¿Podemos conocer el comportamiento del ser humano?

V.E.Rohen

La Probabilidad como Pronóstico

Ya hemos dicho que la probabilidad es una medida de incertidumbre, y esa medida la podemos usar para pronosticar un valor futuro de alguna variable aleatoria o para predecir el comportamiento de ésta bajo circunstancias específicas.

Podemos entonces 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.

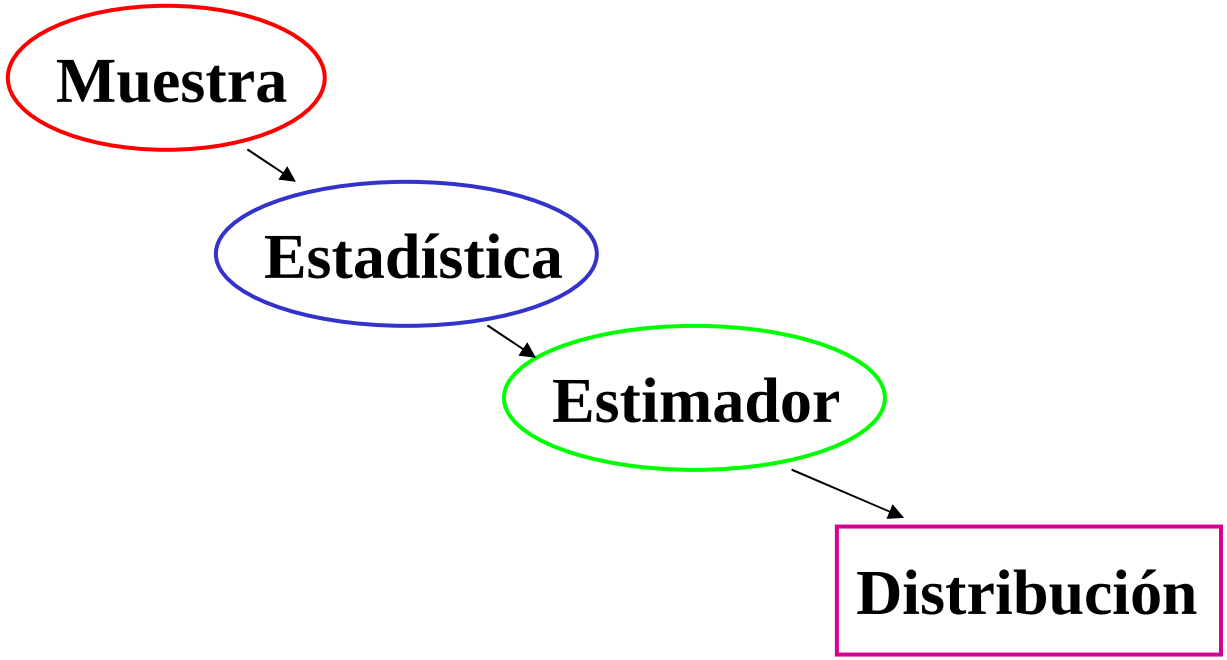
Sin embargo, es una buena aproximación y la podemos utilizar para la estimación de las características propias de dicha población.

Podemos dar además una medida de dicha incertidumbre, es decir la probabilidad de equivocarnos al hacer dicha estimación:

p – value

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

Teoría de Muestreo

Repasemos algunos conceptos:

