Probabilidad y Estadística (Finanzas Cuantitativas)

2022

MFIN – Universidad Torcuato Di Tella

Prof. Sebastián Auguste

sauguste@utdt.edu

CAPITULO 6

1. Intro
2. Muestreo
3. Estimadores
4. Intervalo de Confianza

1. INTRO

Modelo

Datos

Probabilidades

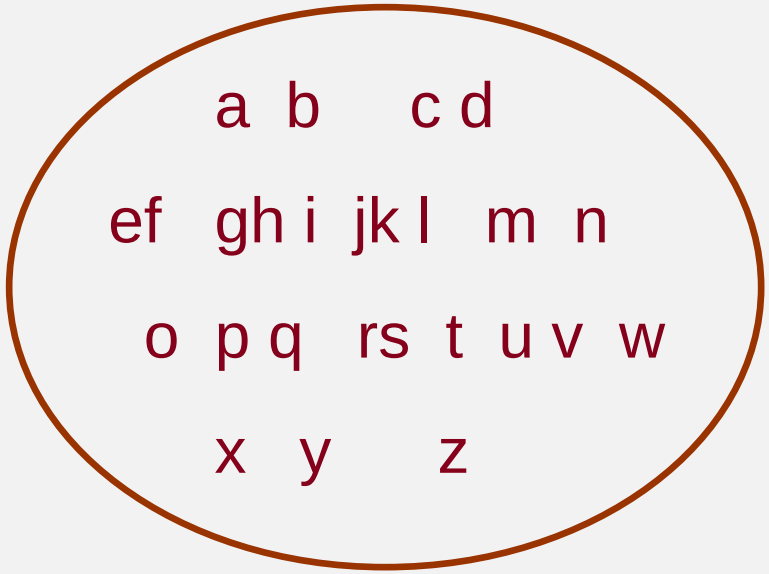
Estadísticas Descriptivas

INFERENCIA

Sacar conclusiones acerca de una población basados solo en datos muestrales

POBLACIÓN VS. MUESTRA

Población



Las medidas usadas para describir una población se denominan parámetros

Muestra

Las medidas calculadas en una muestra se denominan estadísticos

EJEMPLOS DE MODELOS:

- Retorno del Citi: $N(\mu, \sigma^2)$ (pero desconozco μ y σ^2)

- Quiero estimar la Probabilidad de impago de los clientes de Falabella: Prob(cliente no pague)

- Multifactor picing (APT)

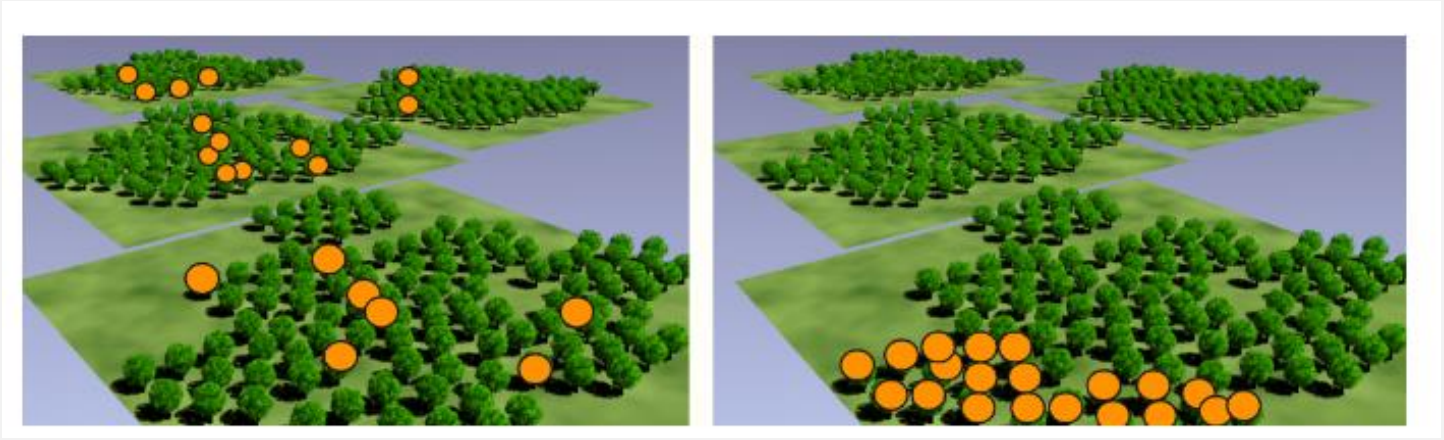
$$E(Ri)=Rf+b1*F1+b2*F2+b3*F3$$

ISSUES

- Cómo obtengo los datos que voy a usar (muestreo)
  - Cómo selecciono la muestra (técnica de muestreo)
  - Cómo pregunto (diseño de cuestionario)
  - Qué tamaño muestral necesito
- Cómo con esos datos obtengo una estimación (definición del estimador)
- Computar el error muestral (intervalo de confianza)
- Hacer hipótesis respecto al parámetro desconocido (test de hipótesis)

11. MUESTREO

POBLACIÓN - MUESTRA



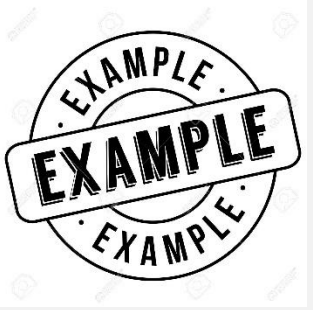
¿CÓMO SE MIDE?

- **Censo:** se recolecta información de toda la población
- **Datos administrativos:** registros e.g. datos de ventas de acciones, datos de clientes. No hay muestreo.
- **Muestreo**
  - **Probabilístico:** se elige a un subgrupo (muestra) de la población con un criterio aleatorio
  - **No Probabilístico:** informantes calificados, focus group, mistery shopping
- **Datos experimentales:** e.g. A/B testing

Técnica de Muestreo: conjunto de técnicas estadísticas que estudian la forma de seleccionar una muestra representativa de la población,

FUNDAMENTAL… SABER CÓMO SE OBTUVIERON LOS DATOS

- Entender sus limitaciones
- Ver si es consistente (¿hay outliers, missing, nonfrequent trading o errores de medición?)
- Entender si podemos decir algo válido sobre el universo o población (extrapolar)
- IDEAL Diseñar la muestra teniendo en cuenta las estimaciones que quiero realizar



EJEMPLO

- Epidemiólogo de Stanford Ioannidis "sospecho que se está exagerando con el Coronavirus porque se hacen ratios de mortandad sobre la gente reportada enferma, y seguro hay muchos enfermos que ni se enteran. Es mucho mejor tomar como referencia el caso del crucero varado, donde la mortandad fue tan sólo de 1%. Para mi la mortandad es mucho más baja como una gripe normal". Infobae, 17 de marzo 2020
- ¿Podría criticar esta afirmación desde le punto meramente estadístico? Está a favor o en contra?

TÉCNICAS DE MUESTREO NO PROBABILÍSTICO

- Muestreo de conveniencia: se decide qué individuos de la población pasan a formar parte de la muestra en función de la disponibilidad de los mismos (proximidad con el investigador, amistad, etc.).
- Muestreo discrecional: la selección es realizada por un experto que indica qué individuos son los que más pueden contribuir al estudio.

TÉCNICAS DE MUESTREO PROBABILÍSTICO

1. Aleatorio simple: cada elemento tiene igual probabilidad de ser seleccionado
2. Por cuotas: impongo cuotas por categorías
3. Estratificado
4. Sistemático
5. En etapas múltiples

TÉCNICAS DE MUESTREO PROBABILÍSTICO

1. Aleatorio simple: cada elemento tiene igual probabilidad de ser seleccionado
2. Por cuotas: impongo cuotas por categorías
3. Estratificado
4. Sistemático
5. En etapas múltiples

Si se conocen las características de la población a estudiar, se elegirán los individuos respetando siempre ciertas cuotas por edad, género, zona de residencia, entre otras que habrán sido prefijadas.

TÉCNICAS DE MUESTREO PROBABILÍSTICO

1. Aleatorio simple: cada elemento tiene igual probabilidad de ser seleccionado
2. Por cuotas: impongo cuotas por categorías
3. Estratificado
4. Sistemático
5. En etapas múltiples

- La población que se quiere estudiar se divide en subgrupos o "estratos" y luego se muestrea en forma independiente en cada estrato, con distintos tamaños muestrales en cada uno. Se usa para esto información previa que nos dice que estrato es más homogéneo y cuál más heterogéneo. Se sobremuestrea en los más heterogéneos y se submuestrea en los más homogéneos. Esto se hace para abaratar el estudio y hacerlo más eficiente

En el ejemplo de la figura, imaginemos tenemos barrios idénticos de tamaño, que son cada uno de los 7 vasos. Los M&M son las personas. Tenemos 6 barrios muy homogéneos y uno muy heterogéneo. Si tomo una muestra al azar de 70 lo más probable es que tenga 10 de cada barrio. Pero ¿para que quiero 10 de un barrio donde sé que son todos rojos? Con tomar un elemento ya sé el color del barrio

- Ponderador en muestras estratificadas: cuando el muestreo es estratificado, todas las medidas que hagamos deben ser "ponderadas" para revertir la sub o sobre ponderación.

