BASES ESTADÍSTICAS EN LA ESTIMACIÓN Y LA PROYECCIÓN PRODUCTIVA

M.Sc. Fabio A. Blanco Rojas

Berufweiss S.A.

“Porque la tierra será llena del conocimiento de Jehová, como las aguas cubren el mar”

Isaías 11:9

Temas a considerar

- **Principios de muestreo**
- Métodos de proyección del rendimiento
- Estimación

DISEÑOS DE MUESTREO

- Los diferentes métodos para seleccionar la muestra son llamados diseños de muestreo.
- El objetivo principal de un diseño de muestreo es proporcionar indicaciones para la selección de una muestra que sea representativa de la población bajo estudio, proporcionando información a un costo mínimo.

Términos técnicos

- Elemento: Es un objeto en el que se toman mediciones. En general corresponde a individuos, plantas, animales, etc.
- Población de inferencia: Es aquella colección de elementos acerca de las cuales se aplican los resultados del estudio.
- Población objetivo: Aquella de la cual se tomará la muestra.

Términos técnicos

- Unidades de muestreo UM: Colecciones no traslapadas de elementos que abarcan completamente la población objetivo.
  - Ej. Personas, animales, plantas, parcelas, etc.
- En un muestreo del área sembrada de caña de azúcar, las UM constituyen áreas de cierta extensión. Definir la UM es el paso más difícil del estudio de muestreo junto con la preparación del marco muestral.

Términos técnicos

- Marco muestral: Lista de todas las unidades de muestreo que constituyen la población objetivo.
- Muestra: Conjunto de UM del marco muestral seleccionadas para su estudio.

Mapa de toda el área sembrada

- Lotes o parcelas de caña de azúcar identificadas
  - Ubicación y área
  - Actualizarse periódicamente.
- Mediciones de área hechas con teodolito, fotogrametría o imágenes satelitales.
- Mapa representa la población de inferencia.
- Sobre él puede definirse la población objetivo.
- Restar las áreas no sembradas (caminos, ríos, etc.).
- Areas de riesgo (generalmente inundaciones) .

Mapa de toda el área sembrada

- Es la base para el marco muestral
- Definir la UM es tarea difícil. Pueden ser útiles:
  - El área total de la población de inferencia,
  - los tamaños de los lotes,
  - otros aspectos de índole práctica.
- Clasificar los lotes según su potencial productivo actual es básico para estimar y estratificar
- Diseño de muestreo puede considerar dos o más puntos de muestreo en cada UM.

Área cosechada Vs área sembrada

- Algunas áreas sembradas puede estar destinadas a fines distintos de la producción de azúcar.
- Puede ser que sucedan eventos agro-climáticos (pestes, inundaciones, etc.) que reduzcan el área a cosechar.

Métodos de selección de la muestra:

- **Aleatoria o al azar:** el más recomendado. Cada elemento de la población tiene una probabilidad conocida de ser incluido en la muestra.
- **Intencional o de juicio:** se debe conocer bien la población, se producen sesgos de muestreo. Sesgo: error sistemático, no son cuantificables.
- **Por conveniencia:** se escogen las unidades estadísticas más disponibles o más fáciles de conseguir.

Algunos diseños de muestreo aleatorio

a. Simple aleatorio
b. Al azar estratificado
c. Sistemático
d. Al azar en varias etapas (muestreo por conglomerados)

MUESTREO SIMPLE AL AZAR

- Se caracteriza porque cada posible muestra, de un tamaño especificado, tiene la misma probabilidad de ser seleccionada.
- Se puede usar cuando:
  a. El marco muestral explicita la lista de las unidades de muestreo.
  b. Las unidades de muestreo se identifican por localización en un lugar del espacio o del tiempo.

MUESTREO ESTRATIFICADO AL AZAR

- Estratos: Se clasifican los elementos de la población en grupos mutuamente excluyentes, llamados estratos o subpoblaciones.
- Se estudia una muestra irrestrictamente aleatoria en cada estrato.

MUESTREO ESTRATIFICADO AL AZAR

- Se puede usar si:
  - Es posible identificar subpoblaciones con promedios diferentes.
  - Hay interés en estudiar las subpoblaciones.
  - Por razones logísticas o de costo es más conveniente dividir la población en grupos más manejables.

