Institución Universitaria

Estadística Básica

jessicarojas5708@correo.itm.edu.co

Capítulo 1

Estadística como ciencia

1.1. Introducción

Iniciamos con la definición de algunos conceptos elementales y básicos, para una comprensión intuitiva y real de lo que es la estadística. Son muchas las definiciones que existen de lo que es la estadística como ciencia del conocimiento, entre estas se destacan las siguientes:

- Ciencia de las matemáticas que se encarga de la selección, recolección, tabulación, presentación y análisis de la información que se utiliza en la toma de decisiones.
- Conjunto de métodos para efectuar decisiones adecuadas frente a la incertidumbre.
- Operación de análisis matemático, que permite estudiar con el máximo de presición, los fenómenos incompletamente conocidos.

Las aplicaciones más importantes en el campo de la estadśitica se relacionan con:

- Recolección de datos.
- Registro y presentación de la información.
- Formulación de modelos.
- Pruebas de hipótesis.
- Diseños de experimentos.

Por lo anterior, se tiene que la estadśitica constituye una herramienta auxiliar en las investigaciones, para planificar la obtención de la información, analizar esta información y extraer conclusiones válidas en términos de probabilidad y así de esta forma tomar decisiones. La estadística se divide en dos grandes ramas: estadística descriptiva y estadística inferencial.

Estadística descriptiva Comprende los procesos de consolidación, resumen y descripción de los datos recopilados. tablas, gráficos o índices que permiten un análisis referido exclusivamente a los datos coleccionados.

Estadística inferencial Incluye procedimientos que permiten la extrapolación y generalización sobre características que tipifican a todos los elementos de la población. Puede decirse que es el proceso de hacer afirmaciones o predicciones sobre toda la población, tomando como base, sólo a la información recolectada a través de una muestra.

1.2. Algunos conceptos básicos

A continuación se presentan los principales conceptos básicos de estadística

Población

Conjunto de elementos que son de interés en un estudio (poseen características comunes acerca de los cuales se desea tener información). Usualmente a dichos elementos se les denomina individuos, observaciones o mediciones. La población puede ser finita o infinita.

Muestra

Es una parte de la población. Para estudios estadísticos, se requieren muestras que nos den información real de la población. El **Muestreo** es el proceso mediante el cual se seleccionan los elementos de una población.

Variables estadísticas

Son características o atributos de interés que pueden ser observadas en los elementos poblacionales. Algunos ejemplos de variables son presión sanguínea diastólica, frecuencia cardiaca, estatura de varones adultos, peso de niños en edad preescolar y la edad de los pacientes que consultan a un odontologo.
Las variables se clasifican en variables cualitativas y variables cuantitativas.
Las cualitativas son aquellas que describen cualidades de los elementos. Algunos ejemplos de estas variables son tipo de sangre, cuyas modalidades o categorías son: O, A, B, AB, estado civil, documento de identificación, filiación política, tipo de religión, raza, tipo de suelo,etc.
Las cuantitativas: son aquellas que generalmente resultan de un proceso de medición. Pueden ser *discretas* o *continuas*.

Discretas:

resultan de conteos y el resultado es un número entero. Ejemplos: *Numero de hermanos*, cuyas modalidades o categorías son: $0, 1, \ldots, N$, *número de pacientes que llegan a un centro de salud un intervalo de tiempo dado, número de sillas en un salón de clases, número de horas que un estudiante dedica semanalmente a sus asignaturas.*

Continuas:

el resultado es un subconjunto de los números reales. Ej: *Tiempo de espera en una parada de bus, velocidad de un vehiculo en una autopista, ingreso económico del jefe de hogar en una familia.*

Escalas de medición

La medición hace referencia a la asignación de números a las características objeto de estudio.

Escala nominal:

Es la más baja de las escalas de medición. Identifica las categorías de la variable de interés y se pueden diferenciar las categorías una de la otra haciendo uso de dígitos. Ej: *Estado civil* (soltero - casado - viudo - unión libre - separado), *cédula de ciudadanía*, *género musical*, *tipo de sangre*, *estado del paciente*.

Escala ordinal:

Identifica las categorías de la variable y pueden ser clasificadas por grados de acuerdo a algún criterio. La función de los dígitos asignados a datos ordinales es la de ordenar. Ej: *Grado de escolaridad* (ninguno - primaria - secundaria - profesional - postgrado), *rangos militares*, *grados de desnutrición*, *tipo de quemadura.*

Escala de intervalos:

La escala de intervalos es una escala más especializada que la nominal y la ordinal, en el sentido de que no solo es posible ordenar las mediciones, sino que también se conoce la distancia entre las observaciones cualesquiera. Aquí no hay un punto cero único. Ej: la escala en la que se mide la *temperatura;* no es posible decir que 30° es doble de frio que 60°, ya que depende de la escala (grados celsius o Fahrenheit). Otros ejemplos son *pérdida auditiva en decibeles* y *coeficiente intelectual en puntaje.*

Escala de razón:

Es el nivel más alto de las escalas de mediciones y se caracteríza por el hecho de que se puede determinar tanto la igualdad de razones como la de intervalos. Existe un punto cero único. Por ejemplo: *altura, peso, longitud, velocidad, área, volúmen.*

Formas de presentación y organización de la información

Existen dos formas básicas para la representación de la información recolectada, a través de tablas o cuadros estadísticos y a través de un gráfico.

Una tabla o cuadro estadístico es una representación en forma ordenada de la variación de un fenómeno, clasificado bajo uno o más variables. Puede ser simple (clasificación bajo una variable) o compuesto (clasificación bajo dos o más variables). A continuación se presenta una serie de **términos relacionados para tablas estadísticas**

Frecuencia absoluta n_i

Sea X una variable estadística cuyos valores son $X_1, \ldots, X_k$; de una muestra de tamaño n, $(k \leq n)$. La frecuencia absoluta corresponde al número de veces que se repite cada valor de la variable.

Ejemplo 1. *El número de vehículos que llegan a un taller automotor en un día dado, es una variable de tipo estadístico que se observó durante un período de 25 días y se obtivieron los siguientes datos:*

8	6	7	9	8
7	8	10	4	10
8	7	9	8	7
6	5	10	7	8
5	6	8	10	11

Se puede definir la variable X como: X : número de vehículos que llegan al taller, Tipo: cuantitativa discreta,

Ejemplo 2. *Una encuesta realizada a 30 fumadores para determinar el número de cigarrillos que encienden (fuman) en un dia corriente arrojó los siguientes resultados:*

Cuadro 1.1: Distibución de frecuencias para el número de vehículos que llegan a un taller automotor en un día dado

Valor de X_i	Frec. Abs. n_i	Frec. Relativa. h_i	Frec. Abs. Acum. N_i	Frec. Relativa Acum. H_i
4	1	0.04	1	0.04
5	2	0.08	3	0.12
6	3	0.12	6	0.24
7	5	0.12	11	0.44
8	7	0.28	18	0.72
9	2	0.08	20	0.80
10	4	0.16	24	0.96
11	1	0.04	25	1.00
Total	25	1	-	-

3	*7*	*5*	*10*	*8*	*4*
5	*8*	*10*	*8*	*8*	*4*
5	*3*	*10*	*5*	*7*	*10*
8	*5*	*5*	*12*	*8*	*4*
4	*3*	*5*	*8*	*12*	*10*

Se puede definir la variable X como: X : número de cigarrillos que encienden un fumador, Tipo: cuantitativa discreta,

Valor de X_i	Frec. Abs. n_i	Frec. Relativa. h_i	Frec. Abs. Acum. N_i	Frec. Relativa Acum. h_i
3	3	0.100	3	0.100
4	4	0.133	7	0.233
5	7	0.233	14	0.467
7	2	0.067	16	0.533
8	7	0.233	23	0.767
10	5	0.167	28	0.933
12	2	0.067	30	1.000
Total	25	1	-	-

Ejemplo 3. *A continuación se presentan los datos sobre el octanaje del combustible para motores de varias marcas de gasolina*

88.5	*89.8*	*89.9*	*90.6*	*93.4*	*90.7*	*90.1*
94.7	*91.6*	*98.8*	*92.2*	*96.1*	*88.6*	*89.3*
84.3	*90.3*	*88.3*	*87.7*	*89.6*	*88.3*	*91.1*
90.1	*90.0*	*90.4*	*91.1*	*90.4*	*94.2*	*83.4*
89.0	*91.5*	*91.2*	*86.7*	*91.6*	*85.3*	*93.2*

Se puede definir la variable X como: X : octanaje del combustible, Tipo: cuantitativa continua,

Ejemplo 4. *Los siguientes datos corresponden a un muestreo de ruido ambiental del nivel de presión sonora (LP), medida en decibeles (dB) en diferentes estaciones de la ciudad de Cali durante el día*

63.7	*75.0*	*74.1*	*69.4*	*64.6*	*71.6*
66.9	*76.3*	*73.7*	*76.5*	*60.5*	*72.1*
66.8	*75.0*	*71.0*	*57.3*	*65.1*	*62.3*
75.3	*77.4*	*56.1*	*71.6*	*55.3*	*72.3*
70.8	*71.4*	*69.0*	*67.2*	*71.3*	*70.5*

Se puede definir la variable X como: X : ruido ambiental del nivel de presión sonora y el tipo de variable es cuantitativa continua.

