Maestría en Ciencia de Datos

Estadística: Muestras y experimentos no sesgados

5/5/2020

Actividad previa

Lectura para la clase del 5 de mayo http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-31952008000300004 para comentar en la clase. (20 minutos)

Muestras y experimentos no sesgados

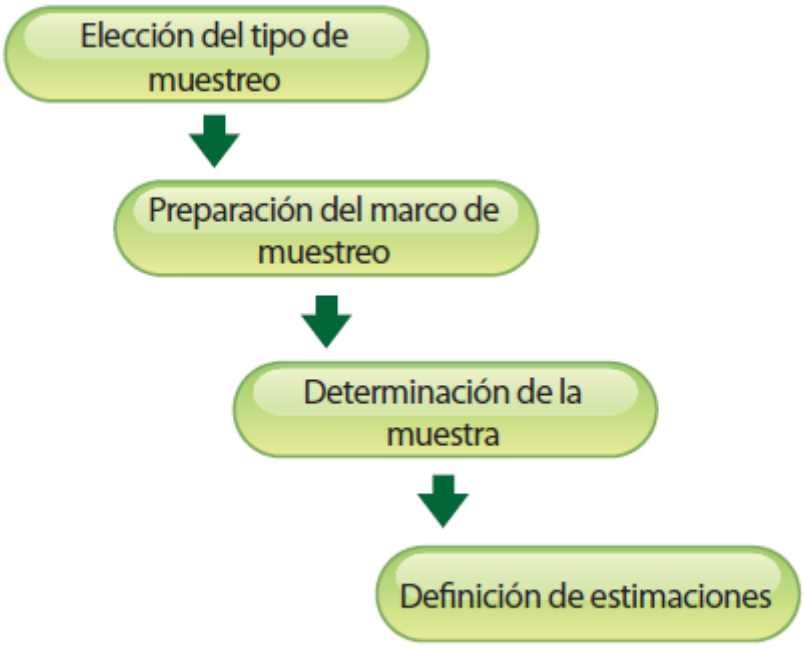

Figure 1: Pasos a seguir para la realización del diseño del proceso de muestreo probabilístico

Consideraciones básicas

- Una población consta de la totalidad de las observaciones en las que estamos interesados.
- Una muestra es un subconjunto de una población.
- Cualquier procedimiento de muestreo que produzca inferencias que sobreestimen o subestimen de forma consistente alguna característica de la población está sesgado.
- Para eliminar cualquier posibilidad de sesgo en el procedimiento de muestreo es deseable elegir una muestra, esto implica que las observaciones se realicen de forma independiente y al azar.

Tamaño de la muestra

A continuación se presenta la forma más simple para calcular el tamaño muestral. No se consideran factores como la tasa de no contestados, un coeficiente por tipo de diseño o el coeficiente de variación.

Si no se conoce el tamaño de la población N

$n = \frac{Z^2 * p * q}{d^2}$

Z = nivel de confianza
p = probabilidad de éxito
q = probabilidad de fracaso
d = error máximo admisible (%)

Si se conoce el tamaño de la población:

$n = \frac{N * Z^2 * p * q}{d^2 * (N-1) + Z^2 * p * q}$

N =tamaño de la población

Los valores de Z utiizados con más frecuencia son:

Valor de Z	Nivel de confianza
1.28	80%
1.65	90%
1.69	91%
1.75	92%
1.81	93%
1.88	94%
1.96	95%

Es importante considerar que altos niveles de confianza y bajo margen de error no significan que la encuesta sea más confiable, antes es preciso minimizar la principal fuente de error que generalmente tiene lugar en la recogida de datos.

Muestreo Probabilístico

Tiene como fin contar con elementos que tienen la misma probabilidad de ser elegidos. Así, los elementos muestrales tendrán valores muy parecidos a los de la población, de manera que las mediciones del subconjunto darán estimados precisos del conjunto mayor.