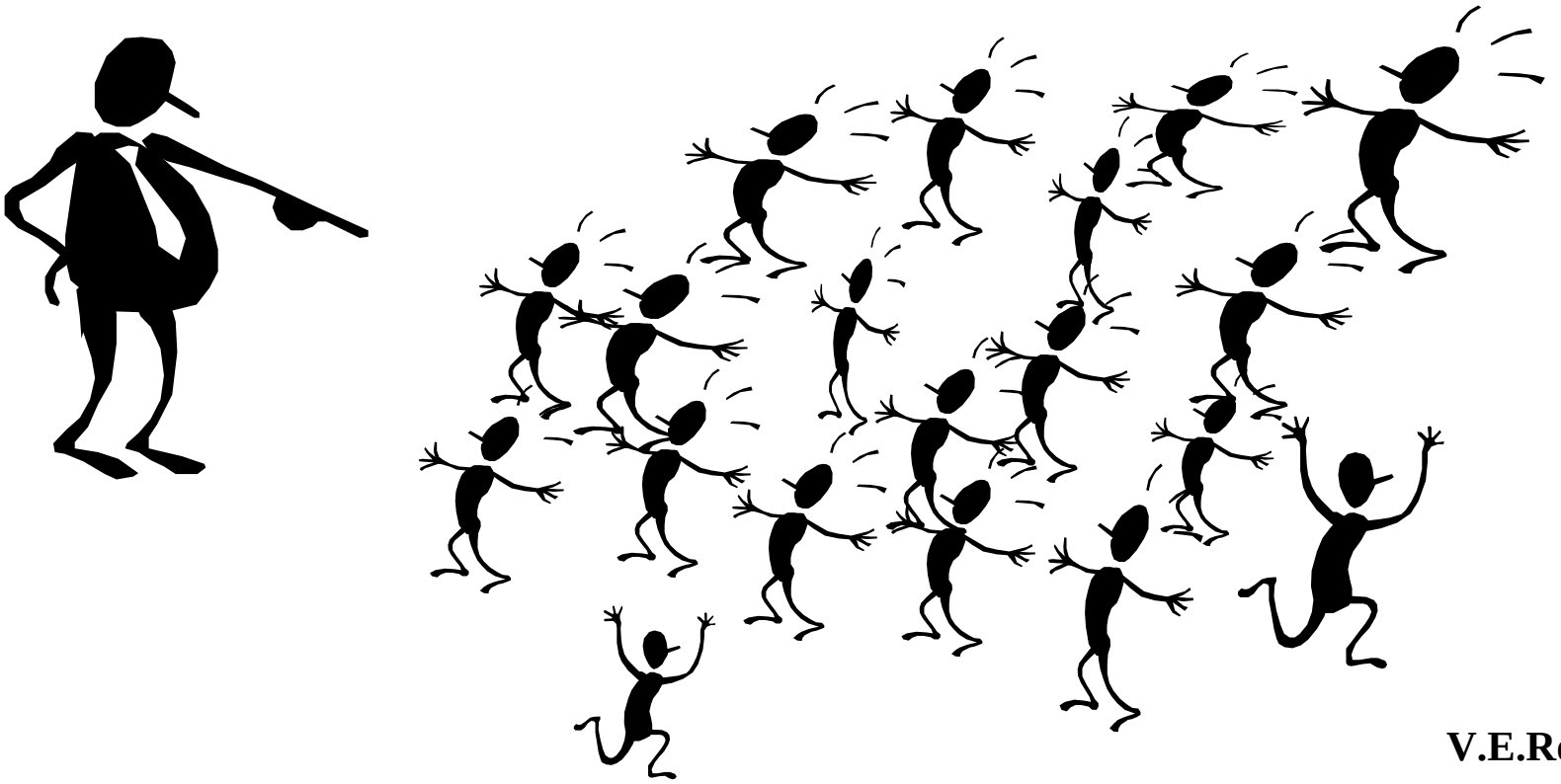
Una población consiste de todas las posibles observaciones de un fenómeno dado.

Una muestra es una parte de la población.

- Cada unidad tiene la misma oportunidad de ser elegida

- La selección de una unidad no tiene influencia sobre la elección de otra unidad

Muestreo Aleatorio



V.E.Rohen

Razones para muestrear:

- Algunas poblaciones son muy grandes y no pueden ser examinadas en su totalidad.

- Puede resultar muy caro censar la población.

- Puede llevar mucho tiempo hacerlo.

- Se puede destruir el objeto examinado.

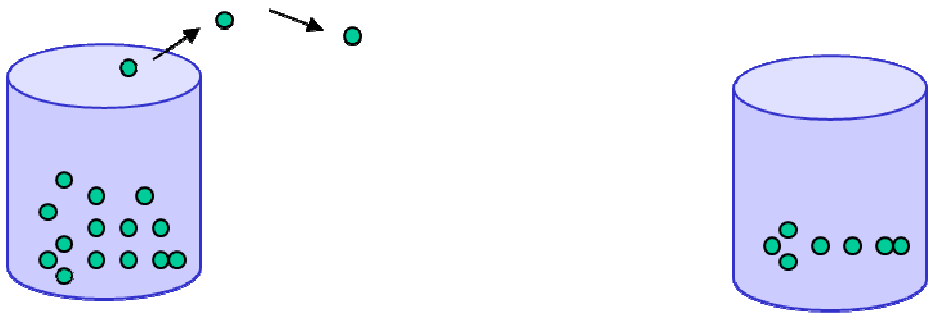
- Es mas seguro tomar una muestra valiéndonos de una persona apta para manejar información que tomar un censo valiéndonos de personas no aptas para el propósito.