El principio de estratificación

- Para maximizar la precisión del estimador del parámetro de interés:
  a) Los estratos deben tener medias tan diferentes como sea posible.
  b) Los estratos deben ser internamente tan homogéneos (varianzas bajas) como sea posible.

Temas a considerar

- Principios de muestreo
- **Métodos de proyección del rendimiento**
- Estimación

MÉTODOS DE PROYECCIÓN DEL RENDIMIENTO

- Modelos empíricos de regresión.
- r Modelos de simulacion
- M Métodos no paramétricos

Modelos de regresión

Ejemplos de proyecciones:

- **No. tallos a la cosecha = A + B * (No. tallos antes de la cosecha)**
- **Peso a la cosecha = A + B (altura antes de la cosecha)**
- **Mediciones se hacen:**
  - **A la cosecha.**
  - **Antes de la cosecha**

Mediciones a la cosecha

- Peso y número de tallos cosechados .
  - Puede haber una diferencia entre el peso de caña cosechada y lo que efectivamente se procesa.

    Rendimiento Neto= Rendimiento Bruto – Pérdidas
- Brix.
- Rendimiento de azúcar.
  - Ecuación que relacione grados brix y azúcar realmente extraído.
- Cualquier otro dato de interés.

Mediciones antes de la cosecha

- mediciones no destructivas y conteos en diferentes puntos del desarrollo, anteriores a la cosecha, que sean de fácil identificación.

Sede Central: Av. La Molina N° 1915 - Lima 12 - La Molina, Lima-Perú
Central Telefónica: (051)-(1)-313 3300 Fax: (051)-(1)-313 3315

Proyecciones con datos históricos y con datos del periodo actual

- Promedios históricos antes de que se disponga de mediciones actuales.
- Una vez que se tienen datos actuales, se aplican a modelos de años previos.
  - Ejemplo:
    - No. tallos a la cosecha = A + B * (No. tallos actuales antes de la cosecha)
    - Los parámetros A y B se toman de los años anteriores (3 a 5 años previos).
- Tales ecuaciones deben desarrollarse para cada punto del desarrollo identificable.

Modelos de simulación

- Para cada UM seleccionada en la muestra se colectan datos de variables de clima, suelo, variedad, manejo, etc. para representar el crecimiento de una planta mediante un modelo que converja con precisión al rendimiento al final del ciclo de crecimiento. Se emplea un simulador estocástico de tiempo para simular el tiempo climático del ciclo de cultivo.
- Ventaja: Depende de información actual y no de años anteriores.
- Puede modelarse bajo escenarios diferentes: desfavorable, normal, favorable.

Desarrollo de indicadores de producción

Escala usada por USDA para cereales:

2. Muy pobre: Pérdida extrema de potencial productivo, pérdida total o casi total.
3. Pobre: Pérdida fuerte del potencial productivo, debido a exceso de humedad, sequía, pestes, etc.
4. Regular: Inferior a la condición normal. Posible bajo rendimiento.
5. Bueno: Expectativa de rendimiento normal. Niveles de humedad son adecuados y daños causados por insectos y enfermedades y presión de malezas son menores.
6. Excelente: Expectativas de rendimiento están sobre lo normal. El cultivo experimenta poco estrés o ninguno. Daños causados por insectos y enfermedades y presión de malezas son insignificantes.

Fackler, P. L., and B. Norwood. 1999. "Forecasting Crop Yields and Condition Indices." Proceedings of the NCR-134 Conference on Applied Commodity Price Analysis, Forecasting, and Market Risk Management. Chicago, IL. [http://www.farmdoc.uiuc.edu/nccc134].

Proyección basada en indicadores de rendimiento

- Suponiendo que cada clase tiene un rendimiento promedio yi; i= 1, . . .,5.
- Y la fracción de área cultivada de cada clase es respectivamente C1, . . ., C5, de modo que C1+ . . . +C5=1.
- El rendimiento promedio se calcula como:

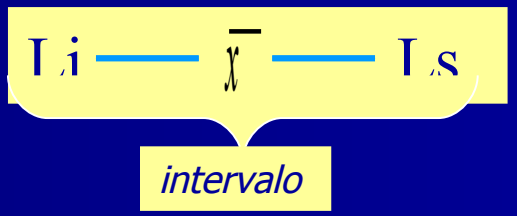
$$\text{Rendimiento promedio} = \sum_{i=1}^{5} y_i c_i$$

