Estadística

Clase 1

Rodrigo Del Rosso

08 de Abril de 2022

Viernes 08 de Abril de 2022

- Principales definiciones y conceptos
- Medidas que resumen información
- Tratamiento de los datos con tabla de frecuencias (discreta y continua)
- Condicionales con Rango Percentilar. Momentos Absolutos y Centrados. Relación entre los mismos. Ejemplo numérico.
- Propiedades de la Media Aritmética y Varianza.

¿Es necesario estudiar estadística?

Data Science without Statistics is like owning a Ferrari without brakes. You can enjoy sitting in Ferrari, show off your newly owned car to others, but you can't enjoy the drive for long because you would crash land soon!

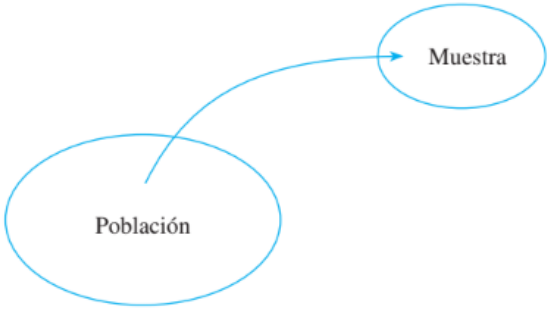
1. *¿Qué es Estadística?*
2. *¿Conocieron alguna vez a un experto en Estadística?*
3. *¿Saben qué hace?*

La Estadística es una rama de las matemáticas que tiene aplicaciones en cada faceta de nuestra vida.

Es un lenguaje nuevo y poco conocido para casi todas las personas, pero, al igual que cualquier idioma nuevo, la estadística puede parecer agobiante a primera vista.

Charla TED – Walter Sosa Escudero

En el lenguaje de la estadística, uno de los conceptos más elementales es el muestreo. En casi todos los problemas de estadística, un número especificado de mediciones o datos, es decir, una muestra, se toma de un cuerpo de mediciones más grande llamado población.



Las palabras ***muestra*** y ***población*** tienen dos significados para la mayoría de personas.

Por ejemplo, Uds. leen en los periódicos que una encuesta Gallup realizada en Estados Unidos estuvo basada en una muestra de 1823 personas.

Presumiblemente, a cada persona entrevistada se le hace una pregunta particular y la respuesta de esa persona representa una sola medida de la muestra.

¿La muestra es el conjunto de las 1823 personas, o es las 1823 respuestas que dan?

Cuando usamos lenguaje de la estadística, distinguimos entre el conjunto de objetos en el cual las mediciones se toman y las mediciones mismas.

Para experimentadores, los objetos en los que las mediciones se toman se denominan ***unidades experimentales***.

El estadístico que estudia las muestras las llama ***elementos de la muestra***.

Cuando primero se le presenta a usted un conjunto de mediciones, ya sea una muestra o una población, necesita encontrar una forma de organizarlo y resumirlo.

La rama de la estadística que presenta técnicas para describir conjuntos de mediciones se denomina estadística descriptiva.

Han visto estadísticas descriptivas en numerosas formas:

gráficas de barras, gráficas de pastel y gráficas de líneas presentadas por un candidato político; tablas numéricas en el periódico; o el promedio de cantidad de lluvia informado por el pronosticador del clima en la televisión local.

Las gráficas y resúmenes numéricos generados en computadoras son comunes en nuestra comunicación de todos los días.

Definición La **estadística descriptiva** está formada por procedimientos empleados para resumir y describir las características importantes de un conjunto de mediciones.

Si el conjunto de mediciones es toda la población, sólo es necesario sacar conclusiones basadas en la estadística descriptiva. No obstante, podría ser demasiado costoso o llevaría demasiado tiempo enumerar toda la población.

Quizá enumerar la población la destruiría, como en el caso de la prueba de "tiempo para falla". Por éstas y otras razones, quizá Ud. tenga una muestra de la población que, al verla, desee contestar preguntas acerca de la población en su conjunto.