V.E.Rohen

En un muestro con reemplazo el individuo observado puede volver a observarse, y la probabilidad de seleccionar a un objeto en especial no cambia y la selección es independiente de las selecciones anteriores

V.E.Rohen

En un muestreo sin reemplazo el individuo observado no puede volver a tomar parte en la selección, y la probabilidad de seleccionar algún otro individuo se ve afectada por la elección de los anteriores al disminuir el tamaño de la población de donde se hace la selección



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"

V.E.Rohen

Un estimador es una función de la información contenida en la muestra

Una estimación es un valor particular del estimador basada en una muestra particular

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

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

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

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.

Supongamos que tomamos una muestra sin reemplazo de tamaño 3 de una población de tamaño 6, cuyo valores son {*1,2,3,4,5,6* }. Tenemos entonces 20 posibles muestras.

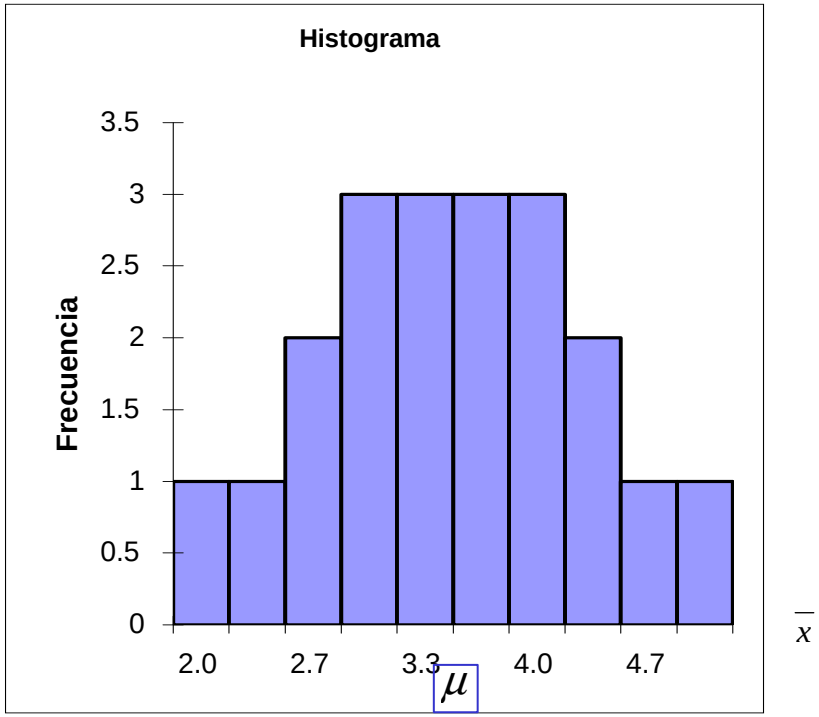
La media poblacional es $\mu = \frac{1}{6}\sum X_i = 3.5$

Si obtenemos el promedio de los números obtenidos en cada una de las 20 muestras obtenemos los siguientes resultados:

Muestra			$\overline{X}$	Muestra			$\overline{X}$
1	2	3	**2.00**	2	3	4	**3.00**
1	2	4	**2.33**	2	3	5	**3.33**
1	2	5	**2.67**	2	3	6	**3.67**
1	2	6	**3.00**	2	4	5	**3.67**
1	3	4	**2.67**	2	4	6	**4.00**
1	3	5	**3.00**	2	5	6	**4.33**
1	3	6	**3.33**	3	4	5	**4.00**
1	4	5	**3.33**	3	4	6	**4.33**
1	4	6	**3.67**	3	5	6	**4.67**
1	5	6	**4.00**	4	5	6	**5.00**

LGN

Si realizamos el histograma de frecuencias vemos que los promedios están alrededor de la media poblacional $\mu = 3.5$



V.E.Rohen

Esto quiere decir que el promedio de todas las medias muestrales posibles 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 el 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