Una de las principales ventajas de este tipo de muestreo es que puede medirse el tamaño de error de las predicciones y en consecuencia, reducir al mínimo el error estándar.

Nota: La variabilidad de las medias muestrales se puede medir por su desviación estándar. Esta medida se conoce como el error estándar y tiende a disminuir cuando aumenta el tamaño de la muestra.

Muestra aleatoria

Sean X_1, X_2,..., X_n variables aleatorias independientes n, cada una con la misma distribución de probabilidad $f(x)$. Definimos X_1, X_2,..., X_n como una **muestra aleatoria de tamaño** n de la población $f(x)$ y escribimos su distribución de probabilidad conjunta como $f(x_1, x_2, ..., x_n) = f(x_1)f(x_2)...f(x_n)$.

Muestro Aleatorio Simple

Se asigna un número a cada individuo de la población y a través de algún medio mecánico (tablas de números aleatorios, números aleatorios generados con una calculadora o computadora, etc.), se eligen tantos sujetos como sea necesario para completar el tamaño de muestra. Este procedimiento, tiene poca utilidad si la población que se va a estudiar es muy grande. Cada elemento tiene una probabilidad de inclusión igual y conocida de n/N.

Ejemplo

Se desea seleccionar una muestra de tamaño n=2 de una población que contiene N=4 objetos. Si los cuatro objetos están identificados con los símbolos x_1,x_2,x_3 y x_4 hay cuatro pares diferentes que pueden seleccionarse. Si las muestra de n=2 observaciones se selecciona de modo que cada una de las muestras tenga la misma probabilidad de ser seleccionada (1/6), entonces la muestra resultante se llama **muestra aleatoria simple ó muestra aleatoria**.

Muestra	Pares que pueden seleccionarse
1	x_1,x_2
2	x_1,x_3
3	x_1,x_4
4	x_2,x_3
5	x_2,x_4
6	x_3,x_4

Muestreo aleatorio estratificado

En ocasiones será conveniente estratificar la muestra según ciertas variables de interés. Para ello se debe conocer la composición estratificada de la población objetivo al hacer un muestreo. Si una muestra está formada or dos o más subpoblaciones o estratos, el muestreo que asegura que cada estrato está representada en la muestra.

En el muestreo estratificado, con frecuencia los resultados se requieren para ciertos estratos de la población y el error deseado se establece para cada uno de ellos. Se debe calcular por separado el tamaño en cada grupo y el tamaño de muestra final será la suma de las establecidas para cada estrato.

Se debe tener cuidado de asignar muestra a todos los estratos contemplados en el marco de muestreo, en caso de que el tamaño de muestra no lo permita, se deben hacer los ajustes (uniones de estratos) necesarios para no dejar a ninguno sin representación.

Distribución de igual número de muestras de cada estrato Esta opción es la más sencilla de aplicar y asume que los estratos presentan las varianzas, los costos y sus tamaños iguales (Nh). En caso contrario, redunda en estimaciones pobres. Sin embargo, puede ser útil cuando se requiere obtener resultados con precisiones semejantes en los diferentes estratos. Para obtener la distribución se aplica la expresión:

$n_h = \frac{n}{L}$

n_h = tamaño de muestra en los diferetes estratos
L = número de estratos

Distribución proporcional Se usa si los estratos presentan varianzas iguales, costos iguales y sus tamaños son distintos. La distribución se obtiene con la expresión:

$n_h = \frac{N_h}{N} n$

N_h = tamaño de cada estrato

Distribución óptima Se emplea si se tienen costos muy diferentes por estrato, las varianzas son distintas y los tamaños de los estratos también son diferentes. La distribución se calcula con la expresión:

$$n_h = \frac{N_h S_h / \sqrt{C_h}}{\sum_{h=1}^{L} N_h S_h / \sqrt{C_h}} n$$

S_h = varianza de cada estrato
C_h = Costos asociados a captar un cuestionario de cada uno de los estratos

Muestreo sistemático

Se numeran todos los elementos de la población, pero en lugar de extraer n números aleatorios sólo se extrae uno. Se parte de ese número aleatorio para elegir, a intervalos constantes, todos los demás hasta completar la muestra.

