Inferencia Estadística

Población y muestra

Conceptos y definiciones

Muestra Aleatoria Simple (MAS)

Consideremos una población, cuya función de distribución esta dada por ***F(x),*** la cual está constituida por un número infinito de posibles valores, de una cierta característica medible ***X***.

Esta característica puede ser, por ejemplo:

- el tiempo de espera para recibir un servicio.
- el valor de las ventas de un determinado producto.

Entonces para seleccionar una muestra aleatoria de tamaño ***n*** de esta población, se diseña un experimento. De tal manera que la primera realización de ese experimento nos proporciona la observación X_1, de la característica medible ***X***, **repitiendo** sucesivamente el experimento bajo las <u>mismas condiciones</u>, para todos los factores controlables, tendremos las ***n*** observaciones:

Elementos de la Muestra Aleatoria Simple

$$X_1,\ X_2,\cdots X_n$$

Muestra Aleatoria Simple (MAS)

Cada observación X_i correspondiente a la repetición *i-ésima* del experimento es una variable aleatoria cuya distribución de probabilidad es idéntica a la de la población con característica *X*, para todo i = 1, 2, ..., n.

Si la población consta de un número finito de elementos, por ejemplo, personas, viviendas, establecimientos comerciales, etc., y realizamos un muestreo aleatorio con reemplazamiento, reiterando el proceso *n* veces tendríamos las *n* observaciones:

$$X_1, X_2, \cdots X_n$$

Variables aleatorias, independientes y uniformemente distribuidas

de la característica medible *X* de la población, que constituyen la muestra aleatoria simple. Cada una de estas observaciones, X_1, X_2, ..., X_n, también es una variable aleatoria cuya función de probabilidad es idéntica a la de la población, ya que cada selección de algún elemento que da lugar a una observación procede de la población original.

Muestra Aleatoria Simple (MAS)

Si en la población con un número finito de elementos, se seleccionan análogamente *n* elementos sin reemplazamiento, tendríamos una muestra aleatoria sin reemplazamiento de observaciones de la característica *X* que estamos investigando.

$$X_1,\ X_2,\cdots X_n$$

Variables aleatorias, **no** independientes y uniformemente distribuidas

Definición de MAS

Una muestra aleatoria simple de tamaño *n* de una población *X* está constituida por un conjunto de *n* variables aleatorias X_1, X_2, ..., X_n, independientes y uniformemente distribuidas a la población *X*, es decir, está constituida por un conjunto de observaciones muestrales independientes e idénticamente distribuidas.

Cuando el experimento se realiza, a cada una de las variables aleatorias se le asignará un valor numérico.

$$X_1 = x_1 \cdots X_n = x_n$$

Es decir, tendremos la realización de la muestra y diremos que ha sido seleccionada una muestra.

Ejemplo 1

Consideremos la población formada por los 100 alumnos de una clase de la universidad. Sus edades aparecen en la siguiente tabla.

N	edad	N	edad	N	edad	N	edad	N	edad	N	edad
1	20	19	21	37	22	55	22	73	22	91	20
2	22	20	21	38	22	56	19	74	22	92	20
3	19	21	19	39	22	57	20	75	21	93	19
4	21	22	19	40	21	58	20	76	21	94	22
5	19	23	22	41	22	59	21	77	20	95	21
6	19	24	22	42	19	60	20	78	22	96	21
7	20	25	19	43	19	61	22	79	22	97	20
8	19	26	19	44	21	62	21	80	20	98	22
9	21	27	22	45	21	63	21	81	20	99	20
10	19	28	21	46	19	64	22	82	21	100	21
11	19	29	22	47	22	65	22	83	22		
12	20	30	20	48	20	66	21	84	21		
13	22	31	19	49	22	67	21	85	19		
14	20	32	20	50	19	68	20	86	20		
15	20	33	19	51	20	69	22	87	21		
16	19	34	20	52	21	70	22	88	21		
17	19	35	22	53	21	71	20	89	21		
18	22	36	22	54	21	72	20	90	19		

Hemos aprendido que la muestra debe ser aleatoria, entonces obtendremos 6 números aleatorios, para obtener una muestra aleatoria de 6 estudiantes, por lo que comenzamos generando estos números.