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

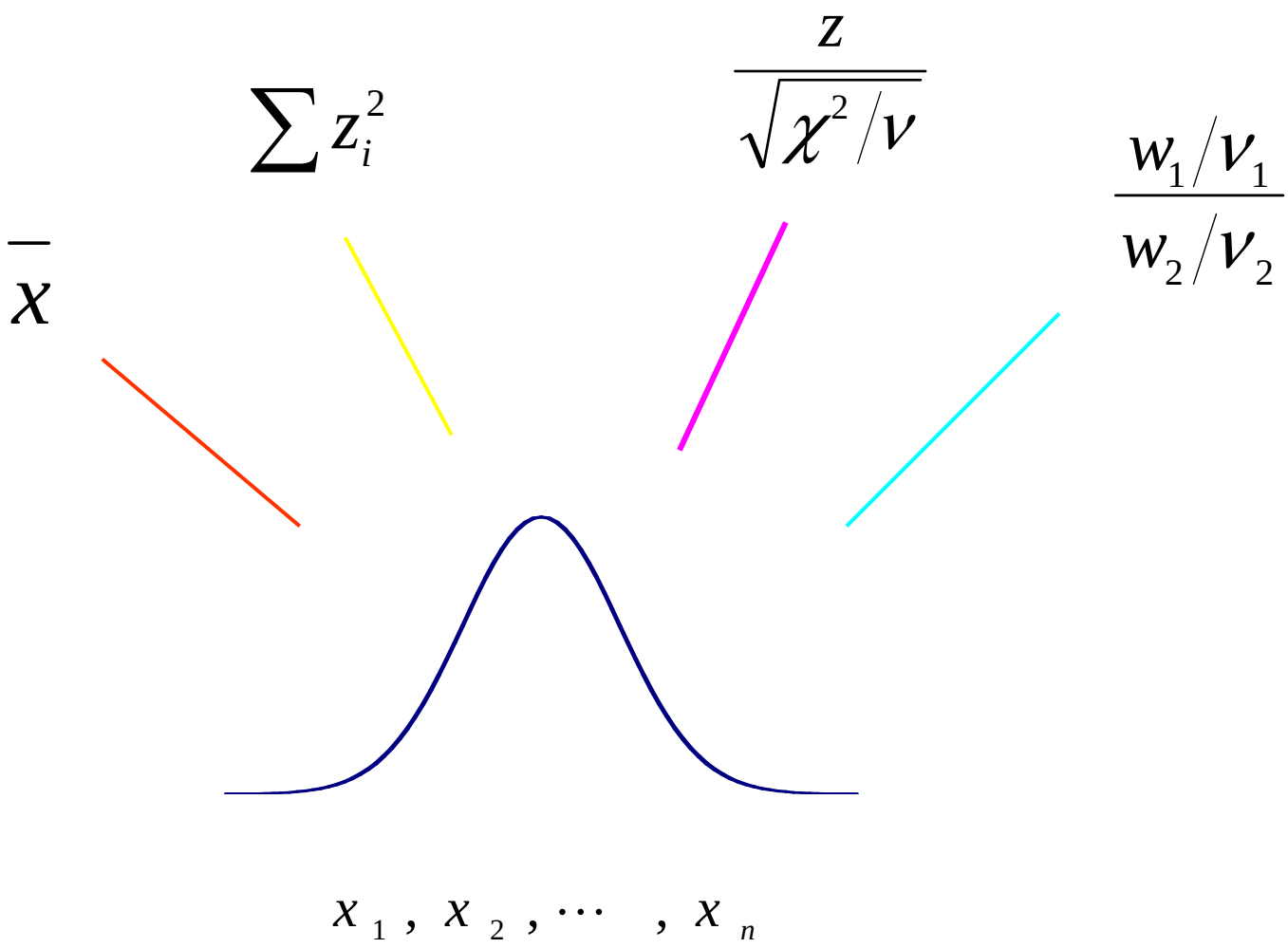
Se puede observar que si el tamaño de la muestra aumenta, el error estándar disminuye.

¿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.

TCL



Distribuciones de Muestreo

V.E.Rohen

DISTRIBUCIÓN DE LA MEDIA MUESTRAL

$\overline{X}$

Recordemos que la media muestral 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