Temas a considerar

- Principios de muestreo
- Métodos de proyección del rendimiento
- **Estimación**

Estimación de la media de una población mediante un intervalo de confianza

$$\bar{x} \pm Z_{1-\alpha} * \sigma_{\bar{x}}$$



$$\bar{x} \pm Z_{1-\alpha} * \frac{\sigma}{\sqrt{n}}$$

$1-\alpha$ = Coeficiente de confianza

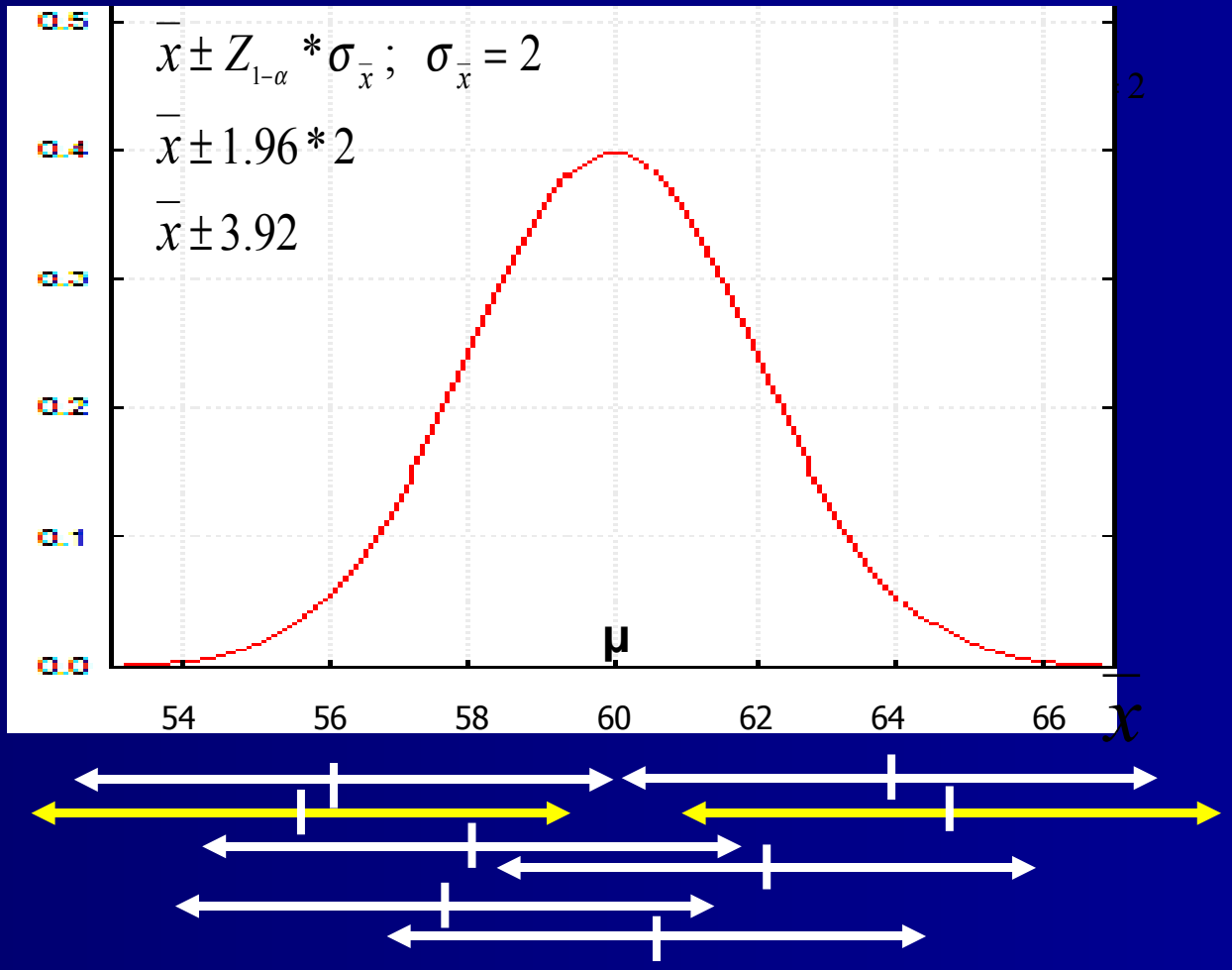
$Z_{1-\alpha}$ = Valor de la variable normal estándar

$\sigma_{\bar{x}}$ = Error estándar de una media

σ = Desviación estándar de población

n = Numero de elementos en la muestra.