Muestreo 10
X1,…,X10

Muestreo 1
X11

- En el ejemplo, si muestreo 10 del barrio de colores y 1 del barrio naranja, si luego hago un promedio simple, el resultado va a estar sobre-representado por el barrio de colores. Debo hacer un promedio ponderado

- Siguiendo con el ejemplo de sólo los dos barrios, la media ponderada sería:

$$\overline{x_w} = \frac{x_1 + \cdots + x_{10} + 10 \times x_{11}}{20}$$

- Antes de estratificar con los dos barrios idénticos de población iba a muestrear igual cantidad en ambos, pero a propósito tomé solo uno del naranja. Ahora para revertir eso multiplico a este valor por 10 (como si inflara mi muestra para que haya 10 del barrio naranja)

TÉCNICAS DE MUESTREO PROBABILÍSTICO

1. Aleatorio simple: cada elemento tiene igual probabilidad de ser seleccionado
2. Por cuotas: impongo cuotas por categorías
3. Estratificado
4. Sistemático
5. En etapas múltiples

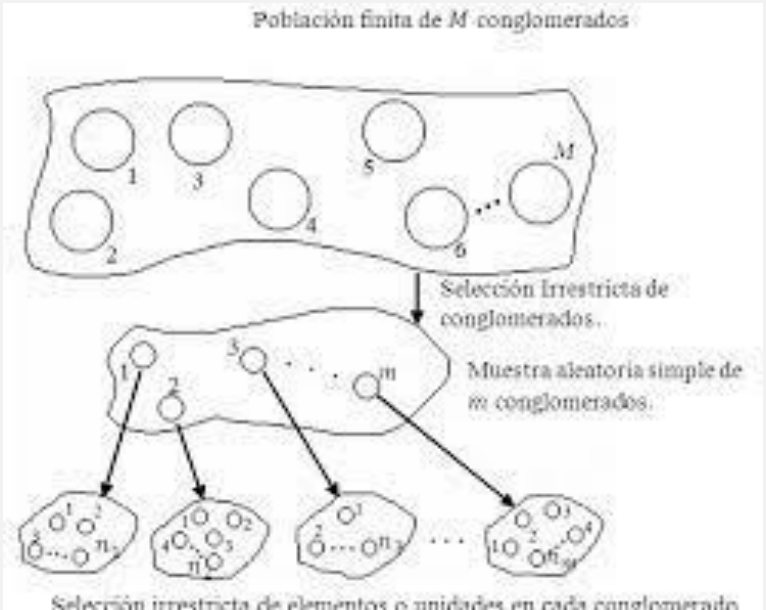
Muestreo sistemático

- El muestreo sistemático es muy similar al muestreo aleatorio simple. Se determina una regla o sistema para elegir gente al azar. Por ejemplo, si todos tuvieran un número del 1 al 60, en vez de elegir 12 al azar, podría decir, elijo al uno al azar del 1 al 5, y luego voy eligiendo el que está ubicado 5 más arriba en la numeración del que salió (en el ejemplo, salió el 2, y luego elijo el 7, 12, y así)

TÉCNICAS DE MUESTREO PROBABILÍSTICO

1. Aleatorio simple: cada elemento tiene igual probabilidad de ser seleccionado
2. Por cuotas: impongo cuotas por categorías
3. Estratificado
4. Sistemático
5. En etapas múltiples

- El muestreo en etapas múltiples consiste en empezar a muestrear por algo que no constituye el objeto de la investigación (unidades primarias), y obtener una muestra dentro de cada una de ellas (unidades secundarias).
- Por ejemplo, elijo primero al azar ciudades de la Argentina con menos de 100 mil habitantes, y luego para las ciudades elegidas, elijo una muestra de hogares para encuestar.
- Podría seguir una tercer etapa, cuando llego al hogar elijo al azar un miembro de la familia para encuestar.
- En cada etapa se pueden usar técnicas de muestreo distintas.
- El objeto final del estudio es al que encuestamos en la última etapa



ELABORACIÓN DEL CUESTIONARIO

- Algo no menor es qué preguntas incluyo, en qué orden van, y cómo pregunto. La idea es armar la pregunta para que el entrevistado lo entienda lo mejor posible, tenga ganas de contestarlo genuinamente y evitar generar sesgos o errores en las respuestas
- ¿hago preguntas abiertas o cerradas? T*rade off* entre minimizar errores de carga y esfuerzo en el análisis con la riqueza de los datos.
- ¿cómo pregunto cuestiones sensibles? "¿usted se droga?"
  - Indirect evidence (e.g. gasto en lugar de ingreso)
  - Randomized Response (RR) technique
- ¿Que hago cuando la población es muy heterogénea? (Muestreo Estratificado o por cuotas)

EJEMPLO DE ERRORES POR WORDING EN CUESTIONARIOS

- **Double-barreled question: ¿**cuán satisfecho está usted con su salario y las condiciones laborales?
- **Leading question:** guía a la persona a la respuesta (de que color son las nubes, de que color es el liquid paper, de qué color es el algodón, ¿qué toma la vaca?)
- **Loaded questions:** contienen supuestos implícitos (e.g. con qué papel arma usted su cigarro? Asume que fuma)
- **One-sided question**: incluye solo algunas alternativas en las respuestas
- **Ambiguous question**

- **Muy importante siempre hacer Prueba Piloto**

ESTUDIOS EXPERIMENTALES

- Los datos surgen de un experimento controlado (a diferencia de los estudios observacionales donde se buscan datos de la realidad para saber qué paso)
- Existe una tercera clase, los experimentos naturales o cuasi-experimentos, que son como experimentos pero no son controlados por el investigador sino que surgieron en la realidad.

En los estudios experimentales, para cuantificar impacto:

- Importante: grupo de control y grupo de tratados
- e.g. quiero saber si poner una promoción de mi producto en el supermercado incrementa las ventas. Debería escoger varios supermercados, luego separarlos al azar en grupo tratado y grupo de control, al grupo tratado le hago la promoción, al de control no, y luego comparo los resultados obtenidos en ambos casos.

FUENTES COMUNES DE ERROR EN CUALQUIER STUDIO EMPÍRICO

- Error de muestreo (Sampling Error), surge porque la muestra difiere de la población, por lo que el resultado de la muestra va a tener un error respecto al que se quiere estimar de la población
- Sesgo de medición (Measurement Bias) –preguntas pobremente fraseadas
- Self-Selection Bias –participantes que rechazan participar
- Response Bias –respuestas incorrectas o inverosímiles
- Non-response bias – sesgo cuando deciden no responder alguna pregunta en particular

CONCLUSIÓN

- Antes de hacer nada se debe conocer el alcance y las limitaciones de los datos que vamos a usar.
- Saber cómo se obtuvo la muestra me indica que tipo de problemas puedo tener.

III. ESTIMADORES

ESTIMADOR

- Un estimador es una función matemática que dice cómo usar los datos (muestrales) para obtener un valor para el parámetro desconocido (estimación).
- Se requiere tener una muestra.
- Cuando la muestra es aleatoria, el estimador también
  - Como el estimador es una variable aleatoria, tendrá una distribución (no siempre me va a dar el mismo valor, depende de la muestra que use)

EJEMPLO

- Quiero estimar el ingreso esperado de un nuevo alumno del MFIN Ejecutivo: E(I)

- ¿Cuál sería un buen estimador de E(I)?
- ¿Qué cosas haría que mi estimador funcione mejor?

Parámetro	Estimador (puntual)
• Probabilidad	• .....
• Media poblacional o esperanza matemática	• .....
• Varianza poblacional	• .....

Principio de analogía: reemplaza los momentos poblacionales con su momento muestral. Ejemplo:

momento poblacional: $E(x) = 0 \rightarrow$ momento muestral: $\frac{1}{n}\sum_{i=1}^{n} x_i = 0$

momento poblacional: $E(xy) = 0 \rightarrow$ momento muestral: $\frac{1}{n}\sum_{i=1}^{n} x_i y_i = 0$

¿QUÉ HAGO LUEGO DE ESTIMAR?

1. Calcular el error muestral de mi estimador
2. Hacer un "Intervalo de confianza" que me dice en qué rango se encuentra el parámetro desconocido con alta probabilidad (confianza)
3. Test de hipótesis: Sirve para hacerse preguntas y testearla con los datos con cierta confianza. E.g. encuesta política dice el político A tiene apoyo de 30%,. Este es el porcentaje que surgió en la encuesta, ahora nos queremos preguntas ¿es posible que el verdadero apoyo sea del 40% y así que gane en primera vuelta la elección?

