Muestreo y Distribuciones en el Muestreo

Daniel Paredes Moreno

Estadística I
Departamento de Estadística-FACES-ULA

03 de Abril de 2013

Introducción al Muestreo

En algunas ocaciones es posible y práctico examinar a cada individuo en el **Universo Estadístico** que deseamos describir. En este caso, se habla de una investigación por **censo o enumeración completa**. Cuando no es posible observar o medir a todos los individuos del Universo entonces debemos recurrir al **muestreo**; es decir, recoger una **muestra** de dicha población.
Recordemos que en cualquier caso, el objetivo de la estadística es describir las principales características de la **Población Estadística**. Estas características de interés, generalmente estan resumidas o representadas por **medidas descriptivas numéricas**, como por ejemplo: La media, la mediana, la moda, los percentiles, la desviación estándar o porcentajes, entre muchos otros. Cuando la medida está calculada en base a datos muestrales se le denomina **estadístico** mientras que cuando se calcula en base a los datos de una población completa se conoce como **Parámetro**.

Estadísticos y Parámetros

Un **Estadístico** *es cualquier función calculada empleando los datos en una muestra.*
Un **Parámetro** *es cualquier función calculada empleando los datos en una población completa.*

Por ejemplo, medidas como la media, mediana, moda, desviación estándar y coeficiente de asimetría pueden ser tanto estadísticos como parámetros dependiendo de los datos utilizados para su cálculo.

Estadísticos y Parámetros

En una investigación por censo o por muestreo, el objetivo de la estadística es determinar el valor o los valores de estos parámetros que describen a la población. Sin embargo, en una investigación por muestreo, se debe usar la información muestral para **inferir** los valores de los parámetros, mientras que en una por censo, solo deben calcularse directamente.
Debe notarse, además, que en dicho proceso de inferencia existe incertidumbre y por lo tanto, puede haber errores en los valores inferidos.

Ejemplos

- Supongamos que la estatura promedio de todos los alumnos de noveno grado de Venezuela es de 152 cm. Si interesan todos los alumnos de noveno grado de Venezuela, entonces este es un parámetro. Si la estatura promedio de los alumnos de una sección de noveno grado en el Liceo Libertador de Mérida es de 154 cm, y manteniendo el interés en los alumnos de toda Venezuela, entonces este promedio es un estadístico.
- Si por el contrario, no se cuenta con los recursos de tiempo, dinero y personal para medir a todos los alumnos de noveno grado en Venezuela, pero considerando que el promedio de los alumnos de la sección en el Liceo Libertador en Mérida si puede calcularse, podemos entonces emplear este **estadístico** para inferir el promedio poblacional usando métodos estadísticos inferenciales.

Tipos de Muestreo

- **Muestreo no aleatorio o muestreo de juicio** En este caso se emplea el conocimiento y la opinión personal o de un panel de expertos en el área para identificar a los elementos del Universo que se medirán para conformar la muestra. A menudo, una muestra de juicio se emplea como guía para decidir como debe tomarse una muestra aleatoria mas adelante.
- **Muestreo aleatorio o probabilístico** En este tipo de muestreo, se conocen las posibilidades de que un individuo sea seleccionado o no para conformar la muestra. De esta manera, puede evaluarse las propiedades de representatividad de la muestra y pueden medirse los errores al realizar inferencia sobre algún parámetro poblacional.

Tipos de Muestreo Aleatorio

Tipos de muestreo Aleatorio

- *Muestreo Aleatorio Simple.*
- *Muestreo Sistemático.*
- *Muestreo Estratificado.*
- *Muestreo por Conglomerado.*

Muestreo Aleatorio Simple

- Muestreo Aleatorio Simple

En el muestreo aleatorio simple, se seleccionan muestras mediante métodos que aseguran que cada posible muestra tenga la misma probabilidad de ser seleccionada. Observe los siguientes casos:

Población Finita e Infinita

- *Población Finita será aquella donde se puede determinar exactamente cuantos elementos existen en ella, es decir, existe un entero N que indica cuantos elementos hay en una población.*
- *Población Infinita será aquella donde es teoricamente imposible medir a todos los elementos del Universo. Aunque muchas poblaciones pueden ser excesivamente grandes, no existe una población realmente infinita de objetos físicos en un instante de tiempo dado. Aun asi, consideraremos infinita una población que surge de medir objetos que no pueden ser evaluados en un tiempo razonable.*

Muestreo Sistemático

- Muestreo Sistemático.