Intervalos de 95% confianza de que contengan a la media verdadera de la población μ.
μ=60, σ=10, n=25.



Interpretacion del intervalo 95% de confianza para estimar μ

- "Se tiene 95% de confianza de que la media verdadera está dentro del IC"

- "Si la misma población se muestrea repetidamente y en cada ocasión se estima un IC95%, entonces aproximadamente el 95% de los IC así calculados contienen la verdadera media de la población. "

Estimación por intervalo para muestras grandes o σ conocido:

Coef. de confianza $(1-\alpha)$	α	$Z_{\alpha/2}$	Lim. Inferior	Lim. superior
0.90	0.10	1.645	$\bar{x} - 1.645 * \frac{\sigma}{\sqrt{n}}$	$\bar{x} + 1.645 * \frac{\sigma}{\sqrt{n}}$
0.95	0.05	1.96	$\bar{x} - 1.96 * \frac{\sigma}{\sqrt{n}}$	$\bar{x} + 1.96 * \frac{\sigma}{\sqrt{n}}$
0.99	0.01	2.58	$\bar{x} - 2.58 * \frac{\sigma}{\sqrt{n}}$	$\bar{x} - 2.58 * \frac{\sigma}{\sqrt{n}}$

Desviación estándar y error estándar de una media

- La desviación estándar (σ) es una medida de la variación de los elementos (x) de una población.
- El error estándar de la media $\sigma_{\bar{x}}$ describe la variación de las medias de muestras de n elementos. Una vez estimado a partir de una muestra, constituye una medida de la precisión de la media estimada.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}, \text{ depende de } \sigma \text{ y de n}$$

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}, \text{ error estándar estimado}$$

de una media

Ejercicio: calcule e interprete un IC 95% para μ, usando Z con los siguientes datos

33.0	30.0	27.7
22.2	31.6	36.8
31.0	35.8	19.4
40.2	44.8	35.2
34.6	38.8	48.0
24.6	33.6	32.6
41.8	27.6	26.0
23.2	25.6	31.4
34.0	26.3	28.2
33.4	27.8	48.0

n=30
media=32.4
s=7.23

Solución: $e=1.96*7.23/\sqrt{30}=2.6$

$Li=32.4-2.6=29.8$

$Ls=32.4+2.6=35.0$

Interpretación: Se tiene una confianza de 95% de que la media de la población esté entre 29.8 y 35.0.

Fundamento: El 95% de las muestras aleatorias de tamaño n=30, producen intervalos de 95% de confianza que contienen la media de población.

Cálculo de un tamaño de muestra n adecuado

Para diseño de muestreo simple al azar:

$$\bar{x} \pm Z_{1-\alpha} * \frac{\sigma}{\sqrt{n}};$$

$$\bar{x} \pm e$$

$$e = Z_{1-\alpha} * \frac{\sigma}{\sqrt{n}}$$

$$n = \left(Z_{1-\alpha} * \frac{\sigma}{e} \right)^2$$

Ejercicio:

Calcule el n necesario para obtener un IC 95% para μ, de modo que el error de estimación sea de 2.0 o menos. La desviación estándar *s* obtenida de una muestra de 40 elementos fue s=10. El diseño de muestreo que se usará es simple al azar.

Solución:

$$n = \left(Z_{1-\alpha} * \frac{\sigma}{e} \right)^2$$

$$n = (1.96*10/2)^2 = 96,$$

Tamaño de muestra *n* para poblaciones finitas

- Si la población es finita, fracción de muestreo es moderada a grande (digamos mayor de 15%), y se muestrea sin remplazo, se requiere un paso adicional para determinar el tamaño final de muestra:

$$n_{final} = \frac{n}{1 + n/N}, \quad donde:$$

$n = tamaño\ de\ muestra\ estimado\ para\ poblaciones$ infi*nitas*.

