ANÁLISIS ESTADÍSTICO

MUESTREO DE POBLACIONES FINITAS

Jorge Fallas
jfallas56@gmail.com

2010

Temario

- **¿Porqué muestrear?**
- **¿Para qué muestrear?**
  - Estimar parámetros de población
  - Prueba de hipótesis
  - Exploratorio
- **Poblaciones, muestras y estimadores**
  - Infinitas
  - Finitas
  - Propiedades de estimadores
- **Esquemas de muestreo**
  - Probabilístico
    - Al azar
    - Simple al azar
    - Diseños más complejos
  - No Probabilístico
    - Sistemático
    - Experto
    - Conveniencia
- **Herramientas**
  - XLStats
  - Instat

¿Porqué muestrear?

- No es posible medir la totalidad de los elementos
  - Dinero, tiempo, herramientas
  - No es necesario
- Existen métodos estadísticos que nos permiten realizar inferencias a partir de muestras
- Deducción: Investigador(a) no posee datos sobre la totalidad de la población y lo que hace es generalizar la conclusión obtenida a partir de una muestra a aquellos individuos no medidos o estudiados

Población y muestra

- **Población: DE elementos, De observaciones**
- **Infinita** (Elementos no se pueden contar)
  - Árboles de Mundo
  - Estrellas del Universo
  - Tiempo de respuesta a un estímulo
- **Finita** (Elementos se pueden contar)
  - Árboles en campus UNED
  - No. estudiantes curso estadística
  - Venados cola blanca en San Lucas
  - Pobladores de San José de la Montaña

PARÁMETROS Y ESTIMADORES

- **Parámetro** es una propiedad de la población de observaciones (Ej. media, desviación estándar). Es una constante desconocida. Representado con letras griegas: α, ß, τ y σ
- **Estimador** es una función que permite estimar el valor del parámetro a partir de los datos de una muestra. Representado con letras latinas: a, b, r y s.
- El estimador es una variable aleatoria- Su valor cambia de muestra en muestra y por tanto posee una distribución de probabilidades conocida como Distribución muestral del estimador
  - Insesgado
  - Variancia mínima
  - Consistente
- **Propiedades**
  - Insesgado
  - Eficiente
  - Suficiente
  - Consistente

Estimadores: Propiedades

- **Insesgado: valor esperado = parámetro**
  - **Sesgo=** $E(\hat{\Theta}) - \Theta$ **= Diferencia entre el valor esperado del estimador y el valor del parámetro**
- **Eficiencia – Mínimo Cuadrado medio del error**

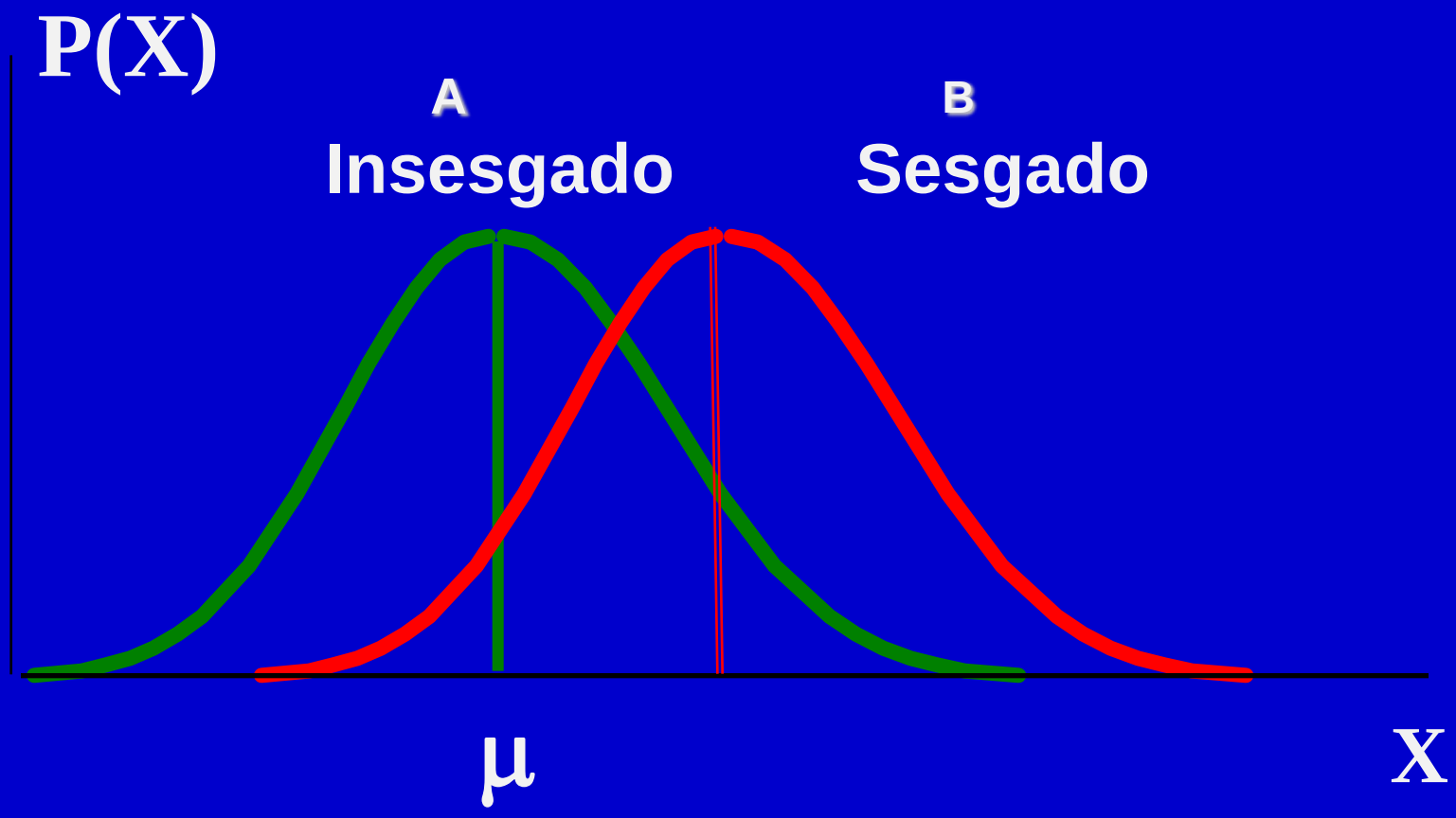
$$\text{M.S.E.} = E(\hat{\Theta} - \Theta)^2$$