Medidas descriptivas

Son valores que caracterizan las observaciones de un conjunto de datos. Estas medidas de resumen pueden ser de *centralidad, dispersión o variabilidad, posición, asimetría y apuntamiento*

Medidas de centralidad

Son valores que representan un valor central hacia el cual tiene tendencia a concentrarse el conjunto de datos.

Media aritmética Es la medida más utilizada en un conjunto de datos, es un valor central que toma en cuenta todos los valores que aparecen en el conjunto de datos y las distancias relativas a estos valores. Los valores tienen la misma importancia en el grupo de datos.

Sean $x_1, x_2 \ldots, x_n$ los valores de una variable X, de una muestra de tamaño n. La media aritmética $\overline{x}$ se define como:

$$\overline{x} = \frac{1}{n}(x_1 + x_2 \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Ejemplo 5. *Para los datos del ejemplo 1. La media aritmética es*

$$\overline{x} = \frac{1}{25}(1 \times 4 + 2 \times 5 + \cdots + 1 \times 11) = 7{,}68$$

Ejemplo 6. *Para los datos del ejemplo 3. La media aritmética es*

$$\overline{x} = \frac{1}{35}(3 \times 84{,}65 + 5 \times 87{,}25 + \cdots + 1 \times 97{,}65) = 90{,}22$$

Mediana Es la segunda medida más utilizada después de la media aritmética, y es útil para estimar el centro de un conjunto de datos. La mediana es el elemento central del conjunto de datos, es una medida de posición y hay el mismo número de observaciones a la derecha y a la izquierda del valor de la mediana.
La mediana se calcula como sigue, si la variable X tiene n valores diferentes, $x_1, \ldots, x_n$, entonces la mediana se escribe como:

$$\text{Me} = \begin{cases} X_{\left(\frac{n+1}{2}\right)}, & \text{si } n \text{ es impar;} \\ \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2}, & \text{si } n \text{ es par.} \end{cases}$$

Moda Representa el valor o valores que tienen la mayor frecuencia dentro del conjunto de datos. La moda puede o no existir; en el evento en que exista, puede no ser única, ya que una distribución puede eventualmente tener una o varias modas

Medidas de dispersión

Permiten generar criterios sobre el grado de homogeneidad o heterogeneidad del conjunto de datos que se está analizando, en relación con una medida de centralidad, o con respecto a datos entre sí.

Rango diferencia entre al valor máximo y el valor mínimo del conjunto de datos y mide la longitud en la cual se encuentran los datos, en general a mayor longitud mayor dispersión de los datos.

$$R = X_{(n)} - X_{(1)}$$

Varianza La varianza mide las variaciones del conjunto de datos con respecto a su media aritmética y se define comola media aritmética de los cuadrados de las desviaciones de cada dato a la media aritmética.
Si la variable X tiene n valores diferentes, $x_1, \ldots, x_n$, entonces la varianza se escribe como:

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

La expresión anterior pueden ser alternativamente escrita como:

$$S^2 = \frac{1}{n}\left[\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right]$$

Desviación estándar Una de las limitaciones de la varianza son sus unidades al cuadrado. Para superar esto se usa la raíz cuadrada de la varianza, dando origen al concepto de desviación estándar.

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

Coeficiente de variación Permite estimar la relación porcentual entre el valor de la media y la desviación estándar. A medida que se presenta mayor heterogeneidad en el conjunto de datos, el valor del coeficiente de variación es mayor

$$CV = \frac{S}{\overline{x}} \times 100\,\%$$

Ejercicios resueltos

Se conduce un estudio relacionado con los efectos de fumar sobre los patrones de sueño. La medición que se observa es el tiempo, en minutos que toma quedar dormido. Al finalizar el estudio se obtienen los siguientes resultados:

No fumadores	28.6	25.1	26.4	34.9	29.8	28.4
	38.5	30.2	30.6	31.8	41.6	21.1
Fumadores	69.3	56.0	22.1	47.6	53.2	23.2
	48.1	52.7	34.4	60.2	43.8	13.8

Se procede a caracterizar la información anterior, utilizando las medidas de tendencia central y las medidas de dispersión.
Como tenemos dos grupos diferentes **No fumadores** y **Fumadores** se realizará el cálculo de dichas medidas por grupo en forma independiente, iniciando con el grupo de los **No fumadores.**

El tiempo que transcurre (en minutos) hasta que un individuo que dormido, es una variable aleatoria cuantitativa continua.

Por definición, **la media aritmética**

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

con x_i los diferentes valores que toma el tiempo en minutos que transcurre hasta que el individuo quede dormido, para $i = 1, 2, \ldots, 12$, el número de observaciones n es igual a 12, $\sum_{i=1}^{n} x_i = 367$ y por tanto,

$$\overline{X} = \frac{1}{12}\sum_{i=1}^{12} X_i = \frac{367}{12} = 30{,}583$$

Interpretación: el valor promedio del tiempo que transcurre (en minutos) hasta que un individuo que dormido en el grupo de **No fumadores** es igual a 30.583 minutos.

Para determinar la **mediana** ordenamos los datos de menor a mayor, obteniendo los siguientes resultados, 21.1 25.1 26.4 28.4 28.6 29.8 30.2 30.6 31.8 34.9 38.5 41.6

Como el número de observaciones es un número par, entonces la mediana corresponde a el valor en la serie de datos que está en la posición

$$\begin{aligned} Me &= \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2} \\ &= \frac{X_{\left(\frac{12}{2}\right)} + X_{\left(\frac{12}{2}+1\right)}}{2} \\ &= \frac{X_{(6)} + X_{(7)}}{2} \\ &= \frac{29{,}8 + 30{,}2}{2} \\ &= \frac{60}{2} \\ &= 30 \end{aligned}$$

Interpretación: como la mediana es igual a 30, este valor, estabece que el 50 % de los valores del tiempo en minutos que transcurre ordenados en forma creciente están por debajo de 30 minutos y el 50 % de los valores estantes están por encima de 30 minutos.

Al examinar los valores, se encuentra que no existe un valor que se registre más de una vez, por tanto, no existe la moda y en esté caso se dise que la distribución de frecuencia es amodal.

Continuando se procede a calcular las medidas de dispersión:

- **Rango:**=$X_{\text{máx}} - X_{\text{mín}} = 41{,}6 - 21{,}1 = 20{,}5$. Es decir, la longitud en que se encuentra un dato del otro es de aproximadamente de 20.5 minutos.
- **Varianza:**

$$\begin{aligned} S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right) \end{aligned}$$

  Para los datos del grupo de los **No Fumadores** tenemos que:

$$\begin{aligned} S^2 &= \frac{1}{12-1}\left(\sum_{i=1}^{12} X_i^2 - 12(30{,}583)^2\right) \\ &= \frac{1}{11}\left(\sum_{i=1}^{12} X_i^2 - 12(935{,}319)\right) \\ &= \frac{1}{11}\left(\sum_{i=1}^{12} X_i^2 - 11223{,}828\right) \\ &= \frac{1}{11}\left(11575{,}2 - 11223{,}828\right) \\ &= \frac{1}{11}\left(351{,}372\right) \\ &= 31{,}942 \end{aligned}$$

  Así, la varianza es igual a 31.942

- **Desviación estándar:** $S = \sqrt{S^2} = \sqrt{31{,}942} = 5{,}651$, es decir, la variación del que trascurre hasta que se quede dormido difieren de su media en aproximadamente en 5.6517 unidades.
- **Coeficiente de variación:**
$$CV = \frac{S}{\overline{X}} = \frac{5{,}651}{30{,}583} = 0{,}184$$
Al multiplicar por 100 %, tenemos 18.4; lo cual indica que la variación de los valores del tiempo transcurrido en minutos.