VARIABLE ALEATORIA

- Como el estimador es una función matemática al cuál le metemos una muestra aleatoria, todo estimador es en sí mismo una variable aleatoria también.
- Esto es, el estimador puede dar muchos resultados distintos dependiendo de qué muestra se utilice.
- Para medir el error muestral es clave entender qué distribución tiene el estimador y de qué depende esa distribución.
- Al desvío estándar de un estimador se lo llama "error estándar" para diferenciarlo.

1. ESTIMADOR DE LA MEDIA POBLACIONAL

- Sea X una v.a. con E(X)=μ (media poblacional) desconocida

$$\mu = \sum_{i=1}^{\infty} x_i p_i$$

- Un estimador de μ es la "media muestral" (n tamaño muestral, muestra aleatoria iid)

$$\hat{\mu} = \overline{X} = \sum_{i=1}^{n} X_i \frac{1}{n}$$

DISTRIBUCIÓN DEL ESTIMADOR

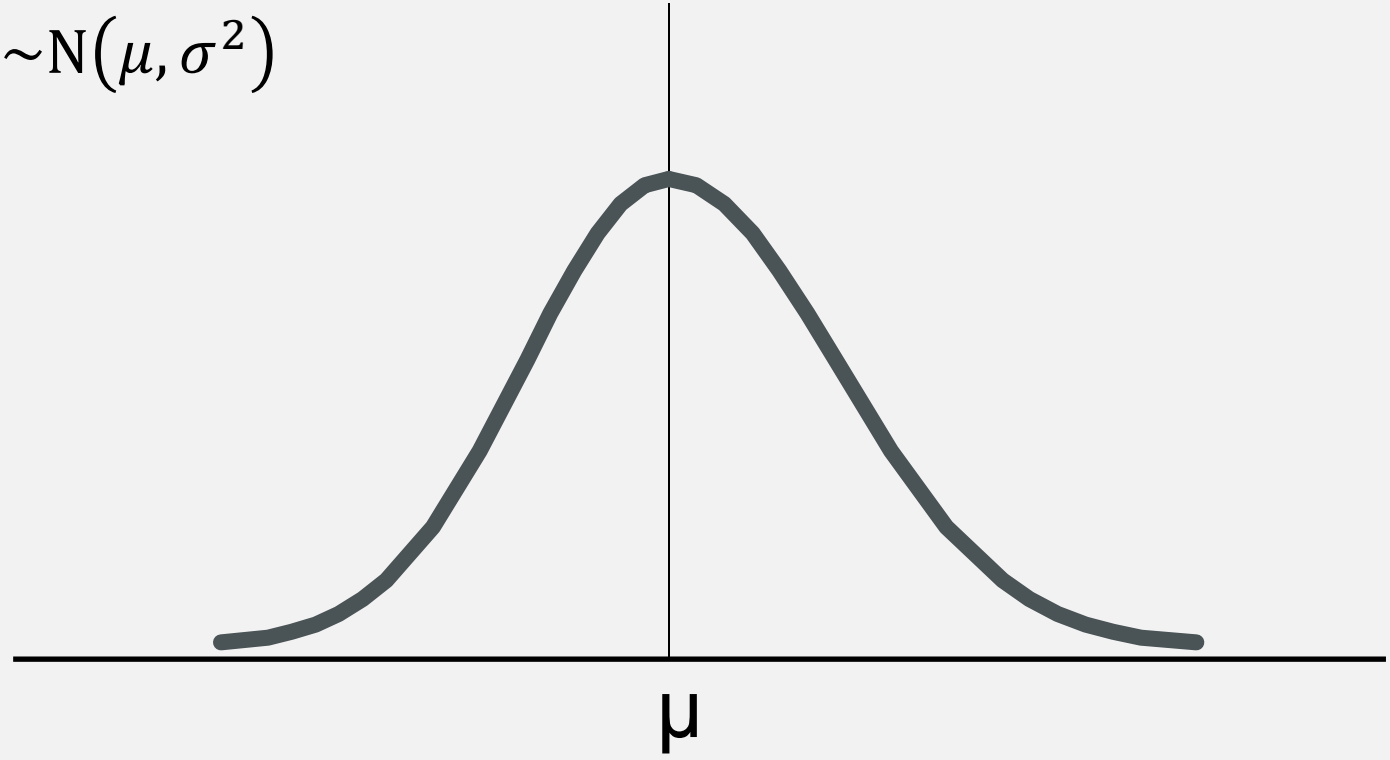
- Como la muestra es aleatoria, el estimador también será una variable aleatoria
- Cuál es su distribución?
- Cuál es su media?
- Cuál es su varianza?

Asumamos que la población que quiero encuestar tiene una distribución Normal con media μ y varianza σ^2

Por ejemplo, el ingreso de los potenciales alumnos del MFIN

Desconozco la media de esta población y hago una encuesta para saberlo

$$x \sim N(\mu, \sigma^2)$$



- Tomo una muestra de tamaño N, con lo cuál cada elemento de esta muestra proviene de la distribución Normal que queremos estudiar.

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

- Cómo cada X viene de esta Normal, cada uno sigue una distribución Normal (X raya es una suma de Normales!)

$$X_i \sim \mathrm{N}(\mu, \sigma^2)$$

- Sólo nos falta saber cuál es la media y varianza de mi estimador X raya.

El valor esperado de X raya es ….

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

$$E\left[\overline{X}\right] = E\left[\sum_{i=1}^{n} X_i \frac{1}{n}\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{n\mu}{n} = \mu$$

μ!!!!! Justo lo que quiero estimar, esto quiere decir que en promedio, si tomara muchas muestras distintas, y computo muchos X raya distintos, el promedio de esto (promedio de largo plazo) coincide con el valor que quiere estimar. Esta propiedad (cuando la esperanza del estimador es igual al parámetro que se quiere estimar) se llama INSESGADO (no tiene sesgos)

- La varianza de X raya es….

$$\text{var}\left[\overline{X}\right] = \text{var}\left[\sum_{i=1}^{n} X_i \frac{1}{n}\right]$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} \text{var}[X_i] + 2\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} covar\left[X_i, X_j\right]$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$$

- Las covarianzas son todas cero porque los X son independientes entre sí
- A la raíz cuadrada de esta varianza se la conoce como "error estándar" que en el caso de X raya es

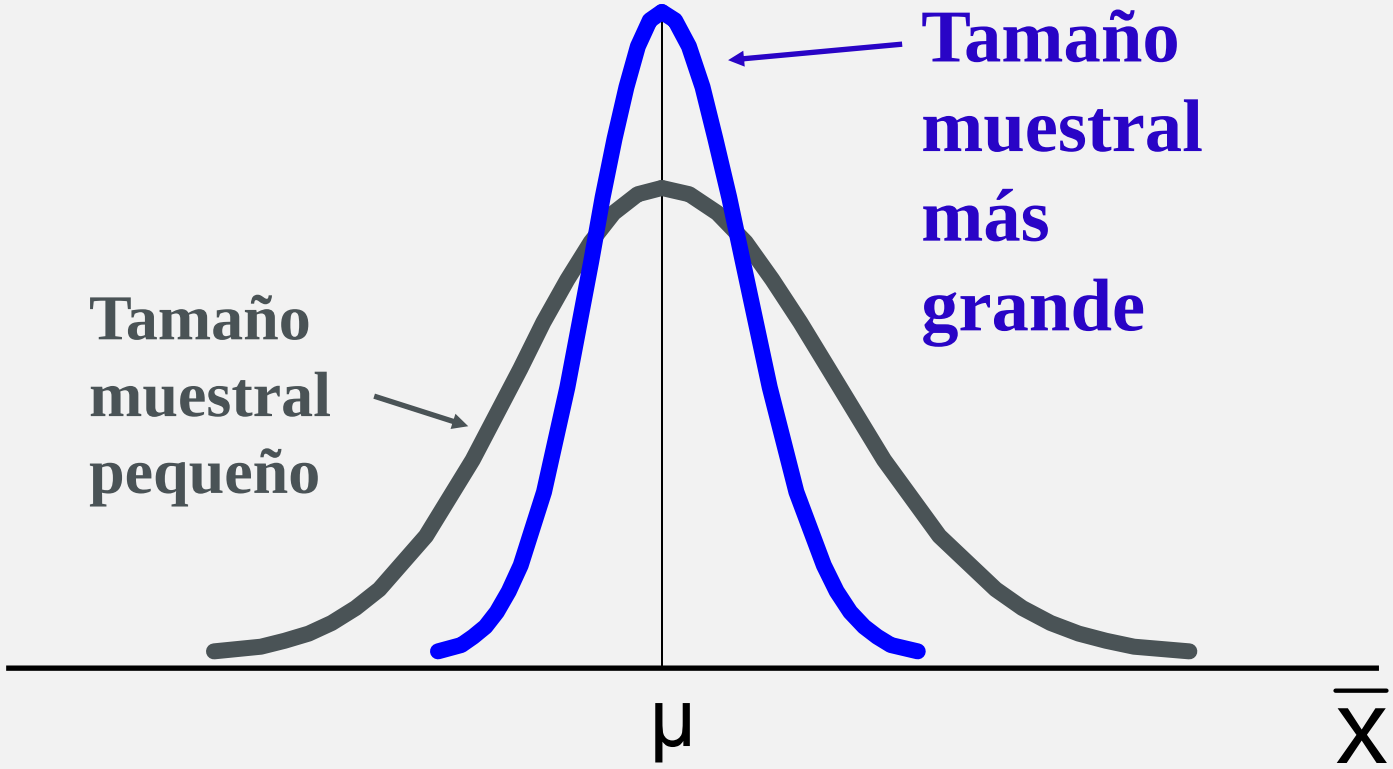
$$\frac{\sigma}{\sqrt{n}}$$

- Notar que el error estándar de X raya se achica cuanto mayor es el tamaño muestral, y si hago tender a infinito el tamaño muestral (incluyo a toda la población en mi muestra) el error estándar converge a cero, se elimina. Esto quiere decir que no voy a tener variación en los resultados del estimador, y me va a dar exactamente μ
- Esta propiedad se llama "consistencia" . Decimos que X raya es un estimador "consistente" de μ

$$\overline{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{\mathrm{n}}\right)$$

Note que la varianza de la media disminuye a medida que el tamaño muestral aumenta

A $\frac{\sigma}{\sqrt{n}}$ se lo llama "error estándar"



- Conclusión: Si la muestra de tamaño n es extraída (en forma independiente) de una población Normal con media μ y varianza σ^2 :

$$\overline{X} \sim \mathrm{N}(\mu, \frac{\sigma^2}{\mathrm{n}})$$

- O bien:

$$\frac{\overline{X} - \mu}{\sqrt{\sigma^2 / n}} \sim \mathrm{N}(0,1)$$

DISTRIBUCIÓN DE XRAYA CUANDO LA POBLACIÓN NO ES NORMAL

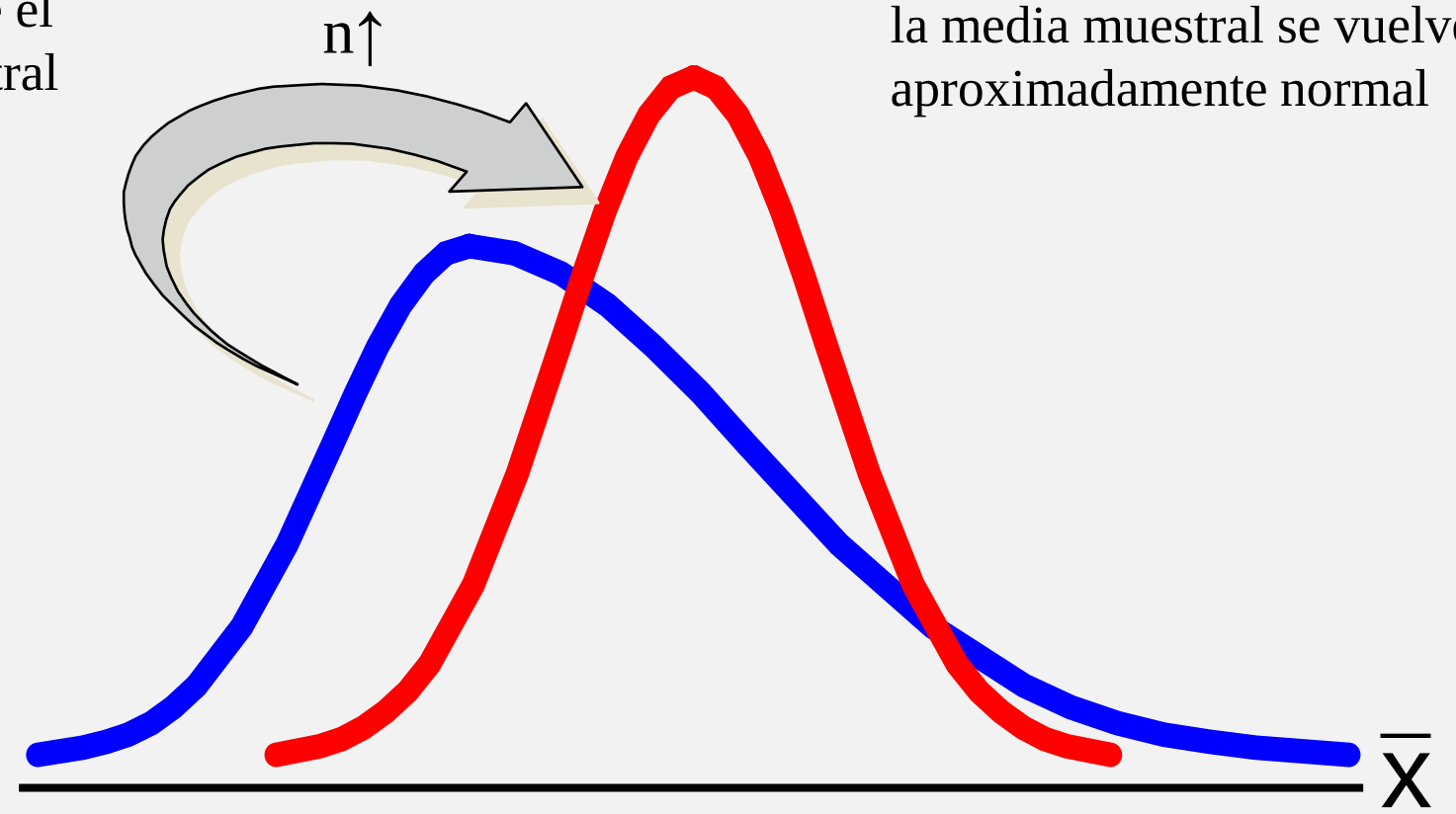
- Como todos los X de la muestra son independientes y están igualmente distribuidos, X raya es la suma de N de estos X, y por Teorema Central del Límite sabemos que si una variable (en este caso X raya) surge de una suma muy grande de variables aleatorias independientes e idénticamente distribuidas, entonces esa variable tiene distribución Normal, aunque no sepamos la distribución de la población, Podemos asumir que Xraya tiene distribución Normal si la muestra es grande:

$$\overline{x}_n \overset{asy}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

- "asy" quiere decir asintóticamente, es decir que es cierta la relación si el N tiende al infinito.

Moraleja. Si la muestra es muy grande X-raya sigue siempre una distribución aproximadamente Normal aunque la muestra no provenga de una población Normal

A medida que el tamaño muestral aumenta…



La distribución muestral de la media muestral se vuelve aproximadamente normal

CONCLUSIÓN

- Si proviene de una población Normal $X_i \sim N(\mu, \sigma^2)$

entonces $\bar{x}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

- Si proviene de una población desconocida $X_i \sim ?(\mu, \sigma^2)$

pero la muestra es grande, entonces $\bar{x}_n \overset{aprox}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$

3. ESTIMADOR DE UNA PROPORCIÓN

- Sea P la proporción de gente en la población que tiene cierta característica, un estimador de dicha proporción es la frecuencia muestral para una muestra de tamaño N
-

$$Fr = \frac{1}{N} \sum_{\substack{i=1 \\ si\ i\ tiene \\ característica}}^{\infty} X_i$$

- Es decir, cuento en la muestra la cantidad de casos que cumplen la característica y lo divido por N
- Ejemplo: quiero estimar la proporción de gente que votaría al candidato X

3. ESTIMADOR DE LA VARIANZA POBLACIONAL

- Sea X una v.a. con var(X)=σ^2 (varianza poblacional) desconocida

$$\sigma^2 = \sum_{i=1}^{\infty} (X_i - \mu)^2 \, p_i$$

- Un estimador de σ^2 es la "varianza muestral"