$N = Tamaño\ de\ la\ población$

$$\text{Si } N = 500 \rightarrow n_{final} = \frac{96}{1 + 96/500} = 81$$

Otras fórmulas para *n*, diseño simple al azar

$$n = \left[\frac{z_{1-\alpha}\,\sigma}{r\,\bar{x}} \right]^2$$

$$n = \left[\frac{z_{1-\alpha}\,CV}{r} \right]^2$$

En ambos casos *e* equivale a una fracción de la media, ej. si $r = 0.05$ entonces $e = 5\%$ de la media

Además, a partir de un valor de $\sigma_{\bar{x}}$:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}},\ n = \frac{\sigma^2}{\sigma_{\bar{x}}^2}.$$

Intervalo de confianza de proporciones binomiales en poblaciones infinitas

Los límites inferior y superior de un IC para estimar p quedan determinados por los resultados de la expresión siguiente:

$$\bar{p} \pm Z_{\alpha/2} * \sqrt{\frac{\bar{p} * (1 - \bar{p})}{n - 1}}$$

Y el tamaño de muestra se obtiene por la fórmula:

$$n = [Z_{\alpha/2} / e]^2 * p*(1-p)$$

Lo anterior funciona si p no es muy pequeño (Ej. P>0.05)

Ejercicio: calcule un tamaño de muestra para estimar el porcentaje de racimos con dedo guápil, con un IC de 95%, de forma que la amplitud máxima sea e=3%, suponiendo que *p* es de alrededor de 5%.

Solución:

Fórmula: $n = [Z_{\alpha/2} / e]^2 * p*(1-p)$

$n=[1.96/0.03]^2*0.05*0.95=203$

Si la población fuera de solo N=500,

$$n_{final} = \frac{n}{1+n/N} = \frac{203}{1+203/500} = 144$$

Estimación con diseño de muestreo estratificado al azar

$$\bar{y} \pm Z_{1-\alpha} * \sigma_{\bar{y}}$$

$$\bar{Y}_{st} = \frac{N_1 \bar{y}_1 + N_2 \bar{y}_2 + \ldots + N_L \bar{y}_L}{N}$$

$\sigma^2_{\bar{y}}$ se calcula como :

$$V(\bar{y}_{st}) = \frac{1}{N^2} * \sum_{i=1}^{L} \left[N_i^2 (N_i - n_i) / N_i * (\frac{s_i^2}{n_i}) \right]$$

L= número de estratos.
Ni= tamaño del estrato i-mo.
N= tamaño de la población;
N1+N2+. . .+NL=N
*n*i= tamaño de muestra del estrato *i-mo*
S_i^2= varianza estimada en el estrato *i-mo*.

Cálculo de n para diseño estratificado, para un valor *e* dado.

$$n = \frac{\sum_{i=1}^{L} N_i^2 \sigma_i^2 \Big/ w_i}{N^2 D + \sum_{i=1}^{L} N_i \sigma_i^2},$$

en donde :

$N =$ número de unidades de muestreo de la población.

$N_i =$ Número de unidades de muestreo del estrato i - mo.

$L =$ Número de estratos.

$\sigma_i^2 =$ Varianza en el estrato i - mo.

$w_i =$ Fracción de unidades de muestreo asignadas al estrato i - mo.

$D = \frac{e^2}{Z^2}$, si $Z = 1.96$, Z^2 aprox. 4.

Ejemplo de cálculo de n, diseño estratificado, valor *e* dado.

- Se quiere que el error de estimación *e* no exceda 2 unidades.
- Fracción de muestra para cada estrato sea $w_i = 1/3$.
- Información de cada estrato:

ESTRATO	N_i	S_i^2	w_i
1	150	200	1/3
2	90	100	1/3
3	60	50	1/3
N	300		

Ejemplo de cálculo de n, en diseño estratificado valor *e* dado.

$$n = \frac{\sum_{i=1}^{L} N_i^2 \sigma_i^2 \Big/ w_i}{N^2 D + \sum_{i=1}^{L} N_i \sigma_i^2},$$

$\sum_{i=1}^{L} N_i^2 \sigma_i^2 \Big/ w_i$ $=150^2*200/(1/3)+90^2*100/(1/3)+60^2*50/(1/3)$
$=16470000.$

$\sum_{i=1}^{L} N_i \sigma_i^2$ $=150*200+90*100+30*50=42000$

$N^2 D$ $=300^2*2^2/4=90000$; aquí Z=2, e =2.

n=16470000/(90000+42000)=125

Para cada estrato, n_i= 125/3,prox. 42

Criterios para repartir la muestra total entre los estratos

1. El número total de elementos en cada estrato. Estratos más numerosos deben tener más representación.
2. La variabilidad de los elementos en cada estrato. Muestras mayores se requieren cuando la variabilidad es mayor.
3. El costo de obtener una observación en cada estrato. Muestras más caras deben ser menos numerosas.