Pasos a seguir:
* Conseguir un listado de los N elementos de la población.
* Determinar tamaño muestral n.
* Definir un intervalo K=N/n.
* Elegir un número aleatorio, r, entre 1 y k (r=arranque aleatorio).
* Seleccionar los elementos de la lista.

Muestreo por conglomerados

La unidad muestral es un grupo de elementos de la población que forman una unidad, que es llamada conglomerado. Consiste en seleccionar aleatoriamente un cierto número de conglomerados (el necesario para alcanzar el tamaño muestral establecido) y posteriormente investigar todos los elementos pertenecientes a los conglomerados elegidos.

El muestreo no probabilístico o determinístico

En el muestreo no probabilístico, también conocido como determinístico, el cálculo del tamaño y selección de la muestra se basan en juicios y criterios subjetivos, por esto se desconoce la probabilidad de selección de las unidades de la población bajo estudio y no es posible establecer la precisión respecto a niveles de confianza predefinidos.

A pesar de esto, el muestreo determinístico representa una alternativa viable cuando la aplicación del muestreo probabilístico resulta demasiado costosa o si existe seguridad en que la información recabada bajo este tipo de muestreo es suficientemente útil para los fines de la investigación.

Actividad 1

Investigar las ventajas y desventajas de cada tipo de muestreo, subir un documento de una cuartilla a Moodle (20 minutos).**Importante: Citar las fuentes utilizadas**.

Actividad 2

Correr el siguiente código en R

```
library(datasets)
data(iris)
summary(iris)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

```
indices <- sample( 1:nrow( iris ), 60 )
# sample takes a sample of the specified size from the elements of x using either with or without repet
iris.muestreado <- iris[ indices, ]
summary(iris.muestreado)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.400   Min.   :2.200   Min.   :1.200   Min.   :0.100
##  1st Qu.:5.200   1st Qu.:2.800   1st Qu.:1.500   1st Qu.:0.275
##  Median :6.000   Median :3.100   Median :4.500   Median :1.450
##  Mean   :5.992   Mean   :3.145   Mean   :3.863   Mean   :1.252
##  3rd Qu.:6.500   3rd Qu.:3.400   3rd Qu.:5.425   3rd Qu.:2.000
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :21
##  versicolor:15
##  virginica :24
##
##
##
```

```
#muestreo aleatorio simple con repetición
indices_cr <- sample( 1:nrow( iris ), 60, replace = TRUE )
iris.muestreado_cr<-iris[indices_cr,]
summary(iris.muestreado_cr)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.400   Min.   :2.200   Min.   :1.300   Min.   :0.100
##  1st Qu.:5.350   1st Qu.:2.700   1st Qu.:1.675   1st Qu.:0.375
##  Median :5.800   Median :3.000   Median :4.200   Median :1.350
##  Mean   :5.842   Mean   :3.065   Mean   :3.695   Mean   :1.153
##  3rd Qu.:6.400   3rd Qu.:3.400   3rd Qu.:5.025   3rd Qu.:1.800
##  Max.   :7.700   Max.   :4.400   Max.   :6.700   Max.   :2.300
##        Species
##  setosa    :21
```

```
##  versicolor:18
##  virginica :21
##
##
##
```

```
#Muestreo aleatorio simple sin reemplazo
library( sampling )
```

```
## Warning: package 'sampling' was built under R version 3.6.3
```

```
estratos <- strata( iris, stratanames = c("Species"), size = c(20,20,20), method = "srswor" )
iris.muestreado <- getdata( iris, estratos )
table(iris.muestreado$Species)
```