En este caso, se seleccionan los elementos dentro de un intervalo uniforme que se mide con respecto al tiempo, al orden o al espacio. Para esto, se escoge un elemento cada $k - esimo$ elemento y para escoger el primer elemento, se elige un numero aleatorio entre 0 y k.
De esta manera, cada elemento de la población tiene la misma probabilidad de ser seleccionado pero cada posible muestra no tiene la misma probabilidad de ser escogida.
Usando este método de muestreo puede dar origen a una **muestra sesgada** pero tiene la ventaja de reducir los costos en tiempo y dinero de seleccionar la muestra.

Mmuestreo Estratificado

- Muestreo Estratificado.

En este caso, se divide la población en grupos relativamente homogéneos llamados **estratos**. Luego de esto,se puede proceder de dos maneras:

1. Seleccionar aleatoriamente en cada estrato una cantidad de elementos proporcional al tamaño del estrato para conformar la muestra.
2. Seleccionar aleatoriamente la misma cantidad de elementos de cada estrato y después se ponderan los resultados según el tamaño de cada estrato.

Esto resulta apropiado cuando la población ya está dividida naturalmente en grupos de diferentes tamaños y deseamos aprovechar esta condición. Este tipo de muestreo asegura que todos los elementos de la población tienen posibilidad de conformar la muestra y cuando se aplica adecuadamente se refleja de manera mas precisa las características de la población.

Muestreo por Conglomerados

- Muestreo por Conglomerados

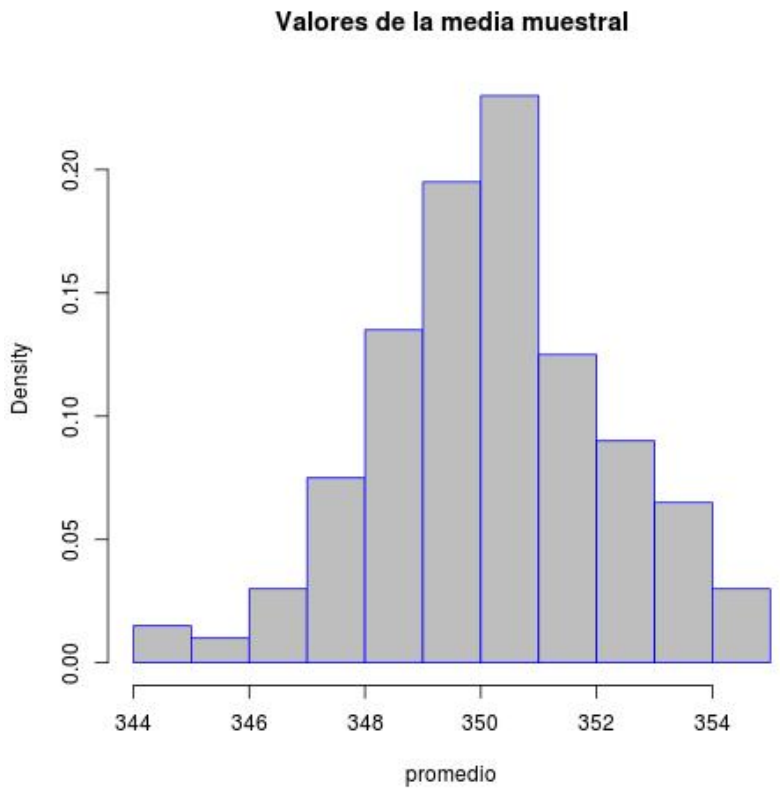
En este caso se divide también la población en grupos o **conglomerados** pero en este caso, en lugar de ser homogéneos buscamos que sean lo mas heterogéneos posibles, de modo que cada grupo por si solo sea representativo de la población. De modo que, para conformar la muestra se seleccionan al azar varios conglomerados y se incluyen todos los elementos en ellos.
En este caso, los conglomerados son muy parecidos entre si, a diferencia de los estratos que son diferentes entre ellos.