Ejemplo

Al utilizar Excel los números aleatorios se obtienen escribiendo en la celda la siguiente fórmula:

= ALEATORIO()

Esta función generará un número de forma aleatoria que se encuentra entre 0 y 1. Si el número que queremos generar debe encontrarse entre 0 y 100, lo único que tenemos que hacer es multiplicar por el factor 100

= ALEATORIO()*100

Sin embargo, como la cantidad que buscamos es un número entero, se puede dar también dar la instrucción de entero para que el número que nos proporcione Excel sea entero, entonces:

= ENTERO(ALEATORIO()*100)

Número aleatorio

Al aplicar estas reglas para generar lo seis números tenemos:

Número de alumnos

1	65
2	22
3	87
4	73
5	51
6	17

Ejemplo

Que corresponde a los seis estudiantes de la población cuyas edades están dadas por:

Número de alumnos

Tamaño de la muestra

1	65	22
2	22	19
3	87	21
4	73	22
5	51	20
6	17	19

Edad del alumno

Experimento, medición de la propiedad *X*

Número aleatorio

Proceso aleatorio

Población de 100 estudiantes

Muestra aleatoria de tamaño seis

Valores observados de las variables aleatorias

22
19
21
22
20
26

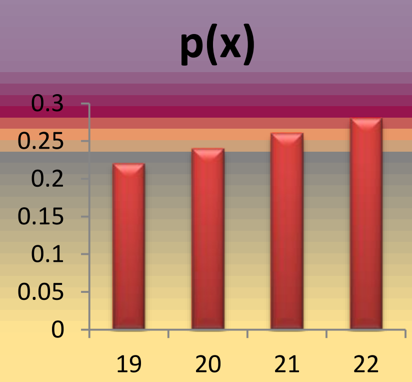
Ejemplo 1

Como estamos interesados en la edad del estudiante, consideramos la variable aleatoria:

X: edad del estudiante seleccionado

Se procede de manera análoga para otras variables aleatorias como: estatura, peso, temperatura, calificación, etc.

Para el ejemplo visto, la distribución de probabilidades de la variable aleatoria ***X*** se obtiene de la siguiente manera: de la población sabemos el número de estudiantes cuyas edades son igual a 19, 20, 21 y 22 años, entonces, al ser ***N=100*** el número total de la población, la probabilidad de que algún estudiante tenga 19, 20, 21 o 22 años esta dada por:



valores de la variable aleatoria X	Probabilidad P(X=x)
19	0.22
20	0.24
21	0.26
22	0.28

Si seleccionamos una muestra con reemplazamiento, de seis estudiantes de la población, para observar la edad de cada uno, entonces definiremos seis variables aleatorias:

X_1: edad del primer estudiante seleccionado.
X_2: edad del segundo estudiante seleccionado,
$\vdots$
X_6: edad del sexto estudiante seleccionado

Cada variable aleatoria tendrá una distribución de probabilidad asociada. Así pues, la distribución de la variable aleatoria X, será exactamente la misma que la distribución de la variable aleatoria X sin reemplazamiento ya que ambas variables aleatorias se refieren a la edad de un estudiante seleccionado al azar.

