}}, \overline{X} + t_{(\alpha/2),n-1} \frac{s}{\sqrt{n}} \right)$$

si *n* es muy grande se puede aproximar la *t* por medio de la normal

Similarmente, un intervalo de confianza del 100($1-\alpha$)% para la proporción p de una población estará dado por

$$\boxed{\left(\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)}$$

donde $\hat{p} = \dfrac{X}{n}$

Siempre y cuando cuando $np > 5$ **y** $n(1-p) > 5$

De manera similar podemos construir intervalos de confianza para la varianza poblacional

Usaremos el hecho de que $\frac{(n-1)S^2}{\sigma^2}$ **tiene una distribución** χ^2_ν**,**

de donde es fácilmente verificable que el intervalo de confianza tiene la forma

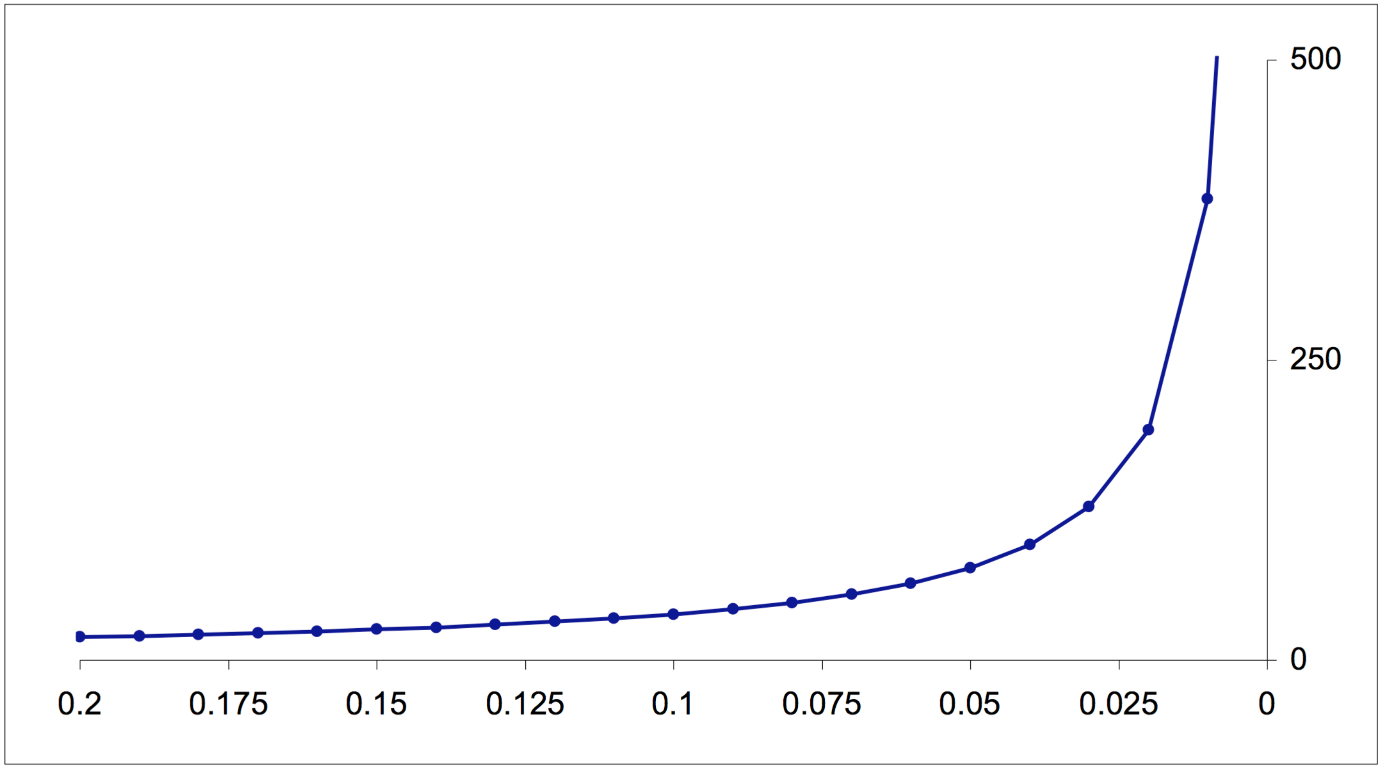
$$\boxed{\left(\frac{(n-1)S^2}{\chi^2_{(1-\alpha/2),n-1}}, \frac{(n-1)S^2}{\chi^2_{(\alpha/2),n-1}} \right)}$$

Tamaño de Muestra

Si queremos que nuestro error de estimación sea a lo más ε , entonces

$$\boxed{\varepsilon = Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)} \longrightarrow \boxed{n = Z_{\alpha/2}^{2}\frac{\sigma^{2}}{\varepsilon^{2}}}$$

Para un nivel de cofianza fijo, un tamaño de error pequeño incrementará el tamaño de muestra.



Aumento del tamaño de muestra para un nivel de confianza del 95%, y una varianza de 1, cuando el error de estimación disminuye.

V.E.Rohen

Referencias:

http://www.hrc.es/bioest/M_docente.html

Zar, Jerrold H.- Biostatistical Analysis.- 4rd ed.- Prentice Hall, Inc

Rosner, B.- Fundamentals of Biostatistics. 6th Ed. Brooks/Cole Publishing Co., 2006

V.E.Rohen