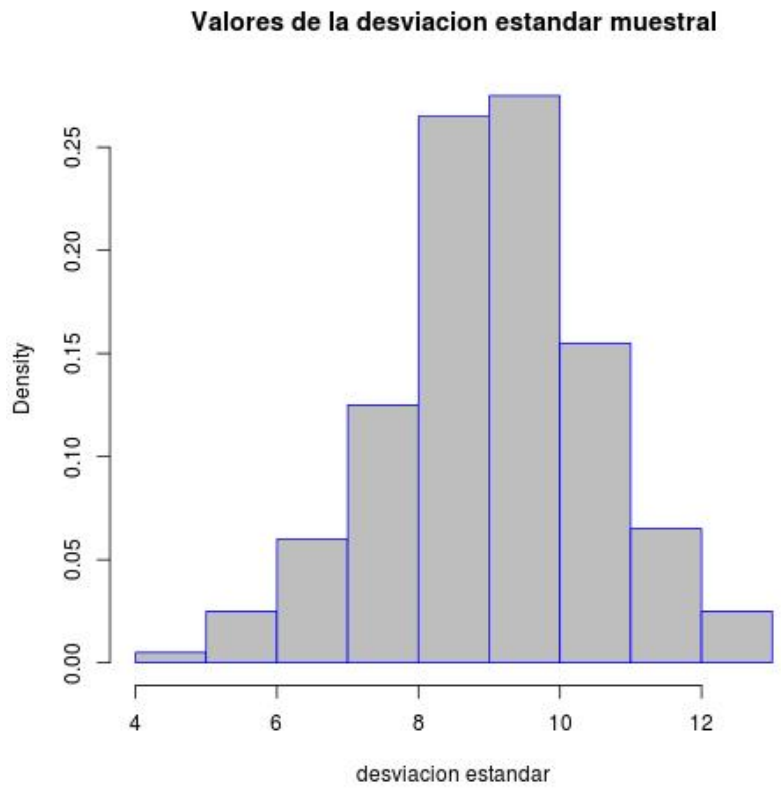
- Estadísticos como variables aleatorias.

Suponga que tenemos una población a ser muestreada con muestras de tamaño n, donde se quiere conocer un parámetro θ, y se quiere calcula un estadístico $\hat{\theta}$ a partir de los datos muestrales. Observe que para cada posible muestra de tamaño n, podemos obtener diferentes valores del estadístico, que depende de la muestra finalmente seleccionada. En este sentido, se tiene incertidumbre respecto al valor que va a tomar efectivamente el estadístico y de esta manera puede ser tratado como una variable aleatoria.
Así, un estadístico, como cualquier variable aleatoria, tendrá una distribución de probabilidad asociada, a la cual se le llama **Distribución en el muestreo del estadístico**.

Tomemos el siguiente ejemplo:
Una empresa vende refresco. Una de sus presentaciones es en latas de 350 ml. La máquina que llena las latas está calibrada para que, en promedio, coloque 350 ml de refresco en la latas, con una desviación estándar de 9 ml. También se sabe que la cantidad de refresco colocado en las latas por la máquina sigue una distribución normal.
Cada dia deben seleccionarse una muestra aleatoria de 20 latas para verificar el buen funcionamiento de la máquina y su correcta configuración. Entre otros cálculos, interesa el contenido promedio de las latas seleccionadas y su desviación estándar.
A continuación, se muestra un histograma con el valor de dicho promedio en 200 días consecutivos.

En este gráfico, se muestra un histograma con los diferentes valores que puede tomar en esos 200 dias consecutivos, tanto la media como la desviación estándar de la muestra.
Estos pueden ser vistos como aproximaciones a la **distribución en el muestreo para la media y la desviación estándar**





- Error Estandar

Para los datos del ejemplo anterior, donde se tienen 200 valores de promedios muestrales y 200 valores de desviaciones estándar muestrales, podríamos calcular a su vez medidas descriptivas numéricas para estos datos.
Por ejemplo, la **media** de los promedios es de 350,1924 ml y la **desviación estándar** de los promedios es de 2,0152 ml.
Del mismo modo, la **media** de las desviaciones estándar es de 9,0542 y la **desviación estándar** de las desviaciones estándar es de 1,5112.
En este caso, la desviación estándar del estadístico **media** se le conoce como **error estándar de la media** y a su vez la desviación estándar del estadístico **desviación estándar** se le conoce como **error estándar de la desviación estándar**