```
##
##     setosa versicolor  virginica
##         20         20         20
```

```
#muestreo aleatorio simple con reemplazo
estratos_cr <- strata( iris, stratanames = c("Species"), size = c(20,20,20), method = "srswr" )
iris.muestreado <- getdata( iris, estratos_cr )
table(iris.muestreado$Species)
```

```
##
##     setosa versicolor  virginica
##         20         20         20
```

Distribución muestral

La distribución muestral de un estadístico es la distribución de probabilidad de los posibles valores de la muestra.

El principal propósito al seleccionar muestras aleatorias consiste en obtener información acerca de los parámetros desconocidos de la población.

La variabilidad en la muestra refleja cómo se dispersan las observaciones a partir del promedio.

Es posible tener dos o más conjuntos de observaciones con las mismas media o mediana que difieran de manera considerable en la variabilidad de sus mediciones sobre el promedio.

Estimación de los parámetros de la población

Para estimar el valor de un parámetro poblacional se usa información obtenida de la muestra por medio de un estimador. Los estimadores se calculan usando información de las observaciones muestrales.

El estimador elegido debe satisfacer las siguientes características: • Ser insesgado. • Ser consistente. • Tener varianza mínima. • Ser fácil de obtener y calcular.

Si se tiene una poblacipon formada por $N = 5$ números: 3,6,9,12,15 y una muestra aleatoria de tamaño $n = 3$ se selecciona sin reemplazo, encuentre las distribuciones muestrales para la media muestral $\overline{x}$ y la mediana m

La probabilidad para cada númeo está dada por $p(x) = 1/5$

Muestra	Valores muestrales	$\overline{x}$	m
1	3,6,9	6	6
2	3,6,12	7	6
3	3,6,15	8	6

Muestra	Valores muestrales	$\overline{x}$	m
4	3,9,12	8	9
5	3,9,15	9	9
6	3,12,15	10	12
7	6,9,12	9	9
8	6,9,15	10	9
9	6,12,15	11	12
10	9,12,15	12	12

```
x<-c(3,6,9,12,15)
media<-mean(x)
mediana<-median(x)
media
```

```
## [1] 9
```

```
mediana
```

```
## [1] 9
```

Actividad 3

Calcule las distribuciones muestrales para la media muestral y la mediana muestral y haga un histograma para cada caso. Subirlo a Moodle en el espacio correspondiente.

$\overline{x}$	$p(\overline{x})$
6	0.1 (frecuencia/10)

m	$p(m)$
6	0.3 (frecuencia/10)

¿Qué estimador elegiría?

Si se toma la Muestra 7, al usar m como estimador: 9-12=-3,9-6=3 con probabilidad de 0.3 El error de estimación es 3, con probabilidad 0.6

Si se usa $\overline{x}$ un error=3 sólo ocurriría con probabilidad de 0.2

Si muestras aleatorias de n observaciones se sacan de una población no normal con media finita μ y desviación estándar σ, entonces, cuando n es grande la distribución de muestreo de la media muestral $\overline{x}$ está distribuida de forma aproximada con media μ y desviación estándar $\sigma/\sqrt{n}$

La aproximación se hace más precisa cuando n crece.

Estimador puntual
El objetivo de la estimación puntual es utilizar una muestra para calcular un número que representa una buena suposición del valor verdadero del parámetro de interés.

La estimación puntual de algún parámetro de la población θ es un solo valor $\hat{\theta}$ de un estadístico.

Se dice que un estadístico $\hat{\theta}$ es un estimador insesgado del parámetro θ si $\mu_{\hat{\theta}} = E(\hat{\theta}) = \theta$

Si se consideran todos los posibles estimadores insesgados de algún parámetro θ, al que tiene la menor varianza lo llamamos estimador más eficaz de θ.

Ejemplo Se ha utilizado una secuencia de 25 choques controlados para probar una defensa automotriz que se supone que absorbe impactos. Sea X es número de choques que no provoca daños visibles en el automóvil. El parámetro que se quiere estimar es p (proporción de choques que no provocan daños en los automóviles). Alternativamente, $p = P$ ningún daño en el choque.Si se observa que X es x=15, el estimador más razonable es:

$\hat{p} = X/n$ la estimación será $x/n = 15/25 = 0.6$

Este es un caso en el sólo hay un estimador, normalmente esto no sucede.

2. Suponga que un conjunto de datos consta de las siguientes observaciones acerca del voltaje de ruptura dieléctrica de piezas de resina epóxica:

0.32 0.53 0.28 0.37 0.47 0.43 0.36 0.42 0.38 0.43

Se asume un comportamiento normal para los datos, enuentre la mediay su estimación, la vida útil mediana de la muestra y su estimación.

