



Seminario sobre Técnicas de Muestreo

Junta de Planificación de Puerto Rico
Programa de Planificación Económica y
Social

Dr. Julio Quintana Díaz
Septiembre 2013

Objetivos

Al finalizar el seminario los participantes:

- ❖ Reconocerán la importancia de seleccionar científicamente una muestra.
- ❖ Se familiarizarán con la terminología estadística básica asociada a los procesos de muestreo.
- ❖ Podrán diferenciar entre muestras probabilísticas y no probabilísticas.
- ❖ Podrán definir los métodos principales de muestreo y determinar cuando es apropiada la aplicación de cada uno.

Validez , confiabilidad, precisión y representatividad

- La **validez** de un instrumento de medida es el grado en el cual dicho instrumento efectivamente mide lo que reclama que mide.
- La **confiabilidad** busca que la aplicación repetida de un instrumento provea siempre resultados consistentes.
- La **precisión** busca minimizar la diferencia entre un valor característico de una variable y su estimado.

Validez , confiabilidad, precisión y representatividad

- La **representatividad** de una muestra no tiene que ver con su tamaño, sino con la capacidad que tenga de reproducir en pequeña escala las características de la población.
- Un mito muy generalizado es que mientras mayor sea el tamaño de una muestra, mayor es su representatividad.

Muestras grandes no son necesariamente representativas

- La encuesta Gallup para las elecciones presidenciales de EU en 1948 fue de 3,250 votantes y falló en predecir el ganador, Truman, porque se utilizó muestreo por cuota. Actualmente con una muestra de 1,500 votantes, se ha logrado pronosticar en las últimas 17 elecciones quién será el ganador, con un margen de error promedio de 1.5%. Lo importante es la **ALEATORIEDAD** y un **DISEÑO MUESTRAL APROPIADO**.

Muestras de probabilidad

- Son las que requieren el uso de procedimientos aleatorios (tablas de dígitos aleatorios y generación computadorizada o electrónica) para la selección de la muestra. Esto permite que las propiedades de los estimados puedan evaluarse probabilísticamente. Por ejemplo, hacer una afirmaciones como la siguiente: “con un 95% de confianza, se estima que el 5% de los individuos que son más ricos en PR tienen casi el 25% de los ingresos” (Fuente: PR Community Survey 2009. Ver Tabla)

Ejemplo de inferencias a partir de una muestra probabilística

	United States		Puerto Rico	
	Estimate	Margin of Error	Estimate	Margin of Error
Quintile Share of Aggregate Income:				
Lowest Quintile	3.3	+/-0.01	1.7	+/-0.19
Second Quintile	8.7	+/-0.01	7.1	+/-0.19
Third Quintile	14.7	+/-0.05	13.0	+/-0.25
Fourth Quintile	23.0	+/-0.01	22.8	+/-0.31
Highest Quintile	50.3	+/-0.04	55.3	+/-0.65
Top 5 Percent	22.2	+/-0.04	24.9	+/-0.87

Muestras no probabilísticas

- No son aleatorias
- No son representativas
- Permiten la subjetividad
- No permiten hacer inferencias sobre la población
- Los estimados son sesgados
- No se puede establecer ni la confiabilidad ni la precisión en la estimación
- Su utilidad es en estudios exploratorios
- Tipos de muestras no probabilísticas: incidentales, por cuota y por referencia

Términos técnicos: En muestreo ...

- Un “**elemento**” es un objeto o un individuo sobre el cual se hace una medición.
- Una “**población**” es un conjunto de elementos sobre los cuales se desea hacer inferencias.
- “**Unidades de muestreo**” son subconjuntos que no traslapan formados por elementos de la población y que cubren la población completa.

Términos técnicos: En muestreo ...

- El “**marco muestral**” es un subconjunto de unidades muestrales.
- Una “**muestra**” es un conjunto de unidades muestrales que se obtiene de un marco o varios marcos muestrales.
- “**Parámetro de una población**” es una característica numérica y constante de la población. En poblaciones muy grandes, sus parámetros generalmente son desconocidos.