La rama de la estadística que se ocupa de este problema se llama estadística inferencial.

Definición La **estadística inferencial** está formada por procedimientos empleados para hacer inferencias acerca de características poblacionales, a partir de información contenida en una muestra sacada de esta población.

El objetivo de la estadística inferencial es hacer inferencias (es decir, sacar conclusiones, hacer predicciones, tomar decisiones) acerca de las características de una población a partir de información contenida en una muestra.

Definición Una **variable** es una característica que cambia o varía con el tiempo y/o para diferentes personas u objetos bajo consideración.

Definición Una **unidad experimental** es el individuo u objeto en el que se mide una variable. Resulta una sola **medición** o datos cuando una variable se mide en realidad en una unidad experimental.

Definición Una **población** es el conjunto de mediciones de interés para el investigador.

Definición Una **muestra** es un subconjunto de mediciones seleccionado de la población de interés.

De entre todos los alumnos de una gran universidad se selecciona un conjunto de cinco estudiantes y las mediciones se introducen en una hoja de cálculo, como la que se muestra.

Identifique los diversos elementos comprendidos en la generación de este conjunto de mediciones.

↓	C1	C2	C3-T	C4-T	C5-T	C6
	Student	GPA	Gender	Year	Major	Number of Units
1	1	2.0	F	Fr	Psychology	16
2	2	2.3	F	So	Mathematics	15
3	3	2.9	M	Su	English	17
4	4	2.7	M	Fr	English	15
5	5	2.6	F	Jr	Business	14

Hay diversas variables en este ejemplo.

La unidad experimental en la que se miden las variables es un alumno del plantel en particular, identificado en la columna C1.

Se miden cinco variables para cada estudiante: promedio de calificaciones (GPA), género, año en la universidad, curso de maestría y número actual de unidades en las que está inscrito.

Cada una de estas características varía de un estudiante a otro. Si consideramos las GPA de todos los estudiantes de esta universidad como la población de interés, las cinco GPA de la columna C2 representan una muestra de esta población.

Si se hubiera medido el GPA de cada estudiante de la universidad, hubiéramos generado toda la población de mediciones para esta variable.

La segunda variable que se mide en los estudiantes es el género, en la columna C3-T. Esta variable puede tomar sólo dos valores: Masc (M) o Fem (F). No es una variable que tenga valor numérico y, por lo tanto, es un poco diferente del GPA. La población, si pudiera ser enumerada, estaría formada por un conjunto de letras M y F, una para cada estudiante de la universidad. Análogamente, las variables tercera y cuarta, año y especialidad, generan datos no numéricos.

El año tiene cuatro categorías (primero, segundo, pasante y graduado) y la especialidad tiene una categoría para cada especialidad en el plantel.

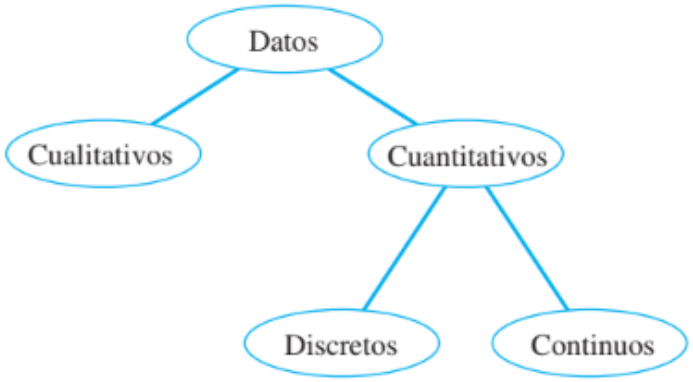
La última variable, el número actual de unidades en que está inscrito, es de valor numérico y genera un conjunto de números en lugar de un conjunto de cualidades o características. Aun cuando hemos examinado cada una de las variables en forma individual, recuerde que hemos medido cada una de estas cinco variables en una sola unidad experimental: el estudiante.

