MUESTREO Y DISTRIBUCIONES DE MUESTREO:

1. Introducción.

2. Muestreo Aleatorio.

3. Diseño de Muestras.

4. Muestreo sistemático.

5. Muestreo Estratificado.

6. Muestreo por conglomerados.

7. Distribuciones muestrales.

8. El error estándar de la media

9. El teorema Central del límite.

1.− Introducción:

La estadística trabaja sobre poblaciones, extrae conclusiones sobre la base de un análisis de un muestrario de datos de una población. Hay muchas maneras de tomar una muestra de una población. Además las conclusiones que se extraen acerca de la población dependen de como se selecciona la muestra, deseamos que la muestra sea representativa de la población.

Vamos a concluir esta introducción con un ejemplo hipotético y algo extraño que ilustra como las conclusiones sacadas de una muestra pueden ser muy diferentes de la realidad.

Supongamos que una nave espacial del planeta Marte llega por primera vez a la Tierra, y aterriza por casualidad en el centro de Africa. Después de haber aterrizado en un claro de la selva, los marcianos recogen muestras de la vegetación que los rodea, toman nota de la constitución, temperatura, presión y humedad del ambiente y capturan como muestra a tres niños pigmeos. Vuelven a Marte y pasan varias semanas analizando los datos recogidos, haciendo, mas tarde, el informe para sus superiores. Éste da a entender que la Tierra está cubierta de selva, envuelta en un ambiente de aire caliente y húmedo y poblada de gente de piel negra, casi sin ropa cuya estatura media no es superior a un metro.

Del ejemplo observamos que hay varias cosas a tener en cuenta cuando se procede a tomar una muestra. Hay que elegir el tamaño de la muestra y esto dependerá no solamente de la cantidad de información que se quiere conseguir, y el grado de certeza deseada, sino también del costo del muestreo. Cualquiera sea el método elegido, el requisito más importante es que la muestra obtenida proporcione una imagen tan real como sea posible de aquella población que sé ha sometido al muestreo.

Describiremos a continuación los metodos de muestreo mas importantes, que introducen el azar y que aseguran la representatividad de las muestras.

2.− Muestreo Aleatorio:

Empezaremos definiendo ***Población,*** como un conjunto de individuos que se pueden identificar por separado.

Se puede pensar en una población concreta que realmente existe, como en una conceptual que no exista ni que existirá jamás. En ambos casos, el interés se centrará casi exclusivamente en las poblaciones números. Una población puede ser ***discreta*** o ***continua***, dependiendo de que el conjunto de números referidos sea discreto o continuo.

Una población es discreta si consta de un número finito o fijo de elementos, medidas u observaciones. Por ejemplo los pesos netos de 20 latas de atún.

A diferencia de las poblaciones discretas, las poblaciones continuas contienen una infinidad de elementos. Este es el caso de cuando observamos una variable continua y hay una infinidad de resultados distintos. También es el caso de las alturas de los estudiantes de la Universidad.

Un método para obtener una muestra sencilla aleatoria de una población es el siguiente: el empleo de una tabla de números aleatorios. Estas tablas son listas de cifras del 0 al 9, colocados de tal manera que si se elige al azar una posición cualquiera de la tabla, cada dígito tiene una posibilidad igual de aparecer en dicha posición. Es fácil seleccionar una muestra al azar de un conjunto de números, empleando estas tablas como se muestra en el siguiente ejemplo:

Obtener una muestra sencilla aleatoria de tamaño 5 de una clase de 30 estudiantes. Suponer que los estudiantes están numerados del 1 al 30 en la listad de la clase.

Solución : vamos a una tabla de números aleatorios, y escogemos un punto de comienzo. Entonces leemos a lo largo de la fila desde este punto, tomando las cifras por parejas (o de una columna de dos cifras hacia abajo), obteniendo los números así:

01, 53, **25**, 73, 49, 82, 35, **15**, **10**, 32, 97, **08**

En la serie elegimos sólo los números comprendidos entre el 1 y el 30, ignorando los otros

****

Para ver la idea de muestreo aleatorio en una población finita de tamaño N, primero veamos cuantas muestras distintas se pueden tomar de tamaño n. El número de muestras distintas es $\binom{N}{n}$

Por ejemplo si N=12 y n= 2 $\binom{12}{2} = \frac{12 \ \ 11}{2!} = 66$

muestras distintas.

Con base en el resultado de que hay $\binom{N}{n}$

muestras distintas de tamaño n de una población finita de tamaño N, podemos definir como ***muestra aleatoria o muestra aleatoria simple*** de una población finita:

Una muestra de tamaño n de una población finita de tamaño N es una variable aleatoria si se selecciona de manera tal que cada una de las $\binom{N}{n}$

muestras posibles tienen la misma probabilidad $\frac{1}{\binom{N}{n}}$

de ser seleccionada.

Por ejemplo si una población consistente en lo N= 5 elementos *a,e,i, o, u* (que podrían ser los ingresos anuales de cinco personas, los pesos de 5 vacas,.....) hay $\binom{5}{3} = 10$

muestras posibles de tamaño n = 3 . estas constan de los elementos:

aei	*aeo*	*aeu*	*aio*	*aiu*	*aou*	*eio*	*eiu*	*eou*	*iou*

si seleccionamos una de esas muestras de forma que esta muestra tenga probabilidad 1/10 de ser elegida, decimos que dicha muestra es aleatoria.

En la práctica el describir todas las posibles muestras seria complicado si N y n son grandes. Por ejemplo si n = 4 y N = 200 tendríamos 64,684,950 muestras distintas.

Por suerte podemos realizar una muestra aleatoria, sin necesidad de describirlas todas. Basta con numerar los N elementos de la población y retirar una a una hasta completar los n− elementos de la muestra. Este procedimiento también da una probabilidad de $\frac{1}{N}$

n

de ser seleccionada la muestra por los que sería aleatoria.

Ahora bien si la población es infinita: diremos que:

Una muestra de tamaño n de una población infinita es aleatoria si consta de valores de variables aleatorias independientes que tienen la misma distribución.

Por ejemplo si lanzamos un dado 12 veces y obtenemos 2, 5, 5, 3, 3, 3, 5, 1, 6, 1,4, 1. Estos números constituyen una variable aleatoria si son valores aleatoria independientes que tienen la misma distribución de probabilidad f(x) = 1/6 para x= 1,2,3,4,5,6

3− Diseños de muestras:

La única clase de muestras estudiadas hasta ahora son las aleatorias, y no hemos considerado siquiera la necesidad de que en ciertas condiciones pueda haber muestras que sean mejores (digamos más fáciles de obtener, más económicas o mas formativas) que las aleatorias, y no hemos entrado en detalles sobre la pregunta de cuando un muestreo aleatorio es imposible.

En estadística un ***diseño de una muestra*** es un plan definitivo, determinado por completo antes de recopilar cualquier dato, para tomar una muestra de una población de referencia.

Vamos a estudiar las mas comunes:

4.− Muestreo Sistemático:

En algunos casos la manera más práctica de realizar un muestreo consiste en seleccionar, un primer elemento al azar y luego ir escogiendo cada x−término de una lista, o dejar pasar a x− individuos y preguntar al que sigue y así sucesivamente. Aunque un muestreo sistemático puede no ser aleatorio de acuerdo con la definición, a menudo es razonable tratar las muestras sistemáticas como si fueran aleatorias.

El riesgo de los muestreos sistemáticos es el de las periodicidades ocultas. Supongamos que queremos testear el funcionamiento de una máquina, para lo cuál vamos a seleccionar una de cada 15 piezas producidas. Si ocurriera la desgracia de que justamente 1 de cada 15 piezas fuese defectuosa y el error de la máquina fuera defectuoso periódicamente, tendríamos dos posibles resultados muestrales:

− Que falla siempre

− Que no falla nunca.