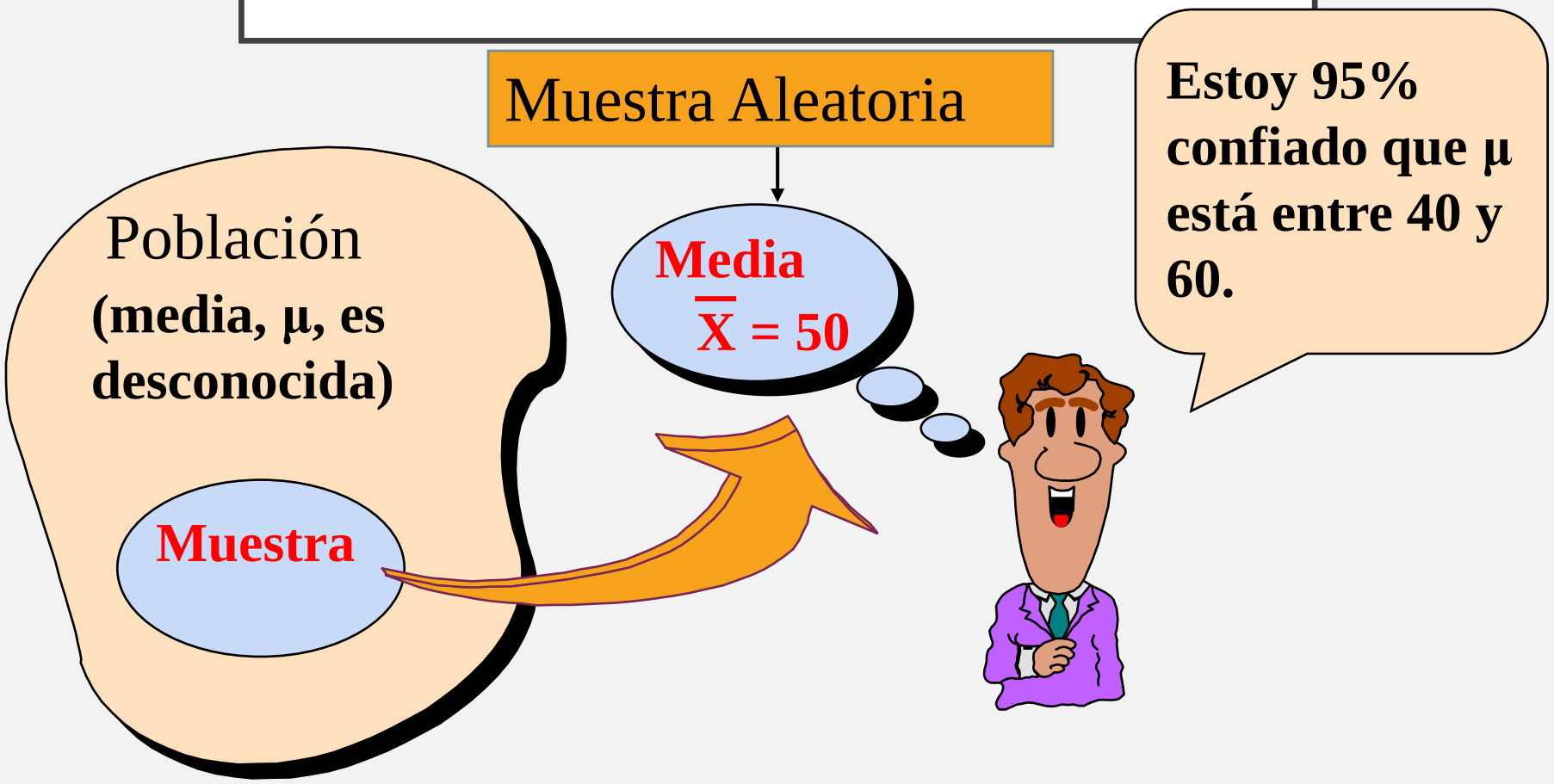
$$S = \frac{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2}{N - 1}$$

IV. INTERVALO DE CONFIANZA

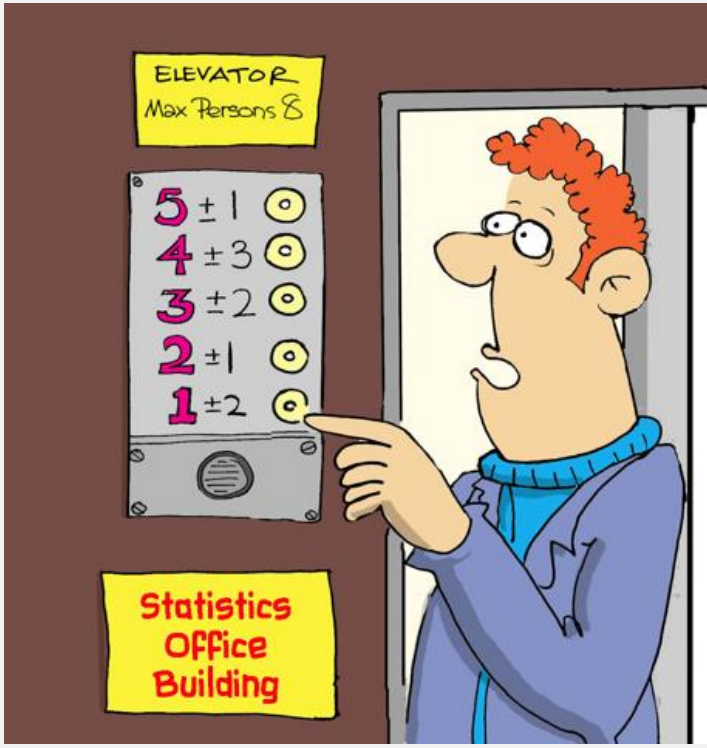
INTERVALOS DE CONFIANZA

- X raya es un estimador "punctual" del parámetro poblacional μ, es decir nos da un único valor.
- ¿Cuál es la probabildiad de que X raya acierte al parametro μ? Si es una variable aleatoria continua, CERO!!!
- Algo más inteligente entonces podría ser tener un estimador interval que me de un rango de valores con alta probabilidad o confianza de acertar para el valor μ. Esto es el Intervalo de confianza.

PROCESO DE ESTIMACIÓN



FÓRMULA GENERAL



- La fórmula general para todos los intervalos de confianza es:

Estimador Puntual ± Error Muestral

- Cada Estimador tiene su propia formula para el Intervalo de Confianza.

"Air Traffic Control, I'm going 400 miles an hour and need to land this thing on a floating narrow runway, so I'd prefer something better than 95% confidence in those coordinates."

1. INTERVALO DE CONFIANZA PARA LA MEDIA (CON VARIANZA CONOCIDA Y POBLACIÓN NORMAL)

- Si conozco σ sé la distribución de X raya salvo por μ

$$\bar{x}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- El intervalo de confianza al 1-α de confianza es:

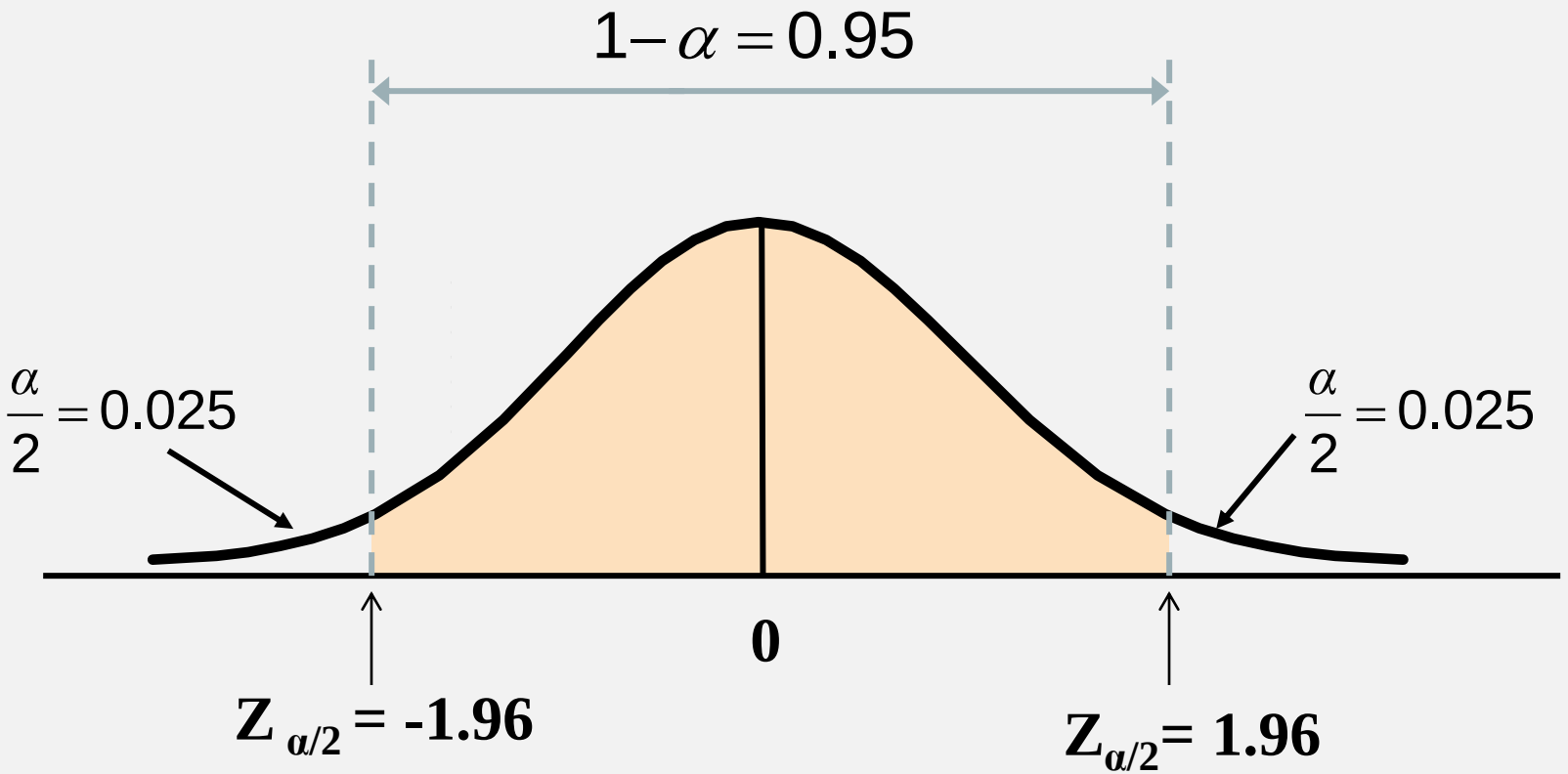
$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- Para una Normal (0,1)

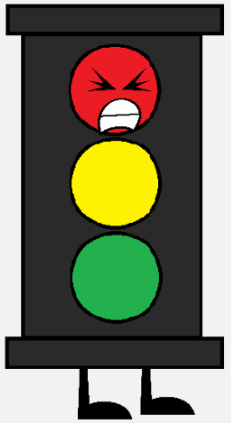
$$P(|Z| \leq |Z_{\alpha/2}|) = 1-\alpha$$

- E.g. Considere un 95% de confianza:

$$Z_{\alpha/2} = \pm 1.96$$



$$\bar{x}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad y \quad \frac{\bar{x}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$



- Por lo que

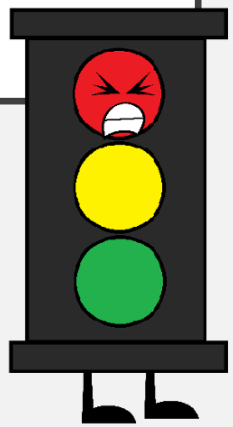
$$P(|Z| \leq |Z_{\alpha/2}|) = 1-\alpha$$

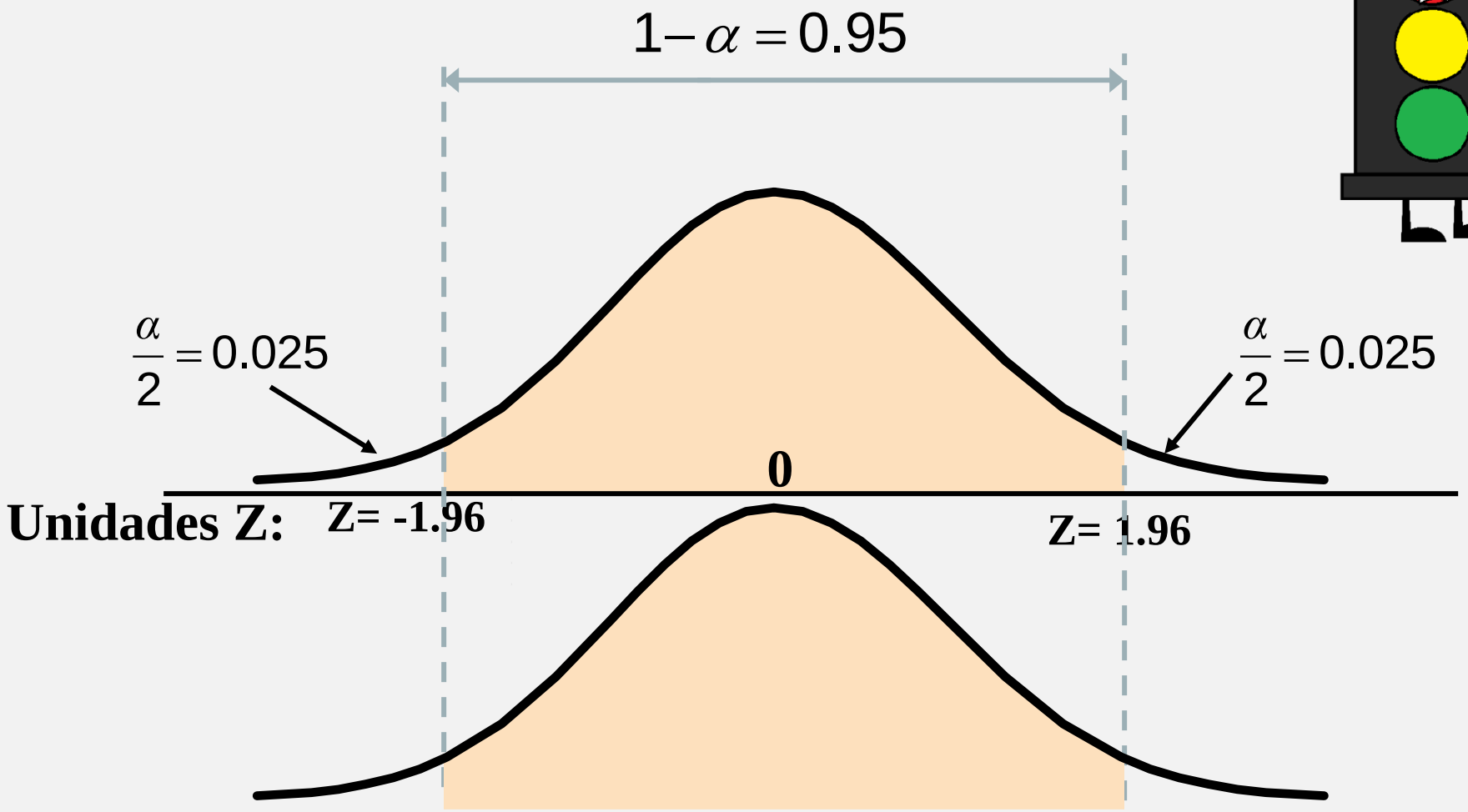
$$P(-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}) = 1-\alpha$$

$$P(-Z_{\alpha/2} \leq \frac{\bar{x}_n - \mu}{\sigma/\sqrt{n}} \leq Z_{\alpha/2}) = 1-\alpha$$

$$P(\bar{x}_n - Z_{\alpha/2}\,\sigma/\sqrt{n} \leq \mu \leq \bar{x}_n + Z_{\alpha/2}\,\sigma/\sqrt{n}) = 1-\alpha$$

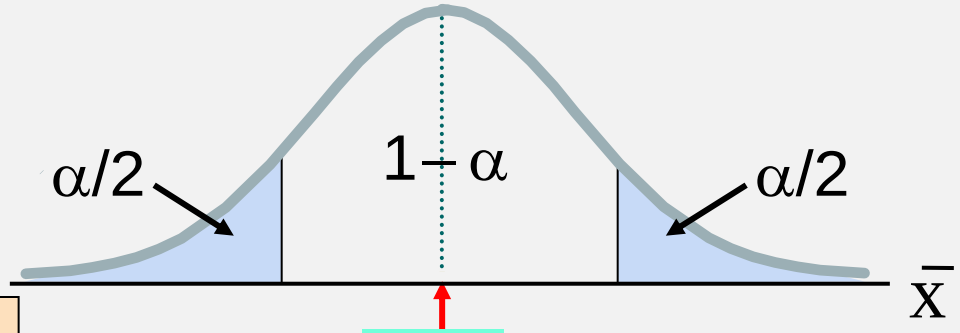
ANALOGÍA




$1-\alpha = 0.95$
$\frac{\alpha}{2} = 0.025$
$\frac{\alpha}{2} = 0.025$
0
Unidades Z:
Z= -1.96
Z= 1.96
Unidades X
$-Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
$\overline{X}$
$Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

INTERVALOS Y NIVEL DE CONFIANZA

Distribución Muestral de la Media



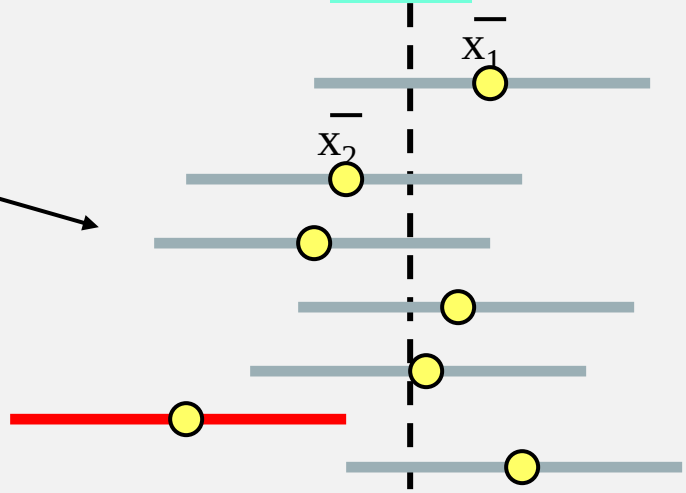
$\mu_{\bar{x}} = \mu$

Intervalos se extienden desde

$$\bar{X} + Z\frac{\sigma}{\sqrt{n}}$$

hasta

$$\bar{X} - Z\frac{\sigma}{\sqrt{n}}$$



(1-α)x100% de los intervalos construidos contienen a μ;

(α)x100% no.