Por lo tanto, en este ejemplo, una “medición” en realidad está formada por cinco observaciones, una para cada una de las cinco variables medidas. Por ejemplo, la medición tomada en el estudiante 2 produce esta observación: (2.3, F, So, Matemáticas, 15)

Definición Resultan **datos univariados** cuando se mide una sola variable en una sola unidad experimental.

Definición Resultan **datos bivariados** cuando se miden dos variables en una sola unidad experimental. Resultan **datos multivariados** cuando se miden más de dos variables.

Definición Las **variables cualitativas** miden una cualidad o característica en cada unidad experimental. Las **variables cuantitativas** miden una cantidad numérica en cada unidad experimental.



Definición Una **variable discreta** puede tomar sólo un número finito o contable de valores. Una **variable continua** puede tomar infinitamente muchos valores correspondientes a los puntos en un intervalo de recta.

Identifique cada una de las siguientes variables como cualitativas o cuantitativas:

1. El uso más frecuente de su horno de microondas (recalentar, descongelar, calentar, otros)

2. El número de consumidores que se niegan a contestar una encuesta por teléfono

3. La puerta escogida por un ratón en un experimento de laberinto (A, B o C)

4. El tiempo ganador para un caballo que corre en el Derby de Kentucky

5. El número de niños en un grupo de quinto grado que leen al nivel de ese grado o mejor

GRÁFICAS PARA DATOS CATEGÓRICOS

1. ¿Qué valores de la variable han sido medidos?

2. ¿Con qué frecuencia se presenta cada uno de los valores?

Se puede construir una tabla estadística que se puede usar para mostrar los datos gráficamente como una distribución de datos.

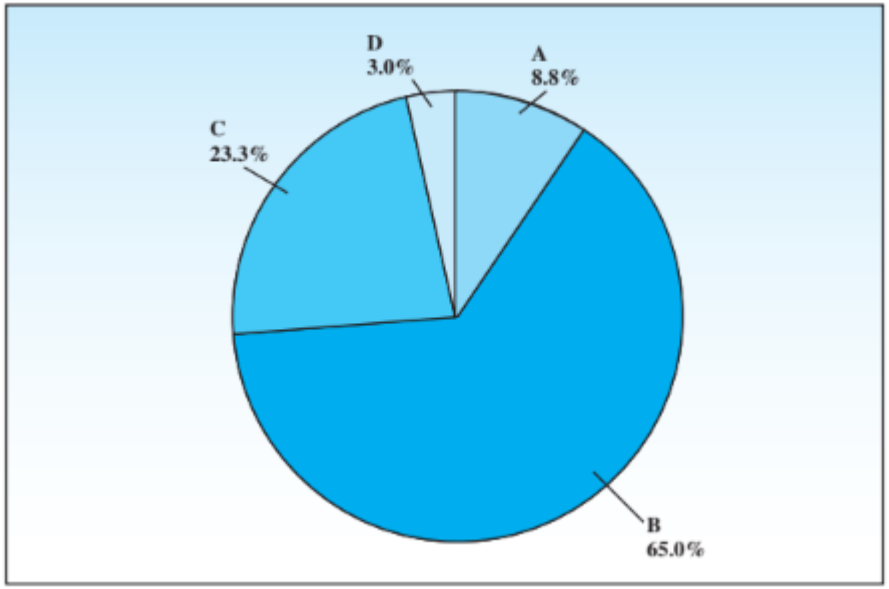
El tipo de gráfica que se escoja depende del tipo de variable que se haya medido.