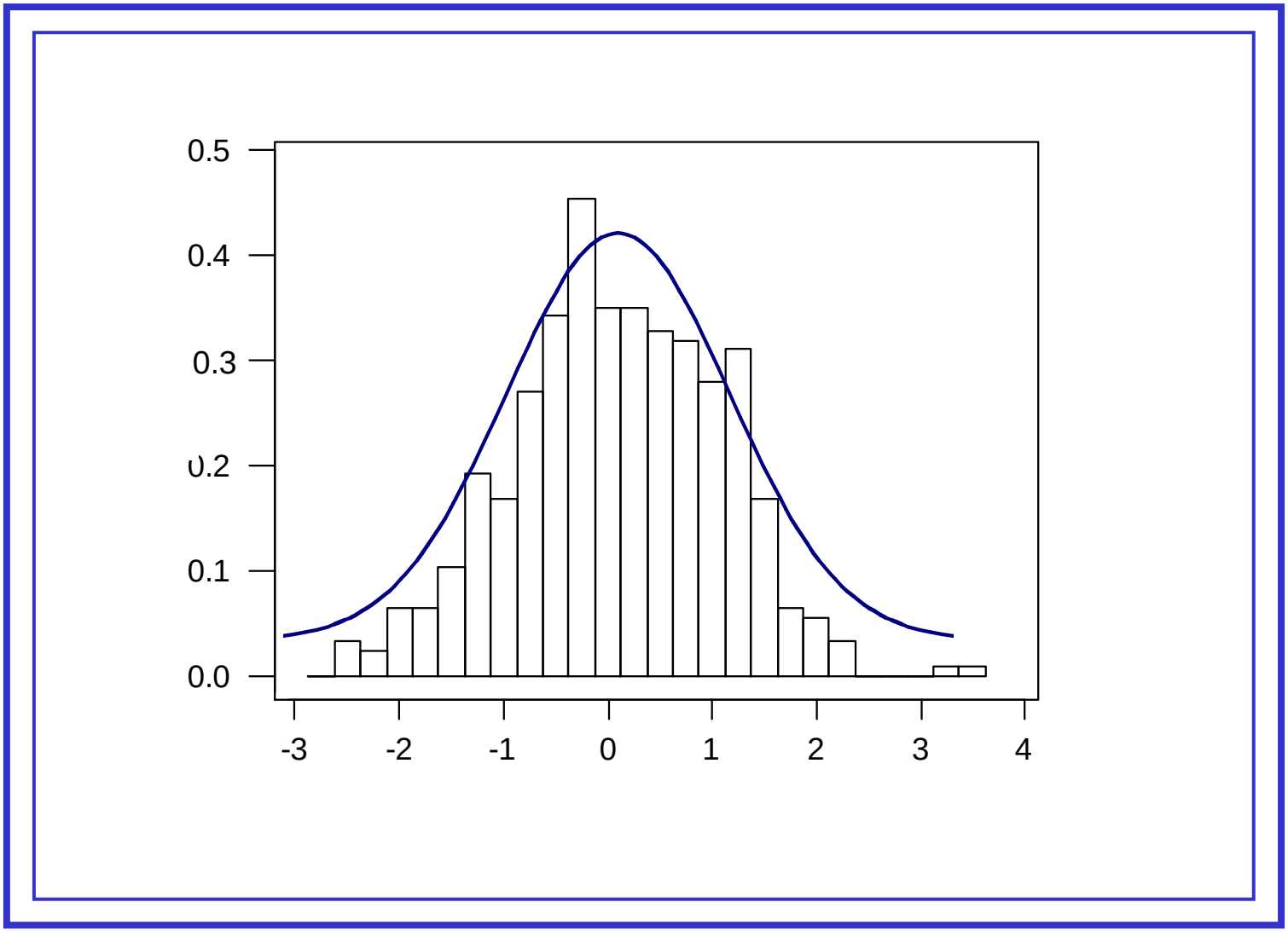
$\bar{x}$

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

DISTRIBUCIÓN DE LA DIFERENCIA DE MEDIAS MUESTRALES

$$\overline{X}_1 - \overline{X}_2$$

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 $\overline{X}$ de cada una de las poblaciones y observamos la diferencia de las medias muestrales $\overline{X}_1$ y $\overline{X}_2$.

Como cada $\overline{X}_i$ es una variable aleatoria normalmente distribuida, entonces $\overline{X}_1 - \overline{X}_2$ es también una variable aleatoria normalmente distribuida con media $\mu_1 - \mu_2$ y con varianza $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

DISTRIBUCIÓN DE LA PROPORCIÓN MUESTRAL

$$\hat{p} = \frac{X}{n}$$

V.E.Rohen

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, la proporción de éxitos en esta muestra puede ser estimada como

$$\hat{p} = \frac{X}{n}$$

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)***

DISTRIBUCIÓN DE LA DIFERENCIA DE PROPORCIONES MUESTRALES

$$\hat{p}_1 - \hat{p}_2$$

V.E.Rohen

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 $\hat{p}_1 - \hat{p}_2$

es aproximadamente normal con media $p_1 - p_2$

y varianza

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

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

V.E.Rohen

Algunas distribuciones que se derivan de la distribución normal

Si $Z \sim N(0,1)$ **entonces** $Z \sim \chi^2_1$

Si $Z_i \sim N(0,1)$ **para** $i=1,\ldots,n$, **entonces** $\sum_{i=1}^{n} Z_i^2 \sim \chi^2_n$