Intervalos de Confianza

NIVELES DE CONFIANZA COMUNMENTE USADOS

- Los niveles de confianza más usados son 90%, 95%, y 99%

Nivel de Confianza	*Coeficiente de Confianza,* $1-\alpha$	*Valor Z*
80%	**0.80**	**1.28**
90%	**0.90**	**1.645**
95%	**0.95**	**1.96**
98%	**0.98**	**2.33**
99%	**0.99**	**2.58**
99.8%	**0.998**	**3.08**
99.9%	**0.999**	**3.27**

EJEMPLO

- Una muestra de 11 conexiones a Internet tomada de una población Normal tiene una duración promedio de 2.20 horas. Sabemos de información pasada que el desvío estándar poblacional es de 0.35 horas.

- Determine un intervalo del 95% de confianza para el verdadero valor de la duración promedio de conexión en la población.

- Solución:

- Nos dice que la distribución es Normal y nos da la varianza conocida de la población, la formula a usar es:

$$\bar{X} \pm Z \frac{\sigma}{\sqrt{n}}$$

$$= 2.20 \pm 1.96\,(0.35/\sqrt{11})$$

$$= 2.20 \pm 0.2068$$

$$1.9932 \leq \mu \leq 2.4068$$

Con un 95% de confianza el verdadero valor de la duración promedio de las conexiones a internet en la población está entre 1.9932 y 2.4068 horas

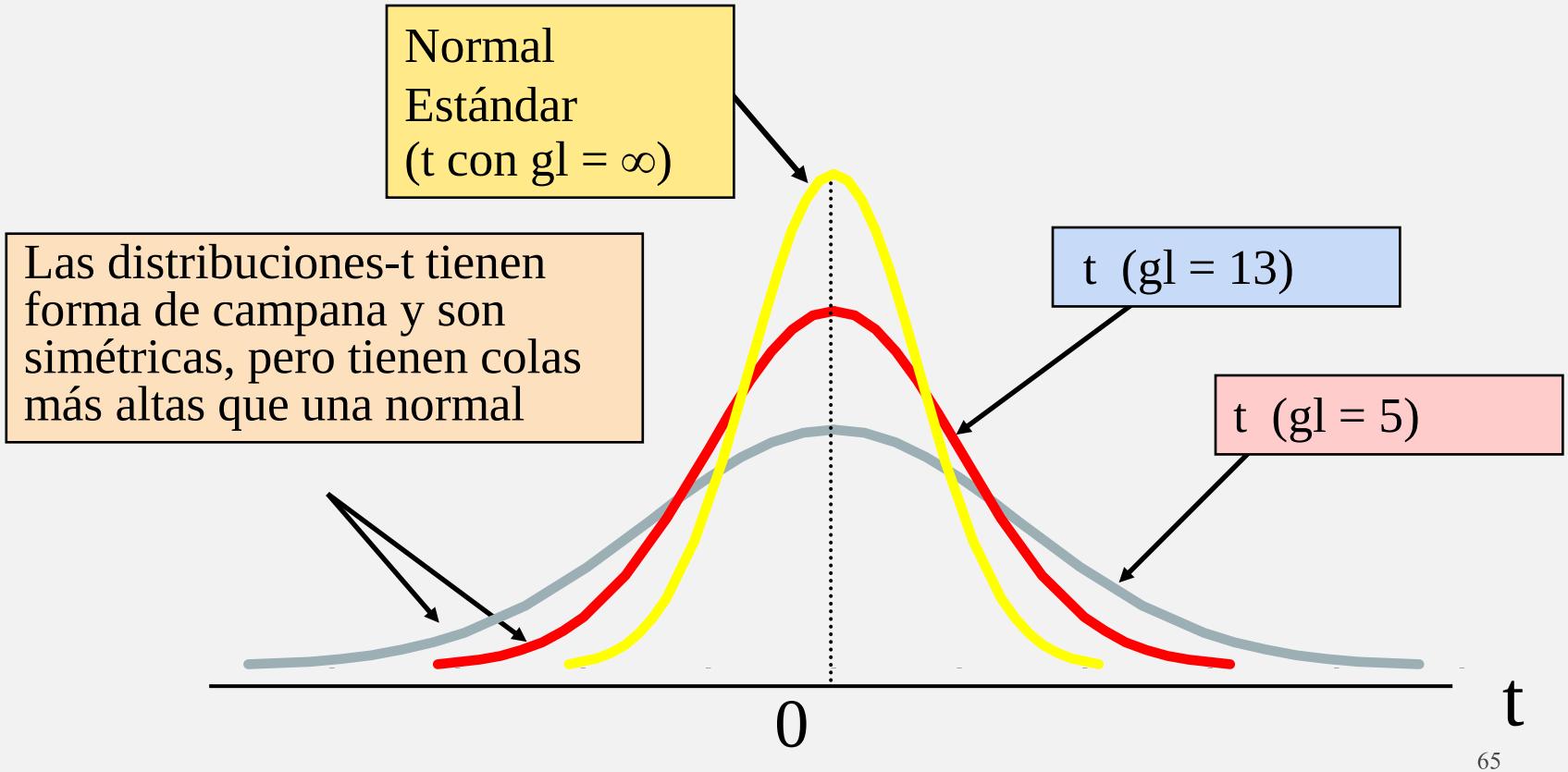
INTERVALO DE CONFIANZA PARA LA MEDIA CON VARIANZA DESCONOCIDA PERO POBLACIÓN NORMAL)

- Si el desvío estándar poblacional σ es desconocido podemos substituirlo por el desvío estándar muestral, S
- Esto introduce cierta incertidumbre ya que S varía de muestra en muestra
- Técnicamente (y si la muestra es chica) se usa la distribución t en lugar de la distribución normal.

$$\overline{X} \pm t_{\alpha/2,\, n-1} \frac{S}{\sqrt{n}}$$

DISTRIBUCIÓN T-STUDENT

Note que: t $\longrightarrow$ Z cuando n crece



FORMALMENTE...

Teorema: Si $\overline{x}$ es el valor para le media de una muestra aleatoria de tamaño n de una población normal y varianza σ^2 conocida, entonces:

$$\overline{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

es un intervalo de confianza al $(1-\alpha)\,\%$ para la media de la población

Teorema: Si $\overline{x}$ y s son los valores para le media y desvío estándares de una muestra aleatoria de tamaño n de una población normal, entonces:

$$\overline{x} - t_{\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}} < \mu < \overline{x} + t_{\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}}$$

es un intervalo de confianza al $(1-\alpha)\,\%$ para la media de la población

INTERVALO DE CONFIANZA PARA LA MEDIA CON POBLACIÓN Y VARIANZA DESCONOCIDA

- Si la muestra es grande:

$$\overline{X} \pm Z_{\alpha/2} \frac{S}{\sqrt{n}}$$

INTERVALOS DE CONFIANZA PARA LA MEDIA MUESTRAL RESUMEN

- Poblacion Normal, varianza conocida

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- Poblacion Normal, varianza deconocida

$$\bar{x} \pm t_{\alpha/2,n-1} \frac{S}{\sqrt{n}}$$

- Cualquier población pero Muestra grande

$$\bar{x} \pm z_{\alpha/2} \frac{S}{\sqrt{n}}$$

PROBLEMA 1.

- Para estimar la producción media diaria de los envases de metal por un pequeño fabricante, el encargado de la tienda examinó una muestra aleatoria de 56 registros diarios de producción, obteniendo una media de 32,5 contenedores y una desviación estándar de 2,9. Con una confianza del 99 por ciento. ¿Qué se puede decir sobre el error máximo en la estimación de que la verdadera producción diaria promedio es de 32.5 contenedores?

$n = 56$ containers

$\bar{x} = 32.5$ containers

$s = 2.9$ containers

$\alpha = 0.01$

$$E = z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

$$= 2.575 \times \frac{2.9}{\sqrt{56}} \quad : \text{(in case n large, } n \geq 30, \text{ to replace } \sigma \text{ with } s)$$

$$= 0.9978 \cong 1.0 \text{ containers}$$

With 99 percent confidence, error is at most 1.0 container.

PROBLEMA 2

- En un estudio de los costos anuales de apartamentos de alquiler en una ciudad del este, una muestra aleatoria de 36 apartamentos tiene un costo promedio de alquiler de 11.535 dólares por año y una desviación estándar de $ 875.

  Construya un intervalo de confianza del 99 por ciento para el verdadero costo promedio anual de alquiler de apartamentos.