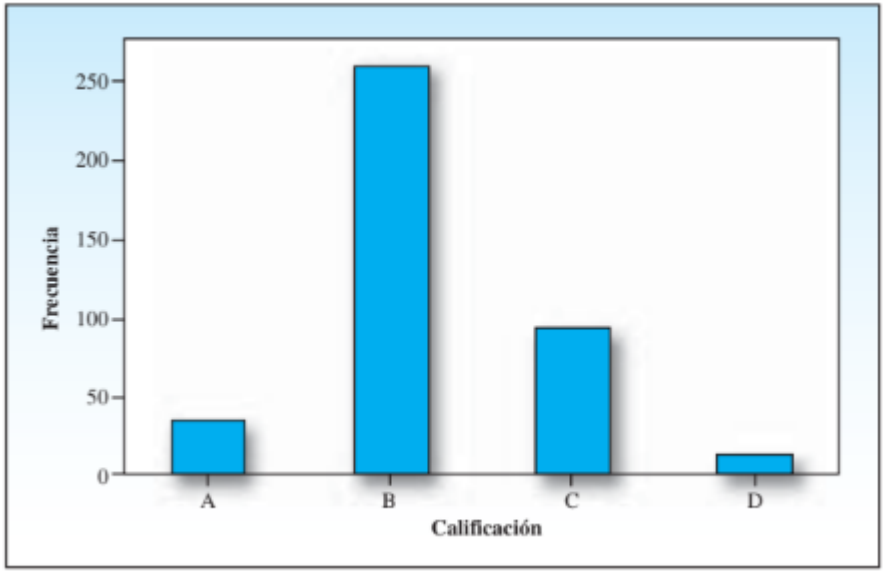
Se puede medir “la frecuencia” en tres formas diferentes:

1. La frecuencia o número de mediciones en cada categoría

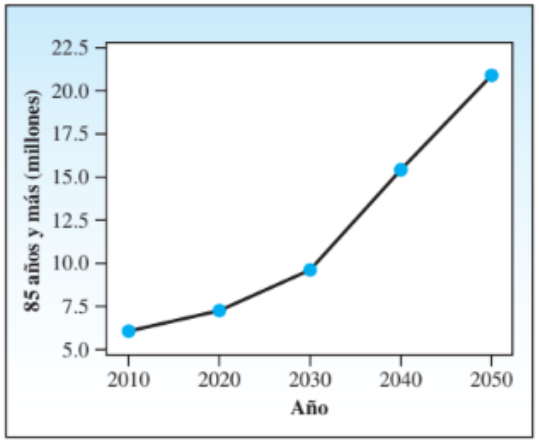
2. La frecuencia relativa o proporción de mediciones en cada categoría

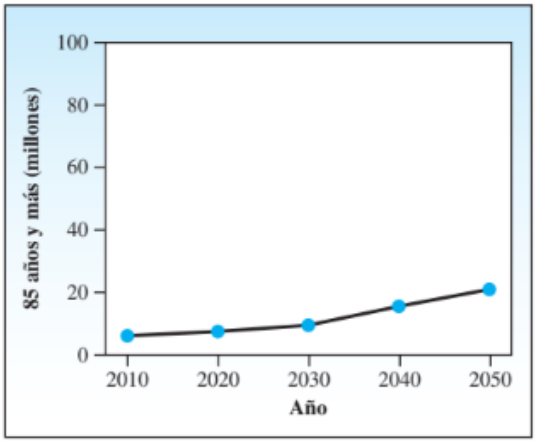
3. El porcentaje de mediciones en cada categoría