Si $Z \sim N(0,1)$ **,** $W \sim \chi^2_n$ **y** Z **y** W **son independientes, entonces**

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

Si $W_1 \sim \chi^2_n$ **y** $W_2 \sim \chi^2_m$ **y** W_1 **y** W_2 **son independientes, entonces**

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

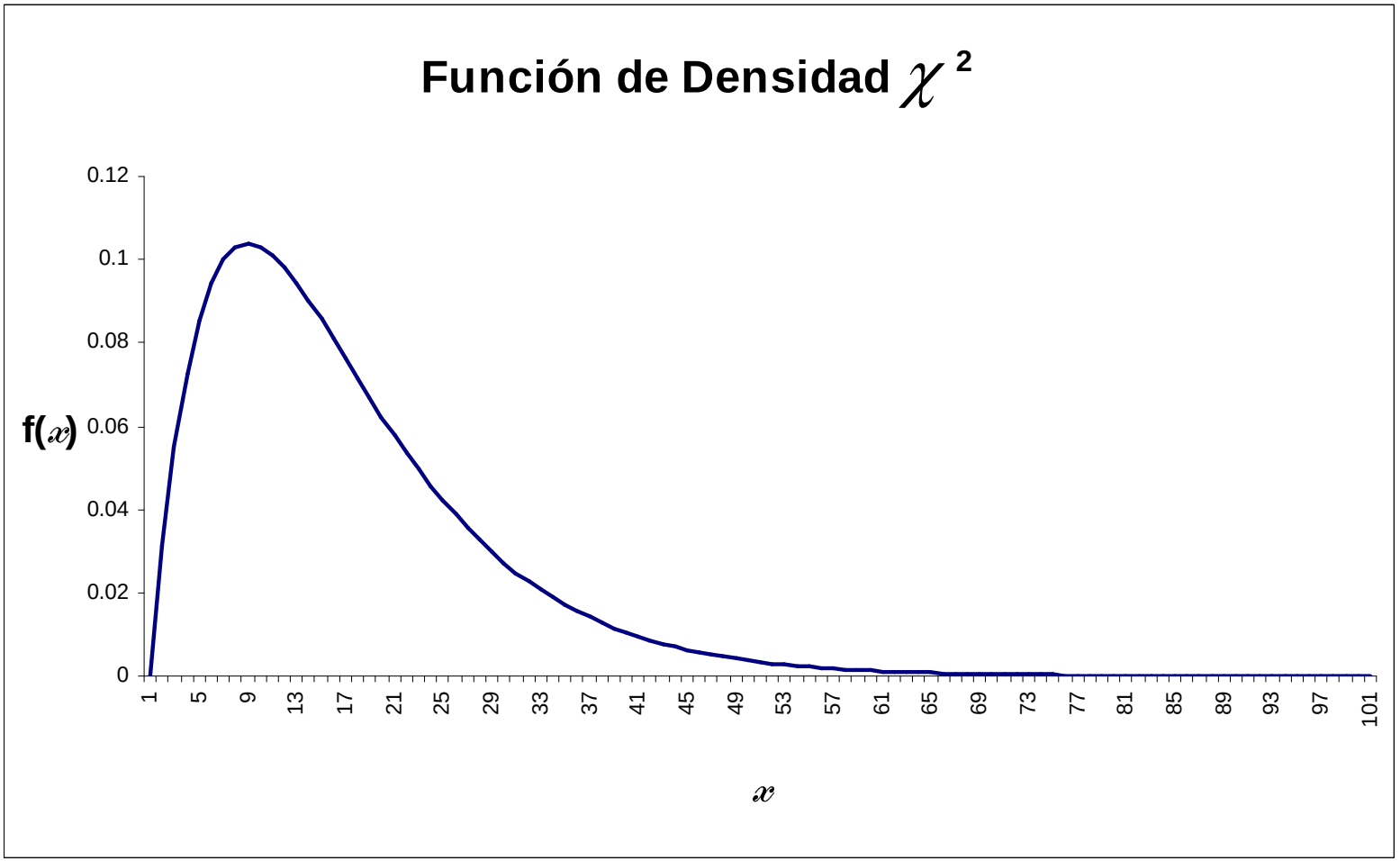
Si nuestro interés es sobre la medida de variación, tendremos que hacer uso de la expresión

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

donde S^2 es la varianza muestral.

