Unidad 1

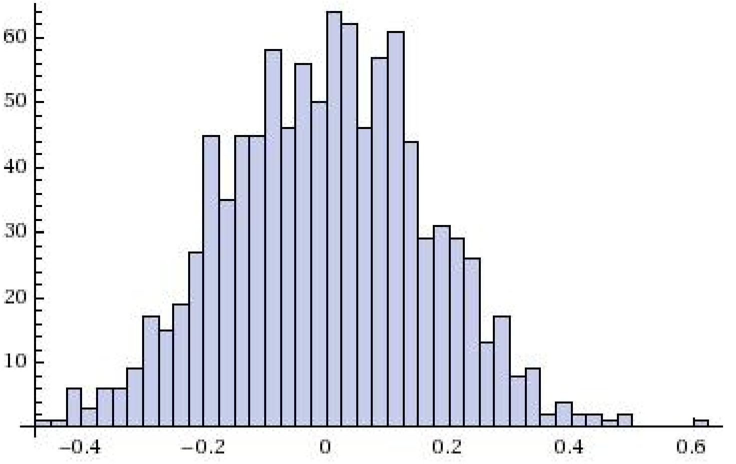
a. Probabilidades y Estadística

CLASE N° 6

IN3401

SEMESTRE OTOÑO, 2012

Teoremas sobre Límites



Desigualdad de Chebychev

- Sean $X_1, X_2, \ldots, X_n$ v.a. i.i.d. (una *muestra aleatoria*) de una cierta distribución.

- Sea $T : IR^n \to IR$. Entonces: $\forall \varepsilon > 0$

$$IP(|T(X_1, X_2, \ldots, X_n) - IE(T(X_1, X_2, \ldots, X_n))| \geq \varepsilon) \leq \frac{V(T(X_1, X_2, \ldots, X_n))}{\varepsilon^2}$$

- En particular si $T(X_1, X_2, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i \equiv \overline{X}_n$ (la *media muestral*), entonces:

$$IP(|\overline{X}_n - IE(X)| \geq \varepsilon) \leq \frac{V(X)}{n\varepsilon^2}$$

Ley de los Grandes Números

- Sean $X_1, X_2, \ldots, X_n$ v.a. i.i.d. (una *muestra aleatoria*).

- La media muestral converge casi seguramente a la media poblacional:

$$\overline{X}_n \xrightarrow[n\to\infty]{c.s.} IE(X)$$

- Este resultado es la base para la estimación puntual de parámetros poblacionales

Teorema Central del Límite

- TCL: indica que, en condiciones muy generales, la distribución de la suma de v.a. tiende a una distribución normal cuando la cantidad de variables es muy grande.

- Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución con media μ y varianza σ^2. Entonces, si n grande, la v.a.:

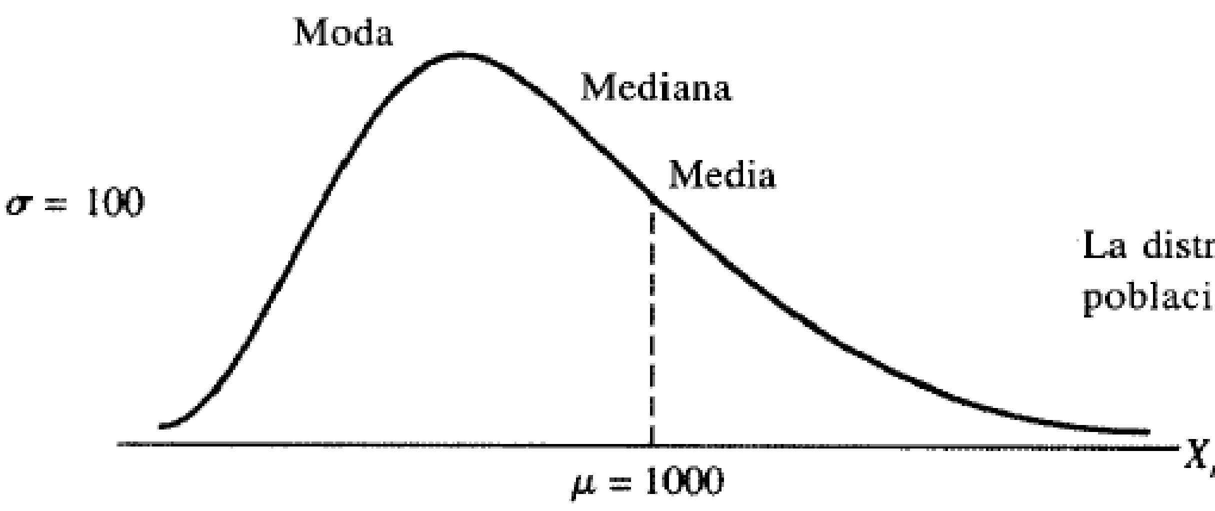
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

tiene aproximadamente una distribución normal con $\mu_{\bar{X}} = \mu$ y $\sigma^2_{\bar{X}} = \sigma^2/n$

Teorema Central del Límite(2)

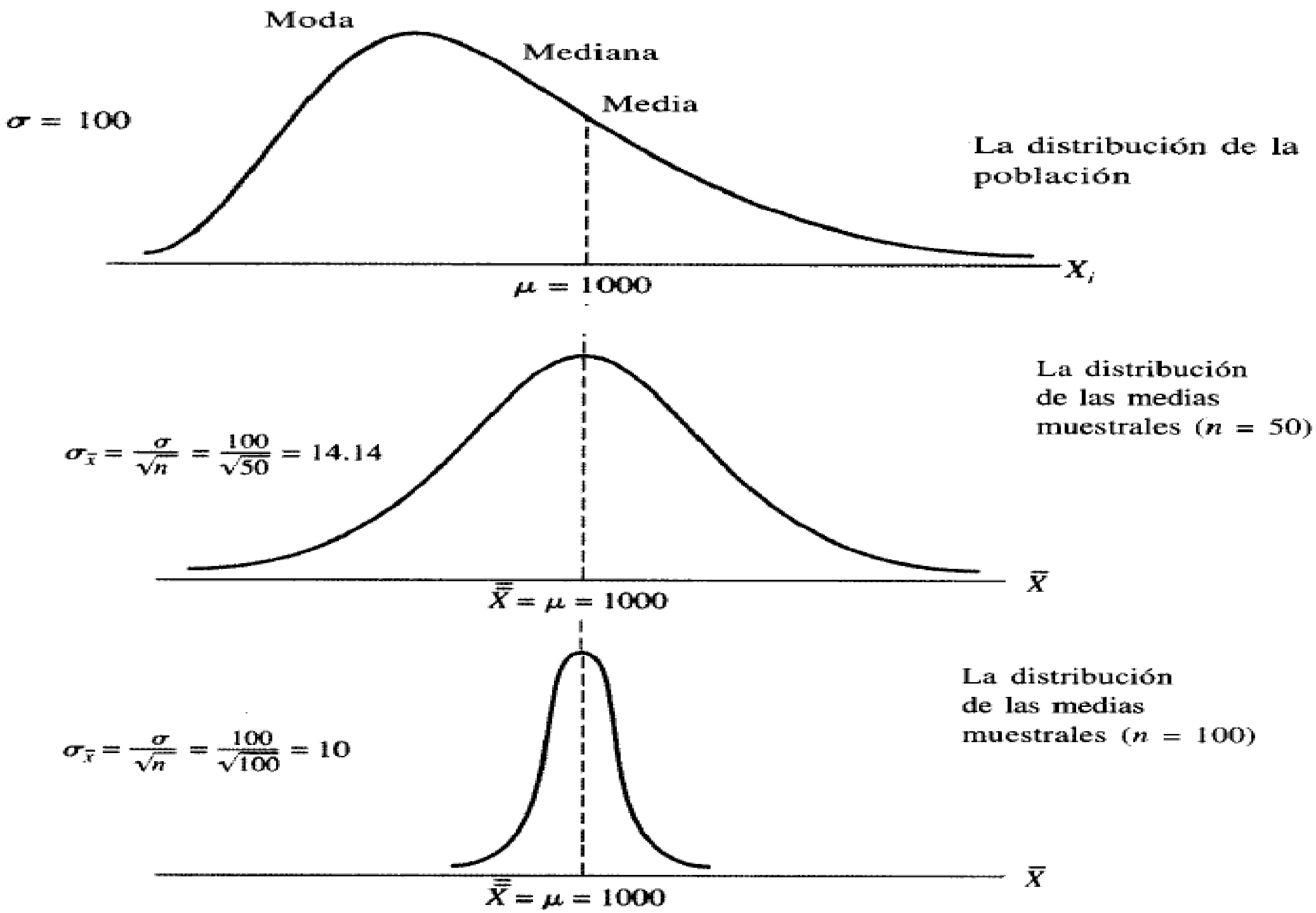
- En particular esto se cumple para la distribución de medias muestrales, aún cuando la distribución original no es normal