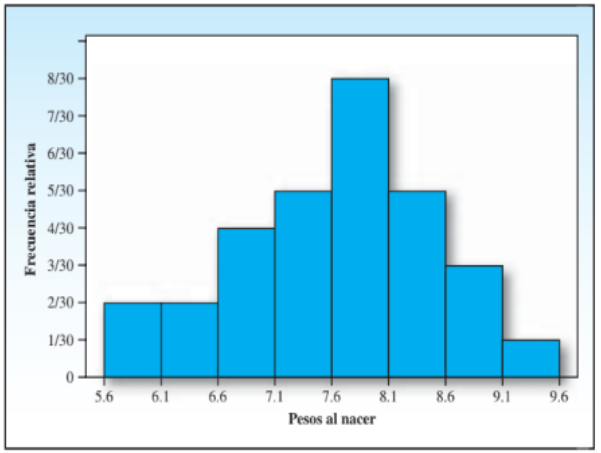

D
3.0%
A
8.8%
C
23.3%
B
65.0%


250
200
150
100
50
0
Frecuencia
A
B
C
D
Calificación

GRÁFICAS PARA DATOS CUANTITATIVOS







Tratamientos de los datos con tablas de frecuencias

Variable Discreta

X = Cantidad de hijos por familias

X_i	f_i
0	4
1	5
2	3
3	2

Una tabla de frecuencias es un ordenamiento de las observaciones de una variable en una tabla, donde para cada valor de variable se indica la cantidad de veces que se repite un valor de la variable

El valor 4 de la variable 0, indica que hay 4 familias que no tienen hijos

Si quisiéramos calcular el promedio aritmético simple, ¿cómo podríamos hacerlo con una tabla de frecuencias?

X_i	f_i	$X_i * f_i$
0	4	0
1	5	5
2	3	6
3	2	6

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

$$\bar{X} = \frac{0+5+6+6}{4+5+3+2} = \frac{17}{14} = 1{,}21\ hijos\ por\ flia$$

Significa que en promedio cada familia tiene 1,21 hijos

Si nos solicitan completar una tabla de frecuencias, nos estarían solicitando disponer de las frecuencias absolutas y relativas, tanto simples como acumuladas.

X_i	f_i	$X_i * f_i$	F_i	fr_i	Fr_i
0	4	0	4	4/14	4/14
1	5	5	9	5/14	9/14
2	3	6	12	3/14	12/14
3	2	6	14	2/14	14/14

De la tabla anterior se observa que el valor de variable X = 1, acumula 9 observaciones. Es decir, posee como mucho (máximo) dicha cantidad de observaciones.

Si nos preguntan, cuál es el valor que se repite con más frecuencia nos estarían preguntando que indiquemos el **Modo/a** de la distribución de datos.

En este ejemplo, el Modo es Mo(X) = 1, porque es el valor de variable que se repite con mayor frecuencia.

Si nos preguntan, cuál es el valor de la variable que se encuentre a la mitad de los datos. Es decir, aquel valor que supera y es superado por igual porcentaje de datos nos estarían solicitando **la Mediana**

Cuando trabajamos con una tabla de frecuencias, hay que localizar aquel valor de variable que es superado por el 50% de los datos.

X_i	f_i	$X_i * f_i$	F_i	fr_i	Fr_i
0	4	0	4	4/14	4/14
1	5	5	9	5/14	9/14
2	3	6	12	3/14	12/14
3	2	6	14	2/14	14/14

$n = 14 \quad Cantidad\ de\ Observaciones$

$OAM = \frac{n}{2} = 7\ Orden\ Absoluto\ Mediano$

$Me(x) = 1\ hijo\ por\ familia$

¿Qué ocurre si ahora existe una Frecuencia Absoluta Acumulada que coincide con el OAM? ¿Qué valor asume la Mediana?

X_i	f_i	$X_i * f_i$	F_i	fr_i	Fr_i
0	4	0	4	4/14	4/14
1	**3**	3	7	5/14	9/14
2	5	10	12	3/14	12/14
3	2	6	14	2/14	14/14

$n = 14$ $\quad Cantidad\ de\ Observaciones$

$$OAM = \frac{n}{2} = 7\ Orden\ Absoluto\ Mediano$$

$$Me(x) = \frac{1+2}{2} = 1{,}5\ hijos\ por\ familia$$

¿Cómo hago para calcular una medida de variabilidad?