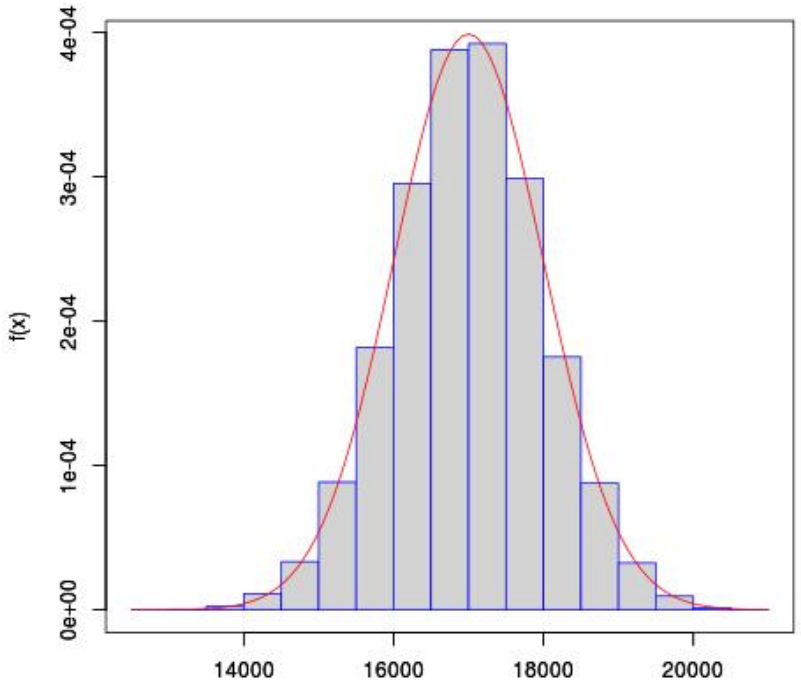
Muestreo de Poblaciones Normales

Ahora nos concentraremos en el caso donde realizamos una investigación por muestreo donde la población muestreada tiene distribución normal.

- Supongamos que se realiza un estudio sobre el salario mensual de todos los trabajadores del sector de empresas básicas en un país.
- Supongamos que la distribución de la variable aleatoria X: Salario mensual de los empleados en empresas básicas en un país sigue una distribución normal con media de 17000 Bs y desviación estándar de 1000 Bs. Es decir;
  $X \sim Normal(17000, 1000)$.

Muestreo de Poblaciones Normales

- **Qué quiere decir esto en la práctica?** Si pudieramos realizar el censo para todos y cada uno de los empleados del país en empresas básicas y obtener su salario mensual, al hacer un histograma de probabilidad este debería comportarse como una distribución normal con los parámetros antes mencionados. Esto sería algo como lo siguiente:



Distribución de la media en el muestreo normal

Sea $X_1, X_2, \ldots\ldots, X_n$ *una muestra aleatoria de tamaño n de una población normal con media* μ *y varianza de* σ^2. *Entonces bajo estas condiciones se sabe que:*

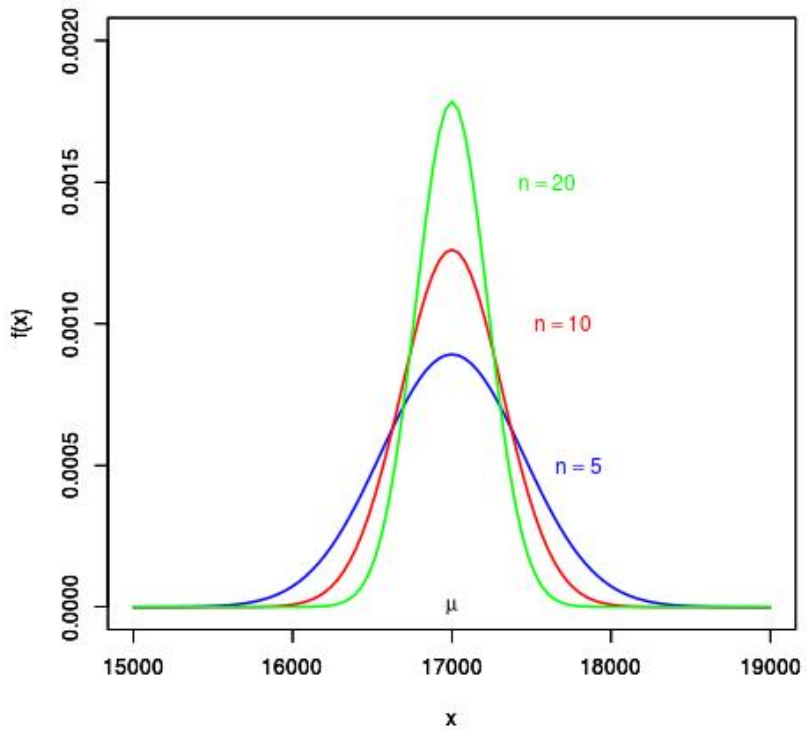
$$\bar{X} = \sum_{i=1}^{n} X_i$$

$$\bar{X} \sim Normal\left(\mu, \frac{\sigma^2}{n}\right)$$

Esto implica que: $E\left[\bar{X}\right] = \mu$, $V\left[\bar{X}\right] = \frac{\sigma^2}{n}$ *y* $EE\left[\bar{X}\right] = \sqrt{\frac{sigma^2}{n}}$
Donde $EE\left[\bar{X}\right]$ *es el* **Error Estándar de la media**.