Pero como el muestreo se ha realizado con reemplazamiento, se puede ver lo siguiente:

- ***La variable aleatoria X_2 , tiene la misma distribución de probabilidades que* X o que *X_1.***
- ***X_2 , y X , son independientes. De igual forma, las variables* aleatorias *X_3, X_4, X_5, y X_6 , tienen la misma distribución que X, y en consecuencia* la sucesión de variables aleatorias *X_1, X_2, ..., X_6 , son independientes e* idénticamente distribuidas.**

Definición MAS

*Sea **X** la variable aleatoria correspondiente a una población con función de distribución **F(x)**. Si las variables aleatorias X_1, X_2, ..., X_n, son independientes y tienen la misma función de distribución, **F(x)**, que la de la distribución de la población, entonces las variables aleatorias X_1, X_2, ..., X_n, forman un conjunto de variables aleatorias independientes e idénticamente distribuidas que constituyen una muestra aleatoria simple de tamaño **n** de la población F (x)*

Al ser las variables aleatorias X_1, X_2, ..., X_n, ***independientes,*** resulta que la función de distribución conjunta será igual al producto de las funciones de distribución marginales, es decir:

Regla (□)

Regla (□)

$$F(x_1,\ x_2,\cdots x_n)=\prod_{i=1}^{n}F\left(x_i\right)$$

Si la población de partida es de tipo discreto y la función de probabilidad de la población es:

$$p_i=P(X=x_i)\qquad\forall\ i=1,\ 2,\cdots,n$$

entonces la función de probabilidad de la muestra será:

$$P(X_1 = x_1, X_2 = x_2, \cdots X_n = x_n) = \prod_{i=1}^{n} P(X = x_i) = \prod_{i=1}^{n} p_i$$

Si la muestra aleatoria simple procede de una población de tipo continuo con función de densidad f(x), entonces la función de densidad de la muestra será:

$$f(x_1, x_2, \cdots x_n) = \prod_{i=1}^{n} f(x_i)$$

Ejemplo 2

Sea $X_1,\ldots X_n$, una MAS de n variables, de un a población cuya distribución de probabilidad es exponencial con densidad dada por

$$f(x,\theta)=\frac{1}{\theta}e^{\frac{-x}{\theta}}\quad \forall\ \ 0<x<\infty$$

Cuando X_1 se observa se registra la realización x_1

$$f(x_1,\theta)=\frac{1}{\theta}e^{\left\{-x_1/\theta\right\}}\quad \forall\ \ 0<x_1<\infty$$

Se observa X_2 se registra la realización x_2. Pero dado que son estadísticamente independientes y tienen la misma densidad

$$f(x_2\mid x_1)=f(x_2)=\frac{1}{\theta}e^{\left\{-x_2/\theta\right\}}\quad \forall\ \ 0<x_2<\infty$$

La función de densidad conjunta X_1y X_2 es

$$f(x_1,x_2,\theta)=f(x_1,\theta)\,f(x_2,\theta)=$$

$$=\frac{1}{\theta^2}e^{\left\{-(x_1+x_2)/\theta\right\}}\quad \forall\ \ 0<x_i<\infty$$

Y se deduce entonces la extensión a X_1 … X_n es:

Parámetros Poblacionales

- Estadísticos muéstrales

Definición

Los parámetros poblacionales son las características numéricas de la población. Un parámetro es una caracterización numérica de la distribución de la población. El conocimiento del parámetro permite describir parcial o totalmente la función de probabilidad de la característica que estamos investigando. Por ejemplo, si la característica a investigar sabemos que sigue una distribución exponencial de parámetro *a* su función de densidad está determinada.

$$f(x) = \begin{cases} ae^{-a} & x < 0 \\ 0 & x \leq 0 \end{cases}$$

Dicha función de densidad no estará totalmente descrita hasta que no se dé el valor del parámetro *a*, y entonces será cuando podremos formular preguntas concretas sobre esa distribución, es decir, podremos calcular las diferentes probabilidades.

Si la característica a investigar sigue una distribución normal, $N(\mu,\sigma)$, la función de densidad contiene dos parámetros, para describir totalmente la función de densidad tendremos que dar valores a μ y a σ, pues si damos valor sólo a un parámetro entonces diremos que está descrita parcialmente.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Comprendo

En la mayoría de los modelos probabilísticos nos encontraremos parámetros cuyos valores tendremos que fijar para especificar completamente el modelo y poder calcular las probabilidades deseadas. Uno de los problemas centrales en estadística se presenta cuando deseamos estudiar una población con función de distribución $F(x, \theta)$, donde la forma de la función de distribución es conocida, pero depende del parámetro θ desconocido, ya que si θ fuese conocido tendríamos totalmente especificada la función de distribución. Si el parámetro θ no se conoce entonces se selecciona una muestra aleatoria simple $(X_1,..., X_n)$ de tamaño n de la población, y se calcula para las observaciones de la muestra el valor de alguna función $g(x_1, ..., x_n)$, que representa o estima el parámetro desconocido θ. El problema es determinar qué función será la mejor para estimar el parámetro θ.