En la primera clase presencial hemos visto que la suma de los desvíos de cada valor de la variable respecto al valor promedio se compensan. En términos formales,

$$\sum_{i=1}^{n} (\boldsymbol{X_i} - \bar{X})\boldsymbol{f_i} = 0$$

Pero podrían elevarse al cuadrado las desviaciones y serían positivas. Esta suma al dividirla por la cantidad de observaciones arroja un valor promedio que se conoce como Varianza (Variancia)

$$S_x^2 = \frac{\sum_{i=1}^{n} (\boldsymbol{X_i} - \bar{X})^2 \boldsymbol{f_i}}{n}$$

Trabajar con esta expresión es un poco complicada para tablas de frecuencias, por lo que se recomienda desarrollar el binomio del numerador de dicha expresión,

$$S_x^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 f_i}{n} = \sum_{i=1}^{n} \frac{X_i^2 f_i}{n} - \bar{X}^2$$

X_i	f_i	$X_i * f_i$	F_i	fr_i	Fr_i	$X_i^2 * f_i$
0	4	0	4	4/14	4/14	0
1	5	5	9	5/14	9/14	5
2	3	6	12	3/14	12/14	12
3	2	6	14	2/14	14/14	18

$$S_x^2 = \frac{0 + 5 + 12 + 18}{14} - (1{,}21)^2 = 1{,}0359\ hijos\ por\ familia^2$$

¿Qué problema tiene este resultado respecto a la unidad de medida de la variable?

Que la variable no se encuentra expresado en la misma unidad de medida que la variable de estudio.

En nuestro ejemplo, son serán hijos por familia sino hijos por familia al cuadrado**.**

Por lo tanto, lo que podemos hacer es extraerle la raíz cuadrada positiva a la varianza, obteniendo lo que se conoce como desvío estándar o típico.

$$\boldsymbol{S_x} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 f_i}{n} = \sum_{i=1}^{n}\frac{X_i^2 f_i}{n} - \bar{X}^2}$$

En nuestro ejemplo, el desvío estándar es, $S_x = \sqrt{1{,}0359} \cong 1{,}02\ hijos\ por\ familia$

Ahora bien, estas medidas de variabilidad son absolutas y no nos permiten comparar con otras variables que presenten distintas unidades de medida.

Por tal motivo, se suele utilizar una medida de variabilidad relativa denominada Coeficiente de Variación (Variabilidad) que se calcula como un cociente entre el desvío estándar y la media aritmética. En términos formales,

$$CV_X = \frac{\boldsymbol{S_x}}{\bar{X}}$$

Si dicha medida es inferior al 10%, es decir si el desvío estándar representa a lo sumo (como máximo) el 10% de la media aritmética, entonces diremos que los datos son homogéneos y por lo tanto la media es representativa de la totalidad de los datos.

En nuestro ejemplo, el $CV_X = \mathbf{0,84} > \mathbf{0,10}$ **, por lo tanto los datos no son homogéneos y se concluye que la media no es representativa.**

Que ocurre si quisiéramos determinar aquel valor de la variable que supera el k% de los datos y es superado por el (100-k)% de los mismos.

Por ejemplo, nos interesa aquel valor de variable que supera el 25% de los datos y es superado por su complemento, es decir el 75%. En esta situación, queremos calcular el percentil de orden 25, o simplemente percentil 25.

X_i	f_i	$X_i * f_i$	F_i	fr_i	Fr_i
0	4	0	4	4/14	4/14
1	5	5	9	5/14	9/14
2	3	6	12	3/14	12/14
3	2	6	14	2/14	14/14

$n = 14$ $\quad Cantidad\ de\ Observaciones$

$$OAP = \frac{25 * n}{100} = \frac{n}{4} = 3{,}5 \quad Orden\ Absoluto\ Percentilar$$

$$P_{25} = 0\ hijo\ por\ familia$$

Tipos Especiales de Percentiles

- Deciles (10,20,…,50,…,100)
- Quintiles (20,40,…,80,100)
- Cuartiles (25,…,75,100)

La Mediana es el Percentil 50. A veces se suele calcular una medida de dispersión con los percentiles cuando la medida de tendencia central elegida es la Mediana.