Muestreo de Poblaciones Normales

Si $\bar{X} \sim Normal\left(\mu, \frac{\sigma^2}{n}\right)$ entonces:



Observese que la distribución de la media tiene el mismo valor promedio que la población y a medida que el tamaño de muestra aumenta, la distribución se hace mas angosta.

- Ejemplo:

Suponga que se toma una muestra aleatoria de tamaño $n = 15$ de los empleados de empresas básicas de un país y se mide el ingreso mensual. Suponga que se conoce de esta caracterítica que sigue una distribución normal, con media de 17000 y desviación estándar de 1000. Calcular la probabilidad de que:

1. $$P\left(16800 \leq \bar{X} \leq 17550\right)$$
2. $$P\left(\bar{X} > 16300\right)$$
3. $$P\left(\bar{X} < \bar{x}\right) = 0{,}025$$

Muestreo de Poblaciones no Normales

Ahora supongamos que estamos realizando una investigación por muestreo sobre una característica, la cual no se distribuye según una normal o simplemente se desconoce la forma de su distribución. Para este caso, la media muestral no tiene una distribución normal. Pero existe un teorema, quizas el mas importante en toda la teoría estadística, que nos permite aún calcular probabilidades sobre la media muestral bajo algunas condiciones:

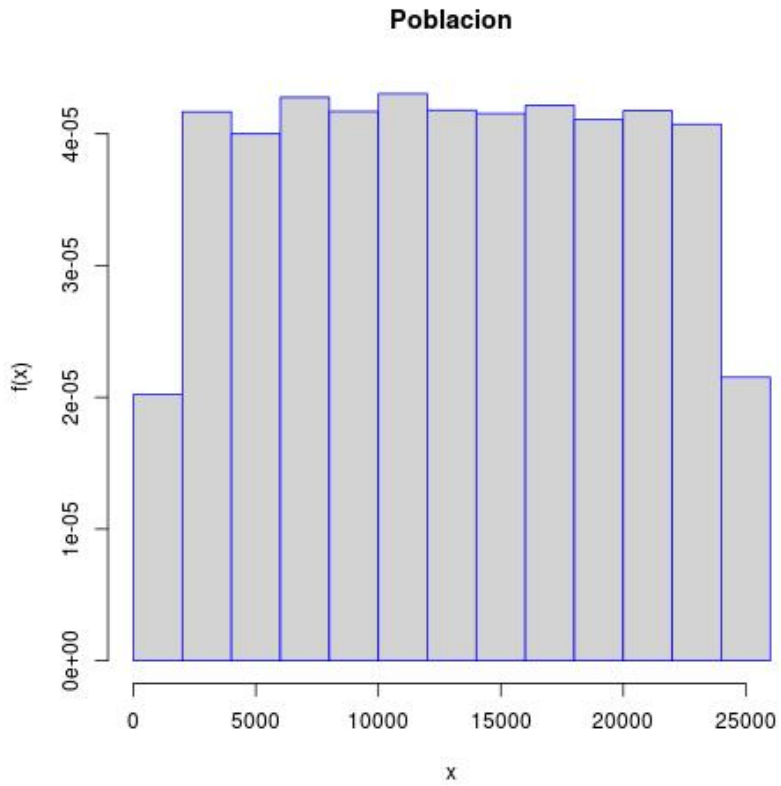
Teorema Central del Límite

Sea $X_1, X_2, , X_n$ *una muestra aleatoria de una población con media* μ *y varianza* σ^2*. Entonces:*

La distribución del estadístico: $\dfrac{(\bar{X}-\mu)}{\sqrt{\frac{\sigma^2}{n}}} \xrightarrow[n\to\infty]{} Normal\,(0,1)$

Muestreo de Poblaciones no Normales

Para ilustrar esto, vamos a suponer que seleccionamos muestras de una población que tiene una distribución como la siguiente:



Muestreo de Poblaciones no Normales

Ahora veamos que efecto tiene el tamaño de la muestra en la forma de la distribución del estadístico: