Probabilidad y Estadística

Clase 6

Gonzalo Blanco

2023

1. Estadística

El campo de la estadística tiene que ver con la recopilación, organización, análisis y uso de datos para tomar decisiones razonables en base a tal análisis.

Estos datos pueden ser cualquier información útil del proceso/grupo/evento de interés que quiero estudiar.

Definimos una **población** como el universo de datos que vamos a analizar. Esta población está formada por la totalidad de las observaciones/datos/resultados del proceso o del experimento de interés.

Por ejemplo, si estoy interesado en realizar algún análisis sobre el poder adquisitivo de los habitantes de argentina, mi población podría ser el valor en pesos del sueldo de cada argentino, tendría entonces alrededor de 45 millones de datos (sueldos).

Definimos una **muestra** como un subconjunto de esa población. Siguiendo el ejemplo anterior, una muestra de esa población podría ser el sueldo de los jubilados.

Si una **muestra es representativa de la población** entonces es posible inferir importantes conclusiones de la población de interés, sin necesidad de contar con los datos/información de toda la población. La parte de la estadística que se trata sobre las condiciones bajo las cuales esta inferencia es válida se la llama **inferencia estadística.** La parte de la estadística que estudia la muestra sin inferir alguna conclusión sobre la población es la **estadistica descriptiva.**

1.1. Estadística Descriptiva

Es la parte de la estadístca que trata sobre los métodos para recolectar, organizar y resumir datos.

Es importante tener en cuenta lo siguiente, no hay una única forma de organizar y visualizar un conjunto de datos, dependerá de cada caso de estudio que método o métodos conviene utilizar.

1.1.1. Distribución de frecuencias e histograma

Supongamos que tenemos el siguiente conjunto de datos: edad de los trabajadores de una empresa.

Esta empresa cuenta con 510 trabajadores, entonces quiero organizar esta información y visualizarla a través de un histograma. Veamos como hacemos esto:

1) Primero buscamos el máximo y mínimo de mis datos, en este caso supongamos que las edades van de 20-55 años.

2) Luego elegimos un intervalo (a,b) que contenga a todos mis datos.

En este caso puedo usar el intervalo [20-55], pero si por ejemplo el mínimo fuera 12,4 y el máximo 61,4 tal vez preferiria usar un números enteros por ej. tomar el intervalo (12,62).

3) Dividimos al intervalo (a,b) en subintervalos que generalmente son de igual longitud, pero pueden no serlo, y contamos cuantas observaciones o datos caen en cada subintervalo, esta cantidad será la **frecuencia del subintervalo.**

Para hacer esto previamente debemos decidir cuantos subintervalos vamos a utilizar, estos subintervalos se llaman intervalos de clase o simplemente clases, nuevamente esto dependerá de cada situación en particular. Pero en general es común definir de 5 a 20 subintervalos.

Supongamos entonces que para nuestro ejemplo decidimos usar 7 subintervalos de igual longitud. Entonces la longitud de cada subintervalo vendrá dada por $(b-a)/7$, es decir, $(55-20)/7 = 5$.

4) Construimos una tabla de frecuencias de manera tal de asociarle a cada sub intervalo o clase su frecuencia, frecuencia relativa, frecuencia acumulada y frecuencia relativa acumulada.

Clase	Frecuencia	Frecuencia relativa	Frecuencia acumulada	Frecuencia relativa acumulada
20-25	90	0.1765	90	0.1765
25-30	150	0.2941	240	0.4706
30-35	120	0.2353	360	0.7059
35-40	70	0.1373	430	0.8431
40-45	50	0.0980	480	0.9412
45-50	20	0.0392	500	0.9804
50-55	10	0.0196	510	1

Figura 1: Tabla de frecuencias

También podemos hacer un gráfico de esta tabla de frecuencias conocido como histograma. El mismo se construye a partir de dos ejes cartesianos, en el eje de las absisas se marcan los límites de cada clase, y para cada clase se construye un rectángulo de tal forma que el área de ese rectángulo sea proporcional a la frecuencia de esa clase. El eje vertical corresponde a la frecuencia, entonces la altura de cada rectángulo representa la frecuencia de cada clase (siempre que cada clase/subintervalo tengan la misma longitud).