Términos técnicos: En muestreo ...

- Una “**estadística**” es un estimado de un parámetro de la población. Este estimado se obtiene en una muestra.
- El “**error estándar**” de una estadística es el valor absoluto de la diferencia entre el estimado y el parámetro.
- Uno de los propósitos fundamentales de la Estadística es minimizar el error estándar en las estimaciones que se realizan, es decir, maximizar la precisión.

Ejercicios:

- Ejercicio I: El Subprograma de Análisis
- Económico interesa estimar el Ingreso Anual Promedio del 2012 por sector económico y el de todas las industrias , usando las Planillas de Ingresos radicadas en el Departamento de Hacienda. Defina el objetivo; la población objetivo; el marco muestral; las unidades muestrales.

Ejercicios:

- Ejercicio # 2: Se desea estimar la tasa anual promedio de desempleo en PR, desde 1970 hasta 2010, usando una muestra de 8 años. Defina la población, un marco muestral apropiado y las unidades muestrales.
- Ejercicio # 3: Un pequeño comerciante desea estimar la cantidad de jugo que puede obtener al exprimir las chinas que hay en un cargamento de 500 que recibió. Defina la población, un marco muestral apropiado y las unidades muestrales.

Ejercicios:

- Ejercicio # 4: Se desea estimar el monto mensual promedio que cargan a su tarjeta de crédito los estudiantes de una universidad. Defina la población, un marco muestral apropiado y las unidades muestrales.
- Ejercicio # 5: Un caficultor desea estimar el número total de arbustos en su finca que están afectados por la rolla del café. Defina la población, un marco muestral apropiado y las unidades muestrales.

Ejercicios...

- Ejercicio 6: Se desea estimar la proporción de adultos de 16 años o más que están desempleados en una comunidad. Hay 25,000 adultos de esa edad allí. La proporción informada oficialmente es 20%. Se toma una muestra aleatoria de 100 adultos de 16 años o más, de los cuales el 18% están desempleados. Indique: a) cuál es el objetivo; b) cuál es la población objetivo; c) cuál es el marco muestral; d) cuáles son las unidades muestrales; e) cuál es el parámetro; f) cuál es la estadística; g) cuál es el error en estimación

Pasos en la planificación de una encuesta (o de un estudio estadístico)

- 1. Establecer los objetivos.
- 2. Definir la población objetivo.
- 3. Definir el marco muestral.
- 4. Definir el diseño muestral.
- 5. Determinar el método de medición.
- 6. Determinar el instrumento de medición.
- 7. Planificar la selección y entrenamiento de los encuestadores.
- 8. Realizar una preprueba (estudio piloto).
- 9. Organizar el trabajo de campo.

Pasos en la planificación de una encuesta (o de un estudio estadístico)

- I 0. Organizar el manejo de los datos.
- I 1. Realizar el análisis de los datos y obtener resultados.
- I 2. Llegar a conclusiones.
- I 3. Preparar el informe.

Tipos de encuesta más comunes

1. Cuestionarios enviados por correo
2. Encuestas telefónicas (uso de cuestionario)
3. Entrevistas (aplicando cuestionarios)
4. Electrónicas (cuestionario en línea)
5. Incidentales

Fuentes de error en encuestas

1. Errores de no observación
2. Errores de observación

- Errores de no observación
 - Pueden ocurrir cuando el marco muestral no coincide perfectamente con la población objetivo. Se dice que es un error de “cubierta”.
 - Surgen cuando las personas en la muestra no responden el cuestionario o la llamada o contestan parcialmente, dejando sin respuesta algunas preguntas, que usualmente son las más importantes, (**Problema de no-respuesta**) El efecto es que se reduce el tamaño de la muestra y aumenta el error estándar.

Fuentes de error...

- Sin embargo, tasas de no-respuesta bajas no implican necesariamente una buena encuesta si la no-respuesta proviene de un segmento de la población que es importante considerar.
- También tasas de no-respuesta altas no implican necesariamente que la encuesta deba descartarse, porque pudiera aun así estar proveyendo muy buena información si los que no responden se parecen a los que responden en todas las características importantes.

Fuentes de error...

- Por ejemplo, si en la muestra seleccionada se escogió la casa X para entrevistar al jefe de hogar y cuando llamaron a la puerta no abrieron, esta casa no puede ser sustituida por la casa vecina o ir a otra en un horario diferente (si es que en el diseño la hora de la entrevista afectaba los estimados).

Precauciones al diseñar un cuestionario

- 1. Orden en que se formulan las preguntas
- 2. Una pregunta sobre un tema puede afectarse por preguntas previas sobre el mismo asunto.
- 3. Cuando se trata de pedirle al encuestado que le asigne prioridad a las alternativas, generalmente la primera que se coloca en la lista es la que tienden a asignarle la primera prioridad, si para el entrevistado la primera luce como una buena selección, sin verificar si es la mejor.

Precauciones al diseñar un cuestionario

- 4. En preguntas de escala donde las alternativas son “fuertemente de acuerdo”, “de acuerdo”, etc., la proporción de los que escogen “fuertemente de acuerdo” es mayor cuando aparece como primera opción y es menor cuando es la última.
- 5. Preguntas abiertas
 - Ventaja: El que responde tiene completa libertad de expresarse sobre el asunto.
 - Desventajas: Son muy difíciles de analizar porque generalmente no permiten cuantificación; es muy difícil hacer comparaciones entre la diversidad de respuestas.

Precauciones al diseñar un cuestionario

- 6. Preguntas cerradas
 - Ventajas: Son fáciles de cuantificar y de analizar. Se pueden observar patrones de respuesta y llegar a conclusiones.
 - Desventaja: Al proveerle un número limitado de alternativas al que responde, obliga a éste a escoger entre una de ellas, aunque quizá opine que es otra que no está en la lista.
- Ejemplo pregunta abierta: ¿Cuál es el problema más importante que confronta PR actualmente? _____

Precauciones al diseñar un cuestionario

- Ejemplo pregunta cerrada: El problema más importante que confronta PR actualmente es:

- La criminalidad
- El cuidado de la salud
- La inflación
- El desempleo

(Nótese que la lista es muy limitada y quizá otros opinen que es la corrupción, el sistema educativo, la violencia doméstica, etc.)

- 7. Fraseo de las preguntas (compare)

- ¿Ud. cree que debieran prohibirse las expresiones públicas de los estudiantes dentro de la UPR?
- ¿Ud. cree que debieran permitirse las expresiones públicas de los estudiantes dentro de la UPR?

Precauciones al diseñar un cuestionario

- **8. Evite los enunciados compuestos:**
 - El profesor explica bien y es organizado
 - El profesor explica bien
 - El profesor es organizado
- **9. Evite los dobles negativos**
 - ¿No apoya la no confrontación?
 - ¿Apoya la no confrontación?
- **10. Formule los enunciados en forma positiva:**
 - ¿No apoya la confrontación?
 - ¿Apoya la no confrontación?

Precauciones al diseñar un cuestionario

- 11. Uso de las escalas Likert. Tenga cuidado con la formulación de la escala de opinión. Generalmente debe ser un número impar de opciones, con una neutral, por ejemplo: “de acuerdo”, “ni de acuerdo ni en desacuerdo”, “en desacuerdo”.
- En las escalas Likert evite proponer opciones ambiguas como “Parcialmente de acuerdo” y “Parcialmente en desacuerdo” .

Precauciones al diseñar un cuestionario

- **Cuidado con los condicionales**

- El enunciado “ si cubro el material hoy les daré un examen mañana” no implica que si no lo cubrió no dará el examen. Lo que único que implica es que “es imposible cubrir el material y no dar el examen. Usted puede dar el examen y no haber cubierto el material sin faltar a su compromiso. De la única forma que Ud. faltaría a su compromiso es si cubriera el material hoy y no diera el examen al día siguiente.

Tipos de muestreo y métodos de estimación más usados

- 1. Simple al azar
- 2. Estratificado al azar
- 3. Sistemático
- 4. Por conglomerados
- 5. Por conglomerados y estratificación
- 5. Por razón
- 6. Por regresión
- 7. Por diferencia
- 8. Estimación del tamaño N de la población (usando estimadores de densidad)
- 9. Por respuesta aleatorizada

Tabla de dígitos aleatorios

- Se utiliza para seleccionar una muestra al azar de tamaño n de un marco muestral de tamaño N (finito)
- Ejemplo: Utilice la Tabla de Dígitos Aleatorios para seleccionar una muestra de tamaño 8 del siguiente marco muestral de 41 tasas de desempleo anuales que se presentan en la hoja que se está distribuyendo.
- Nota: La selección de los números se inicia en la fila 1, columna 1, hacia abajo, de la Tabla de Dígitos Aleatorios.

Procedimiento, uso tabla de dígitos aleatorios

Número seleccionado	Posición que ocupa el dato	Tasas de desempleo seleccionadas en la muestra
10	11	17.0
22	23	16.0
24	25	14.6
37	38	10.9
28	29	13.3
09	10	17.0
07	08	19.9
02	03	11.6

(I) Muestreo simple al azar

- Procedimiento para aplicar este diseño:
 - 1. El marco muestral debe estar muy bien definido. Debe haber una lista exhaustiva de las unidades muestrales.
 - 2. A cada unidad muestral se le asigna un número.
 - 3. Se selecciona la muestra utilizando la tabla de dígitos aleatorios u otro procedimiento que los genere.
 - 4. Se recomienda realizar un estudio piloto con los siguientes objetivos:
 - A) Validar el instrumento y hacer los cambios pertinentes
 - B) Hallar un estimado de la varianza de la estadística de interés, el cual se usará para determinar cuál es el tamaño n de la muestra que se necesita para obtener cierta precisión en la estimación y con cierta confiabilidad.

Muestreo simple al azar: Fórmulas de los estimadores

- A) Para el promedio aritmético estimado:

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$

- B) Para la varianza estimada del promedio aritmético estimado

$$\hat{V}(\bar{y}) = \frac{S^2}{n} \left(\frac{N - n}{N} \right)$$

Muestreo simple al azar: Fórmulas de los estimadores

- C) Para la cota en el error de estimación del promedio (el error estándar)

$$\mathcal{E} = Z_{\alpha/2} \sqrt{\frac{S^2}{n} \left(\frac{N-n}{N} \right)}$$

- D) Para un intervalo de un $(100-\alpha)\%$ de confianza para el promedio de la población a partir del estimado:

$$\left(\bar{y} - \mathcal{E}, \bar{y} + \mathcal{E} \right)$$

Ejemplo con las tasas de desempleo

	Parámetro (población)	Estadística (muestra)
Promedio	15.08	15.04
Desviación estándar	3.70	3.03
Error estándar del estimado		1.07
Intervalo de un 95% de confianza para el promedio poblacional	Límite inferior	Límite superior
	12.50	17.58

Parámetros de la población de tasas de desempleo (1970-2010)

Parámetro de la población	Valor del parámetro
Mean	15.08
Standard Error	0.58
Median	14.6
Mode	11.6
Standard Deviation	3.69
Population Variance	13.68
Kurtosis	-0.63
Skewness	0.58
Range	13.3
Minimum	10.1
Maximum	23.4
Count	41
Confidence Level(95.0%)	1.167

Estadísticas de la muestra de tasas de desempleo (1970-2010)

Estadísticas de la muestra	Valor de las estadísticas
Mean	15.04
Standard Error	1.07
Median	15.3
Mode	17
Standard Deviation	3.03
Sample Variance	9.20
Kurtosis	-0.72
Skewness	0.097
Range	9
Minimum	10.9
Maximum	19.9
Sum	120.3
Count	8
Confidence Level(95.0%)	2.54

Estimación del total τ

- Si N es el tamaño de la población y \bar{y} es un estimado de su promedio, entonces el estimado del total de la variable se calcula por la fórmula:

$$\hat{\tau} = N\bar{y}$$

- Ejemplo: En el ejercicio anterior la sobreestimación promedio fue de \$20.20. Un estimado del monto total de la sobreestimación en el inventario es:

$$\text{Total} = 600 * (20.20) = \$12,120$$

Varianza del total

- La varianza estimada del total estimado está dada por

$$\hat{V}(\hat{\tau}) = \hat{V}(N\bar{y}) = N^2 \left(\frac{S^2}{n} \left(\frac{N-n}{N} \right) \right)$$

Ejemplo: Estimación del total de ventas de materiales de construcción (enero 69 –junio 2011)

- $N =$ número de meses en el período = 390
- $n =$ número de meses en la muestra = 15

$$\bar{y} = \$111,687$$

- Total estimado = $\tau = (390)(111,687) =$
- $= 43,557,930$
- Error en estimación del promedio, con un 95% de confianza = $\varepsilon = 26,333$
- Intervalo de un 95% de confianza para τ :
($\tau - 390 * \varepsilon$, $\tau - 390 * \varepsilon$) =
- (33,288,032 ; 53,827,828)

Estimación del tamaño de la muestra para promedios

- Una pregunta que surge siempre que se va a realizar una encuesta o un estudio es ¿cuál es el tamaño apropiado de la muestra que se va a seleccionar?
- En muestreo simple al azar y para la estimación del promedio, la fórmula es:

$$n = \frac{NS^2}{(N-1)\left(\frac{\varepsilon}{z}\right)^2 + S^2}$$

- Al aplicar esta fórmula se está tomando en cuenta el error máximo que tendría la estimación y su confiabilidad.

Estimación del tamaño de la muestra

- En muestreo simple al azar y para la estimación del total, la fórmula es:

$$n = \frac{NS^2}{(N-1)\left(\frac{\varepsilon}{zN}\right)^2 + S^2}$$

- Ejemplo: En el problema de las tasas de desempleo, suponga que la muestra con $n=8$ sirvió para estimar S^2 y así determinar el tamaño n que se requiere para un error máximo de un 2% en la estimación y una confiabilidad de un 95%. Entonces, la nueva n está dada por:

Ejemplo de estimación de n

$$n = \frac{41 \cdot (3.03)^2}{(41-1) \left(\frac{2}{1.96} \right)^2 + (3.03)^2} \approx 8$$

por lo que el tamaño n de la muestra se mantiene igual al original, pues garantiza que el estimado no difiera en más de 2% de la tasa promedio de desempleo en los 41 meses, con un nivel de confianza de un 95%.

Muestreo simple: estimación e inferencia sobre proporciones

- La proporción P de elementos de una población que tienen cierta característica de interés está dada por la fórmula:

$P = m/N$, donde m es el número de elementos de la población que tienen la característica y N es el tamaño de la población.

- Un estimado de P que se obtiene en una muestra está dado por:

$$\hat{P} = \frac{x}{n}$$

Muestreo simple: estimación e inferencia sobre proporciones

- Donde x = número de elementos en la muestra que tienen la característica y n = tamaño de la muestra.
- La varianza estimada del estimador de P está dada por:

$$\hat{V}(\hat{P}) = \frac{\hat{P}(1-\hat{P})}{n-1} \left(\frac{N-n}{N} \right)$$

El error en estimación viene dado por:

$$\mathcal{E} = Z \cdot \sqrt{\frac{\hat{P}(1-\hat{P})}{n-1} \left(\frac{N-n}{N} \right)}$$