  ¿Qué se puede decir con confianza de 95 por ciento sobre el máximo error si la media muestral de 11.535 dólares se utiliza como una estimación del verdadero costo promedio anual de alquiler de apartamentos?

n = 36 apartments

$\bar{x}$ = \$11,535

s = \$875

α = 0.01

$$\bar{x} - z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

$$\bar{x} - z_{0.01/2} \times \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{0.01/2} \times \frac{\sigma}{\sqrt{n}}$$

(in case n large, n ≥ 30, to replace σ with s)

$$11{,}535 - 2.575\left(\frac{875}{\sqrt{36}}\right) < \mu < 11{,}535 + 2.575\left(\frac{875}{\sqrt{36}}\right)$$

$$\$11{,}159.48 < \mu < \$11{,}910.52$$

PROBLEMA 3

- El gerente de la sucursal 31 del supermercado Kroger encontró, sobre la base de una muestra aleatoria de tamaño n = 60 tomada cuando la tienda estaba llena, que tuvo a los clientes les llevó en promedio 13.5 minutos para pasar por la caja rápida (de diez artículos o menos) –esto es, que le cobren y el embolsado). El desvío estándar de la muestra es S = 3,4 minutos.

(a) ¿Qué se puede afirmar con un nivel de confianza del 99, sobre el máximo error en la estimación de 13.5 minutos?

(b) Construya un IC al 90% de confianza.

$n = 60$ customers

$\bar{x} = 13.5$ minutes

$s = 3.4$ minutes

$\alpha = 0.10$

$$\bar{x} - z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

$$\bar{x} - z_{0.10/2} \times \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{0.10/2} \times \frac{\sigma}{\sqrt{n}}$$

(in case n large, $n \geq 30$, to replace σ with s)

$$13.5 - 1.645\left(\frac{3.4}{\sqrt{60}}\right) < \mu < 13.5 + 1.645\left(\frac{34}{\sqrt{60}}\right)$$

$12.78 < \mu < 14.22$ minutes

PROBLEMA 4

- Una muestra aleatoria de las ventas de combustible a 16 motos en una estación de servicio muestra que las ventas promedio son de 10,2 litros, con un desvío estándar de 2,9 litros. Construya un intervalo de confianza del 95 por ciento para la venta esperada de combustible.

$n = 16$ passenger cars (small samples)

$\bar{x} = 10.2$ gallons, $s = 2.9$ gallons

$\alpha = 0.05$, $t_{\alpha/2} = t_{(0.05/2),(n-1)} = t_{(0.025/2),(16-1)} = 2.131$

$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}$$

$$10.2 - 2.131\frac{2.9}{\sqrt{16}} < \mu < 10.2 + 2.131\frac{2.9}{\sqrt{16}}$$

$8.66 < \mu < 11.75$ gallons

PROBLEMA 5

- Utilizando la información de los FCI de la primera clase, compruebe usando IC si el fondo del BBVA y el del Standard Bank dan un retorno promedio distinto.

DETERMINANDO EL TAMAÑO MUESTRAL

Tamaño Muestral

Para la Media

Error muestral (margen de error)

$$e = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$\overline{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

DETERMINANDO EL TAMAÑO MUESTRAL

Tamaño Muestral

Para la Media

$$e = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Despeje n para obtener

$$n = \frac{Z_{\alpha/2}{}^2 \ \sigma^2}{e^2}$$

TAMAÑO MUESTRAL ÓPTIMO

- A través del concepto de Intervalo de Confianza puedo determinar cuál es el tamaño muestral necesario para tener cierto error muestral en mi estimación.
- Parto de fijar el error muestral que quiero, y luego estimo que tamaño muestral requiero para llegar a dicho error muestral.
- Requiero conocer la varianza poblacional o estimarla, y requiero definir con qué nivel de confianza quiero trabajar (el Z)

PROBLEMA 6

- Empresa de seguridad mantiene un registro diario de la cantidad de avisos de alarma antirrobo que se prenden en su tablero cada día. Suponiendo que es razonable utilizar una desviación estándar de 6,5 advertencias de alarma, ¿cuán grande debe ser la muestra de aleatoria de registros diarios que se requeriría si la empresa desea obtener un intervalo para el valor esperado de avisos diarios con una probabilidad de 0,99 y un Error de a lo sumo 2 avisos?

E = 2 alarm warnings

s = 6.5 alarm warnings

α = 0.01

$$E = z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$, (in case σ unknow , to replace σ with s)

$$n = \left[\frac{z_{\alpha/2} \times \sigma}{E}\right]^2$$

$$= \left[\frac{2.575 \times 6.50}{2}\right]^2$$

$$= 70$$

PROBLEMA 7

- Una tienda de bicicletas quiere saber el tiempo medio que tarda un empleado para montar una bicicleta. Si se sabe que la desviación estándar es de 4 minutos, ¿cuán grande debiera ser la muestra para poder afirmar con una probabilidad de 0,90 de que la media de la muestra se encuentra entre +/- 2 minutos?

$$E = 2 \text{ minutes}$$

$$s = 4 \text{ minutes}$$

$$\alpha = 0.10$$

$$E = z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} \quad , \text{ (in case } \sigma \text{unknow , to replace } \sigma \text{with s)}$$

$$n = \left[\frac{z_{\alpha/2} \times \sigma}{E}\right]^2$$

$$= \left[\frac{1.645 \times 4}{2}\right]^2$$

$$n = 10.8 \cong 11$$

INTERVALO DE CONFIANZA PARA LA PROPORCIÓN

π = la proporción de la población que tiene cierta característica

- Proporción muestral (p) provee una estimación de π:

$$p = \frac{X}{n} = \frac{\text{número of items en la muestra que tienen la característica de interés}}{\text{tamaño muestral}}$$

- $0 \leq p \leq 1$
- p tiene distribución Binomial

(asumiendo muestreo con reemplazo desde una población finita o sin reemplazo desde una población infinita)

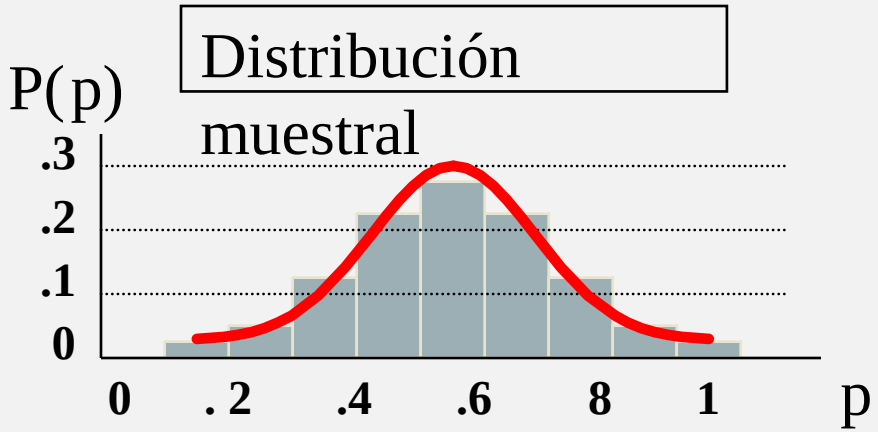
DISTRIBUCIÓN MUESTRAL DE P

- Aproximada por una distribución normal si:

$$np \geq 5$$

y

$$n(1-p) \geq 5$$



donde

$$\mu_p = \pi$$

y

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

(donde π = proporción poblacional)

INTERVALO DE CONFIANZA

$$\mathrm{Z} = p \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{\mathrm{n}}}$$

3. ESTIMADOR DE LA VARIANZA POBLACIONAL

- Varianza Poblacional: $\sigma^2 = E(x-\mu)^2$
- Siguiendo el principio de analogía, podríamos pensar en un estimador:

$$S_n^{\,2} = \sum_{i=1}^{N} \frac{(x_i - \bar{x})^2}{n}$$

- Este estimador es consistente, pero no insesgado por lo que se usa

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

DISTRIBUCIÓN DE LA VARIANZA

- Sea X1 a Xn una secuencia de variables aleatorias iid normales, con media μ y varianza $\sigma 2>0$
- La varianza muestral sigue:

$$(n-1)\frac{s^2}{\sigma^2} \sim \chi^2_{n-1}$$

INTERVALO DE CONFIANZA PARA LA VARIANZA

- Como

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

$$P(\chi^2_{1-\alpha/2,n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-1}) = 1-\alpha$$

- Lo que da un intervalo de confianza para $\sigma 2$:

$$[\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}]$$

EJEMPLO

- En un estudio de colesterol HDL se tomaron muestras a 12 personas, asumidas como una muestra aleatoria. La varianza muestral del indicador fue 0,3918680978. Encuentre un intervalo de confianza al 95% para la varianza poblacional.

- Solución

$$\frac{(n-1)s^2}{\chi^2_{1-(\alpha/2)}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\alpha/2}}$$

$$\frac{11(.3918--)}{21.920} < \sigma^2 < \frac{11(.3918--)}{3.816}$$

$$.196649- < \sigma^2 < 1.129598--$$

$$.4435 < \sigma < 1.0628$$

$\chi^2_{.975} = 21.920$

$\chi^2_{.025} = 3.816$