- Ej:



Teorema Central del Límite(2)

- Ej:



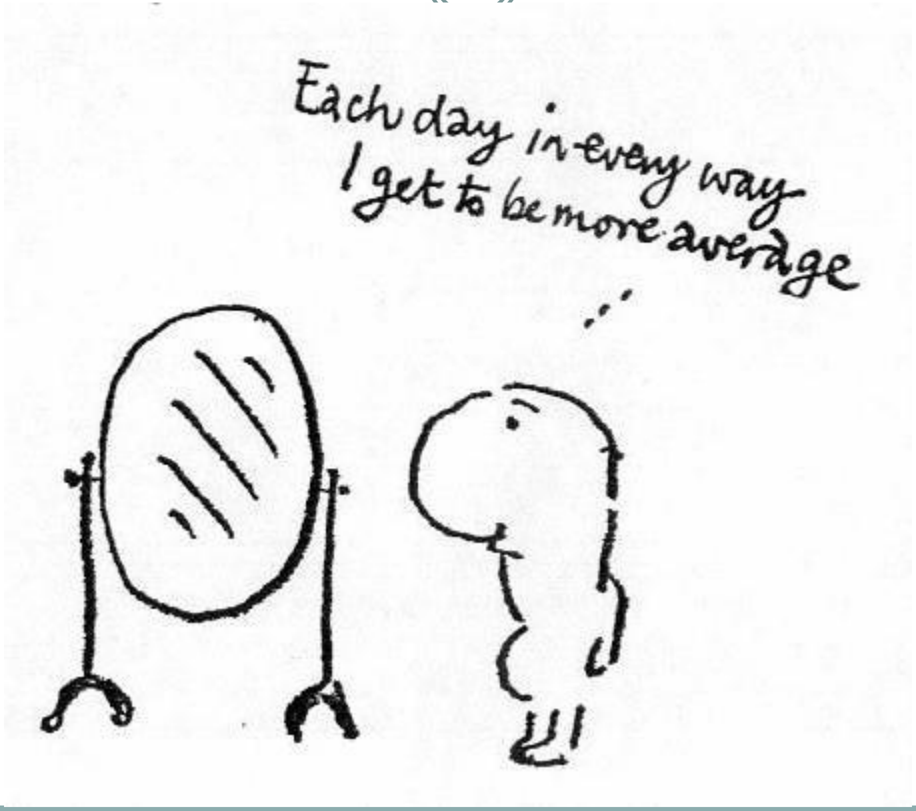
Teorema Central del Límite(3)

- El TCL permite el uso de la transformación Z para distribuciones muestrales:

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

- Esta fórmula es válida cuando la desviación estándar **poblacional** es conocida.

Inferencia Estadística



Estimación puntual

- Sean $X_1, X_2, \ldots, X_n$ v.a. i.i.d. (una *muestra aleatoria*) de una cierta distribución.
- Supongamos que la distribución contiene un parámetro θ
- Un *estimador puntual* de θ es una función

$$\theta^* = \theta^*(X_1, \ldots, X_n)$$

que se usa para aproximar (estimar) θ .

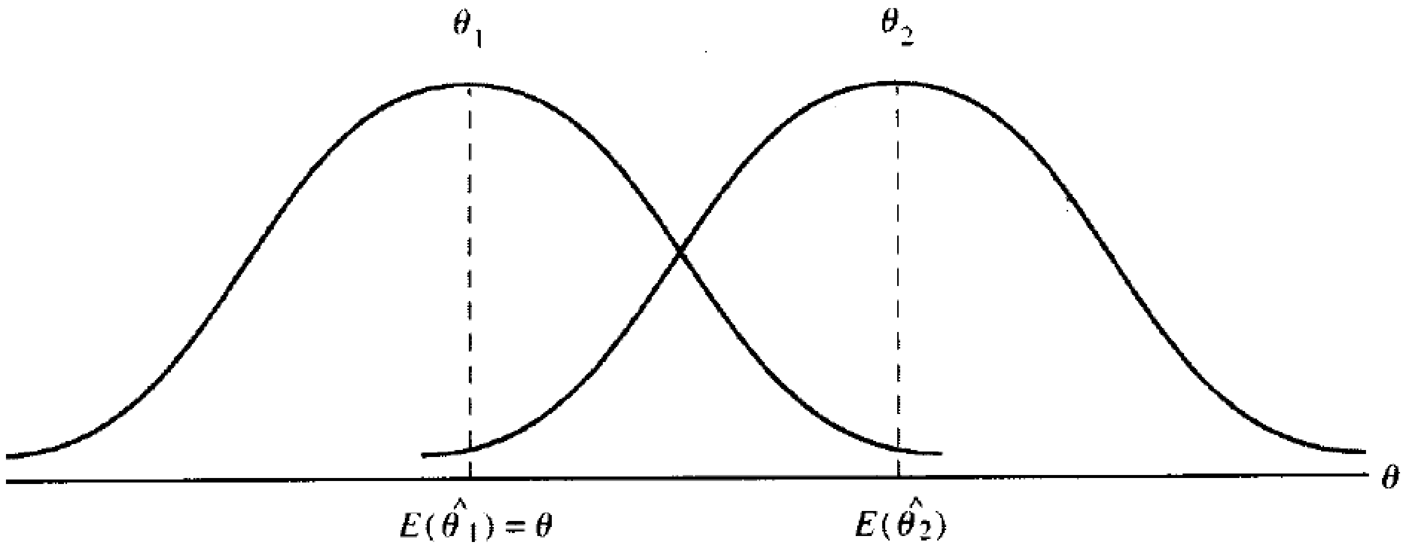
- Entre los métodos de estimación puntual más conocidos (y útiles) se encuentran: el Método de Máxima Verosimilitud y el Método de los Momentos

Propiedades de un buen estimador

- **Insesgadez**: un estimador es insesgado si la media de su distribución muestral es igual al parámetro en estudio.

- Sea $\hat{\theta}$ un estimador de θ . $\hat{\theta}$ Es insesgado ssi:

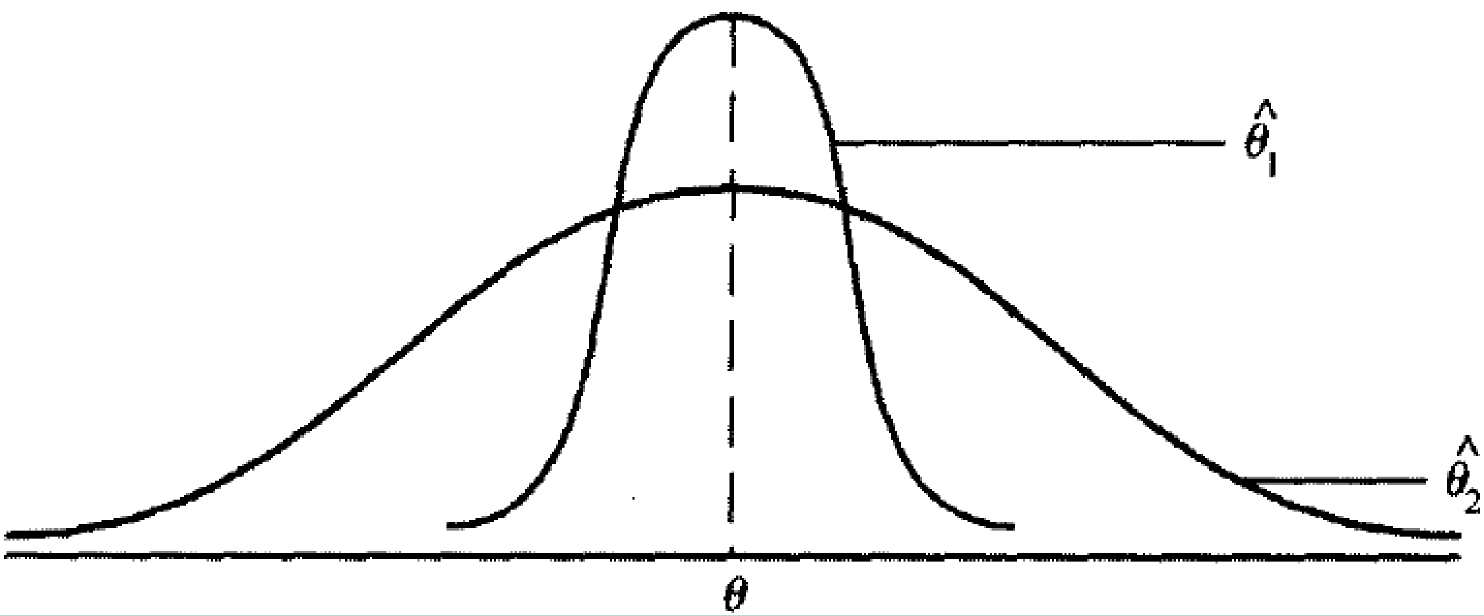
$$E(\hat{\theta}) = \theta$$



Propiedades de un buen estimador(2)

- **Eficiencia**: de todo estimador insesgado, el estimador más eficiente es aquel que tenga menor varianza:

- Sean 2 estimadores insesgados ($E(\hat{\theta}_1) = E(\hat{\theta}_2) = \theta$)



Propiedades de un buen estimador(2)

- **Consistencia**: un estimador es consistente si, a medida que n aumenta, el valor del estadístico se aproxima al parámetro

- Para que un estimador sea consistente debe ser insesgado y su varianza debe aproximarse a cero a medida que la muestra aumente.

- Por ejemplo, la varianza de la media muestral $\sigma_{\bar{X}}^2 = \sigma^2/n$ se aproxima a cero cuando n aumenta, y como $E(\bar{X}) = \mu$ entonces $\bar{X}$ es un estimador consistente.

Intervalos de Confianza

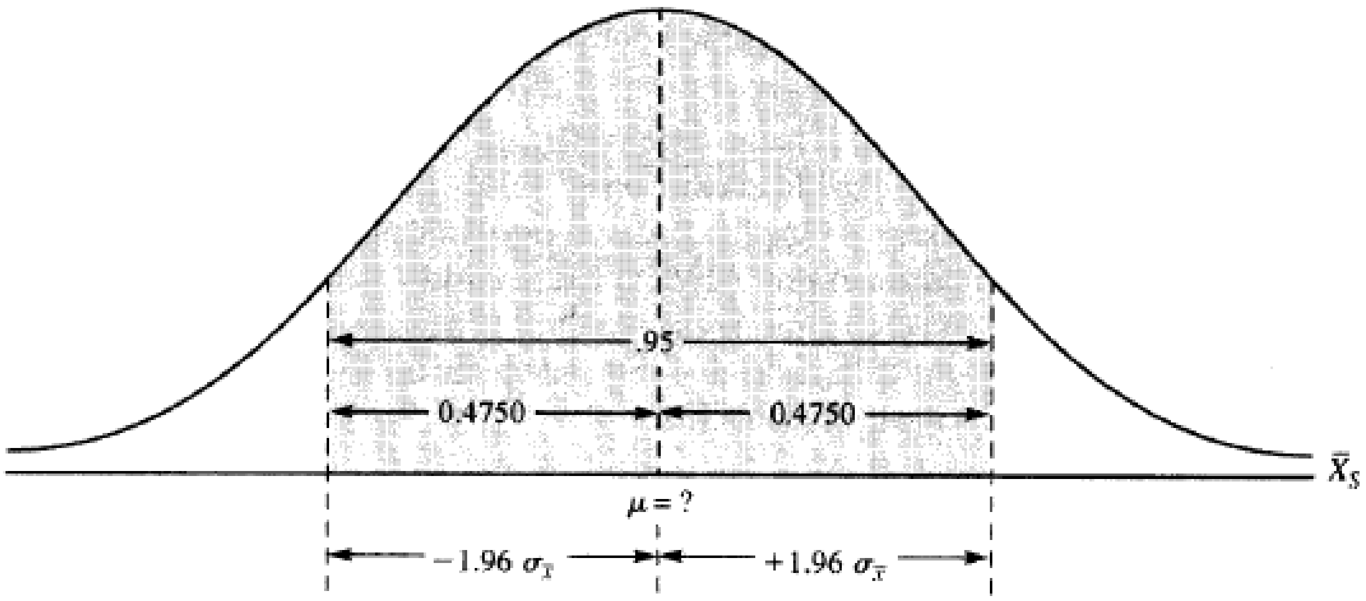
- Un intervalo de confianza denota un rango dentro del cual puede encontrarse un parámetro, y el nivel de confianza que el intervalo considera del parámetro.

- Los niveles de confianza más usados son 90%, 95%, 99%. Se denomina α a la probabilidad de que el intervalo NO contenga el parámetro desconocido.

- Si se desea construir un intervalo para la media muestral de 95%, se debe dividir el área en 2: 47.5% y luego hallar el valor de Z para esta área: Z=1.96.

Intervalos de Confianza(2)

- Este valor representa el número de desviaciones estándar que se permite para la variación de la media muestral, tanto por encima como por debajo.



Intervalos de Confianza – Muestras Grandes*

- Si σ^2 es conocido, el intervalo de confianza para la media muestral μ es $\bar{X} \pm Z\sigma_{\bar{x}}$
- Z=1.96 para un 95% de confianza
- Si σ^2 es desconocido, se utiliza la varianza muestral s^2

$$\text{I.C. para estimar } \mu = \bar{X} \pm Zs_{\bar{x}}$$

*>30 obs.

Intervalos de Confianza – Muestras pequeñas

- El TCL asegura normalidad sólo en muestras grandes

- Cuando las muestras son pequeñas se utiliza la distribución t de Student. Se deben cumplir 3 condiciones:
  - Muestra pequeña
  - σ^2 desconocido (si es conocido se usa Z)
  - Población normal o casi normal (si no se cumple se deben considerar tests no parámetricos)

Intervalos de Confianza – Muestras pequeñas(2)

- Al igual que la distribución Z, t tiene media cero y es simétrica c/r a la media.

- Sin embargo, mientras Z tiene varianza 1, la varianza de t es mayor y depende de la muestra:

$$\sigma^2 = \frac{n-1}{n-3}$$

Intervalos de Confianza – Muestras pequeñas(3)

- El estadístico t se calcula de manera similar a Z:

$$t = \frac{\bar{X} - \mu}{s_{\bar{x}}}$$

- El intervalo de la media muestral de confianza en muestras pequeñas:

$$\text{I.C. para estimar } \mu = \bar{X} \pm (t)(s_{\bar{x}}) = \bar{X} \pm t\frac{s}{\sqrt{n}}$$

Intervalos de Confianza – Muestras pequeñas(4)

- Ej (Wall Street Journal): Una empresa de construcción fue culpada de inflar los comprobantes que registra para los contratos de construcción.

- El contrato estableció que un cierto tipo de trabajo debería promediar US$1150.

- Los directivos de sólo 12 agencias de gobierno dieron testimonio, llegando a una media de US$1275 y desv. est. de US$235. Asumiendo montos normales, ¿un i. de c. de 95% apoyaría el caso legal de la empresa?

Intervalos de Confianza – Muestras pequeñas(5)

- Sol: a un nivel de 95% con 12-1=11 g.l., el valor de t es 2.201

$$\text{I.C. para estimar } \mu = \bar{X} \pm t\frac{s}{\sqrt{n}}$$

$$= 1275 \pm (2.201)\frac{235}{\sqrt{12}}$$

$$= 1275 \pm 149.31$$

$$\text{US\$}1{,}125.69 \leq \mu \leq \text{US\$}1{,}424.31$$

- Como US$1150 está contenido en el intervalo, se fortalece la defensa de la empresa. Notar que el intervalo es más amplio que si se utiliza la normal.

Intervalos de Confianza – Proporción poblacional

- Las decisiones dependen con frecuencia de parámetros que son binarios (ej proporción de clientes que paga sus créditos)

- La distribución de las prop. muestrales se distribuye normal en muestras grandes (*np*, *n(1-p)*>5) con media igual p y desv. estandar:

$$s_p = \sqrt{\frac{p(1-p)}{n}}$$

- El I.C. para estimar $\pi = p \pm Z s_p$

Intervalos de Confianza – Proporción poblacional

- Ej: el gerente de una estación de televisión debe determinar qué porcentaje de casas tiene más de un televisor.

- Una muestra aleatoria de 500 casa revela que 275 tienen dos o + tvs. ¿Cuál es el i.c. del 90% para estimar la proporción de casas con dos o + tvs?

- Dados los datos p=275/500=0.55 y $s_p = \sqrt{\frac{(0.55)(0.45)}{500}} = 0.022$

Intervalos de Confianza – Proporción poblacional

- Ej (cont): de la tabla Z=1.65:

$$\text{I.C. para estimar } \pi = 0.55 \pm (1.65)(0.022)$$

$$= 0.55 \pm 0.036$$

$$0.514 \leq \pi \leq 0.586$$

- El gerente puede tener un 90% de confianza que entre el 51.4% y el 58.6% de las casas de la ciudad tienen más de un televisor.

Test de Hipótesis

- El propósito del análisis estadístico es reducir el nivel de incertidumbre en la toma de decisiones.

- Sólo es posible tomar buenas decisiones cuando se cuenta con suficiente información. La prueba de hipótesis es una herramienta muy efectiva, utilizada en muchas circunstancias:
  - **Un embotellador de bebidas desea determinar si el peso promedio del contenido de sus botellas es medio litro.**
  - **Se desea certificar que la proporción de productos defectuosos es menor al 1% del total.**
  - **Se desea verificar que los costos promedio de producción se encuentran bajo los US$5 por unidad.**

Test de Hipótesis(2)

- Para realizar un test de hipótesis, se realiza alguna inferencia con sentido respecto a la población.

- Ej: La media del contenido en botellas es μ=500cc. Ésta se conoce como hipótesis nula H_0.

- La hipótesis nula se prueba contra la *hipótesis alternativa* H_1, que establece lo contrario: H_1: $\mu \neq 500$cc.

- De acuerdo a los datos muestrales, H_0 puede rechazarse o no rechazarse.

Test de Hipótesis(3)

- Siguiendo el ejemplo, se puede tener una media muestral de 505cc.

- ¿Se puede concluir que la media poblacional no es 500cc.?
- Probablemente no. Esta diferencia podría deberse a un error de muestreo.

- En este caso, esta diferencia se consideraría *estadísticamente insignificante*. El problema radica en identificar que tan grande puede ser una diferencia para considerarla estadísticamente insignificante.

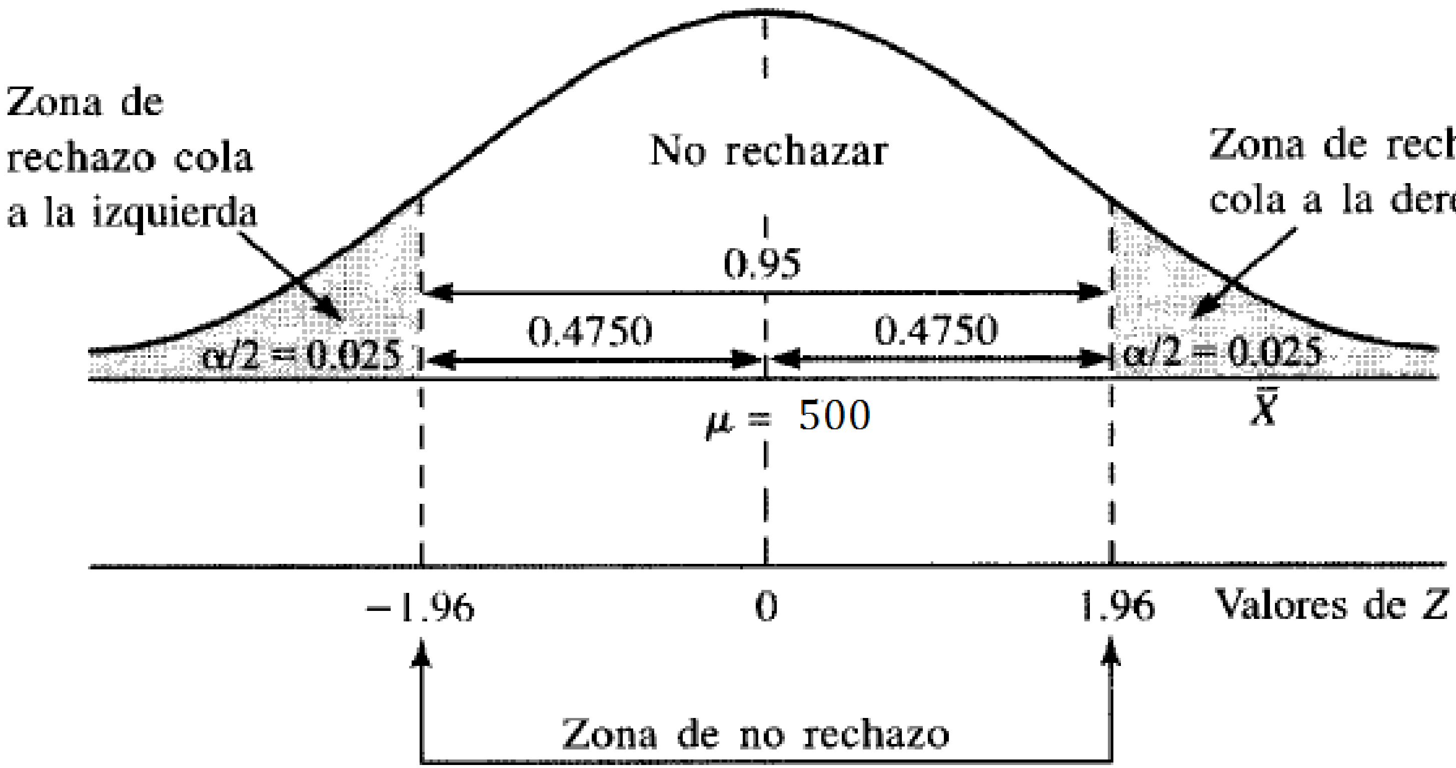
Test de Hipótesis(4)

- Recordando las distribuciones muestrales, podemos aplicar la fórmula Z:

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

- Asumiendo varianza poblacional desconocida, puede utilizarse la varianza muestral. La regla empírica indica que el 95% de las veces la media muestral esta a 1.96 desviaciones de la media poblacional.

Test de Hipótesis(5)



Existe un 95% de probabilidad de que los resultados muestrales puedan caer entre ± 1.96 si la hipótesis nula es verdadera.

Test de Hipótesis(6)

- Valores críticos de $Z \pm 1.96$ permiten establecer una regla de decisión que diga si se rechaza la hipótesis nula o no:

- "No se rechaza H_0 si los valores Z están entre $\pm$ 1.96. Se rechaza si el valor de Z es menor que -1.96 o mayor que 1.96"

- Se conoce como error tipo I el rechazar H_0 cuando es verdadera. Este error equivale al nivel de significancia o valor α.

Test de Hipótesis(7)

- El error tipo II corresponde a no rechazar una hipótesis que es falsa, y se identifica con la letra β.

- El error tipo II es difícil de determinar, pero a través de los costos de los errores permite establecer diferentes valores para α.

- Ej: si es más costoso equivocarse al rechazar la hipótesis nula, yo puedo "exigirle más" al test, disminuyendo α al 1%.

Test de Hipótesis –Prueba de 2 colas

Se deben seguir los siguientes pasos:

- Plantear la hipótesis
- Con base a los resultados de la muestra, calcular el valor del estadístico de prueba Z.
- Determinar la regla de decisión con base a los valores críticos de Z.
- Interpretación y conclusiones.

Test de Hipótesis –Prueba de 2 colas(2)

- Retomando el ejemplo del embotellador:

  H_0: $\mu = 500cc.$

  H_1: $\mu \neq 500cc.$

- Para probar la hipótesis nula se calcula el estadístico de prueba Z, por lo general asumiendo varianza poblacional desconocida:

$$Z = \frac{\bar{X} - \mu_H}{\frac{s}{\sqrt{n}}}$$

Test de Hipótesis –Prueba de 2 colas(3)

- Si el embotellador selecciona una muestra de 50 botellas y la media muestral es 505cc. y la desviación estándar es s=2.3, Z es:

$$Z = \frac{505 - 500}{\frac{2.3}{\sqrt{50}}} = 15{,}7$$

- Utilizando la regla de decisión: “No se rechaza H0 si los valores Z están entre ± 1.96. Se rechaza si el valor de Z es menor que -1.96 o mayor que 1.96”. Se rechaza la hipótesis nula ya que se encuentra en la región de rechazo.

Test de Hipótesis –Prueba de 2 colas(3)

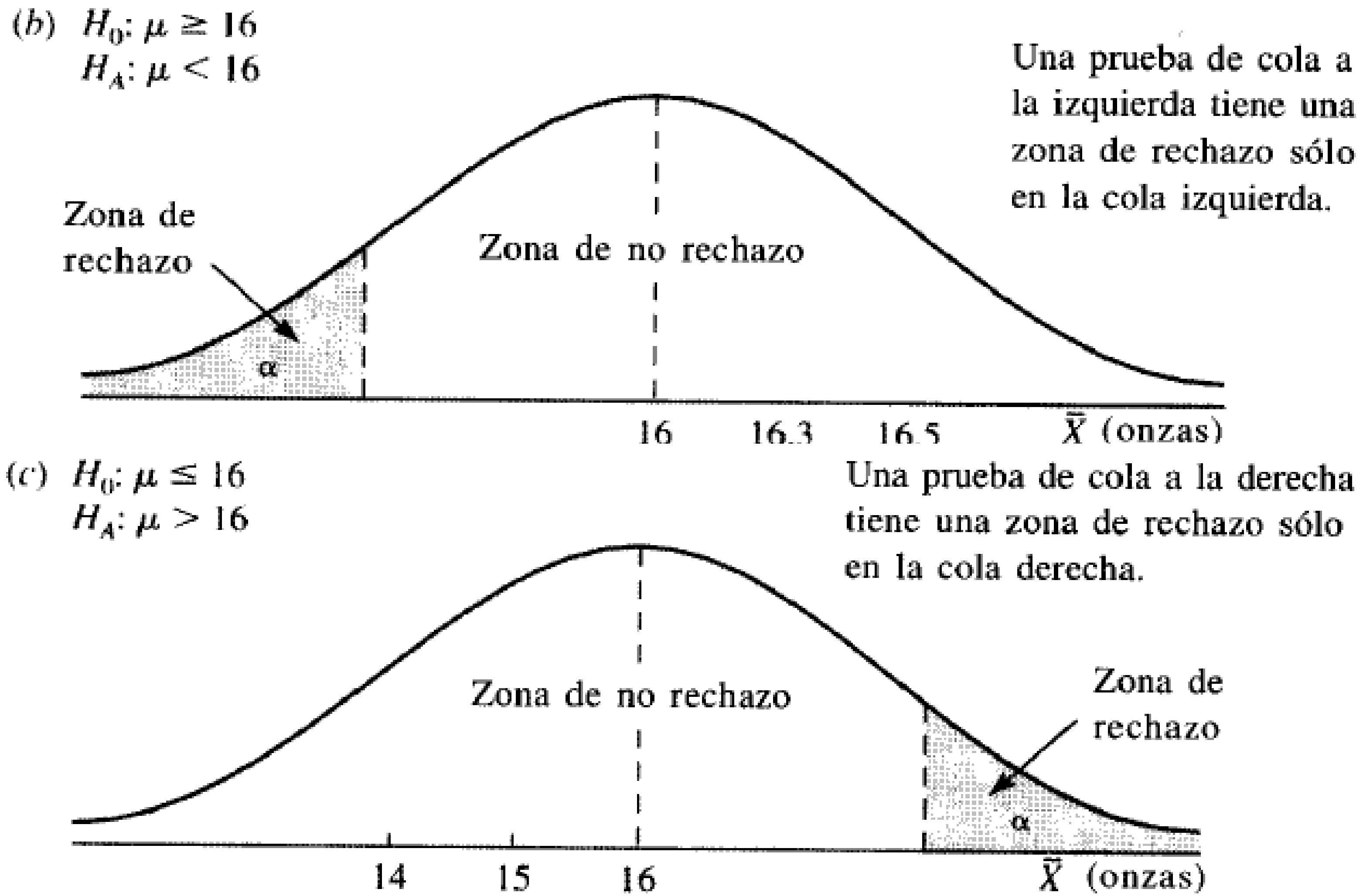
- Esto significa que la media poblacional NO es 500cc.?

- No del todo, la media poblacional podría ser 500cc., sin embargo, la probabilidad de cometer el error tipo I es de sólo un 2,5%.

Test de Hipótesis –Prueba de una cola

- En algunos casos sólo interesa un extremo u otro.
- Un restaurante de comida fresca no se interesa en saber qué tan rápidos son sus proveedores, sólo si toman demasiado tiempo.
- Un minorista se alarmará sólo si sus ingresos son demasiado bajos, ventas altas no son un problema.

Test de Hipótesis –Prueba de una cola(2)



Test de Hipótesis –Prueba de una cola(3)

- En vez de plantear que el contenido de las botellas es igual a 500cc., puede definirse como hipótesis nula que el contenido sea "al menos" 500cc., versus un contenido mayor a 500cc:

$$H_0: \mu \leq 500cc.$$

$$H_1: \mu > 500cc.$$

- En este caso sólo los valores significativamente por arriba pueden ser causa del rechazo de la hipótesis nula. Ahora se considera el 5% superior como zona de rechazo.

Test de Hipótesis –Prueba de una cola(4)

- Ej: En una reunión informativa, el gerente de un hotel reportó que el n° promedio de habitaciones alquiladas por noche es al menos 212. Un funcionario corporativo considera que esta cifra puede estar sobrestimada.

- Una muestra de 150 noches produce una media de 201 hab. Y una desv. de 45.5 hab. Si estos resultados sugieren que el gerente ha inflado su reporte, será amonestado severamente.

- A un nivel de 1% ¿cuál es el destino del gerente?

Test de Hipótesis –Prueba de una cola(4)

- Se utiliza un test de una cola:

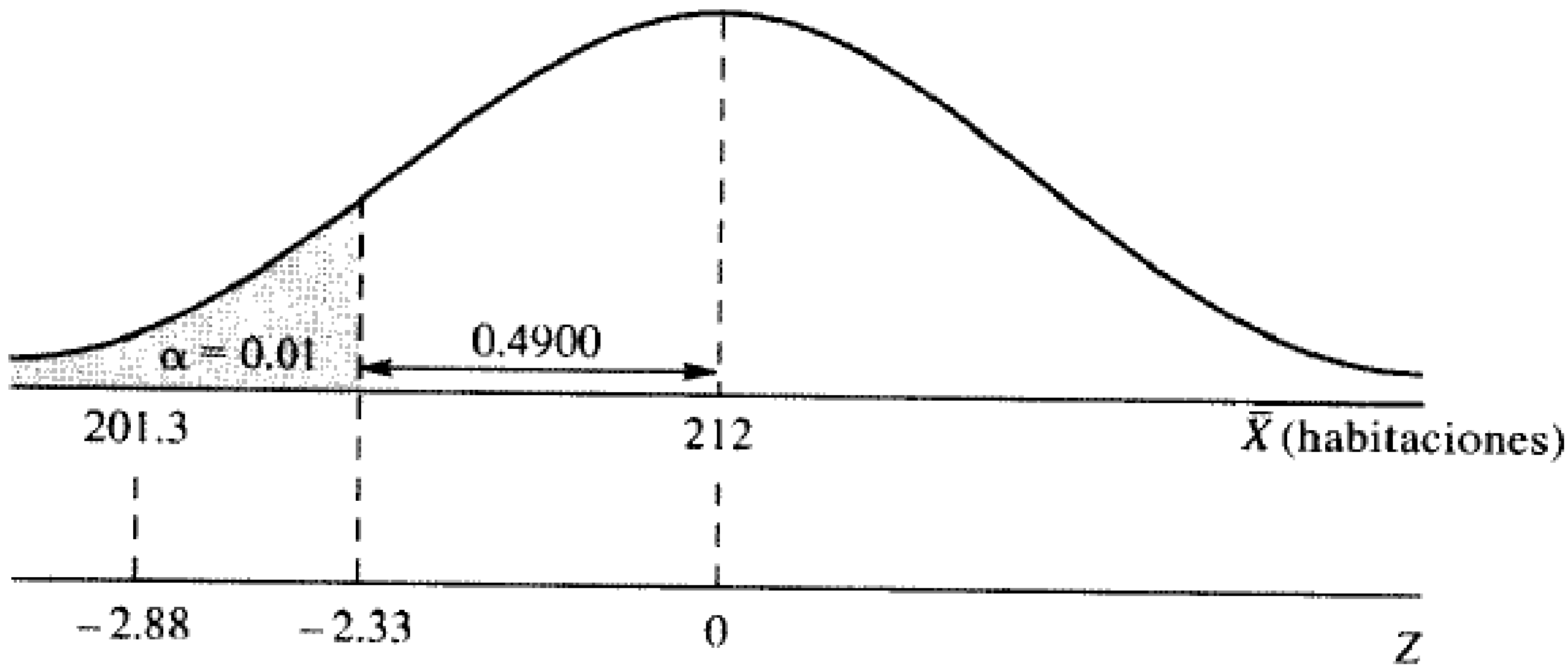
$$H_0: \mu \geq 212 \qquad H_A: \mu < 212$$

- El valor de Z es:

$$Z = \frac{201.3 - 212}{\frac{45.5}{\sqrt{150}}} = -2.88$$

- Como muestra la figura, un α=1% deja un área de 0.49 que, de la tabla Z, requiere un valor crítico de -2.33:

Test de Hipótesis –Prueba de una cola(4)



Test de Hipótesis –Prueba de una cola(5)

- De acuerdo a la regla de decisión:

"No rechazar H_o si $Z \geq -2.33$. Rechazar H_o si $Z < -2.33$"

Z se encuentra en la zona de rechazo, la hipótesis nula no se confirma.

- El gerente probablemente recibirá una reprimenda.

Test de Hipótesis –P_valor

- El p_valor es el área en la cola que está más allá del valor del estadístico para la muestra.

- Es el nivel más bajo de significancia al cual se puede rechazar la hipótesis nula.

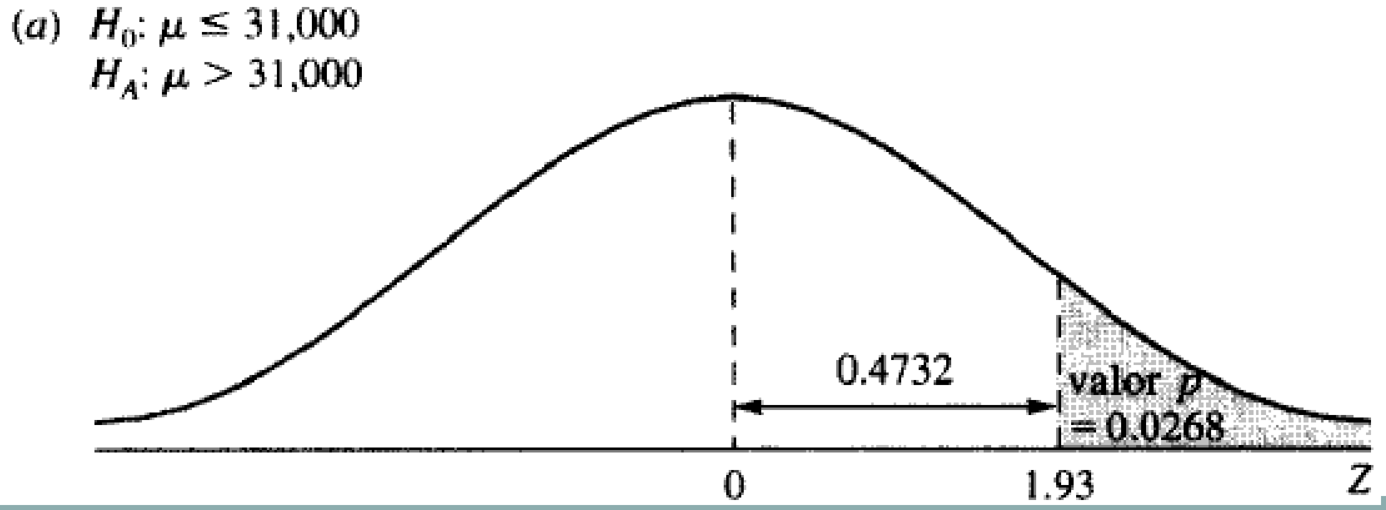
- Ej: El jefe de personal de una empresa desea estudiar si los empleados tienen un promedio de más de US$31000 en sus cuentas de pensiones. Al tomar 100 empleados, se llega a una media y desviación muestral de US$31366 y US$1894 respectivamente.

Test de Hipótesis – P-valor(2)

- Se plantea la hipótesis y se calcula Z:

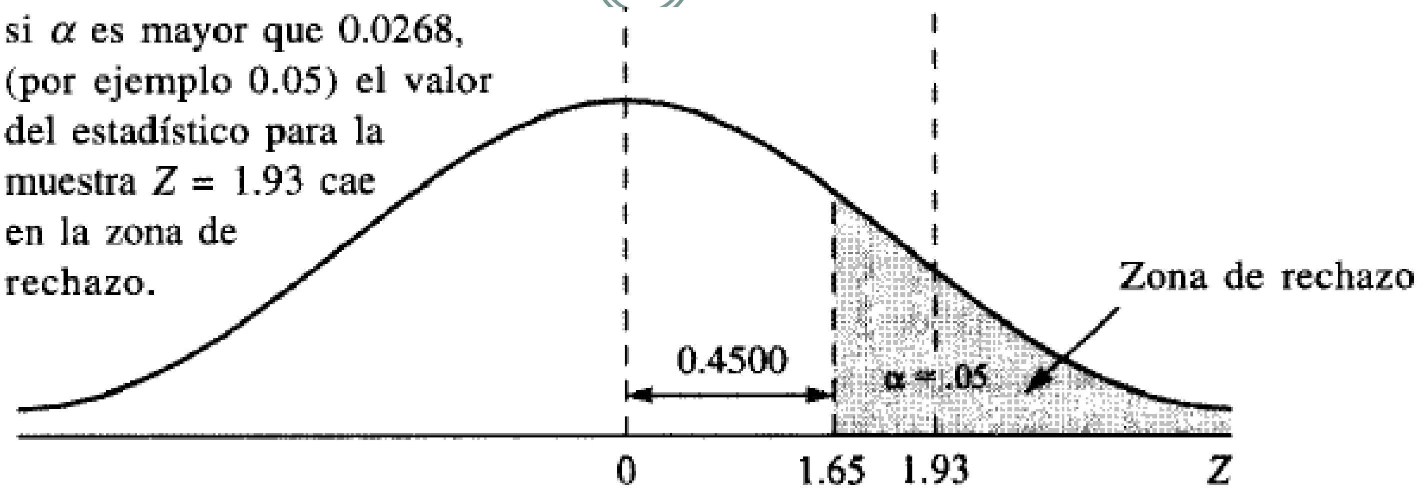
$$H_0\colon \mu \leq 31{,}000 \qquad H_A\colon \mu > 31{,}000$$

$$Z = \frac{31366 - 31000}{\frac{1894}{\sqrt{100}}} = 1.93$$

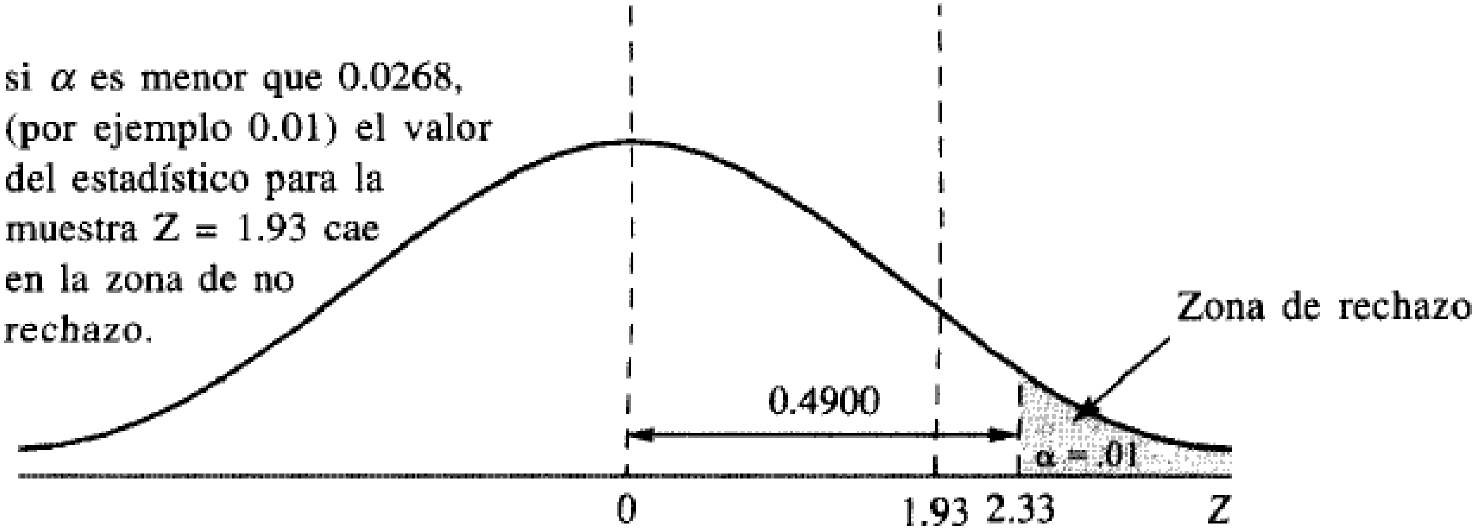


Test de Hipótesis – P-valor(3)

(*b*) si α es mayor que 0.0268, (por ejemplo 0.05) el valor del estadístico para la muestra $Z = 1.93$ cae en la zona de rechazo.



(*c*) si α es menor que 0.0268, (por ejemplo 0.01) el valor del estadístico para la muestra $Z = 1.93$ cae en la zona de no rechazo.



Test de Hipótesis – P-valor(4)

- Para dos colas: El mismo jefe de personal sospecha que los empleados invierten un promedio de US$100 mensuales en el plan de opción de compra de acciones de la compañía.

- Al tomar 100 empleados, llega a una media y desviación de US$106.81 y US$36.60 respectivamente.

- Se define la hipótesis

$$H_0: \mu = 100 \qquad H_A: \mu \neq 100$$

Test de Hipótesis – P-valor(5)

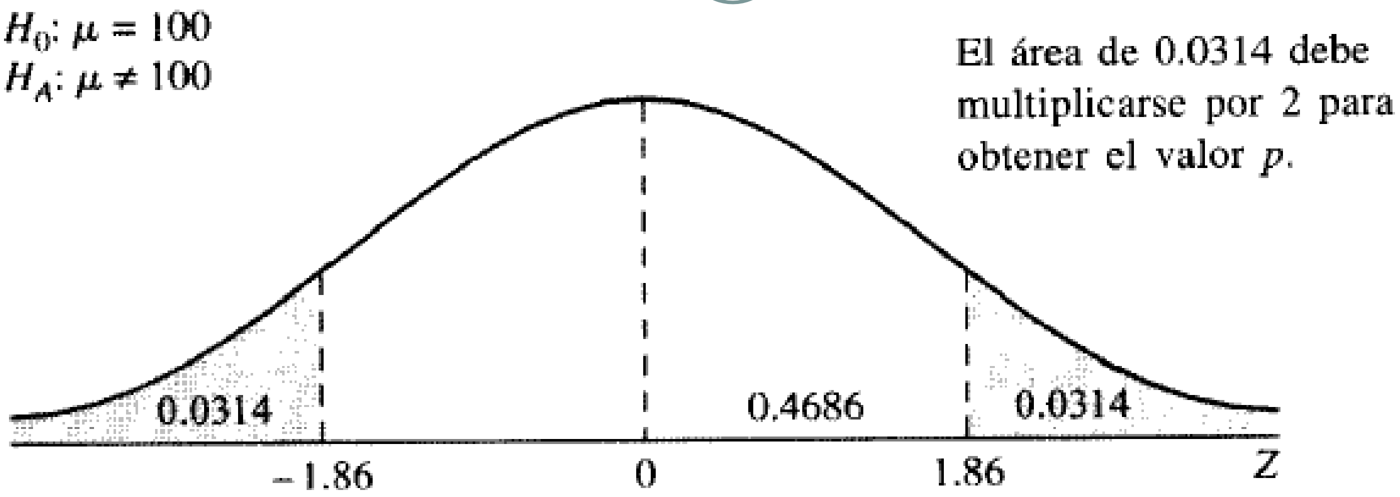
- Se calcula Z:

$$Z = \frac{106.81 - 100}{\frac{36.60}{\sqrt{100}}} = 1.86$$

- Para calcular p, se determina el área en la cola que va más allá del valor del estadístico para la muestra de Z=1.86

- Esta área es 0.0314 :

Test de Hipótesis – P-valor(6)



- Esta área debe multiplicarse por 2 para obtener p, ya que en un test de dos colas α se divide en las dos zonas de rechazo:

- P-valor=0.0314*2=0.0628. Si se fija α en menos que p (por ejemplo 5%) la hipótesis nula no puede ser rechazada.

Test de Hipótesis –muestras pequeñas

- Si la población es normal o casi normal, la muestra es pequeña y la varianza poblacional es desconocida, puede utilizarse la distribución t - Student.

- Ej: los estudiantes cuestionan que las hamburguesas "cuarto de libra" de McDonalds tengan efectivamente 0.25 libras de carne.

- Para probar la afirmación publicitaria cada estudiante compra una hamburguesa y la lleva a clase, donde éstas se pesan. Los resultados de la muestra son $\bar{X}$ =0.22 libras y s=0.09.

- Si hay 25 estudiantes y a=5%, a que conclusiones se llega?

Test de Hipótesis –muestras pequeñas(2)

- Se definen las hipótesis y se calcula t (n<30):

$$H_0: \mu = 0.25 \qquad H_A: \mu \neq 0.25$$

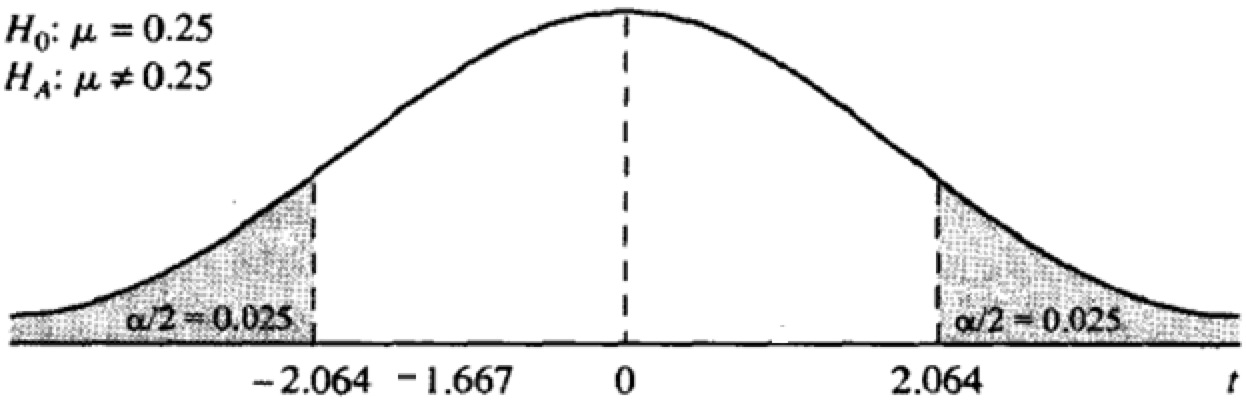
$$t = \frac{\bar{X} - \mu_H}{\frac{s}{\sqrt{n}}} \qquad t = \frac{0.22 - 0.25}{\frac{0.09}{\sqrt{25}}} = -1.667$$

- El valor t de -1.667 se compara con el valor crítico de t con n-1 g.l. =24 al 5%. De la tabla t para 2 colas: $t_{.05,24}$=-2.064.

Test de Hipótesis –muestras pequeñas(3)

- Se considera la regla de decisión:

Regla de decisión: "No rechazar H_o si t está entre ± 2.064. Rechazar H_o si t es menor que –2.064 o mayor que + 2.064"



- Como -1.667 está en el intervalo, no se rechaza la hipótesis nula.

Test de Hipótesis –proporción poblacional

- La prueba de hipótesis es muy similar a de la media poblacional:

$$Z = \frac{p - \pi_H}{\sigma_p}$$

- Se debe considerar la desviación de la distribución muestral de las proporciones muestrales:

$$\sigma_p = \sqrt{\frac{\pi_H(1 - \pi_H)}{n}}$$

Test de Hipótesis –proporción poblacional(2)

- Ej: Una muestra de 800 clientes de una entidad financiera revela que 492 tienen grados universitarios. Ud. considera que el 60% de los clientes se han graduado de la universidad.

- A un nivel de 5%, ¿qué se puede concluir?

- Las hipótesis son:

$$H_0: \pi = 0.60 \qquad H_A: \pi \neq 0.60$$

Test de Hipótesis –proporción poblacional(3)

- Se tiene que p=492/800 y la desviación es:

$$\sigma_p = \sqrt{\frac{0.60\,(1 - 0.60)}{800}} = 0.017$$

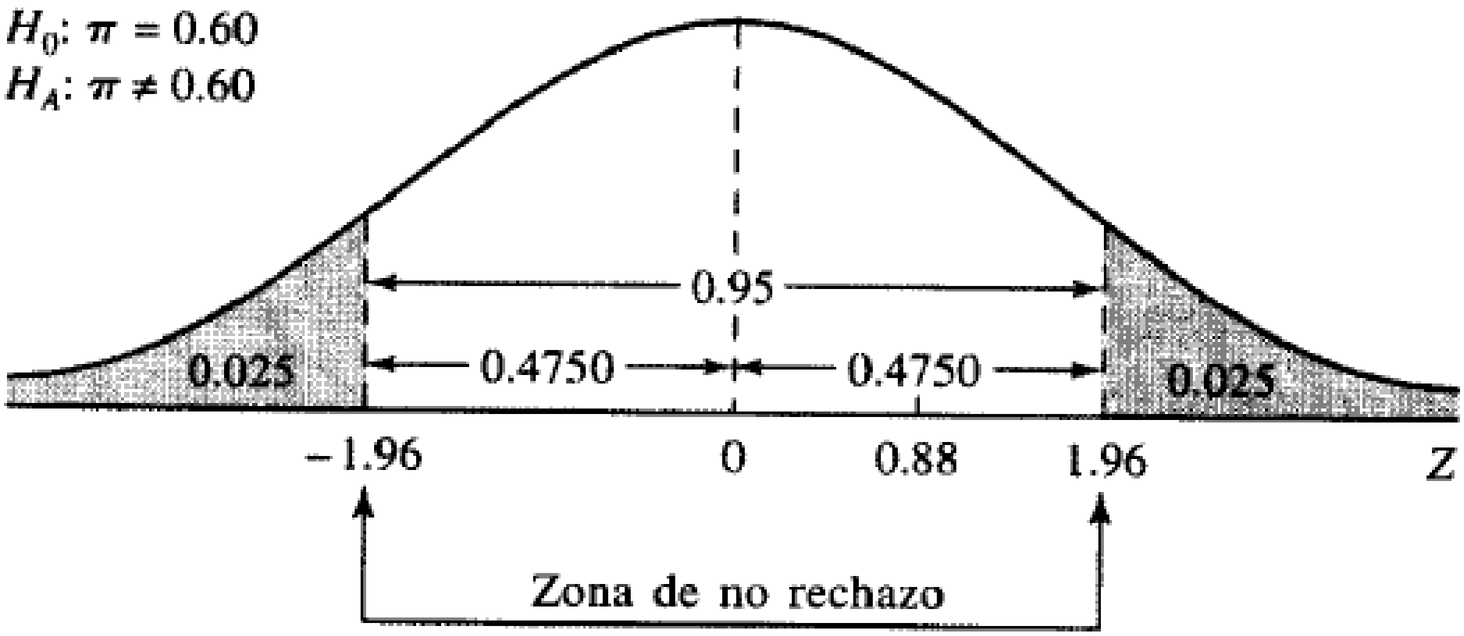
- Z es entonces:

$$Z = \frac{0.615 - 0.60}{0.017} = 0.88$$

- Se consideran dos zonas de rechazo (2.5% c/u), de lo cual se tienen valores críticos de Z ± 1.96. Como el valor muestral está dentro de la zona de no rechazo, no es posible rechazar H_o

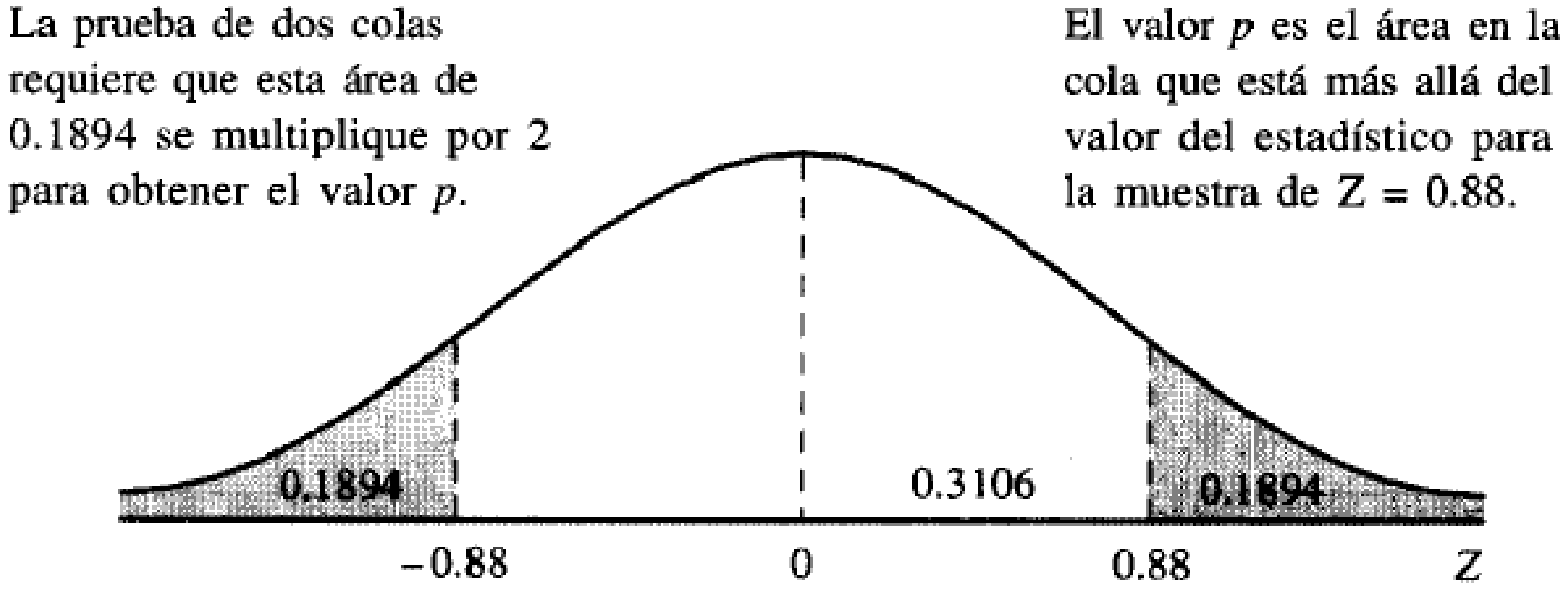
Test de Hipótesis –proporción poblacional(4)

- Gráficamente:



Test de Hipótesis –proporción poblacional(5)

- El p-valor es 0.1894*2=0.3788:



Test de Hipótesis – Resumen

Prueba sobre	**Hipótesis Nula**	**Suposiciones**	**Estadístico de Prueba**	**Fórmula TAREA**
La media	$\mu = \mu_0$	s^2 conocida	Normal	
	$\mu = \mu_0$	s^2 desconocida	T	
Igualdad de medias	$\mu_1 = \mu_2$	$s_1^2 = s_2^2$ conocidas	Normal	
	$\mu_1 = \mu_2$	$s_1^2 = s_2^2$ desconocidas	T	
	$\mu_1 = \mu_2$	$s_1^2 \neq s_2^2$ conocidas	T	
La varianza	$\sigma^2 = \sigma_0^2$	Dist. Normal, N pequeño	χ^2	
	$\sigma^2 = \sigma_0^2$	N grande	Normal	
Igualdad de dos varianzas	$\sigma_1^2 = \sigma_2^2$		F	
Una proporción	$p = p_0$		Normal	
Igualdad de dos proporciones	$p_1 = p_2$		Normal	