Esta medida de dispersión se denomina:

Rango Intercuartilar

$$\text{IQR} = P_{75} - P_{25}$$

Rango Interdecil

$$\text{IDR} = P_{90} - P_{10}$$

Variable Continua

X = Monto pagado en concepto de siniestros (en miles de pesos)

X	f_i
100 - 105	4
105 - 110	5
110 – 115	3
115 – 120	2

Ahora el valor 4 del intervalo [100 – 105) indica que hay 4 siniestros que tienen un monto pagado en ese rango de valores

A diferencia de una variable discreta, no representa la cantidad de veces que se repite un valor de la variable sino la cantidad de datos contenidos en dicho intervalo

Si quisiéramos calcular el promedio aritmético simple, ¿cómo podríamos hacerlo con esta tabla de frecuencias?

X	f_i	X_i	$X_i * f_i$
100 - 105	4	102,5	410
105 - 110	5	107,5	537,5
110 – 115	3	112,5	337,5
115 – 120	2	117,5	235

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

$$\bar{X} = \frac{1520}{14} = 108{,}57\ miles\ de\ \$$$

Significa que en promedio cada se paga por siniestro un monto de $ 108,57 miles

Si nos preguntan, cuál es el valor que se repite con más frecuencia nos estarían preguntando que indiquemos el **Modo/a** de la distribución de datos. En este caso, no hay un valor que se repite con más frecuencia sino que existe un intervalo con la mayor frecuencia (Intervalo Modal).

Sobre este intervalo se aplicará una fórmula de aproximación numérica conocida como interpolación lineal, deviene de determinar una recta entre dos puntos.

Lo primero que hacemos es determinar dicho intervalo y aplicar la siguiente expresión,

$$Mo_X = Li_m + \frac{d_1}{d_1+d_2} * a$$

Donde,

$Li_m = Límite\ inferior\ del\ Intervalo\ Modal$
$a = Amplitud\ del\ Intervalo\ Modal$
$d_1 = f_m - f_{m-1}$
$d_2 = f_m - f_{m+1}$

En este ejemplo, si se reemplazan por los valores de la tabla arroja el siguiente resultado,

Intervalo Modal = [105 – 110)

$Li_m = 105$
$a = 110 - 105 = 5$
$d_1 = 5 - 4 = 1$
$d_2 = 5 - 3 = 2$

$$Mo_X = 105 + \frac{1}{1+2} * 5 \cong \$\ 106{,}67 \text{ miles}$$

Es decir que el monto pagado en concepto de siniestros más frecuente en esta distribución tabular es de $ 106,67 miles.

Ahora bien para determinar la Mediana se procede de igual forma que con el Modo, mediante una fórmula de interpolación lineal. Primero se determinar el OAM (Orden Absoluto Mediano), es decir la mitad de los datos y se ubica a través de las Frecuencias Absolutas Acumuladas en que intervalo se encuentra contenido esta medida de tendencia central.

En nuestro ejemplo, $n = 14$ y $OAM = \frac{n}{2} = 7$. Por lo tanto, la Mediana se ubica en el Intervalo Mediano = [105 – 110) porque la Frecuencia Absoluta Acumulada de este intervalo es el primer valor que supera al OAM. En este intervalo se aplica la siguiente fórmula,

$$Me_X = Li_m + \frac{\frac{n}{2} - F_{m-1}}{f_m} * a$$

Donde,

$Li_m = Límite\ inferior\ del\ Intervalo\ Mediano$
$a = Amplitud\ del\ Intervalo\ Mediano$
$f_m = Frecuencia\ Absoluta\ Simple\ del\ Intervalo\ Mediano$
$F_{m-1} = Frecuencia\ Absoluta\ Acumulada\ del\ Intervalo\ Inmediato\ Anterior\ al\ Mediano$

En este ejemplo, si se reemplazan por los valores de la tabla arroja el siguiente resultado,

Intervalo Mediano = [105 – 110)

$Li_m = 105$
$a = 110 - 105 = 5$
$f_m = 5$
$F_{m-1} = 4$

$$Me_X = 105 + \frac{7-4}{5} * 5 = \$\ 108 \text{ miles}$$

Es decir que el monto pagado en concepto de siniestros hasta el cual se acumula el 50% de los datos en esta distribución tabular es de $ 108 miles. Las tres medidas dieron,

$$\bar{X} = \$\ 108{,}57\ miles$$
$$Mo_X = \$\ 106{,}67 \text{ miles}$$
$$Me_X = \$\ 108 \text{ miles}$$

¿A qué se debe? ¿Con que lo relacionan?