$$= E[\hat{\Theta} - E(\hat{\Theta})]^2 + [E(\hat{\Theta}) - \Theta]^2$$

(varianza + (sesgo)^2

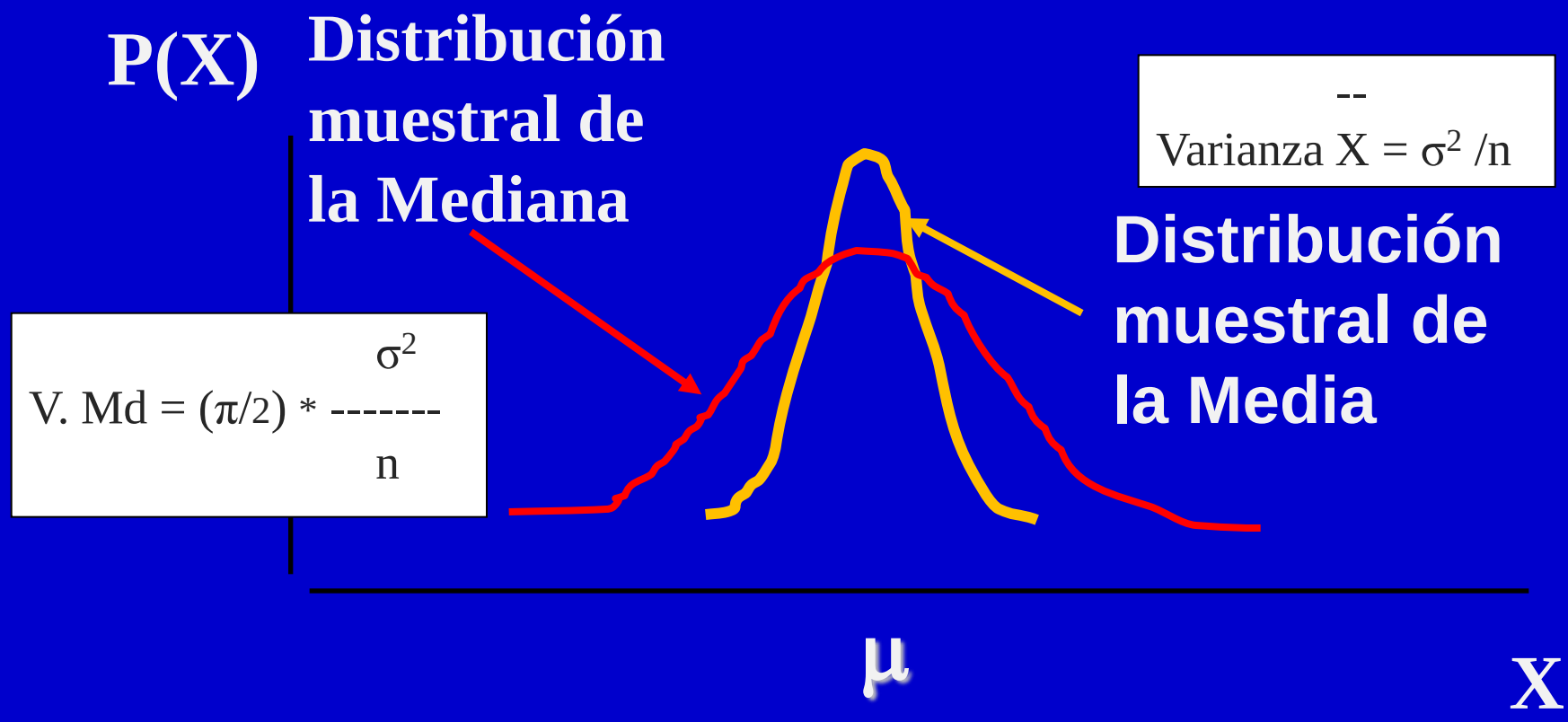
  - **Que tan bueno es el estimador cuando hacemos predicciones.**
  - **Se prefiere aquel estimador que tenga el error medio cuadrático más pequeño alrededor del parámetro**

Insesgado



¿Porqué el estimador en A es insesgado y en B es sesgado?

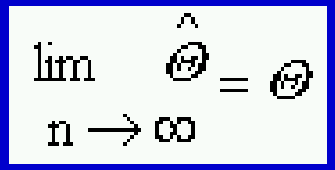
Eficiencia de mediana y media



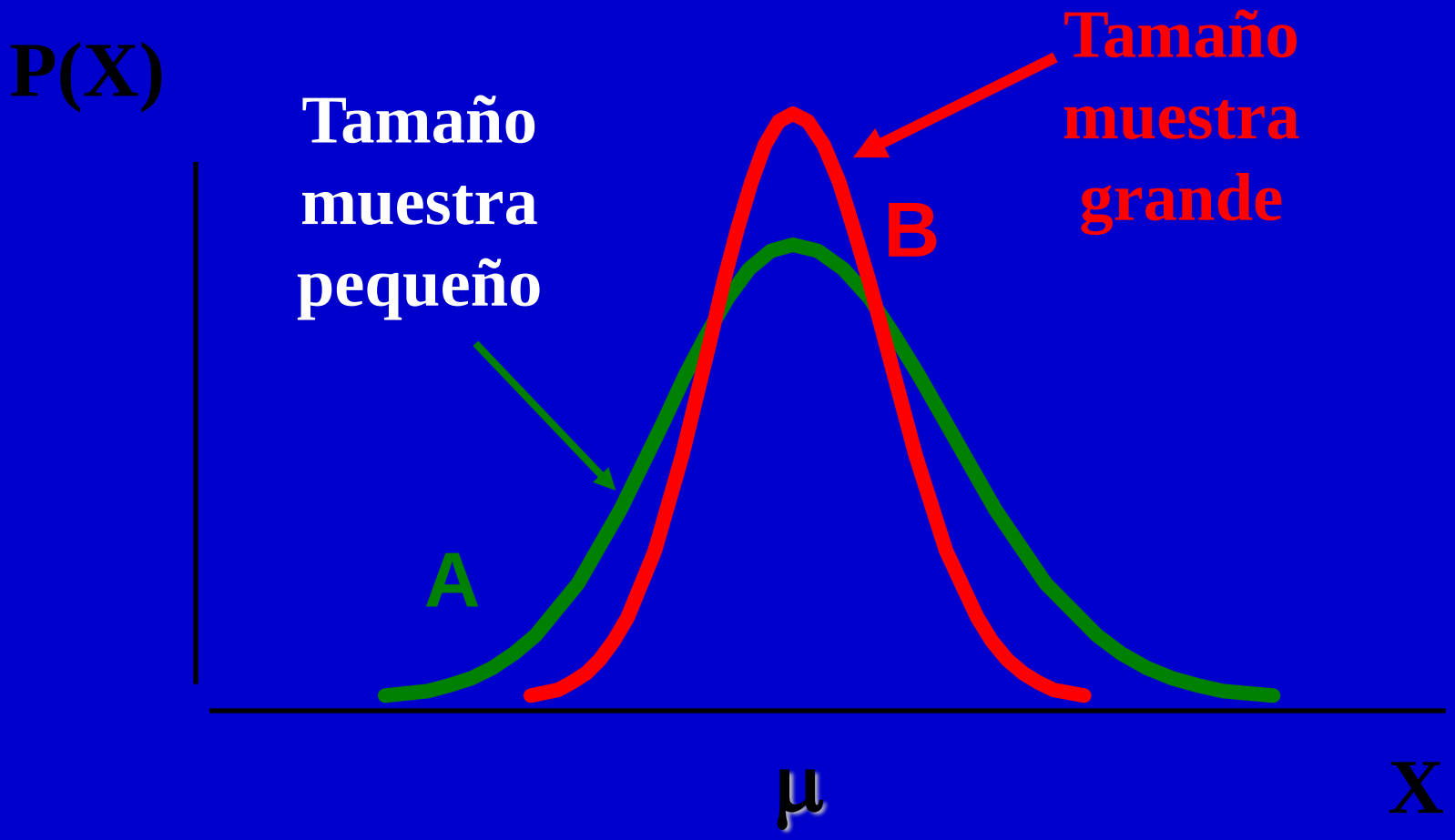
Eficiencia – Mínimo Cuadrado medio del error