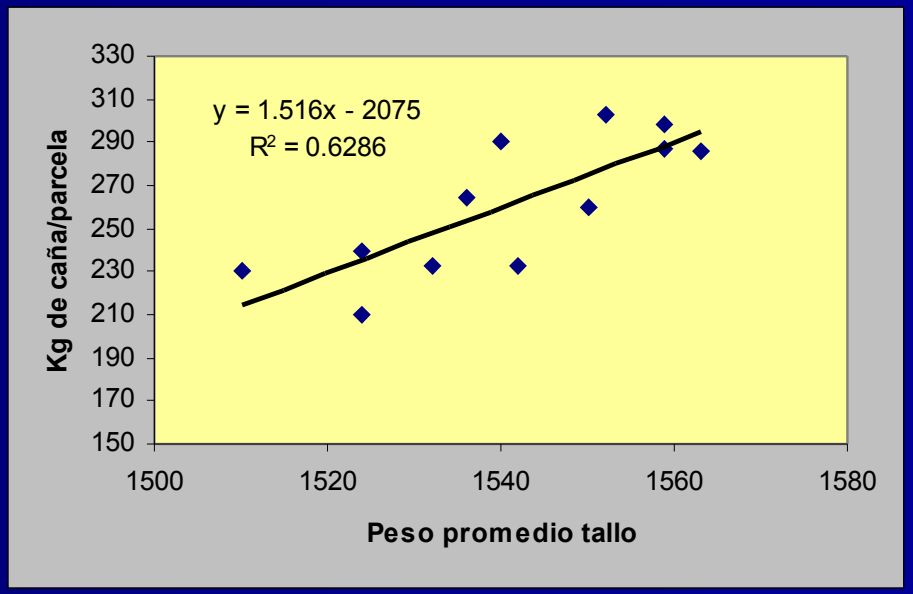
Tres métodos de definir las fracciones de muestra para los estratos

Fracción de muestra $f_i = \frac{n_i}{n}$	
Minimizando el costo para un valor fijo de o minimizando para un costo fijo	$f_i = \frac{N_i \sigma_i / \sqrt{c_i}}{\sum_{i=1}^{L} (N_i \sigma_i / \sqrt{c_i})}$
Asumiendo iguales costos por observación en todos los estratos (asignación de Neyman)	$f_i = \frac{N_i \sigma_i}{\sum_{k=1}^{L} N_i \sigma_i}$
Costos iguales, varianzas iguales en todos los estratos	$f_i = \frac{N_i}{N}$

Estimación en regresión

Ejercicio: ¿Estime la producción media mediante un IC 95%?

	Peso Prom. 10 tallos (g)	Producc. Parcela Kg
	1559	298
	1552	303
	1542	233
	1559	287
	1524	210
	1510	230
	1540	290
	1532	233
	1563	286
	1550	260
	1536	264
	1524	239
Media=	1540.9	261.1



ANOVA

	df	*SS*	*MS*	*F*	*Pr<F*
Regressior	1	6850.951	6850.951	16.92	0.0021
Residual	10	4047.966	**404.7966**		
Total	11	10898.92			

Estimación en regresión

Media estimada de y es $\hat{\mu}_y = 261.1$

Si población es grande :

$$\sigma^2_{\hat{\mu}_y} = \frac{\sigma^2_y}{n} = \frac{404.8}{12} = 33.7 ,$$

$$\text{IC}\,95\% \rightarrow \bar{y} \pm 1.96\,\sigma_{\hat{\mu}_y}$$

$$\text{IC}\,95\% \rightarrow 261.1 \pm 1.96\,\frac{\sqrt{33.7}}{\sqrt{12}}$$

$$\text{IC}\,95\% \rightarrow 261.1 \pm 3.28$$

Si población es pequeña :

$$\sigma^2_{\hat{\mu}_y} = \frac{\sigma^2_y}{n}\left(\frac{N-n}{N}\right)$$

Gracias