Ahora observemos la caracterización de la variable de interés en el grupo de los **Fumadores**

La media aritmética
$$\overline{X} = \frac{1}{12}\sum_{i=1}^{12} X_i = \frac{524{,}4}{12} = 43{,}7$$

Interpretación: el valor promedio del tiempo que transcurre (en minutos) hasta que un individuo que dormido en el grupo de **No fumadores** es igual a 43.7 minutos.
Para determinar la **mediana** ordenamos los datos de menor a mayor, obteniendo los siguientes resultados, 13.8 22.1 23.2 34.4 43.8 47.6 48.1 52.7 53.2 56.0 60.2 69.3
Como el número de observaciones es un número par, entonces la mediana corresponde a el valor en la sere de datos que está en la posición

$$\begin{aligned} Me &= \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2} \\ &= \frac{X_{\left(\frac{12}{2}\right)} + X_{\left(\frac{12}{2}+1\right)}}{2} \\ &= \frac{X_{(6)} + X_{(7)}}{2} \\ &= \frac{47{,}6 + 48{,}1}{2} \\ &= \frac{95{,}7}{2} \\ &= 47{,}85 \end{aligned}$$

Interpretación: como la mediana es igual a 47.85, este valor, estabece que el 50 % de los valores del tiempo en minutos que transcurre ordenados en forma creciente están por debajo de 47.85 minutos y el 50 % de los valores estantes están por encima de 47.85 minutos.

Al examinar los valores, se encuentra que no existe un valor que se registre más de una vez, por tanto, no existe la moda y en esté caso se dise que la distribución de frecuencia es amodal.

Continuando se procede a calcular las medidas de dispersión:

- **Rango:**$=X_{\text{máx}} - X_{\text{mín}} = 69{,}3 - 13{,}8 = 55{,}5$. Es decir, la longitud en que se encuentra un dato del otro es de aproximadamente de 55.5 minutos.
- **Varianza:**
$$\begin{aligned} S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right) \end{aligned}$$

Para los datos del grupo de los **Fumadores** tenemos:

$$\begin{aligned} S^2 &= \frac{1}{12-1}\left(\sum_{i=1}^{12} X_i^2 - 12(43{,}7)^2\right) \\ &= \frac{1}{11}\left(\sum_{i=1}^{12} X_i^2 - 12(1909{,}69)\right) \\ &= \frac{1}{11}\left(\sum_{i=1}^{12} X_i^2 - 22916{,}28\right) \\ &= \frac{1}{11}\left(26068{,}32 - 22916{,}28\right) \\ &= \frac{1}{11}\left(3152{,}04\right) \\ &= 286{,}54 \end{aligned}$$

Así, la varianza es igual a 286.54

- **Desviación estándar:** $S = \sqrt{S^2} = \sqrt{286{,}54} = 16{,}92$, es decir, la variación del que trascurre hasta que se quede dormido difieren de su media en aproximadamente en 16.9274
- **Coeficiente de variación:**

$$CV = \frac{S}{\overline{X}} = \frac{16{,}92}{43{,}7} = 38\,\%$$

Es decir, la variación del tiempo que trascurre hasta que la persona se quede dormida es de aproximadamente el 38,735 %

Medidas de posición

Cuando se desea presentar un análisis con respecto a la posición que ocupa la información que resulta relevante, las medidas de posición son muy utiles.

Las medidas de posición son valores que particionan la población o muestra en varios puntos, dando una descripción más fina, puesto que dan más información del comportamiento de los datos que las medidas de tendencia central (media aritmetica, mediana y moda).

Estas medidas indican que porcentaje de datos dentro de una distribución de frecuencias superan estas expresiones (mitad, 3 partes, 5 partes, diez partes, etc) y facilitan la información sobre la serie de datos que estamos analizando. Entre las medidas de posición más utilizadas encontramos los cuartiles, deciles y percentiles.

Definición 1. ***Cuartiles***
Los cuartiles son tres valores que distribuyen la serie de datos, ordenada de forma creciente o decreciente en cuatro tramos iguales, en los que cada uno de ellos concentra el 25 % de las observaciones. Estos valores on denotados por Q_1, Q_2, Q_3 y establecen las siguientes convenciones:

- Q_1 := *es aquel valor que supera al 25 % de los datos y es superado por el 75 % restante.*
- Q_2 := *supera y es superado por el 50 % de los datos.*
- Q_3 := *supera al 75 % y es superado por el 25 % de los datos restantes.*

Definición 2. ***Deciles***
Los deciles son nueve valores que distribuyen la serie de datos, ordenada de forma creciente o decreciente en diez tramos iguales, en los que cada uno de ellos concentra el 10 % de las observaciones.

Definición 3. ***Percentiles***
Los percentiles son noventa y nueve valores que distribuyen la serie de datos, ordenada de forma creciente o decreciente en cien tramos iguales, en los que cada uno de ellos concentra el 1 % de las observaciones.

De forma general se tiene que, el $100k-$*ésimo percentil* $0 < k < 1$*, denotado por* p_k*, es un valor tal que al menos el* $100k\,\%$ *de las observaciones son menores o iguales que él y al menos el* $100(1-k)\,\%$ *son mayores o iguales que él.*
Para calcular el percentil $100k\,\%$ *se procede de la siguiente forma:*

1. *Ordene los datos en forma creciente, es decir, de menor a mayor.*
2. *Calcule* nk*, donde* n *es el número de datos*
   a) *Si* nk *no es entero aproxímelo al entero siguiente y esa es la posición del percentil* $100k\,\%$.
   b) *Si* nk *es entero, el percentil* $100k\,\%$ *se obtiene promediando las observaciones que ocupan los lugares* nk *y* $nk+1$.

Ejemplo 7. *Los datos que se muestran a continuación corresponden a el peso (en kilogramos) de 25 niños al momento de nacer*

2536	2505	2652	2573	2380
2443	2617	2556	2489	2415
2434	2491	2345	2350	2536
2577	2464	2571	2550	2437
2472	2580	2436	2200	2851

Para los datos anteriores, se procede a calcular las medidas de posición.

- Para obtener el primer cuartil Q_1, determinamos el valor de la observación que se encuentra en la posición $(n = 25)(k = 0{,}25) = 6{,}25 \approx 7$, por tanto $Q_1 = 2436$.
- Para obtener el segundo cuartil Q_2, determinamos el valor de la observación que se encuentra en la posición $(n = 25)(k = 0{,}50) = 12{,}5 \approx 13$, por tanto $Q_2 = 2491$.
- Para obtener el tercer cuartil Q_3, determinamos el valor de la observación que se encuentra en la posición $(n = 25)(k = 0{,}75) = 18{,}5 \approx 19$, por tanto $Q_3 = 2571$.
- Para obtener el valor del percentil $p_{0{,}80}$ determinamos el valor de la observación que se encuentra en la posición $(n = 25)(k = 0{,}80) = 20$ como es un número entero, se procede a promediar el que está en la posición 20 y 21 por tanto $p_{0{,}80} = \frac{2573+2577}{2} = 2575$.
- Para obtener el percentil $p_{0{,}90}$, determinamos el valor de la observación que se encuentra en la posición $(n = 25)(k = 0{,}90) = 22{,}5 \approx 23$, por tanto $p_{0{,}90} = 2612$.
- Para obtener el percentil $p_{0{,}95}$, determinamos el valor de la observación que se encuentra en la posición $(n = 25)(k = 0{,}95) = 23{,}75 \approx 24$, por tanto $p_{0{,}95} = 2652$.

Nota 1. *Una forma util de representar la variabilidad de los datos es de manera gráfica, utilizando el diagrama de cajas o box plot, el cual se construye a partir de los cuartiles.*

Definición 4. *Diagrama de caja y bigotes.*
Este tipo de gráfico también llamado box and whisker plot, o simplemente box plot, facilita la lectura sobre localización, variabilidad, simetría y presencia de datos atípicos (outliers según la literatura estadística inglesa). El box plot consiste en una caja y guiones con una linea a través de la caja que representa la mediana (segundo cuartil Q_2*). El extremo inferior de la caja es el primer cuartil* Q_1

y el superior es el tercer cuartil Q_3. El bigote superior se extiende desde el tercer cuartil hasta la observación más grande que es menor o igual que $Q_3 + 1{,}5x(Q_3 - Q_1)$. El bigote inferior se extiende hasta la observación más pequeña que es mayor o igual que $Q_1 - 1{,}5x(Q_3 - Q_1)$. Las observaciones que están por fuera de estos límites se clasifican como datos atípicos y se ubican en el diagrama.

Para construir el diagrama de caja siga los siguientes pasos:

- *Dibujar y marcar un eje de medida vertical (eje de coordenadas).*
- *Construir un rectángulo cuyo borde inferior se ubica en el cuartil inferior (Q_1) y cuyo borde superior se ubica en el cuartil superior (Q_3).*
- *Dibujar un segmento de recta horizontal dentro de la caja justo en el segundo cuartil (mediana).*
- *Prolongar una recta (el bigote) desde el extremo superior de la caja hasta la observación más grande que es menor o igual que $Q_3 + 1{,}5x(Q_3 - Q_1)$.*
- *Prolongar una recta (el otro bigote) desde el extremo inferior de la caja hasta la observación más pequeña que es mayor o igual que $Q_1 - 1{,}5 \times (Q_3 - Q_1)$*

1.3. Ejercicios propuestos

1. Se realizó un estudio para determinar la eficacia de la vacuna BCG (bacillus-Calmette-Guerín) realizaron un estudio para prevenir la meningitis tuberculosa. Entre los datos recolectados e cada individuo está la medición del estado nutricional (peso expresado como porcentaje del peso esperado para cada estatura real). La siguiente tabla muestra los valores:

73.3	80.5	50.4	50.9
64.8	74.0	72.8	72.0
59.7	90.0	76.9	71.4
45.6	77.5	60.6	67.5
54.6	71.0	66.0	71.0
82.6	70.5		

2. Se efectuó un estudio para investigar si la autotransfusión de sangre extraida del mediastino podia reducir el numero de pacientes que necesitaba transfusiones de sangre homologa y reducir la cantidad de sangre homó1oga transfundida utilizando criterios de transfusion fijos. La siguiente tabla muestra las estaturas en centimetros de varios individuos

1.720	1.710	1.700	1.655
1.730	1.700	1.820	1.810
1.800	1.800	1.790	1.820
1.680	1.730	1.820	1.720
1.790	1.880	1.730	1.560

3. En un estudio de la actividad proliferativa del cancer de seno, Veronese y Gambacorta utilizaron los metodos inmunohistoquimico y de anticuerpos monoclonal Ki-67. Los investi gadores obtuvieron tejido tumoral de 203 pacientes con carcinoma de pecho. Los pacientes tenfan entre 26 y 82 aftos de edad. La siguiente tabla muestra los valores de Ki-67 (expresa dos en porcentajes) para esos pacientes:

10.12	10.15	19.30	33.00
10.80	10.54	27.30	10.15
19.39	16.40	4.40	26.80
9.63	9.31	7.40	9.35
21.42	25.11	12.60	17.96
28.30	19.50	15.92	19.40

4. Realizaron un estudio para investigar las caracteristicas de unión de la imipramina a las plaquetas en pacientes maniacos y comparar los resultados con datos equivalentes de personas sanas y pacientes depresivos. Como parte del estudio, los investigadores obtuvieron los valores máximos de unión a la molecula receptora en estos individuos. Los siguientes valores son de individuos estudiados que fueron diagnosticados con depresión unipolar.

1074	797	485	334
670	510	299	333
303	372	473	797
385	769	768	392
475	319	301	556
300	339	488	306
1113	761	571	306

5. Se compararon dos métodos para colectar sangre para estudios de coagulación. Los siguien tes valores son el tiempo parcial de tromboplastina activada (APTT, siglas en Ingles), de 30 pacientes en cada uno de los dos grupos.

Método 1			
20.7	29.0	46.1	44.8
31.2	20.3	56.6	39.7
24.9	20.9	28.8	22.8
22.9	34.4	33.9	46.1
52.4	28.5	35.5	45.3
26.9	30.1	35.0	54.7
38.3	28.4	22.5	22.1

Método 2					
23.9	23.2	56.2	30.2	27.2	21.8
53.7	31.6	24.6	49.8	22.6	48.9
23.1	34.6	24.2	23.7	56.2	24.6
41.3	21.1	35.7	30.2	49.8	34.1
40.7	29.2	27.2	22.6	26.7	39.8
27.4	21.8	48.9	20.1	21.4	23.2

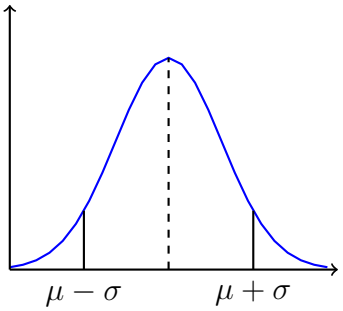
6. Como parte de un proyecto de investigación, los investigadores obtuvieron los siguientes datos sobre los niveles sericos de peróxido lípido (SLP, por las siglas en ingles de serum lipid peroxide), a partir de los informes de laboratorio de una muestra de 10 individuos adultos que recibían tratamiento para la diabetes mellitus: 5.85, 6.17, 6.09, 7.70, 3.17, 3.83, 5.17, 4.31,3.09,5.24. Calcule la media, mediana, varianza y desviación estándar.

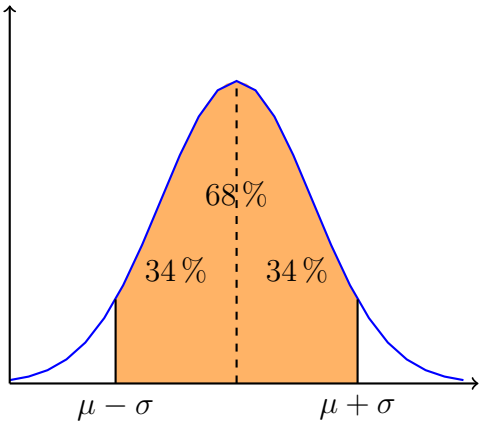
7. Los siguientes valores corresponden a los niveles de SLP que se obtuvieron de una muestra de 10 adultos aparentemente sanos: 4.07, 2.71, 3.64, 3.37, 3.84, 3.83, 3.82,4.21,4.04,4.50. Calcule para estos datos la media, mediana, varianza y desviación estándar. Compare los resultados con los del ejercicio anterior. ¿Qué es 10 que sugieren estos resultados con respecto a los niveles de SLP entre los pacientes con y sin diabetes mellitus? ¿Estos resultados proveen suficientes bases para tomar acci6n medica? Explique su respuesta.

Teorema de Chebyshev

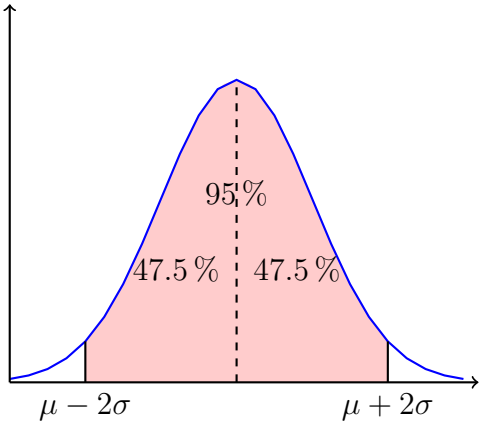
Para todo conjunto de datos, por lo menos $1 - \left(\frac{1}{K^2}\right)$ de las observaciones están dentro de K desviaciones estándar de la media, en donde K es cualquier número mayor que 1.

Regla empírica

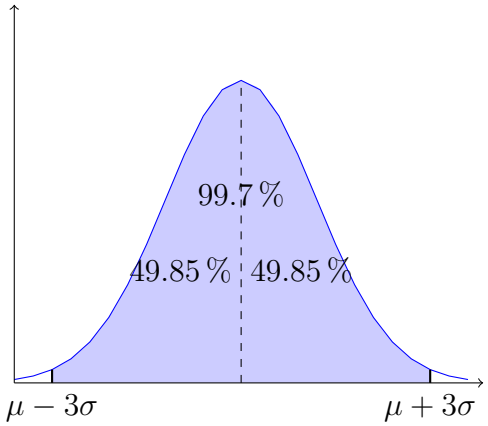




Si se incluyen todas las observaciones que están a dos desviaciones estándar de la media (dos desviaciones estándar por encima de la media y dos desviaciones estándar por debajo de la media) estas serán el 95 % de todas las observaciones.



Si se incluyen todas las observaciones que están a tres desviaciones estándar de la media (tres desviaciones estándar por encima de la media y tres desviaciones estándar por debajo de la media) estas serán el 99,7 % de todas las observaciones.



Sesgo

El sesgo puede medirse mediante el **coeficiente de sesgo de Pearson,** el cual está dado por:

$$\mu_3 = \frac{3(\text{Promedio} - \text{Mediana})}{\text{Desviación}}$$

si $\mu_3 < 0$, los datos están sesgado a la izquierda, si $\mu_3 > 0$ los datos están sesgados a la derecha; si $\mu_3 = 0$ los datos están distribuidos normalmente.

Capítulo 2

Probabilidad

Cotidianamente escuchamos expresiones en donde la probabilidad se asocia con la incertidumbre sobre un suceso de interés. Según Canavos (1988) la probabilidad es un mecanismo por medio del cual pueden estudiarse sucesos aleatorios, cuando estos se comparan con fenómenos deterministicos. La probabilidad juega un papel muy importante en la aplicación de la inferencia estadística, porque una decisión cuyo fundamento se encuentra en la información contenida en la muestra aleatoria, puede ser equivocada.

2.0.1. Elementos de Probabilidad

Un **experimento** es cualquier acción o proceso cuyo resultado está sujeto a la incertidumbre. Estos experimentos se llevan a cabo bajo ciertas condiciones un número definido o indefinido de veces.

Un experimento de dice que es **determinístico**, cuando además de conocer los posibles valores del experimento, también se conoce un resultado particular de él.

Un experimento de dice que es **aleatorio** cuando, puede producir resultados diferentes, aún cuando se repita siempre de la misma manera. Son ejemplos de experimentos aleatorios los siguientes:

- el lanzamiento de un dado no cargado y observar el número que aparece en la cara superior.
- el lanzamiento de una moneda cuatro veces y contar el número total de caras obtenidas.
- la fabicación de artículos en una línea de producción y contar el número de artículos defectuosos producidos en un período de 24 horas.
- Fabricar una bombilla. Luego se prueba su duración conectándola a un portalámparas y se anota el tiempo transcurrido (en horas) hasta que se quema.
- Fabricar artículos hasta producir 10 no defectuosos. Contar el número total de artículos manofacturados.
- Un termógrafo marca la temperatura continuamente en un período de 24 horas. En un sitio y una fecha señalados, leer dicho termógrafo.
- Tiempo empleado por una persona de su casa al trabajo.
- Número de personas que llegan a una oficina bancaria en un período de 10 horas.

Con cada experimento considerado en el ejemplo anterior, definimos el **espacio muestral** como el conjunto de todos los resultados posibles del experimento. Usualmente designamos este conjunto como Ω.

Para cada experimento considerado anteriormente, se describe el espacio muestral asociado como sigue

- $\Omega_1 = \{1, 2, 3, 4, 5, 6\}$
- $\Omega_2 = \{0, 1, 2, 3, 4\}$
- $\Omega_3 = \{0, 1, 2, \ldots, N\}$, donde N es el número máximo de artículos que se pudo construir en 24 horas.
- $\Omega_4 = \{0, 1, 2, \ldots, M\}$, donde M es el número de remaches instalados.
- $\Omega_5 = \{t : t \geq 0\}$
- $\Omega_6 = \{10, 11, \ldots\}$
- $\Omega_7 = \{S : S \geq 0\}$
- $\Omega_8 = \{t : m \leq t \leq M\}$, donde m es la temperatura mínima y M es la temperatura máxima.
- $\Omega_9 = \{t : t \geq 0\}$
- $\Omega_{10} = \{0, 1, \ldots, N\}$

Un **evento** A respecto a un espacio muestral particular Ω, es cualqier recopilación (subconjunto) del espacio muestral Ω. Esto significa que Ω mismo es un evento y también lo es el conjunto $\emptyset$.

Los siguientes son ejemplos de eventos asociados a los experimentos antes anotados: A_i se referirá a un evento asociado con el experimento ε_i

- A_1 : Un número par ocurre; esto es, $A_1 = \{2, 4, 6\}$.
- A_2 : Se obtienen dos o más caras; $A_2 = \{2, 3, 4\}$.
- A_3 : Todos los artículos fueron no defectuosos; $A_3 = \{0\}$.
- A_4 : Se obtienen menos de cuatro remaches defectuosos; $A_4 = \{0, 1, 2, 3\}$.
- A_5 : La bombilla se quema en menos de 10 horas; $A_5 = \{t : 0 \leq t \leq 10\}$.
- A_6 : El número total de artículos manofacturados es inferior a 16; $A_6 = \{10, 11, 12, 13, 14, 15\}$.

Dados los eventos A y B asociados a un experimento aleatorio ε, tales que, $A \cap B = \emptyset$, entonces A y B se denominan eventos excluyentes.

Sean A y B eventos asociados a Ω, tales que $A \cap B = \Omega$ y $A \cup B = \emptyset$, entonces A y B se denominan eventos complementarios. Note que el elemento complementario de A es $B = A'$.

2.0.2. Concepto de probabilidad

Probabilidad clásica

Sea ε un experimento y Ω un espacio muestral asociado con ε. Sea A un evento de ω. La probabilidad de ocurrencia del evento A, denotada $P(A)$, se define como:

$$P(A) = \frac{n(A)}{n(\Omega)}$$

donde $n(A)$ es el número de elementos de A (casos favorables a A) y $n(\Omega)$ es el número de elemetos del espacio muestral Ω (total de casos posibles).

Probabilidad axiomática

Sea ε un experimento y Ω un espacio muestral, el objetivo de la probabilidad es asignar a cada elemento de A, un número $P(A)$, llamado la probabilidad del evento A, el cual dará una medida precisa de la oportunidad de que A ocurra. Todas las asignaciones deberán satisfacer lkos siguientes axiomas (propiedades básicas) de probabilidad

1. Para cualquier evento A en el espacio muestral Ω, $0 \geq P(A) \leq 1$
2. $P(\Omega) = 1$
3. si $A_1, A_2, , A_3, \ldots$ es un conjunto de eventos mutuamente excluyentes, es decir, $A_i \cap A_j = \phi$ para todo $i \neq j$, entonces

$$P(A_1 \cup A_2 \cup A_3 \cup \cdots) = \sum_{i=1}^{\infty} P(A_i)$$

Propiedades Con base en los axiomas anteriores y haciendo uso de las principales propiedades presentes en la teoría de conjuntos, se presentan las siguientes reglas operativas muy utiles en el calculo de probabilidades.

1. $P(\emptyset) = 0$, donde $\emptyset$ es el evento nulo
2. Para cualquier evento A, $P(A') = 1 - P(A)$
3. Para cualquier evento A, $P(A) \leq 1$
4. Para dos eventos cualesquiera A y B,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

   Si A y B son mutuamente excluyentes, entonces

$$P(A \cup B) = P(A) + P(B)$$

5. Para dos eventos cualesquiera A y B,

$$P(A - B) = P(A \cap B') = P(A) - P(A \cap B)$$

6. Para tres eventos A, B y C se tiene

$$\begin{aligned} P(A \cup B \cup C) = P(A) + P(B) + P(C) \\ - P(A \cap B) - P(A \cap C) - P(B \cap C) \\ + P(A \cup B \cup C) \end{aligned}$$

   Si A, B y C son mutuamente excluyentes, entonces

$$P(A \cup B \cup C) = P(A) + P(B) + P(C)$$

Permutaciones: es parte de las técnicas de conteo la cual permite determinar cuantos arreglos de tamaño h se pueden realizar en una muestra de n objetos conservando un orden. A continuación se resumen el número de arreglos teniendo en cuenta si los datos son o no tomados con o sin reemplazamiento:

Sin repetición	Con repetición
$\frac{n!}{(n-h)!}$	n^h

Ejemplo: como se pueden subir al auto bus 6 personas, la respuesta a este interrogante es 6!.
Combinaciones: es parte de las técnicas de conteo la cual permite determinar cuantos arreglos de tamaño h se pueden realizar en una muestra de n objetos si conservar un orden espécifico. La formula que se realiza es la siguiente:

$$\binom{n}{h} = \frac{n!}{h!(n-h)!}$$

Valor esperado de una variable aleatoria: la esperanza matemática de una variable aleatoria se determina dependiendo el tipo de variable (discreta o continua).

- Si la variable es discreta, entonces,

$$EX = \sum_x xPr(X = x)$$

- Si la variable es continua, entonces,

$$EX = \int_x xf_X(x)$$

Varianza de una variable aleatoria: la varianza de una variable aleatoria se determina dependiendo el tipo de variable (discreta o continua) y además se puede utilizar una formula alterna que incluye determina el segundo momento de la variables aleatorias.
Así $VarX = EX^2 - (EX)^2$

- Si la variable es discreta, entonces,

$$EX^2 = \sum_x x^2Pr(X = x)$$

- Si la variable es continua, entonces,

$$EX^2 = \int_x x^2f_X(x)$$

Probabilidad condicional

Para dos eventos cualesquiera A y B con $P(B) > 0$, la **probabilidad condicional de A dado que B ha ocurrido** está definida por

$$P(A|B)] = \frac{P(A \cap B)}{P(B)}$$

Regla multipicativa para $P(A \cap B)$

De la definición de probabilidad condicional se obtiene el siguiente resultado, multiplicando ambos miembros de la ecuación anterior por $P(B)$

$$P(A \cap B) = P(B)P(A|B)$$

Para tres eventos A, B y C se tiene

$$P(A \cap B \cap C) = P(A)P(B|A)P(C|A \cap B)$$

Regla de la probabilidad total

Para cualquier par de eventos A y B en S se satisface que:

$$P(B) = P(B \cap A) + P(B \cap A^{'}) = P(B|A)P(A) + P(B|A^{'})P(A^{'})$$

Regla de la probabilidad total para varios eventos

Sean $A_1, A_2, \ldots, A_k$ eventos mutuamente excluyentes y exhaustivos. Entonces para cualquier otro evento B,

$$\begin{aligned} P(B) &= P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + \cdots + P(A_k)P(B|A_k) \\ &= \sum_{i=1}^{k} P(A_i)P(B|A_i) \end{aligned}$$

Teorema de Bayes

Sean $A_1, A_2, \ldots, A_k$ eventos mutuamente excluyentes y exhaustivos con probabilidades *previas* $P(A_1)$ (para $i = 1, 2, \ldots, k$). Entonces para cualquier otro evento B para el cual $P(B) > 0$, la probabilidad *posterior* de A_j dado que B ha ocurrido es

$$\begin{aligned} P(A_j|B) &= \frac{P(A_j \cap B)}{P(B)} \\ &= \frac{P(A_j)P(B|A_j)}{\sum_{i=1}^{k} P(A_i)P(B|A_i)} \end{aligned}$$

Independencia

Los eventos A y B son **independientes** si $P(A|B) = P(A)$, de lo contrario se dice que son dependientes.

De lo anterior se deduce que:

A y B son independientes si y solo si $P(A \cap B) = P(A)P(B)$

Independencia de más de dos eventos

Los eventos $A_1, a_2, \ldots, A_k$ son **mutuamente independientes** si para cada k ($k = 1, 2, \ldots, n$) y cada subconjunto de índices $i_1, i_2, \ldots, i_k$

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k})$$

Capítulo 3

Algunas distribuciones discretas

Definición 5 (**Variables aleatorias**). *Una variable aleatoria es una función que asigna un número real a cada resultado en espacio muestral S de un experimento aleatorio arbirtrario.*
Considerando el experimento de lanzar tres monedas, el interés recae en la variable aleatoria $\boldsymbol{X}$ *que asigna a cada punto del espacio muestral* $\boldsymbol{S}$ *el número de sellos que se obtiene en la realización del experimento.*

3.0.1. Distribuciones uniforme, binomial, y de Bernoulli

Definición 6 (**Distribución discreta uniforme**). *Se dice que una variable aleatoria X tiene una distribución discreta uniforme de parámetro N, donde N es un entero positivo, si su función de densidad está dada por:*

$$f(x) = \begin{cases} \frac{1}{N} \text{ si } x = 1, 2, \ldots, N \\ 0 \text{ en otro caso} \end{cases}$$

En la distribución uniforme discreta la variable aleatoria denota los N resultados diferentes en un experimento donde todos los resultados tienen la misma probabilidad de ocurrir, por ejemplo los juegos de loterias y de azar de forma general.

Teorema 1 (**Propiedades de la distribución discreta uniforme**). *Si X es una variable aleatoria con distribución discreta uniforme de parámetro N, entonces,*

- $EX = \frac{N+1}{2}$
- $Var(X) = \frac{N^2-1}{12}$
- $m_X(t) = \sum_{i=1}^{N} \frac{1}{N} e^{tk}$

Definición 7 (**Distribución Bernoulli**). *Una variable aleatoria con distribución Bernoulli es una variable que tiene solo dos posibles resultados, los cuales se denotan como exito o fracaso, los cuales tiene probabilidad de ocurrencia p y $1-p$ respectivamente. Otra forma de denotar los resultados es asignando 1 cuando el resultado del experimento aleatorio es un éxito y asigna 0 cuando el resultado del experimento aleatorio es un fracaso. La función de densidad de probabilidad para está variable aleatoria está dada por:*

$$f(x) = \begin{cases} p^x(1-p)^{1-x} \text{ si } x = 0, 1 \\ 0 \text{ en otro caso} \end{cases}$$

Teorema 2 (**Propiedades de la distribución Bernoulli**). *Si X es una variable aleatoria con distribución Bernoulli de parámetro p, entonces,*

- $EX = p$

- $Var(X) = 1 - p$
- $m_X(t) = \sum_{i=1}^{N} \frac{1}{N} e^{tk}$

Definición 8 (**Distribución binomial**). *La distribución binomial se usa frecuentemente en la descripción de aquellos experimentos que consta de un número fijo de prueba o ensayos en los que el resultado es la ocurrencia o no ocurrencia de un suceso y la probabilidad del suceso es constantante y no varía de una prueba a otra. La variable aleatoria denota el número de éxitos en cada prueba o ensayos independientes del experimento Bernoulli.*
Se dice que una variable aleatoria X tiene distribución binomial de parámetros n y p si su función de densidad de probabilidad está dada por:

$$f_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} \text{ si } x = 0, 1, \ldots, n \\ 0 \text{ en otro caso} \end{cases}$$

donde n es un entero positivo y $0 < p < 1$.

Teorema 3 (**Propiedades de la distribución binomial**). *Si X es una variable aleatoria con distribución binomial de parámetros n y p, entonces,*

- $EX = np$
- $Var(X) = np(1-p)$
- $m_X(t) = (pe^t + (1-p))^n$

Nota 2. *La distribución binomial está asociada a experimentos de muestreo con repetición o muestreo de poblaciones infinitas (casos en que la probabilidad de éxito no cambia).*

Definición 9 (**Distribución hipergeométrica**). *La distribución hipergeométrica se usa frecuentemente en la descripción de experimentos en los que unicamente interesa el número de elementos con una característica especifica entre n elementos. Las probabilidades a cada uno de los resultados no son constante y cada ensayo o repetición del experiemnto no es independiente de los demás. La variable aleatoria denota el número de objetos (o elementos) de la clase de interés contenidos en la muestra seleccionada de la población, de los cuales K son del tipo requerido. En está distribución el muestreo se realiza sin reemplazamiento y los ensayos no son independientes; el tamaño de la población N es finito, n tamaño de la muestra es fijo y pequeño.*
Se dice que una variable aleatoria X tiene distribución hipergeométrica de parámetros n, R y N si su función de densidad de probabilidad está dada por:

$$f_X(x) = \begin{cases} \binom{R}{x}\binom{N-R}{n-x} \text{ si } x = 0, 1, \ldots, n \\ 0 \text{ en otro caso} \end{cases}$$

donde N, n son enteros positivo, R es un entero no negativo menor o igual a N y $n \leq N$

Teorema 4 (**Propiedades de la distribución hipergeométrica**). *Si X es una variable aleatoria con distribución hipergeométrica de parámetros N, R y n, entonces,*

- $EX = \frac{nR}{N}$
- $Var(X) = n\left(\frac{R}{N}\right)\left(\frac{N-R}{N}\right)\left(\frac{N-n}{N-1}\right)$
- $m_X(t) =$

Nota 3. *La distribución hipergeométrica está asociada a experimentos de muestreo de poblaciones finitas sin repetción (casos en que la probabilidad de éxito cambia).*

Teorema 5 (**Aproximación de la distribución hipergeométrica a la distribución binomial**). *Sea* $0 < p < 1$, *si* $N, R \to \infty$ *de tal forma que* $\frac{R}{N} \to p$, *entonces:*

$$\frac{\binom{R}{k}\binom{N-R}{n-k}}{\binom{N}{n}} \longrightarrow \binom{n}{k} p^k (1-p)^{n-k}$$

Definición 10 (**Distribución Poisson**). *La distribución Poisson se utiliza para describir el número de ocurrencias de cierto evento dentro de un intervalo de tiempo dado conociendo de antemano la tasa media de ocurrencia del parámetro de interés. La variable aleatoria denota el número de eventos independientes que ocurren en un tiempo fijo o un intervalo de tiempo dado o en el espacio. Por ejemplo cuando se desea describir la distribución del:*

- *Número de personas que llegan a un supermercado.*
- *Número de defectos en piezas similares.*
- *Número de bacterias en un cultivo.*
- *Número de solicitudes de seguros procesadas por una compañia en un periodo especifico.*
- *Número de partículas que llegan a un determinado punto en el espacio, durante un periodo de tiempo* t, *y que son emitidas por una sustancia radioactiva.*
- *Número de individuos que llegan a una línea de espera, en un periodo de tiempo* t.

Se dice que una variable aleatoria X *tiene distribución Poisson de parámetro* $\lambda > 0$, *si su función de densidad está dada por:*

$$f_X(x) = \begin{cases} \dfrac{e^{-\lambda}\lambda^x}{x!} \text{ si } x = 0, 1, \ldots \\ 0 \text{ en otro caso} \end{cases}$$

El parámetro λ *denota el número promedio de ocurrencias del evento de interés por unidad de tiempo.*

Teorema 6 (**Propiedades de la distribución Poisson**). *Si* X *es una variable aleatoria con distribución Poisson de parámetros* λ, *entonces,*

- $EX = \lambda$
- $Var(X) = \lambda$
- $m_X(t) =$

Teorema 7 (**Aproximación de la distribución binomial a la distribución Poisson**). *Si* $p(n)$ *es una sucesión con* $0 < p(n) < 1$ *y* $n\,(p(n)) \longrightarrow \lambda$ *cuando* $n \longrightarrow \infty$, *entonces*

$$\binom{n}{k} (p(n))^k (1 - (p(n)))^{n-k} \longrightarrow \frac{e^{-\lambda}\lambda^x}{x!}$$

cuando $n \longrightarrow \infty$

El teorema anterior implica que la distribución Poisson ofrece un modelo probabilístico adecuado para todos aquellos experimentos aleatorios en los que las repeticiones son independientes unas de otras y en los que sólo hay dos posibles resultados: éxito o fracaso, con probabilidad de éxito pequeña, y en los que el interés se centra en conocer el número de éxitos obtenidos al realizar el experimento un número suficientemente grande de veces. Empíricamente se ha establecido que la aproximación se puede aplicar si $n \geq 100$, $p \leq 0{,}01$ y $np \leq 20$

Nota 4. *Si X_t es la variable aleatoria que denota el número de individuos que llegan en el intervalo de tiempo $(0,t]$, entonces, se tiene que X_t tiene una distribución Poisson con parámetro λt.*

Definición 11 (**Distribución binomial negativa**). *Esta distribución se utiliza cuando el interés recae en determinar cual es la probabilidad de tener que repetir el experimento n veces para obtener de manera exacta k éxitos. La variable aleatoria denota el número de ensayos necesarios antes de obtener el k-ésimo éxito o el número de fracasos ocurridos antes de obtener de manera exacta k éxitos.*
Para el primer caso, se dice que una variable aleatoria X tiene distribución binomial negativa de parámetros k y p, si su función de densidad está dada por:

$$f_X(x) = \begin{cases} \binom{x-1}{k-1} p^k (1-p)^{x-k} & si\ x = k, k+1, \ldots \\ 0\ en\ otro\ caso \end{cases}$$

En el segundo caso, cuando se desea conocer la probabilidad del número de fracasos ocurridos antes de obtener de manera exacta k éxitos, la función de densidad está dada por:

$$f_Y(y) = \begin{cases} \binom{k+y-1}{k-1} p^k (1-p)^{y} & si\ y = 0, 1, 2, \ldots \\ 0\ en\ otro\ caso \end{cases}$$

Nota 5. *En el caso especial $k = 1$, se dice que la variable aleatoria tiene distribución geométrica de parámetro p.*

Teorema 8 (**Propiedades de la distribución binomial negativa**). *Si X es una variable aleatoria con distribución binomial negativa de parámetros k y p, entonces,*

- $EX = \frac{k}{p}$
- $Var(X) = \dfrac{k(1-p)}{p^2}$
- $m_X(t) = \left[\frac{pe^t}{1-(1-p)e^t}\right]^k$

Definición 12 (**Distribución geométrica**).

- *Cuando la variable aleatoria cuenta el número de ensayos antes de obtener el primer éxito, la función de densidad está dada por:*

$$f_X(x) = \begin{cases} p\,(1-p)^{x-1} & si\ x = 1, 2, 3, \ldots \\ 0\ en\ otro\ caso \end{cases}$$

- *Cuando la variable aleatoria cuenta el número de fracasos antes de obtener el primer éxito, la función de densidad está dada por:*

$$f_X(x) = \begin{cases} p\,(1-p)^{x} & si\ x = 0, 1, 2, 3, \ldots \\ 0\ en\ otro\ caso \end{cases}$$

Está distribución se usa para representar tiempos de espera y en la inspección sucesiva de artículos para control de calidad.

Teorema 9 (**Propiedades de la distribución geométrica**). *Si X es una variable aleatoria con distribución geométrica de parámetros p, entonces,*

- $EX = \frac{1}{p}$
- $Var(X) = \dfrac{1-p}{p^2}$
- $m_X(t) = \left[\frac{pe^t}{1-(1-p)e^t}\right]$

3.1. Algunas distribuciones continuas

Definición 13 (**Distribución uniforme**). *La variable aleatoria distribuida uniformente en el intervalo $[a, b]$ toma valores que estan igualmente distribuidos sobre dicho intervalo. Se dice que una variable aleatoria X está distribuida uniformemente sobre el intervalo $[a, b]$, con $a < b$ números reales, si su función de densidad está dada por:*

$$f(x) = \begin{cases} \frac{1}{b-a} \ si \ a \leq x \leq b \\ 0 \ en \ otro \ caso \end{cases}$$

Teorema 10 (**Propiedades de la distribución uniforme**). *Si X es una variable aleatoria con distribución uniforme sobre el intervalo $[a, b]$, entonces,*

- $$F(x) = \begin{cases} 0 \ si \quad x < a \\ \dfrac{x-a}{b-a} \quad si \ a \leq x \leq b \\ 1 \quad si \ x > b \end{cases}$$
- $EX = \frac{a+b}{2}$
- $Var(X) = \frac{(b-a)^2}{12}$

Definición 14 (**Distribución normal o gaussiana**). *Se caracteriza por una medida de posición: la media y una medida de dispersión: la varianza o su raíz cuadrada la desviación estándar. Se dice que una variable aleatoria X tiene distribución normal de parámetros $\mu \in \mathbb{R}$ y $\sigma \in \mathbb{R}^+$, si su función de densidad está dada por:*

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]; \ x \in \mathbb{R}$$

Teorema 11 (**Propiedades de la distribución normal**). *Si X es una variable aleatoria con distribución normal de parámetros μ y σ^2, entonces,*

- $$F(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} \exp\left[-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2\right] du$$
- $EX = \mu$
- $Var(X) = \sigma^2$
- $m_X(t) = \exp\left[\mu t + \frac{\sigma^2 t^2}{2}\right]$

La distribución normal que tiene media cero y varianza uno se conoce como la distribución normal estandarizada y se representa por Z. Para estandarizar una variable basta con restarle la media y dividirla por la desviación estándar.
Las distribuciones derivadas de la distribución normal son:

- Ji cuadrado
- t-Student
- F de Snedecor y Fisher

Definición 15 (**Distribución gamma**). *Algunas de las principales características de las variables aleatorias con distribución gamma son las siguientes:*

- *Las variables aleatorias son no negativas, es decir, mayor o igual a 0.*
- *La distribción es segada a la derecha, es decir, la mayor parte del área bajo la curva de la función de densidad, se encuentra cerca del origen y los valores disminuyen gradualmente cuando x aumenta.*
- *Se utilizan para describir los intervalos de tiempo entre dos fallas consecutivas, intervalos de tiempo entre las llegadas de los clientes a las filas de un punto de pago.*
- *La variable aleatoria con distribución gamma denotan el tiempo trascurrido desde el momento en que se inicia la observación hasta que se presenta el evento de interés.*
- *La distribución gamma es la generalización de tres casos particulares de distribución, la distribución exponencial, la distribución Erlang y la distribución Ji-cuadrado.*

Se dice que una variable aleatoria X tiene distribución gamma de parámetros $r > 0$(parámetro de forma) y $\lambda > 0$
(parámetro de escala), si su función de densidad está dada por:

$$f_X(x) = \frac{\lambda}{\Gamma(r)}(\lambda x)^{r-1}\exp\{-\lambda x\}$$

donde $\Gamma(\cdot)$ es la función gamma, esto es,

$$\Gamma(\cdot) := \int_0^\infty t^{r-1}\exp(-t)dt$$

Teorema 12 (**Propiedades de la distribución gamma**). *Si X es una variable aleatoria con distribución gamma de parámetros $r > 0$(parámetro de forma) y $\lambda > 0$(parámetro de escala), entonces,*

- $F(x) = \int_0^x \frac{\lambda}{\Gamma(r)}(\lambda t)^{r-1}\exp\{-\lambda t\}dt$
- $EX = \frac{r}{\lambda}$
- $Var(X) = \frac{r}{\lambda^2}$
- $m_X(t) = \left(\frac{\lambda}{\lambda - t}\right)^r$; *si* $t < \lambda$.

Definición 16 (**Distribución exponencial**).

- *Se utiliza como modelo para describir la distribución de tiempo transcurrido entre las ocurrencias sucesivas de un determinado suceso. Por ejemplo el número de clientes que llegan a una entidad bancaria, las llamadas que entran a una central telefonica, entre otros.*
- *Modela la distribución de la duración de los componentes que no se deterioran, ni mejoran con la edad, esto es, aquellos para los cuales la distribución de la duración restante del componente es independiente de la edad actual.*
- *La función de densidad se obtiene al reemplazar en la función de densidad de la distribución gamma $r = 1$ y $\lambda > 0$ arbirtrario, esto es,*

$$f_X(x) = \lambda\exp\{-\lambda x\}$$

- *La función de distribución para una variable aleatoria X con distribución exponencial de parámetro λ*

$$F_X(x) = 1 - \exp(-\lambda x)$$

- $EX = \frac{1}{\lambda}$
- $Var(X) = \frac{1}{\lambda^2}$
- $m_X(t) = \left(\frac{\lambda}{\lambda - t}\right)$; *si* $t < \lambda$.

Definición 17 (**Distribución Cauchy**)**.** *Se dice que una variable aleatoria* X *tiene distribución Cauchy de parámetros* θ *y* β, $\theta \in \mathbb{R}$ *y* $\beta \in \mathbb{R}^+$, *si su función de densidad está dada por:*

$$f(x) = \frac{1}{\pi\beta} \frac{1}{1 + \left(\frac{x-\theta}{\beta}\right)^2}; \ x \in \mathbb{R}$$

Capítulo 4

Encuestas y estudios por muestreo

En este capítulo se presentan los principales conceptos básicos presentes en el diseño, recolección, procesamiento y análisi de las investigaciones que utilizan como técnica estadística el muestreo.

4.1. Conceptos básicos

El muestreo es un procedimiento que responde a la necesidad de información estadística precisa sobre la población y los conjuntos de elementos que la conforman. La importancia de este radica en que las investigaciones parciales sobre la población apuntan a inferir a la población completa y es un procedimiento que cuesta mucho menos dinero, consume menos tiempo y puede ser más preciso que al realizar una enumeración completa, también llamada censo. Una muestra bien seleccionada de unos cuantos miles de individuos puede representar con gran precisión una población de millones. Es requisito fundamental de una buena muestra que las características de interés que existen en la población se reflejen en la muestra de la manera más cercana posible.

- **Población objetivo:** colección completa de todas las unidades que se quieren estudiar.
- **Muestra:** es un subconjunto de la población.
- **Unidad de muestreo:** es el objeto a ser seleccionado en la muestra que permitirá el acceso a la unidad de observación.
- **Unidad de observación:** es el objeto sobre el que finalmente se realiza la observación.
- **Variable de interés:** es la característica propia de los individuos sobre la que se realiza la inferencia para resolver los objetivos de la investigación.

4.1.1. Marco de muestreo

En investigaciones por muestreo se consideran dos tipos de objetos:

- **Elementos:** las unidades e individuos sobre las que se realiza la medición.
- **Conglomerado:** agrupación de elementos cuya característica principal es que son homogéneos dentro de sí, y heterogéneos entre sí. Todo procedimiento de muestreo probabilístico requiere de un dispositivo que permita investigar, seleccionar y ubicar a todos y cada uno de los objetos pertenecientes a la población objetivo y que participaran en la selección aleatoria. Este dispositivo se conoce con el nombre de **marco de muestreo.**
  Cuando se dispone de un marco de muestreo de elementos, se puede aplicar un diseño de muestreo de elementos. Si no se dispone de un marco de elemntos (o es muy costoso construirlo) se debe recurrir a diseños de muestreo en conglomerados. Por ejemplo, al realizar una

encuesta cuya unidad de observación sean las personas que viven en la ciudad, es muy difícil poder acceder a un marco de muestreo de las personas. Sin embargo, se puede tener acceso a la división sociodemográfica de la ciudad, y así seleccionar algunos barrios de la ciudad, en una primera instancia y luego, seleccionar a las personas de los barrios en una segunda instancia.

Existen dos tipos de marcos de muestreo, a saber:

- **De lsita:** listados físicos o magnéticos, ficheros, archivos de expedientes, historias clínicas que permiten identificar y ubicar los objetos que participaran en el sorte aleatorio.
- **De área:** mapas de ciudades y regiones en formato físico o magnético, fotografías aéreas, imágenes de satélite o similares que permitan delimitar regiones o unidades geográficas en forma tal que su identificación y su ubicación sobre el terreno sea posible.

Tipos de poblaciones objetivos

Muchos autores como Groves, Foler, Couper consideran que los tipos de poblaciones objetivos que se presentan de manera más frecuente en un estudio por muestreo son las siguientes:

- **Hogares y personas:** el marco de muestreo más utilizado en estas poblaciones es de área. Como está basada en zonas geográficas, este tipo de marco requiere la vinculación de los hogares o personas a cada una de las áreas.
  Cuando se requiere seleccionar personas, este tipo de marcos hace necesarias muchas etapas de muestreo.
- **Clientes, empleados o miembros de organizaciones**
- **Organizaciones:** iglesias, prisiones, hospitales, escuelas, etc.
- **Eventos:** matrimonios, nacimientos, fallecimientos, periodos de depresión, tránsito de automóvil en un segmento de vía.
- **Poblaciones poco frecuentes**

4.1.2. Característica de interés y parámetro de interés

El proposito de cualquier estudio por muestreo es estudiar una característica de interés y y que se encuentra asociada a cada unidad de la población. El objetivo de la investigación por muestreo es estimar una función de interés T, llamado parámetro, de la característica de interés en la población.