```
library(modeest)
```

```
## Warning: package 'modeest' was built under R version 3.6.3
```

```
conjunto<-c(0.32, 0.53, 0.28, 0.37, 0.47, 0.43, 0.36, 0.42, 0.38, 0.43)
moda<-mlv(conjunto, method = "mfv")
# una medida de localización o tendencia central en
#una muestra no da por sí misma una indicación clara de la naturaleza de ésta, de manera
#que también debe considerarse una medida de variabilidad en la muestra
moda
```

```
## [1] 0.43
```

```
shapiro.test(conjunto)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  conjunto
## W = 0.98512, p-value = 0.9867
```

```
#Interpretación: Con un p-value mayor de 0.05 no se puede rechazar la hipótesis nula #(hipótesis de nor
media<-mean(conjunto)
mediana<-median(conjunto)
media_rextortada_10<-mean(conjunto,trim=10/100)
prom_vida_util_extremos<-max(conjunto)-min(conjunto)/2
desviacion_st<-sd(conjunto)


# Falta determinar cuál de estos estimadores, si se utiliza en otras muestras de de Xi,tiende a produci
```

Error estándar El error estándar de un estimador $\hat{\theta}$ es su desviación estándar $\sigma_{\hat{\theta}} = \sqrt{V(\hat{\theta})}$. Es una magnitud de una desviación típica o representativa entre una estimación y el valor de θ.

Si el error estpandar implica parámetros desconocidos cuyos valores pueden ser estimados, las sustitución de estas estimaciones en $\sigma_{\hat{\theta}}$ da el error estándar estimado (desviación estándar estimada) del estimador. El error estándar estimado se denota $\acute{\sigma}\theta$ o por $S_{\hat{\theta}}$

. . . continuación del ejercicio anterior

Si el voltaje de ruptura está normalmente distribuido, $\hat{\mu} = \overline{X}$ es la mejor estimación de μ. Si se sabe que $\sigma = 0.072$, el error estándar de $\overline{X}$ es $\sigma_{\overline{X}} = \sigma/\sqrt{n} = 0.072/\sqrt{10}$

Si la distribución del estimador puntual $\hat{\theta}$ es normal de un modo aproximado, lo que sucede con frecuencia cuando n es grande, se puede confiar en que el el valor verdadero de θ queda dentro de aproximadamente dos errores estándar de $\hat{\theta}$.

Si $\hat{\theta}$ no es necesariamente normal pero es insesgado, entonces la estimación se desviará de θ hasta 4 veces el error estándar cuando mucho 6% de las observaciones.

Actividad 4

Para hacer en grupos de dos alumnos. Subir reporte en Moodle en el espacio correspondiente.

Seleccione una base de datos en el paquete (library(datasets)).

Para enlistar las bases de datos disponibles:
library(help=datasets)

Por ejemplo: chickwts, CO2, MU284, swissmunicipalities Es conveniente no elegir series de tiempo.

1. Determine qué tipo de muestro es más conveniente realizar.
2. Encuentre los estimadores adecuados a la base de datos.

En ambos casos, justifique sus respuesta.