5.− Muestreo Estratificado:

Si tenemos información a cerca de una población (es decir de su composición) y esta es importante para nuestra investigación, podemos mejorar el muestreo aleatorio por medio de la ***estratificación***. Este es un procedimiento que consiste en estratificar o dividir la población en un numero de ***subpoblaciones o estratos.*** Y seleccionamos de cada estrato una muestra aleatoria.

Este procedimiento se conoce como ***muestreo aleatorio (simple) estratificado.***

Supongamos una población de tamaño N que se divide en k estratos cuyos tamaños son:

N1, N2, ...,Nk (N1 +N2 +.....+Nk =N) Para obtener una distribución proporcional hemos de tener en cuenta que :

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \ldots\ldots = \frac{n_k}{N_k} = \frac{n}{N}$$

de donde se obtiene que $n_i = \frac{N_i}{N} \quad n$

para y=1,2,3,4,.... k donde n= tamaño de la muestra.

Esta seria una distribución proporcional, pero hay otras formas de distribuir porciones de una muestra entre los distintos estratos, que serían:

− Distribución óptima.

− Estratificación cruzada.

− Muestreo por cuotas.

Distribución óptima:

En la Distribución optima, no sólo se maneja el tamaño del estrato, como en la distribución proporcional, sino que también se maneja la variabilidad (o cualquier otra característica pertinente) del estrato.

La idea de la Distribución óptima, trata de jugar no sólo con el tamaño del estrato, sino que también pretende jugar con la variabilidad del mismo, de forma que parece lógico que los estratos de mayor variabilidad le correspondan muestras mayores. Si , , , ..., k son las desviaciones típicas de los k−estratos podemos explicar tanto los tamaños de los estratos, así como su variabilidad.

$$\frac{n_1}{N_1\ \sigma_1} = \frac{n_2}{N_2\ \sigma_2} = \frac{n_3}{N_3\ \sigma_3} = \ldots\ldots = \frac{n_k}{N_k\ \sigma_{1k}}$$

de donde se obtienen los tamaños muestrales de la distribución óptima o Distribución de Neyman (su inventor) que se obtienen por la fórmula:

$$n_i = \frac{n\ N_i\ \sigma_i}{N_1\ \sigma_1 + N_2\ \sigma_2 + \ldots\ldots + N_k\ \sigma_k}$$

para y=1,2,...., k

n= n1+n2+.....+nk

Estratificación cruzada:

La estratificación no se limita a una variable única de clasificación o una característica y las poblaciones a menudo se estratifican atendiendo a diversos criterios de ordenación o clasificación. Así por ejemplo si queremos realizar un estudio entre los alumnos de distintos centros de EE. MM. podríamos estratificar la muestra atendiendo al nivel de estudios, al sexo, a la especialidad,.... Así parte de la muestra se dedicaría a los alumnos de sexo femenino del 1º de Bachillerato técnico, otra parte a los alumnos de sexo masculino de 1º Bachillerato artístico, y así sucesivamente. Así y hasta cierto punto una estratificación de este tipo, llamada estratificación cruzada, incrementará la precisión de las estimaciones y otras generalizaciones que se usan comúnmente en el muestreo de opinión y las investigaciones de mercado.

Muestreo por cuotas:

En el muestreo estratificado, el costo de la toma de muestras aleatorias de los estratos individuales es tan alto, que a los encuestadores sólo se les dan cuotas que deben cubrir de los diferentes estratos, con alguna restricciones (si no es que ninguna) Por ejemplo si se quiere hacer un sondeo sobre la mejora de los servicios de salud, por ejemplo se le pide que encueste a 10 mujeres de entre 35 y 45 años que sean asalariadas, 20 hombres de entre 30 y 45 años que vivan en pisos de 3 o 4 habitaciones, a 3 hombres de mas de 60 años que estén jubilados.... esto es lo que se determina un muestreo por cuotas y es relativamente económico, lo único es que las muestras resultantes no cumplen las características esenciales de las muestras aleatorias. Por tanto estos muestreos, por cuotas en esencia son muestras de opinión, pero no son válidos para realizar un estudio estadístico formal.