Ejemplo: estimación de una proporción

- En la Encuesta del Viajero, en el mes de septiembre de 2008 salieron de PR 215,273 pasajeros. En la encuesta se encontró que, para ese mes, de 1,502 pasajeros entrevistados 637 eran visitantes. Por lo tanto, la proporción estimada de visitantes es:

- Un estimado del total de visitantes en ese mes de ese año es: $\hat{p} = \frac{637}{1502} = 0.4241$

- $$\hat{T} = N\hat{P} = 215,273(0.4241) \approx 91,298$$

Ejemplo: inferencias sobre una proporción

- Un intervalo de un 95% de confianza para la proporción p de elementos con la característica de interés es:

$$\hat{P} \pm z \cdot \sqrt{\frac{\hat{P}(1-\hat{P})}{n-1} \left(\frac{N-n}{N} \right)}$$

En el caso de la estimación de la proporción de visitantes en septiembre de 2008, el intervalo está dado por

$$.4241 \pm 1.96 \cdot \left(\frac{(.4241)(.5759)}{1502} \right)^{1/2}$$

Ejemplo: inferencias sobre una proporción...

- Al efectuar los cálculos el intervalo de un 95% de confianza para p es, aproximadamente:
(0.40 ; 0.45).
- Un intervalo de un 95% de confianza para el total de visitantes en ese período sería:
(86,110 ; 96,873)

Muestreo estratificado

- ¿Cuándo debe usarse?
 1. Cuando la población está dividida en subpoblaciones distintas (estratos), claramente identificables.
 2. Cuando se desea obtener estimados de los parámetros de la población total, pero también estimados de esos parámetros en las subpoblaciones.
 3. Cuando se supone, para propósitos del estudio o de las variables de interés, que cada subpoblación muestra homogeneidad entre sí, pero difiere de las demás subpoblaciones.
 4. Cuando se desea reducir la variabilidad (reducir el error estándar y aumentar la precisión)

Muestreo estratificado: Procedimiento

1. Segregar la población en estratos que no traslapen.
2. Debe conocerse el tamaño N de la población, así como los tamaños N_i de cada estrato.
3. Seleccionar muestras simples al azar en cada estrato y aplicar las fórmulas usuales para estimar los promedios o las proporciones en cada estrato.
4. Si se desea estimar el promedio de la población utilice la fórmula siguiente

Muestreo estratificado: fórmulas

- Fórmula para el estimado del promedio aritmético:

$$\bar{Y}_{estr} = \frac{N_1(\bar{Y}_1) + N_2(\bar{Y}_2) + \dots + N_k(\bar{Y}_k)}{N}$$

- Fórmula para la varianza estimada del promedio estimado:

$$\hat{V}(\bar{Y}_{str}) = \frac{1}{N^2} \left(\sum_{i=1}^k N_i^2 \cdot \left(\frac{N_i - n_i}{N_i} \right) \frac{S_i^2}{n_i} \right)$$

Fórmulas para muestreo estratificado..

- Desviación estándar del promedio

$$\hat{\sigma}(\bar{Y}_{str}) = \sqrt{\hat{V}(\bar{Y}_{str})}$$

- Error estándar del promedio para un margen de error α

$$\varepsilon = Z_{\alpha/2} \cdot \hat{\sigma}(\bar{Y}_{str})$$

Muestreo estratificado: fórmulas

- Fórmula para el estimado de una proporción

$$\hat{P}_{estr} = \frac{N_1 \hat{P}_1 + N_2 \hat{P}_2 + \dots + N_K \hat{P}_K}{N}$$

- Fórmula de la varianza estimada de la proporción estimada

$$\hat{V}(\hat{P}) = \frac{1}{N^2} \sum_{i=1}^k N_i^2 \left(\frac{\hat{P}_i \hat{Q}_i}{n_i - 1} \right) \left(\frac{N_i - n_i}{N_i} \right)$$

- Fórmula para el error estándar del estimado

$$\varepsilon = Z_{\alpha/2} \cdot \hat{\sigma}(\hat{P}_{str})$$