Eficiencia de Md con respecto a la media para población normal 157%

Propiedades estimadores

- **Consistencia** 
  - **Al aumentar el tamaño de la muestra, la variación del estimador con respecto al valor del parámetro se reduce**
- **Suficiente**
  - **Todos los datos son utilizados en la estimación**

Consistencia



Estimaciones de μ
Media, Mediana, Rango SIC

Insesgados
Media
Mediana

Sesgado
Rango SIC

Consistente
Media
Mediana

Varianza mínima
Media

Inconsistente
Rango SIC

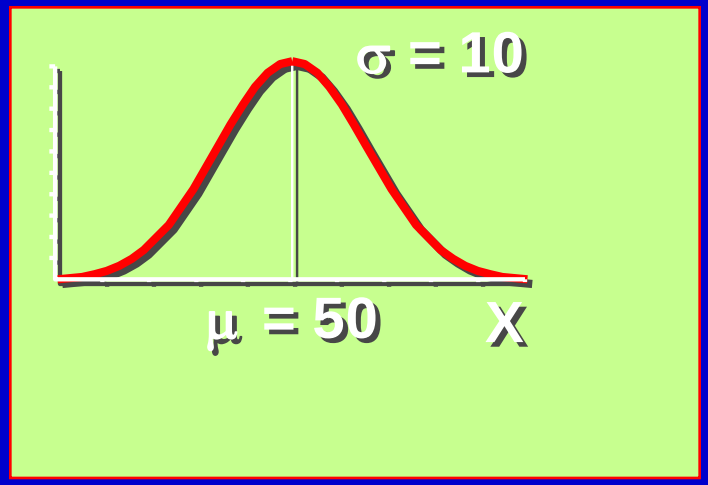
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Error estándar: al aumentar n decrece el error

Media aritmética: estimador insesgado, de varianza mínima y consistente

Población normal → muestra normal

Distribución Población



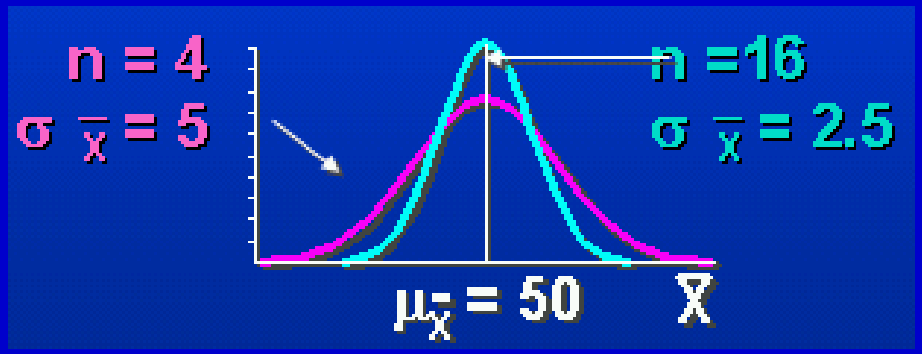
Tendencia Central

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

Variación

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

Distribuciones muestrales



Teorema del límite central

- Conforme el tamaño de muestra incrementa la distribución muestral de los promedios se aproxima una distribución normal con media μ y varianza σ^2/n