6− Muestreo Por Conglomerados:

Para ilustrar esta clase de muestreo, supongamos que una gran empresa quiere estudiar los patrones variables de los gastos familiares de una ciudad como Buenos Aires. Al intentar elaborar los programas de gastos de una muestra de 1200 familias, nos encontramos con la dificultad de realizar un muestreo aleatorio simple, (es complicado tener una lista actualizada de todos los habitantes de una ciudad). Una manera de tomar una muestra en esta situación es dividir el área total (Buenos Aires en este caso) en áreas más pequeñas que no se solapen (Por ejemplo código postal, barrios, manzanas etc..) En este caso seleccionaríamos algunas áreas al azar y todas las familias (o muestras de éstas) que residen en estos códigos postales, barrios o manzanas, constituirían la muestra definitiva.

En este tipo de muestreo, llamado ***muestreo por conglomerados***, se divide la población total en un número determinado de subdivisiones relativamente pequeñas y se seleccionan al azar algunas de estas subdivisiones o conglomerados, para incluirlos en la muestra total. Si estos conglomerados coinciden con áreas geográficas, este muestreo se llama también ***muestreo por áreas***.

Aunque las estimaciones basadas en el muestreo por conglomerados, por lo general no son tan fiables como las obtenidas por muestreos aleatorios simples del mismo tamaño, son más baratas. Volviendo al ejemplo anterior, es mucho más económico visitar a familias que viven en el mismo vecindario, que ir visitando a familias que viven en un área muy extensa.

En la práctica se pueden combinar el uso de varios de los métodos de muestreo que hemos analizados para un mismo estudio.

7.− Distribuciones Muestrales:

Veamos ahora el concepto de ***distribución muestral*** de una estadística, que quizá es el concepto mas importante de la inferencia estadística.

Para introducir el concepto de distribución muestral, elaboraremos la de la media de una muestra aleatoria de tamaño n=2 tomada sin remplazo de la población finita de tamaño N=5, cuyos elementos son: 3,5,7,9,11.

La media de esta población es: $\mu = \frac{3+5+7+9+11}{5} = 7$

y su desviación típica es:

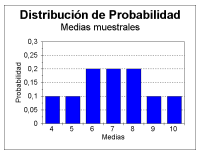
$$\sigma = \sqrt{\frac{(3-7)^2+(5-7)^2+(7-7)^2+(9-7)^2+(11-7)^2}{5}} = \sqrt{8}$$

Ahora si tomamos una muestra aleatoria de tamaño n = 2 de esta población hay $\binom{5}{2} = 10$

posibilidades:

nº nuestra	Muestras		x
1	3	5	4
2	3	7	5
3	3	9	6
4	3	11	7
5	5	7	6
6	5	9	7
7	5	11	8
8	7	9	8
9	7	11	9
10	9	11	10

Media	Probabilidad
4	1/10
5	1/10
6	2/10
7	2/10
8	2/10
9	1/10
10	1/10



Un análisis de esta distribución muestral revela cierta información relacionada con el problema de la estimación de la media de la población de referencia con una muestra aleatoria de tamaño n=2. Por ejemplo para $\bar{x}$
= 6,7 u 8 la probabilidad de que la media población (7) no difiera por más de 1 de la muestral es de 6/10. Sin embargo para $\bar{x}$
= 5,6,7,8 0 9 la media de una muestra no difiera en mas de 2 unidades es 8/10. Por consiguiente si no conociéramos la media de la población de referencia y quisiéramos estimarla con la media de una muestra aleatoria de tamaño n=2, el procedimiento anterior nos da alguna idea del posible tamaño del error.