Esta estadística tiene una distribución

χ^2_{n-1} **con n-1 grados de libertad**


Función de Densidad χ^2
0.12
0.1
0.08
0.06
0.04
0.02
0
f(x)
1
5
9
13
17
21
25
29
33
37
41
45
49
53
57
61
65
69
73
77
81
85
89
93
97
101
x

V.E.Rohen

Cuando desconocemos la varianza poblacional, es preciso estimarla.

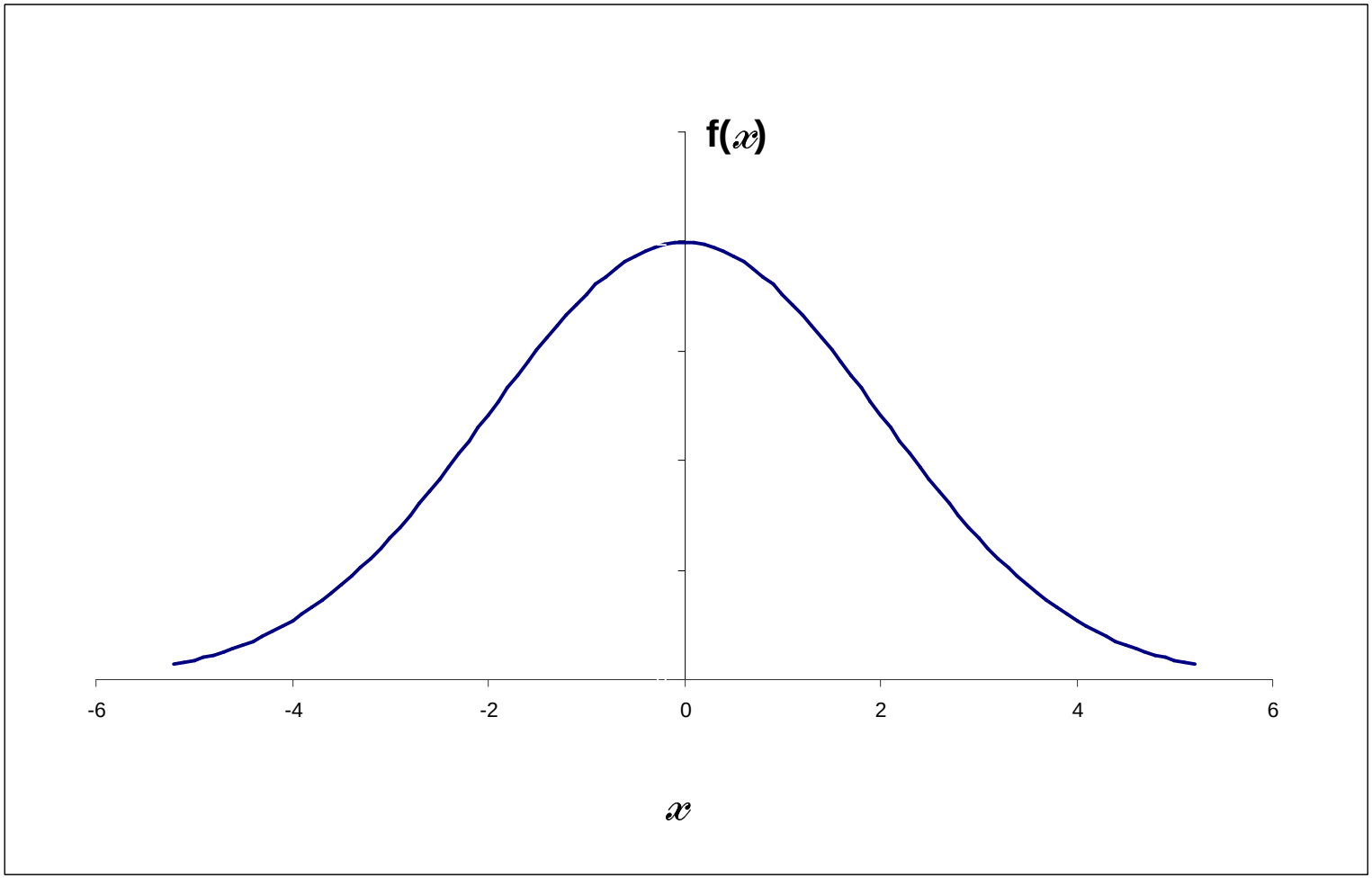
La expresión $$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

tiene que ser sustituida por $$T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$