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

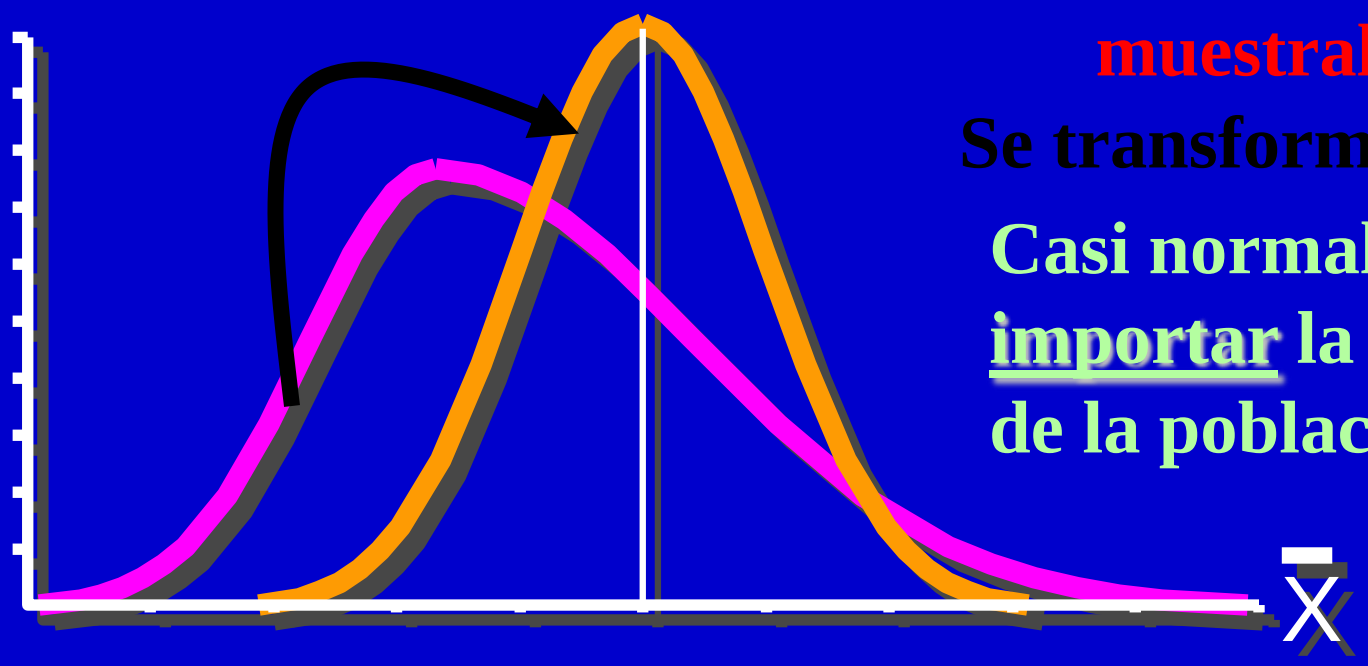
Teorema del límite central

Dado un tamaño de muestra suficientemente grande

La distribución muestral

Se transforma en

Casi normal sin importar la forma de la población

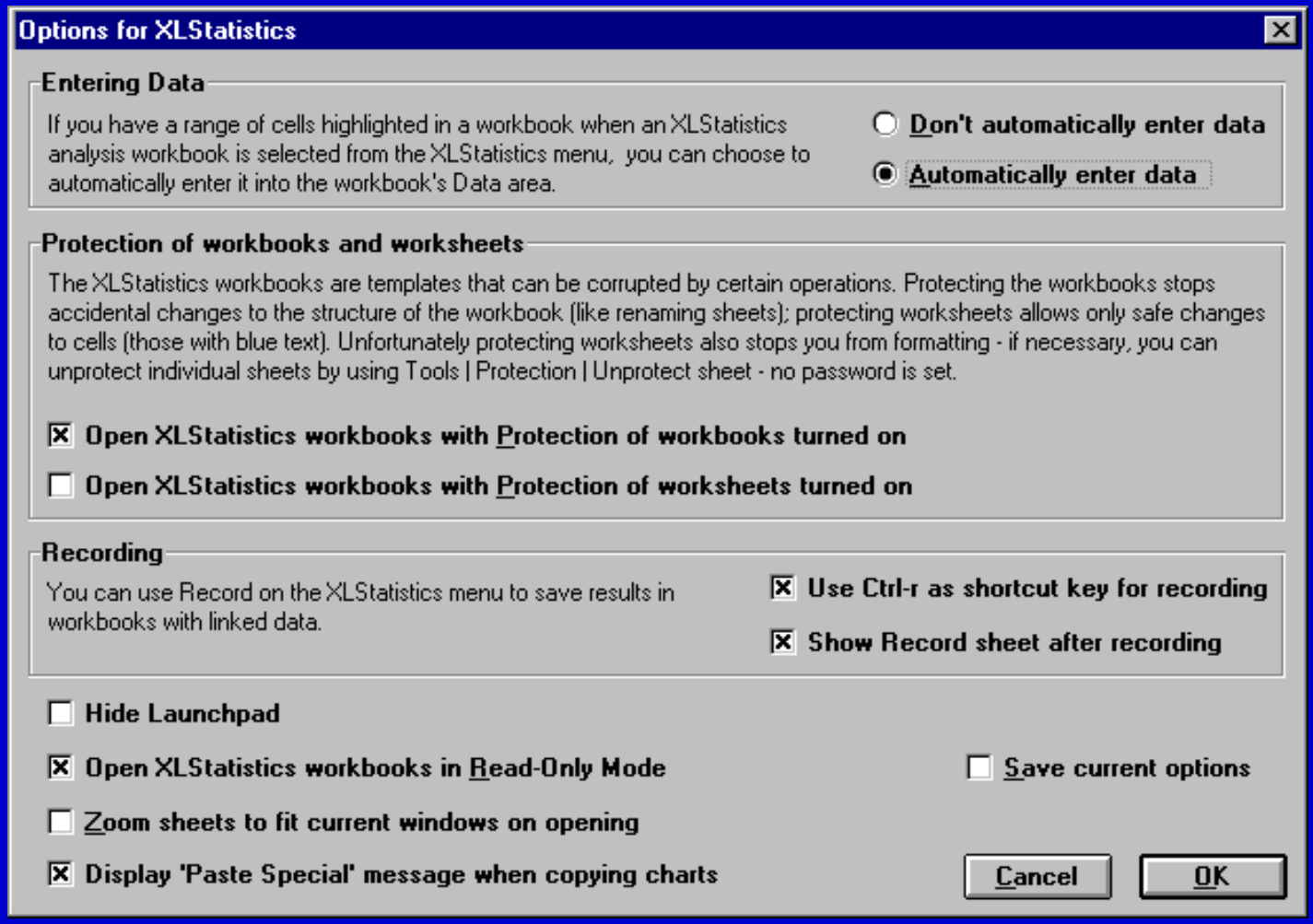


Pruebas de normalidad

- W de Shapiro-Wilk
- Considerada una de las pruebas mas fidedignas o imparciales.
- D de Kolmogorov-Smirnov
- Prueba de bondad de ajuste de la distribución normal
- Lillifors
- Versión modificada de la D de Kolmogorov-Smirnov

- Ho: los datos provienen de una distribución normal
- Ha: los datos no provienen de una distribución normal
- Ver Gráfico de probabilidad normal

XLStat



Para carga el macro haga un doble clic sobre el archivo XLSTAT.xls. XLSTAT es un conjunto de funciones estadísticas y matemáticas que le permiten visualizar y analizar sus datos. Utiliza como interface de entrada y salida las hojas de cálculo Excel.

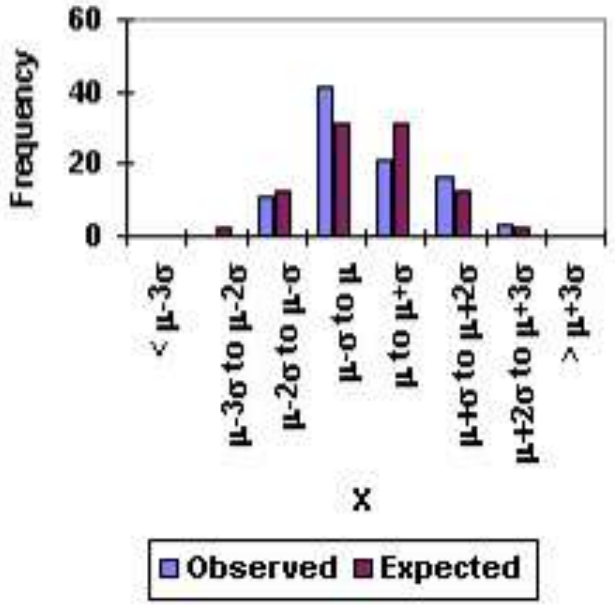
XLStats: prueba normalidad

Goodness-Of-Fit Test for Normality of X

Data

Mean (μ)	96.91304
St Dev (σ)	10.96613

	Proportions		Frequencies	
Range	Observed	Expected	Observed	Expected
< μ-3σ	0	0.00135	0	0.124191
μ-3σ to μ-2σ	0	0.0214	0	1.968822
μ-2σ to μ-σ	0.119565	0.135905	11	12.50327
μ-σ to μ	0.445652	0.341345	41	31.40372
μ to μ+σ	0.228261	0.341345	21	31.40372
μ+σ to μ+2σ	0.173913	0.135905	16	12.50327
μ+2σ to μ+3σ	0.032609	0.0214	3	1.968822
> μ+3σ	0	0.00135	0	0.124191



Hypothesis Test

H_0: Population is normally distributed with the stated Mean and St Dev

H_1: Population is not normally distributed with the stated Mean and St Dev

Chisquare	10.29499
DF	7
p-value =	0.172464

Conclusión: El valor de todos los P calculados es >0.1, por tanto con una confianza 90% se afirma que la variable proviene de una distribución normal

Muestreo probabilístico

- Muestreo probabilístico parte del supuesto de que se conoce cada una de las unidades de muestreo que componen la población y que por lo tanto se pueden listar **todas las posibles muestras**, las cuales se conocen como **espacio o marco muestral**
- Permite evaluar el sesgo y el error estándar de los estimadores así como la eficiencia de diferentes diseños de muestreo.
  - Permite estudiar ventajas, desventajas, limitaciones y alcances de dos o más propuestas de muestreo y seleccionar aquella que mejor se ajuste a las características de la población en estudio

Muestreo al azar y simple al azar

- **Muestra al azar**
  - Muestreo con reemplazo
  - Muestras se obtienen en forma **independiente** de una **población** de unidades **infinitas**. En este esquema de muestreo cada elemento tiene la misma probabilidad de ser seleccionada (P(A)= n/N).
- **Muestra simple al azar**
  - Muestreo sin reemplazo
  - Muestras se obtienen en forma **no independiente** de una **población** de unidades **finitas**. En este esquema de muestreo cada elemento tiene probabilidad P(A/B) = (n-1)/(N-1) de ser seleccionada.
- Muestreo ***con reemplazo*** es ***ineficiente***; ya que el seleccionar un individuo más de una vez no aporta información adicional sobre la población.

Intensidad muestreo

- Dado por n/N en donde n es el número de muestras seleccionadas y N el total de muestras de la población.
- Cuántas muestras debo seleccionar ? ó ¿Qué tan grande debe ser la muestra?

Estimación tamaño muestra

$$n = \frac{t^2 * S^2 * N}{N E^2 + t^2 * S^2}$$

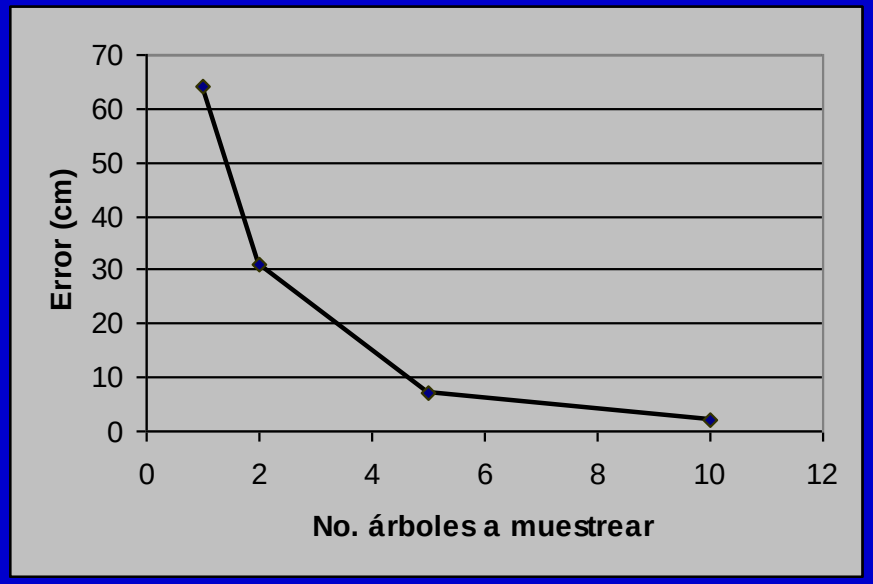
S^2 se puede cambiar por CV%

E: error permisible
S: desviación estándar
t: valor de t de Estudiante para (1- α/2) con n-1 grados de libertad; n > 30 utilizar Z ó 2 como aproximación.
N: Número de muestras o elementos en la población.

Ejemplo

Datos:
E= 2cm N=100 t= 2

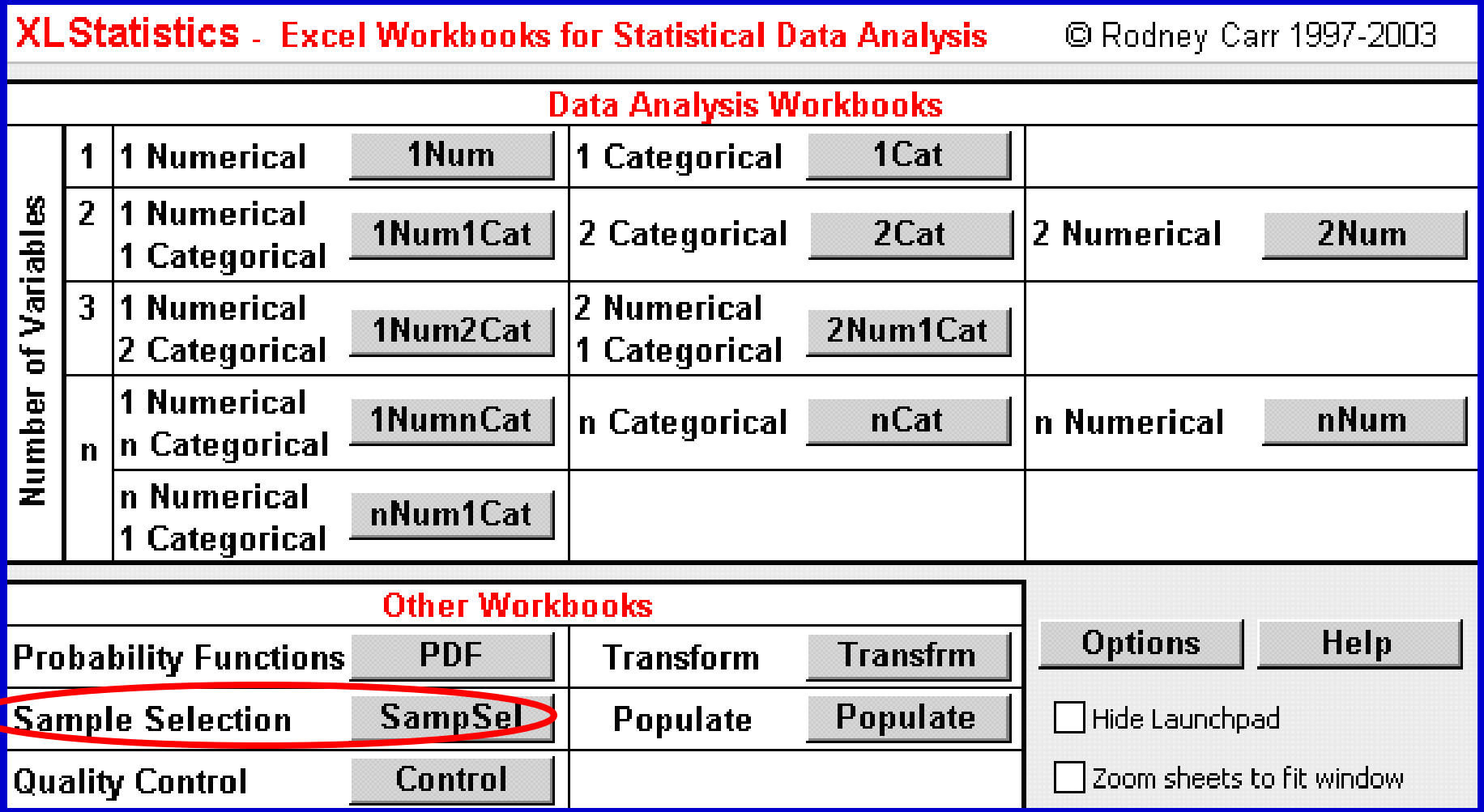
Error (cm)	Tamaño de Muestra
± 1	64
± 2	31
± 5	7
± 10	2



1. **Al menor error mayor amaño de muestra**
2. **A menor variabilidad menor tamaño de muestra para un error dado**

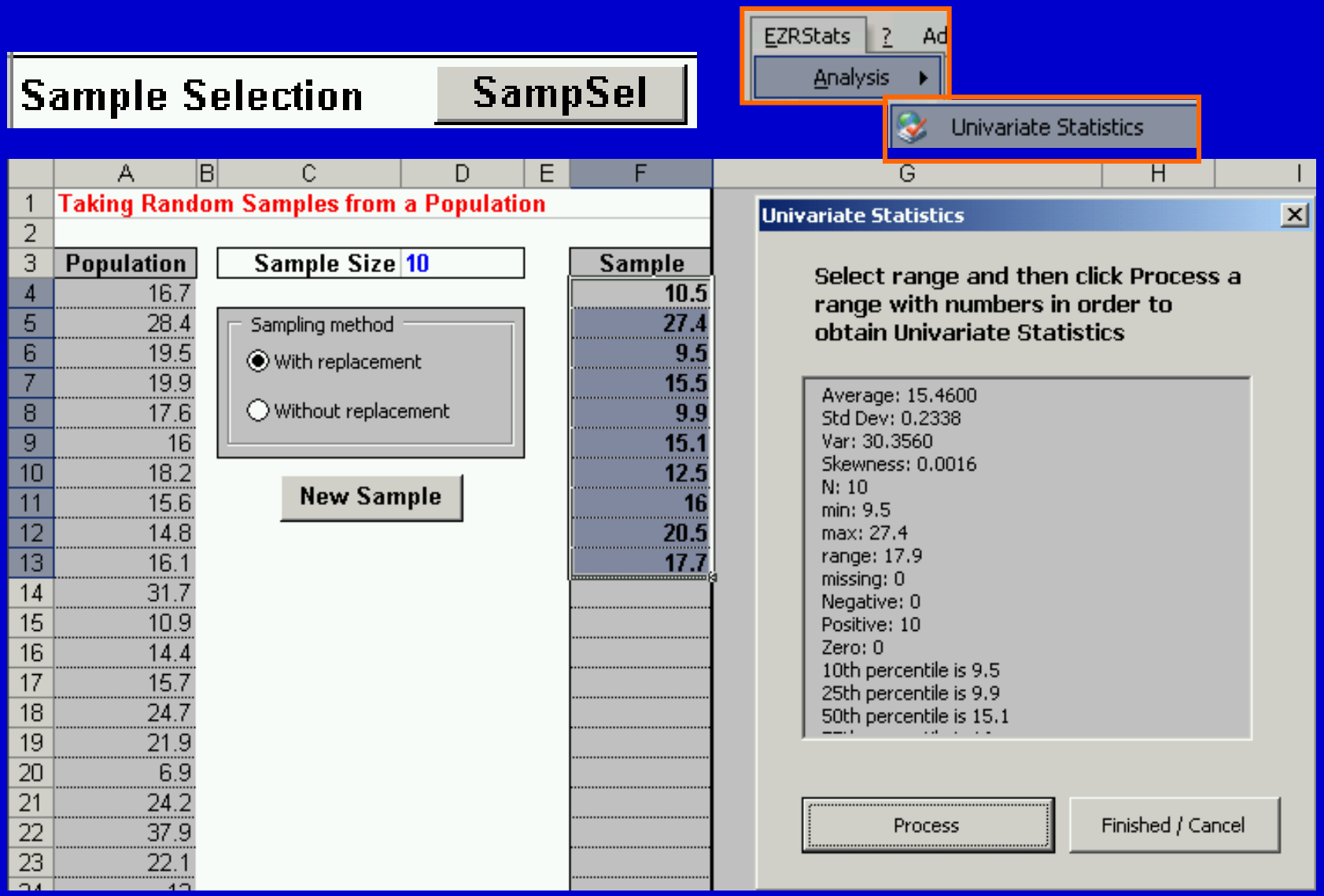
<u>CONCLUSION</u>: TAMAÑO DE MUETRA DEPENDE DE Variabilidad población, Error deseado y nivel de confianza

XLStatistics



Menú general

Muestras al azar: XLStats



Muestreo Estratificado

- Población a muestrear es muy heterogénea pero se puede dividir en estratos homogéneos a su interior pero heterogéneos entre sí.
- Se obtiene una nuestra simple al azar de cada estrato. La muestra agregada se denomina una muestra *aleatoria estratificada*.
- Intensidad de muestreo puede ser diferente para cada estrato y está dado por nh / Nh; en donde nh es el número de elementos seleccionados en el estrato h y Nh es el total de elementos del estrato h.

Ventajas

- Permite utilizar información sobre una o más características de la población para delimitar estratos homogéneos a su interior.
- El muestreo al azar estratificado es más preciso que el muestreo simple al azar siempre y cuando la variabilidad al interior del estrato sea menor comparada con la variabilidad entre estratos.
- El **error de muestreo** está dado por la variabilidad entre unidades de muestreo de cada estrato y no por la variabilidad entre estratos. La ganancia en precisión se manifiesta en una reducción del error de muestreo

Estratos: coberturas

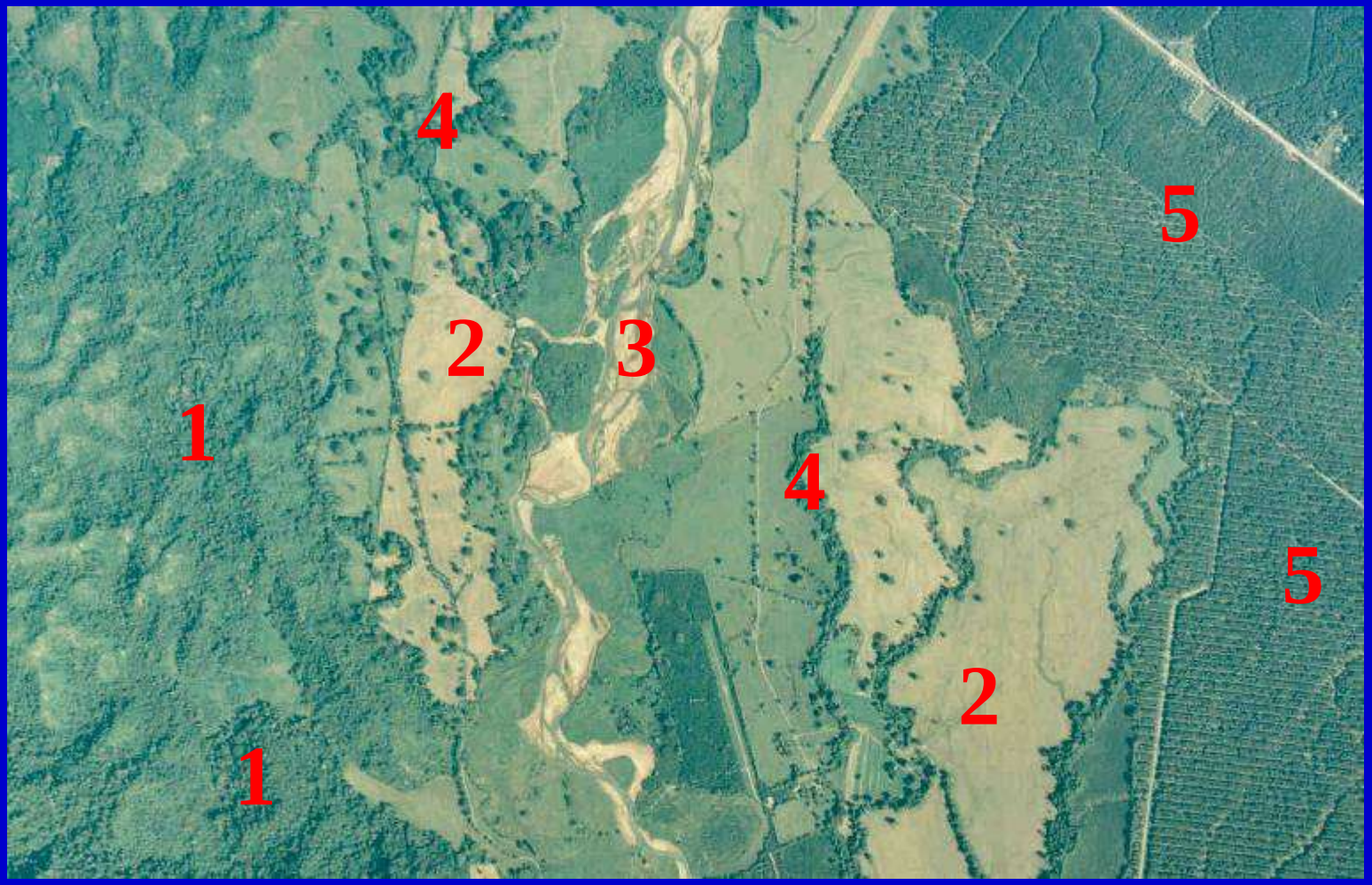
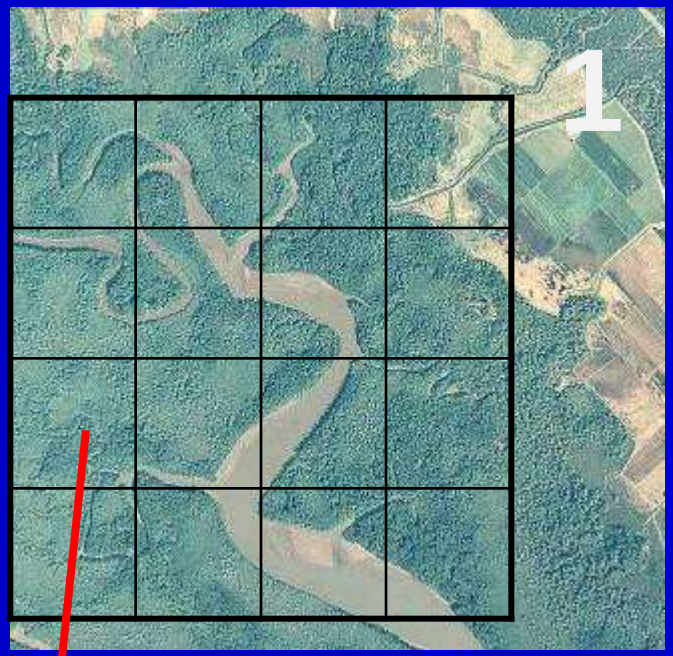


Foto TERRA: 23 dic. 97 Escala 1: 40.000

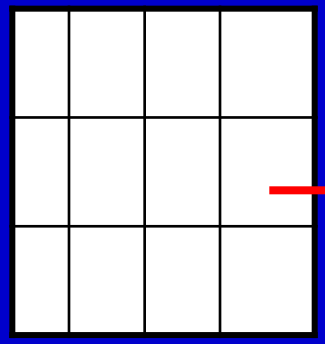
Muestreo Polietápico o Jerárquico

- Las unidades de muestreo de orden uno dan origen a las unidades de orden dos, éstas a su vez a las de orden tres y así sucesivamente hasta llegar al último elemento de la muestra. El diseño de muestreo es apropiado cuando se quiera estudiar una población a diferentes niveles de intensidad o detalle.
- **Ejemplo**: investigar producción de frutos de los aguacatillos. Este esquema de muestreo permite evaluar la variación a nivel nacional, regional, entre localidades en cada región y entre árboles en cada localidad
- El método es excelente para el estudio exhaustivo de poblaciones donde se requiera conocer los patrones de variabilidad natural, como por ejemplo en el área de mejoramiento genético y biosistemática

Ejemplo



Nivel 1: Unidades de orden 1
Nivel 2: Unidades de orden 2
Nivel 3: Unidades de orden 3



Árboles, mantillo

Desventajas

- El tamaño de la muestra es usualmente pequeño en cada nivel o etapa.
- Las estimaciones no alcanzan generalmente la confiabilidad requerida (debido al reducido tamaño de la muestra).
- Se puede aumentar el tamaño de la muestra, aunque por la naturaleza misma del muestreo (jerárquico) esto significa un incremento considerable en los costos.

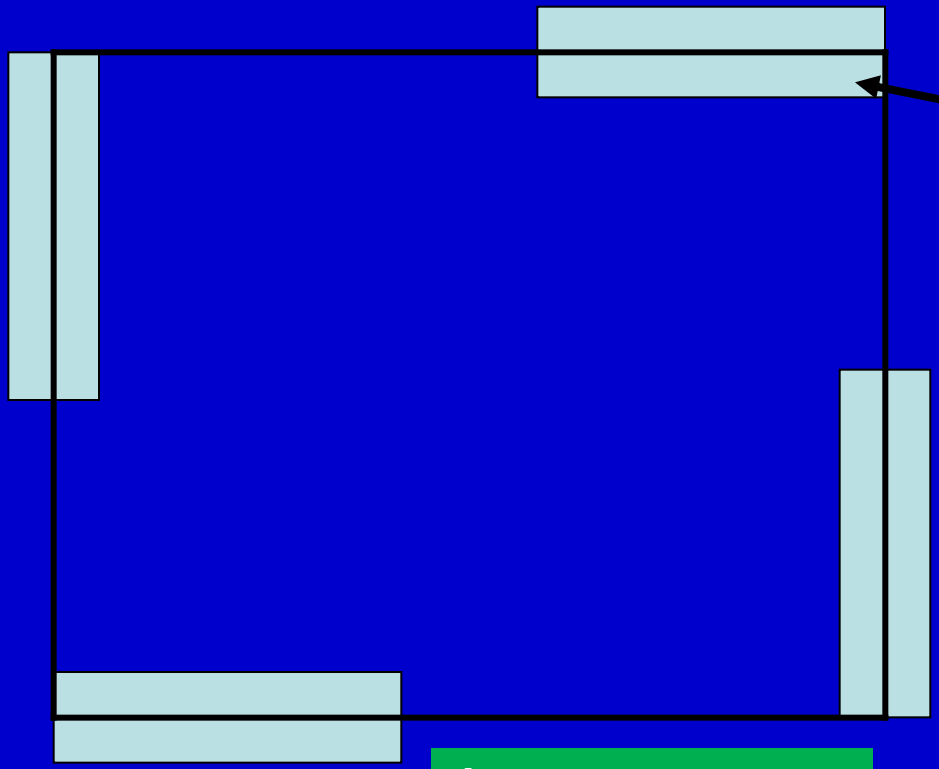
Muestreo Polifásico

- El muestreo polifásico utiliza algunas de las unidades en diferentes fases del muestreo y se utiliza cuando se requiere levantar datos a diferentes niveles de una población.
- Por ejemplo, el muestreo se denomina doble cuando se trabaja con dos fases y triple cuando se utilizan tres fases.
- Las fases pueden ser independientes o dependientes. Un ejemplo de muestreo triple es un inventario de cobertura que utiliza fotos aéreas en el levantamiento de vegetación (fase uno), parcelas en la fase de campo (fase dos) y árboles individuales para estimar producción de flores y frutos (fase tres).

Muestreo Conglomerados

- Unidades de muestreo se establecen en grupos o conjuntos.
- Cada grupo constituye un conglomerado.
- El método es especialmente útil para situaciones donde es difícil o muy costoso obtener una muestra independiente como por ejemplo en áreas boscosas de topografía escarpada y con escasas vías de comunicación.

Ejemplo: conglomerado



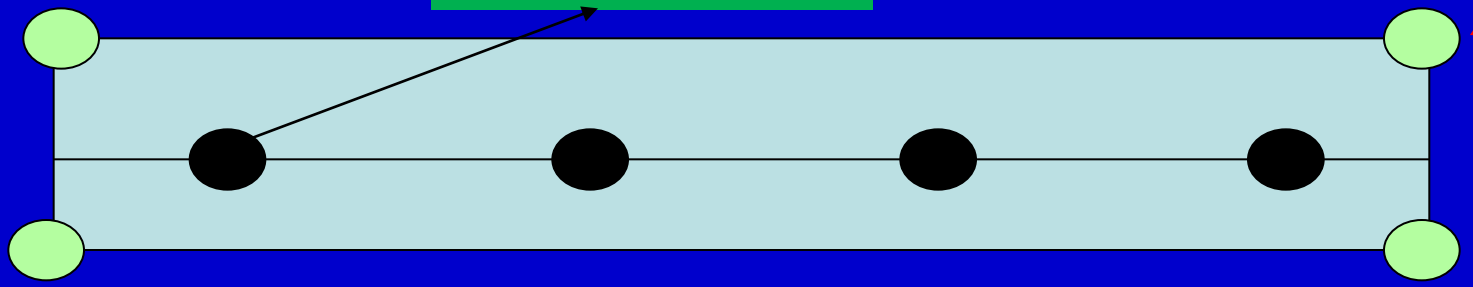
Área de cada subparcela
150 m *20m; árboles >30cm

Conglomerado: 4 subparc.
Área tota: 1.2ha

Se calculan estadísticos para el conglomerado

árboles 10-30 cm

Regeneración



Muestreo sistemático

- Muestreo no probabilístico
- Requiere que el investigador (a) conozca la totalidad de los elementos que conforman la población.
- Seleccionar primera unidad de muestreo al azar y las remanentes siguiendo un patrón fijo y constante; por ejemplo, cada décima unidad.
- **Desventaja**: no permite calcular ni el error de muestreo ni el intervalo de confianza asociado a las estimaciones realizadas.

Ejemplo

	1	2	3	4	5	6	7	8	8	10
1	6,7	28,4	19,5	19,9	17,6	16,0	18,2	15,6	14,8	16,1
2	31,7	10,9	14,4	15,7	24,7	21,9	6,9	24,2	37,9	22,1
3	12,0	14,8	18,1	18,0	9,1	13,7	17,0	18,2	23,4	6,8
4	15,7	17,5	20,7	13,7	16,1	21,5	10,5	10,1	17,5	16,4
5	17,8	13,2	21,3	16,6	7,2	14,0	14,6	22,4	15,1	13,9
6	17,6	21,4	19,0	12,5	7,5	20,5	22,8	14,8	17,7	37,0
7	17,6	10,2	12,6	9,9	19,8	15,5	13,7	14,4	22,3	7,9
8	18,2	27,4	22,0	9,5	15,2	11,6	22,8	22,7	17,6	29,9
9	11,0	11,8	18,8	20,6	9,7	21,5	10,6	16,2	16,1	24,1
10	37,7	11,9	18,4	11,5	22,4	23,9	18,9	19,6	18,5	6,5

Población de observaciones

Selección de 1 en cinco con inicio aleatorio en la observación 2

Muestra: 28.14, 10.8. 14.8, 17.5, 13.2, 21.4.....

Un sitio adecuado para la practica de muestreo!!!!

Jorge Fallas
jfallas56@gmail.com