Si calculamos la media y la desviación típica de la distribución de las medias obtenemos que:
x = 7 y x = $\sqrt{3}$
, luego la media x coincide con la media de la población y la desviación típica ha disminuido.

Evidentemente este proceso realizado con una muestra pequeña no es lo suficientemente explicativo. si tomásemos para n=10 y N=100 sería necesario una lista de mas de 17 billones de muestras.. por lo que para realizar el proceso sería necesario hacer una simulación por computadora.

8.− El error Estándar de la media:

En la mayoría de las situaciones reales, no podremos numerar todas las muestras posibles, o simular una distribución del muestreo para determinar cuánto puede aproximarse la media a la media de la población de la muestra. No obstante normalmente podemos obtener la información que necesitamos a partir de dos teoremas que expresan hechos esenciales sobre las distribuciones en el muestreo de la media:

El primero nos expresa formalmente lo que descubrimos en el ejemplo anterior . La media de la distribución del muestreo es igual a la media de la población y la desviación típica de la distribución del muestreo es menor que la desviación típica de la población.

Esto se puede expresar de la siguiente forma:

En el caso de variables aleatorias de tamaño n tomadas de una población con la media y desviación típica la distribución del muestreo de $\bar{x}$
tiene la media:

Media de la distribución muestral de $\bar{x}$	$\mu_{\bar{x}} = \mu$	
Error estándar de la media (desviación típica de la muestra)	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ ó	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$

dependiendo de que la población de infinita o de tamaño N

Es común referirse a x como el error ***estándar de la media*** donde se utiliza estándar en el sentido de desviación típica de la distribución muestral. Su función es fundamental en la estadística pues mide el grado en el que se puede esperar que fluctúen o varíen las medias de una muestra como consecuencia del azar. si x

es baja, hay buenas posibilidades de que la media de una muestra se aproxime a la media de la población si x alta, es más probable que obtengamos una muestra que difiera considerablemente de la media de la población.

A partir de las dos fórmula anteriores se puede apreciar lo que determina el tamaño de x . Ambas fórmulas demuestran (para poblaciones finitas e infinitas) $\bar{x}$
se incrementa conforme aumenta la variabilidad de la población y que se reduce conforme el tamaño de la muestra es mayor. De hecho es directamente proporcional a e inversamente proporcional a $\sqrt{n}$
(en las poblaciones finitas se reduce aún más rápido ya que aparece el factor $\sqrt{\frac{N-n}{N-1}}$

)

El factor $\sqrt{\frac{N-n}{N-1}}$
de la segunda fórmula de x se conoce como ***factor de corrección de la población finita.*** En la práctica, este se omite a menos de que la muestra constituya al menos un 5% de la población, pues en otro caso se aproxima tanto a 1 que es despreciable (es decir si la muestra no llega al 5% del tamaño de la población, no es necesario usar el factor de corrección)

9− El Teorema Central del Límite:

Antes de introducir este teorema, sin duda de los mas importantes dentro de la estadística moderna, vamos a estudiar un teorema previo. El Teorema de Chebyshev.

El Teorema de Chebyshev.

Para cualquier conjunto de datos (de una población o una muestra) y cualquier constante k mayor que 1, el porcentaje de los datos que debe caer dentro de k−veces la desviación típica de cualquier lado de la media es de por lo menos: $1-\frac{1}{k^2}$

El teorema de Chebyshev se aplica a cualquier tipo de datos, pero sólo nos indica por lo menos que porcentaje debe caer entre ciertos límites. Pero para casi todos los datos, el porcentaje real de datos que cae entre esos limites es bastante mayor que el que especifica el teorema de Chebyshev.

Para las distribuciones que tienen forma de campana puede hacerse una aseveración más fuerte:

(1) alrededor del 68% de los valores caerán dentro de una desviación típica de la media esto es: entre $\bar{X}-\sigma, \bar{X}+\sigma$
;

(2) aproximadamente el 95% de los valores caerán dentro de dos desviaciones típicas de la media, esto es : $\bar{X}-2\sigma, \bar{X}+2\sigma$
;

(3) aproximadamente el 99,7% de los valores caerán dentro de dos desviaciones típicas de la media, esto es : $\bar{X}-3\sigma, \bar{X}+3\sigma$
;

Basándonos en el teorema de Chebyshev con k=2 ¿Qué podemos decir del tamaño de nuestro error, si vamos a usar la media de una muestra aleatoria de tamaño n=64 para estimar la media de una población infinita con

=20?

Sustituyendo n=64 y =20 en la fórmula apropiada para el error estándar de la media, obtenemos que :

$$\sigma_x = \frac{20}{\sqrt{64}} = 2{,}5$$

y por el teorema de Chebyshev podemos afirmar que como mínimo 1 − 1/22 = 0,75 que el error será menor que k· x = 2·2,5= 5.

Es decir que tenemos una garantía de que en el 75% de los casos la media de la población estará entre la media calculada ±5 .

Pero esto no es suficiente, cuando la probabilidad real de este caso puede estar entre 0,98 y el 0,999

Teorema Central del Límite.

Para muestras grandes, se puede obtener una aproximación cercana de la distribución muestral de la media con una distribución normal.

Teniendo en cuenta que ya sabemos la media y desviación típica de la distribución muestral, podemos decir que:

x = y $\sigma_x = \frac{\sigma}{\sqrt{n}}$

para muestras aleatorias infinitas con media y desviación típica y n grande, entonces:

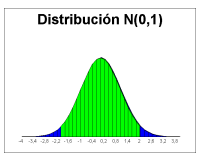
$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

es un valor de una variable N(0,1)

Este teorema es muy importante, puesto que justifica el uso de los métodos de la curva normal en una gran cantidad de problemas. se utiliza para poblaciones infinitas y para poblaciones finitas cuando n a pesar de ser grande representa una porción muy pequeña de la población.

Es difícil señalar con precisión qué tan grande debe ser n de modo que podamos aplicar el Teorema Central del límite, pero a no ser que la distribución sea muy Inusual, por lo general se considera que n =30 es lo suficientemente alto.

Veamos el mismo ejemplo anterior aplicando el Teorema Central del Límite.



La probabilidad se obtiene por medio del área marcada de la zona gris, específicamente por medio del área de la N(0,1) entre:

$$z = \frac{-5}{20/\sqrt{64}} = -2 \; y \; z = \frac{5}{20/\sqrt{64}} = 2$$

lo que consultando en las tablas da una probabilidad de 0,9544. Así sustituimos la afirmación de que la probabilidad es como mínimo 0,75 por una aseveración más firme de que la probabilidad es aproximadamente de 0,95 (de que la muestra aleatoria de tamaño n=64 de la población de referencia difiera de la de la población menos de 5 unidades)

También se puede usar el teorema Central del límite para poblaciones finitas, pero una descripción precisa de las situaciones en que se puede hacer esto, sería más bien complicada. El uso apropiado más común es en el caso en que n es grande y n/N es pequeña. Este es el caso de la mayoría de las encuestas políticas.

Veamos a continuación un ejemplo de la importancia de la selección adecuada de la muestra.

Para ello vamos a suponer una población de tamaño 60 elementos en el que se ha medido una determinada característica. De esta población vamos a realizar 25 muestras aleatorias y vamos a comprobar las diferencias existentes entre los valores estimados y los valores poblacionales.

111	539	216	128	462	283	413	237	193	177
406	257	290	213	325	306	184	168	310	266
279	393	450	92	241	302	319	193	281	313
295	402	183	310	257	257	302	315	353	128
244	116	127	348	418	232	400	166	451	315
335	707	266	91	703	380	618	79	588	199
Media		298,87							
Desviación Típica		139,4278							

A continuación observemos, las muestras obtenidas:

Número de muestras:

Como se puede observar las diferencias con respecto a los valores poblacionales son importantes.

Muestreo y Distribuciones en el Muestreo

página número 9