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

V.E.Rohen

Distribución *t*- Student



V.E.Rohen

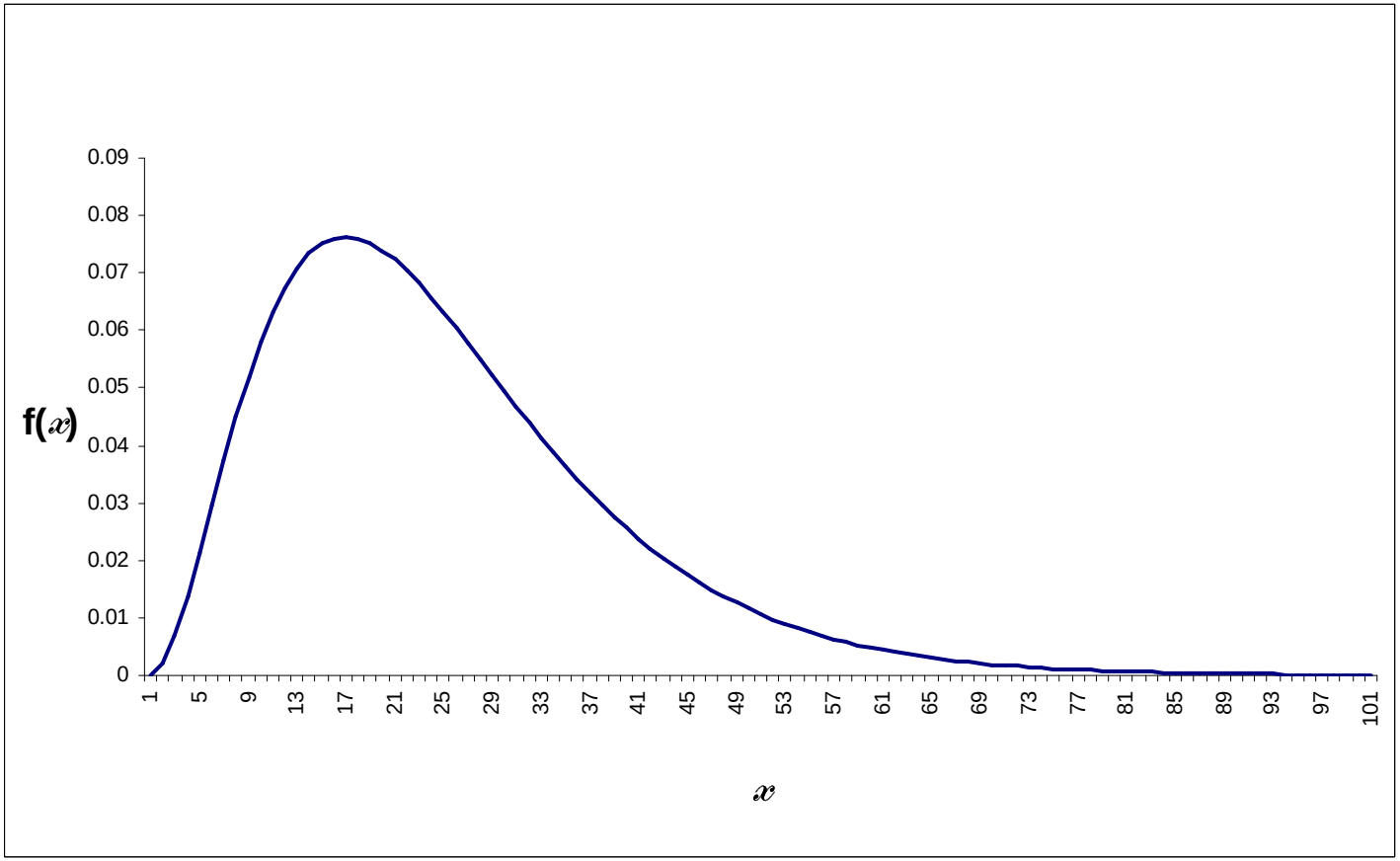
La comparación de dos varianzas poblacionales se realiza por medio del cociente de las mismas.

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

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

que tiene una distribución *F* con $(n_1\text{-}1)$ y $(n_2\text{-}1)$ grados de libertad

Distribución *F*



V.E.Rohen