Definición Estadístico

Un estadístico es cualquier función real de las variables aleatorias que integran la muestra, es decir, es una función de las observaciones muestrales, la cual no contiene ningún valor o parámetro desconocido.

- Identificamos la población con función de distribución ***F(x, θ)***
- Consideramos una MAS constituida por n variables (AIUD) $(X_1, \ldots, X_n)$

Es posible definir algunos estadísticos o funciones de esas variables aleatorias, como por ejemplo:

$$g_1(X_1, \cdots X_n) = \frac{X_1 + X_2 + \cdots X_n}{n}$$

$$g_2(X_1, \cdots X_n) = \frac{X_1^2 + X_2^2 + \cdots X_n^2}{n}$$

$$g_3(X_1, \cdots X_n) = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots (X_n - \overline{X})^2}{n}$$

Los estadísticos se determinan siempre a partir de lobservaciones muestrales.

Definición Estadístico

De forma general al estadístico lo representaremos por la letra ***T***, es decir:

$$T = g(X_1, \cdots X_n)$$

La función g() será la función que describa las observaciones muestrales, asimismo será una variable aleatoria, ya que para cada muestra, el estadístico ***T*** tomará un valor distinto. Entonces para una muestra de n observaciones ***(x_1, ..., x_n)***el estadístico tomará el valor

$$T = g(x_1, \cdots x_n)$$

Se toman distintas muestras y desde luego se obtienen valores diferentes para el estadístico, resultando entonces que efectivamente el estadístico T , también es una variable aleatoria, Ahora bien, si es una variable aleatoria, entonces también pose su función de distribución, la cual se llama Función de Distribución Muestral del Estadístico

Un parámetro y un estadístico son conceptos diferentes, pues el parámetro es una constante Y el estadístico es una variable aleatoria.

Hemos estudiado diferentes medidas numéricas correspondientes a conjuntos de datos, entre otras, estudiamos la media, la desviación estándar etc. Ahora vamos a distinguir entre medidas numéricas calculadas con conjuntos de datos poblacionales y las calculadas con datos muestrales. Si la medida numérica se calcula para el conjunto de datos poblacionales le llamaremos valor del parámetro poblacional y si se calcula para el conjunto, de datos muestrales, le llamaremos valor del estadístico muestral.

Parámetros

En una población finita de tamaño N los parámetros poblacionales media, varianza y proporción poblacional vienen dados por:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$$

$$p = \frac{X}{N} = \frac{\#ENP}{\#P}$$

Estadísticos

Para una muestra aleatoria simple de tamaño n, $(X_1, \ldots, X_n)$ los estadísticos: media, varianza y proporción muestral se definen como:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{N}(X_i - \overline{X})^2$$

$$p = \frac{X}{n} = \frac{\#EnP}{\#P}$$

Si en vez de considerar las n variables aleatorias, independientes e idénticamente distribuidas (X_1, ..., X_n) que constituyen la muestra aleatoria simple, consideramos una muestra concreta (x_1, ..., x_n) entonces los valores de estos estadísticos muestrales tomarían la siguiente forma:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$x^2 = \frac{1}{n-1} \left(x_i - \overline{x}\right)^2$$

$$\hat{p} = \frac{x}{n} = \frac{\# EnP}{\# P}$$

Veremos que efectivamente el estadístico es una función de las observaciones muestrales, y en estos casos asigna a cada muestra observada la media de los valores, la varianza o la proporción, respectivamente.