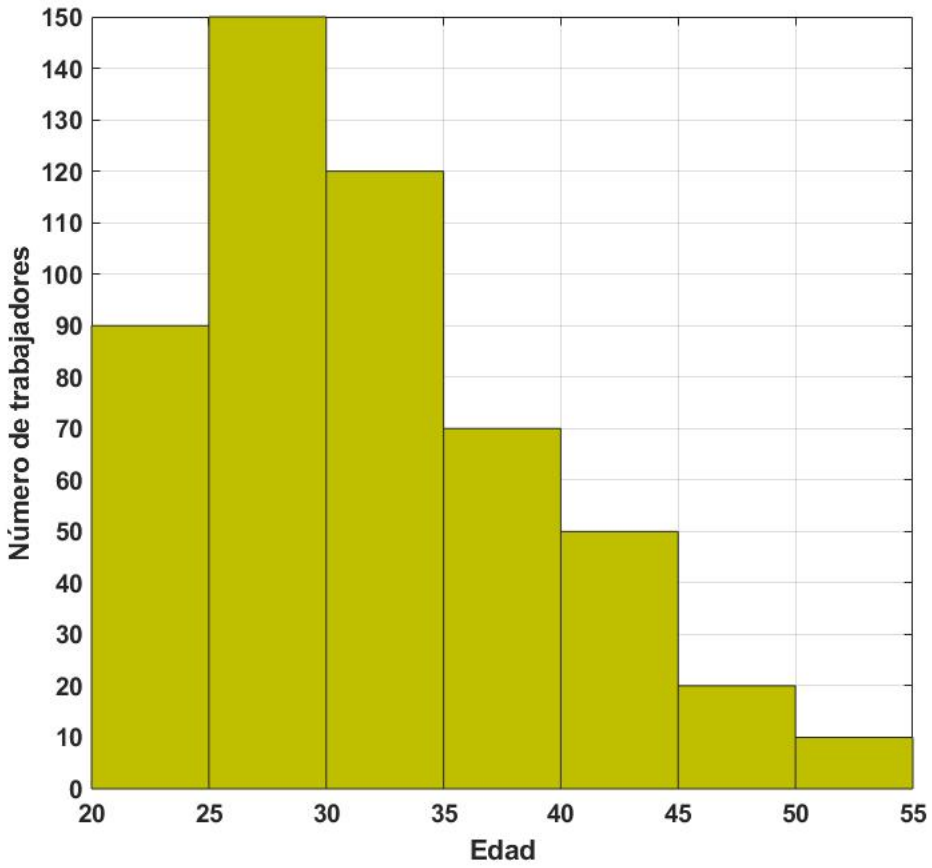

Figura 2: Histograma

También es posible realizar un polígono de frecuencias, al unir los puntos medios de la tapa superior de cada rectángulos por segmentos y agregando un valor nulo en los extremos del histograma.

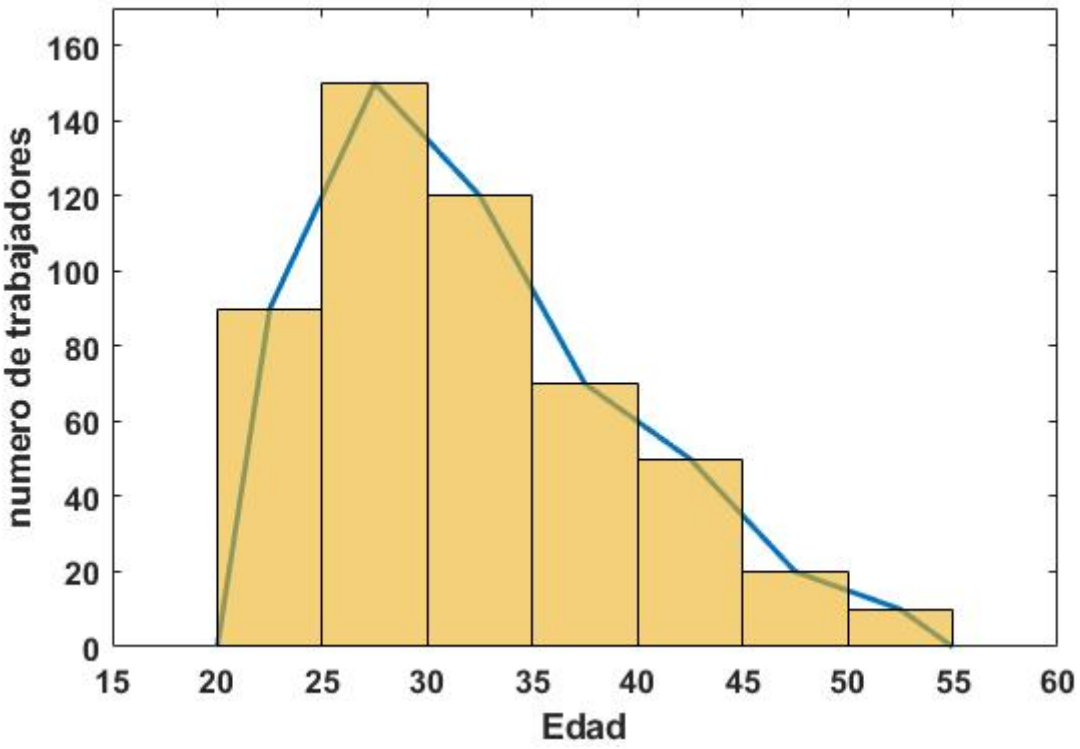

Figura 3: Polígono de frecuencias

Notar que este polígono de frecuencias puede considerarse una primera aproximación a la función de probabilidad que regiría a mis datos. Existen diversas formas y distintos programas para transformar un histograma en una función continua que represente como se comportan mis datos, es decir, es posible obtener a partir de un histograma una función de densidad de probabilidad para mi conjunto de datos.

Observación: la utilización de histogramas es válida también para datos categóricos, por ejemplo si quisieramos visualizar en un histograma la cantidad de autos de la argentina pero organizados según la marca, entonces cada clase (sub-intervalo) del mismo ahora sería una marca de auto y no un intervalo numérico.

1.2. Diagrama de tallo y hoja

Este diagrama es muy util para obtener una representación visual informativa del conjunto de mis datos. Supongamos que tenemos un conjunto de datos $x_1, x_2, x_3, ..., x_n$, donde cada x_i es un número de al menos dos dígitos, entonces la idea es dividir a cada número en dos partes, a cada una de esas partes las llamaremos tallo y hoja, respectivamente. El tallo será la parte o la información del dato que más nos interesa, y la hoja el resto. Veamosló con el mismo ejemplo de arriba. Supongamos uan fábrica con 510 trabajadores y nos interesa hacer una análisis sobre la edad de los mismo, entonces necesitamos visualizar esta información. Supongamos que los queremos separar en grupos dados por las décadas vividas. Usaremos 10 datos a modo de ejemplo

$$22, 24, 25, 31, 33, 34, 36, 37, 38, 44$$

Entonces dividiremos cada edad en dos partes. El tallo que corresponde a la década, 20,30 o 40 en este ejemplo, y a las hojas que corresponden a lo que resta de información de cada edad que sería cuantos años dentro de cada década. Esa información se reduce a la siguiente tabla.

Tallo	Hoja
20	2 4 5
30	1 3 4 6 7 8
40	4

Figura 4: Diagrama de tallo y hoja

Notar que esto es una forma de visualizar la información que es útil para un conjunto pequeño de datos, si la empresa tuviera 10.000 empleados, no sería apropiado tratar de visualizar la información con este método.

1.3. Medidas de localización

Del mismo modo que los gráficos sirven para presentar y analizar los datos, existen medidas numéricas que nos permiten caracterizar a nuestro set de datos.

Supongamos entonces que tenemos una muestra conformada por n datos $\{x_1, x_2, x_3, ..., x_n\}$.

Media muestral: la medida de localización más común o centro de un grupo de datos no es más que la media aritmética o promedio simple, esta media aritmética que llamaremos media muestral viene dada por

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

Mediana: La mediana es aquel valor de mis datos que separa a mi cantidad de observaciones en dos partes iguales. Supongamos que nuestros datos $\{x_1, x_2, x_3, ..., x_n\}$ están en orden creciente, entonces la mediana se define como la observación que está en la posición $\frac{n+1}{2}$, si n es impar, ahora bien si n es par definimos a la mediana como el promedio entre las observaciones $x_{n/2}$ y $x_{n/2+1}$, de tal forma que tanto a la derecha como a la izquierda de la mediana tengo la misma cantidad de datos.

$$\widetilde{x} = \begin{cases} x_{n/2} & si \quad n \ es \ impar \\ \frac{x_{n/2}+x_{n/2+1}}{2} & si \quad n \ es \ par \end{cases}$$

Moda: es la observación que se presenta con más frecuencia en la muestra.

Percentiles y cuartiles: Dijimos que la mediana nos dividía la muestra en dos partes iguales, también es posible dividirla en más de dos partes. Cuando se divide a un conjunto ordenado de datos en cuatro partes iguales, cada uno de estos puntos divisorios se los conoce como **quartil**. El primer quartil, q_1 es aquel valor de la muestra que debajo de él contiene alrededor del 25 % de los datos y el 75 % por delante de él. El segundo quartil q_2 tiene al 50 % de mis datos tanto por encima como por debajo de él, es decir, es la mediana. El tercer quartil q_3 tiene al 75 % de los datos por debajo de él. Como en el caso de la mediana es posible que los quartiles no sean únicos, en caso de que esto suceda tomaremos el promedio de los puntos que cumplan la definición de quartil.

Ejemplo: En 20 automóviles elegidos aleatoriamente se tomaron las emisiones de carbono en partes por millón (ppm).

$$141, 359, 247, 940, 882, 494, 306, 210, 105, 880, 200, 223, 188, 940, 241, 190, 300, 435, 241, 380$$

Primero ordenamos los datos de menor a mayor:

$$105, 141, 188, 190, 200, 210, 223, 241, 241, 247, 300, 306, 359, 380, 435, 494, 880, 882, 940, 940$$

Buscamos el 1er quartil, para eso hacemos $\frac{n}{4} = 5$, es decir el 5to dato corresponde al 25 % de la muestra, lo podemos tomar como quartil o para ser mas rigurosos con la definición y que el 25 % de mis datos se encuentren por debajo del quartil entonces tomo como quartil a $q_1 = \frac{x_5+x_6}{2} = \frac{200+210}{2} = 205$.

Buscamos la mediana o segundo cuartil, como n = 20 y es número par entonces la mediana es el promedio de la observaciones que se encuentran en los lugares $\frac{n}{2} = 10$ y $\frac{n}{2} + 1 = 11$, es decir $q_2 = \frac{247+300}{2} = 273{,}5$.

y el tercer quartil viene dado por $q_3 = \frac{x_15+x_16}{2} = \frac{434+494}{2} = 464{,}5$.

Entonces a mi conjunto de datos ordenados podría dividirlo en 100 partes iguales, definiendo los **percentiles**. El percentil p_k corresponde a aquel valor tal que el $100k$ % de mis datos se encuentran por debajo de él, y el $100(1-k)$ % de datos por encima de él.

$$p_{0,25} = q_1 \quad p_{0,5} = q_2 \quad p_{0,75} = q_3$$

1.4. Medidas de variabilidad

Rango: una medida de variabilidad es el rango de la muestra que viene dado por la diferencia entre la observación más grande y la más pequeña.

$$rango = max(x_i) - min(x_i)$$

Rango intercuartílico: Al igual que las observaciones máxima y mínima de una muestra llevan información sobre la variabilidad, el rango intercuartílico definido como $q_3 - q_1$puede emplearse como medida de variabilidad. Básicamente es el rango que contiene al 50 % de mis muestras centrales.

$$RIC = q_3 - q_1$$

Varianza y desvío estándar muestral: Se define la varianza muestral S^2 como

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

y al desvío estándar muestral como $S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$.

S^2 presenta las unidades de los datos pero al cuadrado, mientras que S tiene las mismas unidades que las observaciones.

Otra forma de escribir a la varianza es $S^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}}{n-1}$.

Coeficiente de variación muestral: viene definido por

$$cv = \frac{S}{\bar{x}}$$

El coeficiente de variación muestral es útil cuando se compara la variabilidad de dos o más conjuntos de datos que difieren de manera considerable en la magnitud de las observaciones.

Un ejemplo de esto es cuando se realiza la misma medida sobre algo utilizando dos instrumentos diferentes. Supongamos que queremos comparar dos termónetros distintos, entonces con el 1ro medidos 10 veces la temperatura de una persona y obtenemos que el promedio y el desvío estandar de esas mediciones dan $\bar{x_1} = 36{,}5$ y $S_1 = 0{,}25$ y con el 2do termómetro obtenemos $\bar{x_2} = 36{,}7$ y $S_2 = 0{,}1$

Entonces los coeficientes de variación muestral son $cv_1 = \frac{0{,}25}{36{,}5} = 0{,}0068$ y $cv_2 = \frac{0{,}1}{36{,}7} = 0{,}0027$, es decir $cv_1 > cv_2$. En consecuencia el 1er termómetro tiene una variabilidad en su medición mayor que la del termómetro 2.

1.5. Diagrama de Caja

El **diagrama de caja** es una presentación visual que describe simultaneamente varias características importantes de un conjunto de datos, tales como el centro, la dispersión, la desviación de la simetría y la presencia de valores atípicos. El diagrama de caja presenta los tres quartiles, y los valores mínimo y máximo de los datos sobre un rectángulo en posición horizontal o vertical.

En general, la dispersión de los cuartos no se ve afectada por las posiciones de las observaciones comprendidas en el 25 % más pequeño o el 25 % más grande de los datos. Por

consiguiente es resistente a valores apartados. La gráfica de caja más simple se basa en el siguiente resumen de cinco números:

$$x_{min} \qquad q_1 \qquad mediana \qquad q_3 \qquad x_{max}$$

Este gráfico se arma de la siguiente forma:

Los quartiles q_1 y q_3 son los valores que definen el tamaño de mi caja.

La mediana se marca con una línea vertical continua dentro de la caja, en el valor que corresponde a la mediana.

Luego se hacen 2 lineas horizontales que nacen desde el centro de los extremos de la caja hasta los valores mínimos y máximos.

Supongamos que tenemos los siguientes datos donde cada observación es la profundidad de picadura más grande en una una placa metálica debido al óxido.

40, 52, 55, 60, 70, 75, 85, 85, 90, 90, 92, 94, 94, 95, 98, 100, 115, 125, 125

El resumen de los cinco parámetros es como sigue:

$x_{min} = 40$ $\qquad$ $q_1 = 72{,}5$ $\qquad$ $mediana = 90$ $\qquad$ $q_3 = 96{,}5$ $\qquad$ $x_{max} = 125$

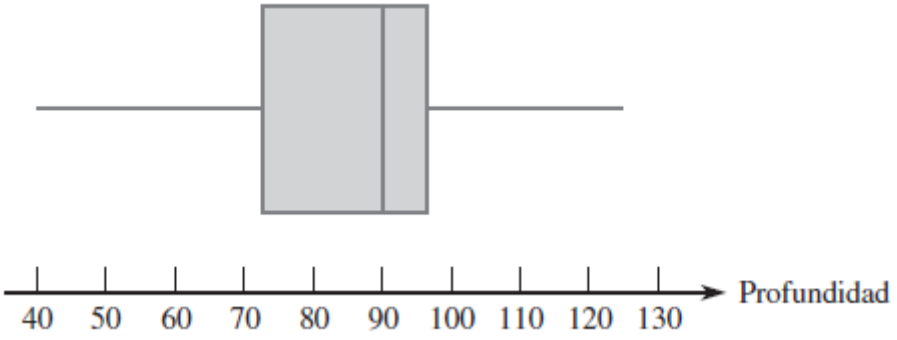

Figura 5: Diagrama de Caja

El lado derecho de la caja está mucho más cerca a la mediana que el izquierdo, lo que indica una asimetría sustancial en la mitad derecha de los datos. El ancho de la caja también es razonablemente grande con respecto al rango de datos (distancia entre las puntas de los bigotes).

Existe otra forma de hacer el gráfico de cajas que es reemplazando el x_{min} y x_{max} con los puntos que estan en los valores $q_1 - 1{,}5RIC$ y $q_3 + 1{,}5RIC$ es decir, aquellos puntos que se alejan de los quartiles en un 1.5RIC (recordar que el RIC es el rango intercuartículico = $q_3 - q_1$ que corresponde al ancho de la caja), en este caso serían $q_1 - 1RIC = q_1 - 1{,}5(q_3 - q_1) = 72{,}5 - 1{,}5(96{,}5 - 72{,}5) = 36{,}5$ y $q_3 + 1{,}5RIC = q_3 + 1{,}5(q_3 - q_1) = 96{,}5 + 1{,}5(96{,}5 - 72{,}5) = 132{,}5$. Entonces ahora tendríamos el siguiente diagrama.

$q_1 - 1{,}5RIC = 36{,}5$ $q_1 = 72{,}5$ $mediana = 90$ $q_3 = 96{,}5$ $q_3 + 1{,}5RIC = 132{,}5$

Cualquier valor por fuera de este rango se lo marcará de forma independiente (con un símbolo a elección, un punto un círculo, etc).Las observaciones que se encuentran entre 1.5 y 3 veces el rango intercuartílico a partir de las aristas del rectángulo reciben el nombre de valores atípicos. Las observaciones que están más allá de tres veces el rango intercuartílico a partir de las aristas del rectángulo se conocen como valores atípicos extremos.

En nuestro ejemplo ningún valor es menor a 36.5 o mayor a 132.5 por lo tanto el diagrama de caja queda como está arriba (con x_{min} y x_{max}) sin valores atípicos o extremos. Abajo se muestra el gráfico de otro conjunto de datos para visualizar estas ideas.

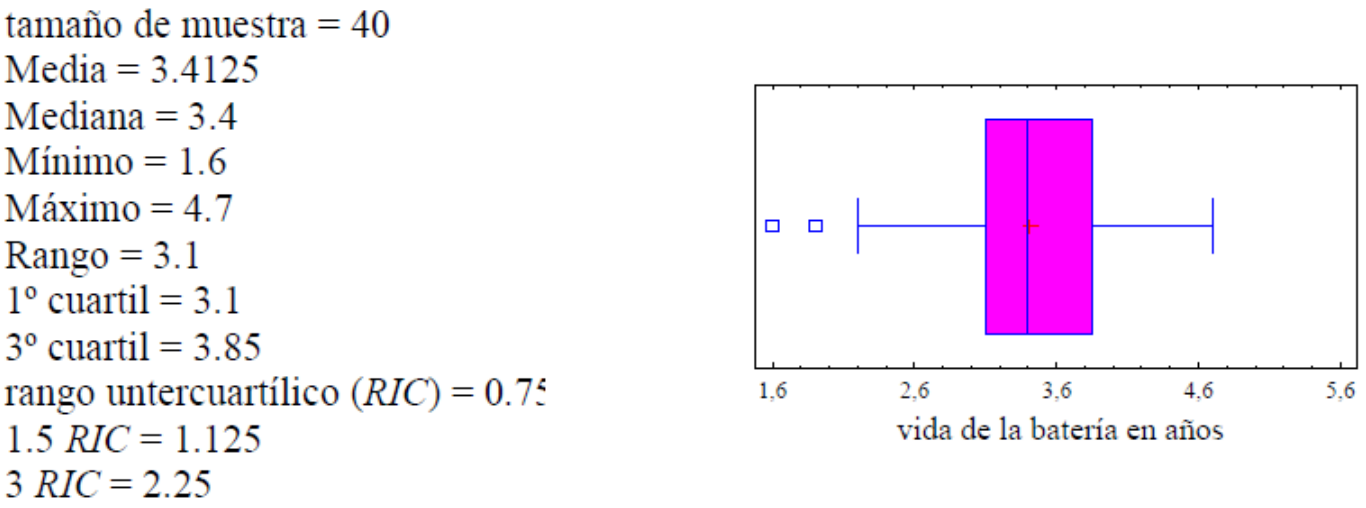

Figura 6: Diagrama de Caja con valores atípicos

2. Teorema del Límite central

Dada una población, si tomamos muestras lo suficientemente grandes (mayores a 30 elementos) entonces la media muestral tiende a seguir una distribución gaussiana, no importa la distribución de probabilidad que siga la población. Además el valor medio de la distribución gaussiana de la media muestral tiende a la media de la población.

En términos de variables aleatorias lo puedo describir de la siguiente forma. Sean $X_1, X_2, ..., X_n$ variables aleatorias independientes con $E[X_i] = \mu$ y $V[X_i] = \sigma^2$ $\forall i$. No importa la distribución de estas variables aleatorias.

$\Longrightarrow \bar{X} = \frac{\sum X_i}{n}$ sigue una distribución gaussiana con $E[\bar{X}] = \mu$ y varianza $Var(\bar{X}) = \frac{\sigma^2}{n}$.

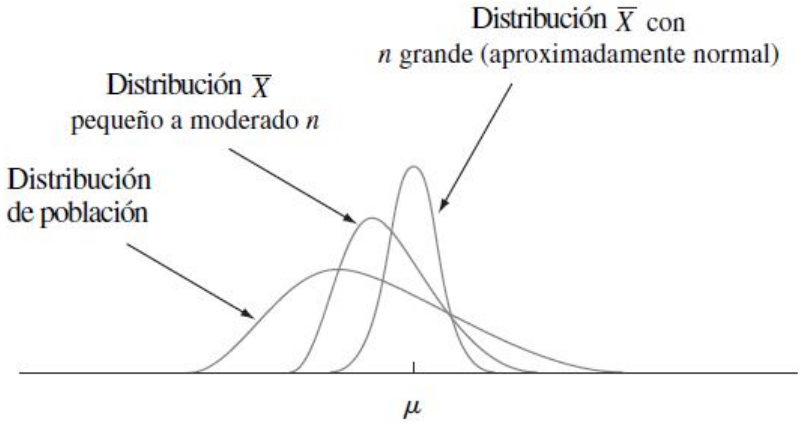

Figura 7: Explicación gráfica Teorema del Límite central

Mientras mayor sea el tamaño de la muestra, más se parecerá la distribución de $\bar{X}$ a una distribución normal.

Por supuesto que si $\bar{X} = \frac{\sum X_i}{n}$ sigue una distribución gaussiana entonces $S_n = \sum X_i$, tambíen seguirá una distribución gaussiana, en este caso con $E[S_n] = n\mu$ y $V[S_n] = n\sigma^2$.

Recordemos que estas variables aleatorias $\bar{X}$ y S_n, que siguen sus respectivas distribuciones gaussianas, las podemos transformar en variables aleatorias que sigan una distribución normal estándar N(0,1).

$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ y $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$, ambas siguen una N(0,1).

Ejemplo: Aplicado a una población.

Supongamos que tengo una población, tomo una muestra que llamo M_1 de 50 elementos seleccionados al azar, y a esta muestra le calculo su media muestral que llamo μ_1 y su varianza muestral σ_1^2. Repito el procedimiento, tomo una muestra M_2 de 50 elementos y calculo μ_2 y σ_2^2.

Supongamos entonces que repito esto 1000 veces, es decir, tome mil muestras y a cada una le estimé su media y su varianza muestral.

Entonces tengo el siguiente conjunto de datos

$M_1 \longrightarrow \mu_1\ ,\ \sigma_1^2$

$M_2 \longrightarrow \mu_2\ ,\ \sigma_2^2$

...

...

...

$M_{1000} \longrightarrow \mu_{1000}\ ,\ \sigma_{1000}^2$

Entonces si ahora hago un histograma de las **medias muestrales**, el mismo va a tener una forma aproximadamente gaussiana con valor medio $\mu = \frac{\sum \mu_i}{1000}$ y varianza $\sigma^2 = \frac{\sum \sigma_i^2}{1000}$.

Además, si tomo una muestra de forma aleatoria a cada elemento de mi muestra lo puedo considerar una variable aleatoria, ya que cuando se "sacó" cada elemento i-ésimo para formar la muestra, este elemento podía ser cualquiera de mi población, por lo tanto a cada elemento x_i de una muestra lo puedo considerar como una varibale aleatoria independiente.

Ejemplo: de aplicación.

Supongamos que tengo 30 instrumentos electrónicos $D_1, D_2, D_3, ..., D_{30}$ que conforman un circuito de la siguiente manera: primero está en funcionamiento D_1, tan pronto como falla se activa D_2 y así sucesivamente hasta D_{30}, cuando falla el último instrumento entonces el circuito deja de operar. Supongamos que el tiempo de falla de cada D_i sigue una distribución exponencial con parámetro $\lambda = 0{,}1$ horas. Sea T : *el tiempo total de operación del circuito.* ¿ Cuál es la probabilidad de que el circuito opere más de 350 horas?

X_i :*tiempo de falla del instrumento i-ésimo.*
$X_i \sim Exp(0{,}1)$

$T = \sum_{i=1}^{30} X_i$
$E[T] = E[\sum_{i=1}^{30} X_i] = \sum_{i=1}^{30} E[X_i] = 30\frac{1}{\lambda} = 300$

$V[T] = V[\sum X_i]$, como los X_i son independientes entre ellos.
$V[T] = V[\sum X_i] = V[T] = \sum V[X_i] = 30\frac{1}{\lambda^2} = 3000$

$\Longrightarrow$ por TLC la variable aletoria $Z = \frac{T-300}{3000} \sim N(0,1)$

$P(T > 350) = P(\frac{T-300}{\sqrt{3000}} > \frac{350-300}{\sqrt{3000}}) = 1 - P(\frac{T-300}{\sqrt{3000}} < \frac{350-300}{\sqrt{3000}}) = 1 - \Phi(\frac{350-300}{\sqrt{3000}}) = 1 - \Phi(0,9128) = 0,1814.$