Si quisiéramos determinar la varianza, desvío estándar y coeficiente de variación es similar de la forma que lo hicimos con una variable discreta. La única diferencia es que utilizamos el punto medio de clase (del intervalo) para representar a las observaciones de un mismo intervalo. Ahora bien, si necesitáramos determinar un percentil, que podríamos hacer?

El procedimiento es similar a calcular la Mediana, recordar que esta medida es el Percentil 50. Por lo tanto, lo primero que deberíamos hacer es identificar el OAP (Orden Absoluto Percentilar) para ubicar el primer intervalo que supera a dicho porcentaje de datos y luego aplicar una fórmula similar a la Mediana. En términos formales,

$$P_k = Li_k + \frac{\frac{k*n}{100} - F_{k-1}}{f_k} * a$$

Donde,

$Li_k = Límite\ inferior\ del\ Intervalo\ Percentilar$

$a = Amplitud\ del\ Intervalo\ Percentilar$

$f_k = Frecuencia\ Absoluta\ Simple\ del\ Intervalo\ Percentilar$

$F_{k-1} = Frec.Absoluta\ Acumulada\ del\ Intervalo\ Inmediato\ Anterior\ al\ Per$centilar

En este ejemplo, si se quisiera determinar el Percentil 75, tal como se expreso previamente primero se determina el OAP asociado al 75% de los datos. Es decir,

$$OAP = \frac{k * n}{100} = \frac{75 * n}{100} = \frac{3}{4} \cdot n = \frac{3}{4} \cdot 14 = 10{,}5$$

Intervalo Percentilar = [110 – 115)

$Li_k = 110$
$a = 115 - 110 = 5$
$f_k = 3$
$F_{k-1} = 9$

$$P_{75} = 110 + \frac{10{,}5-9}{3} * 5 = \$\ 112{,}5 \text{ miles}$$

Es decir que el monto pagado en concepto de siniestros hasta el cual se acumula el 75% de los datos en esta distribución tabular es de $ 112,5 miles.

Realizar lo mismo para el Percentil 25 y determinar el Rango Intercuartilar.

Ahora bien, supongamos que estamos interesados en dado dos valores cualesquiera determinar el porcentaje de datos acumulados. Ahora bien, supongamos que queremos determinar el porcentaje de datos acumulados entre dos valores que figuran en la tabla. Por ejemplo, entre 105 y 110, esto es fácil dado que basta con observar la tabla y expresar su frecuencia relativa simple. En términos formales,

X	f_i
100 - 105	4
105 - 110	5
110 – 115	3
115 – 120	2

$$fr(105 \leq X \leq 110) = \frac{5}{14} \cong 0{,}3571$$

Es decir, que el 35,71% de los montos pagados en concepto de siniestros se encuentra entre $ 105 y $ 110 miles.

Si quisiéramos el porcentaje de datos entre $ 105 y $ 115, se podría proceder de la forma siguiente,

$$fr(105 \leq X \leq 115) = \frac{5+3}{14} \cong 0{,}5714 \quad \text{o} \quad Fr(X = 115) - Fr(X = 105) = \frac{12}{14} - \frac{4}{14} = \frac{8}{14}$$

Es decir, que el 57,14% de los montos pagados en concepto de siniestros se encuentra entre $ 105 y $ 115 miles.

Ahora bien, si quisiéramos determinar el porcentaje de datos entre $ 108 y $ 112,5, se podría calcular como la diferencia de Frecuencias Absolutas Acumuladas a cada valor de la variable,

$$Fr(X = 112{,}5) - Fr(X = 108) = ?$$

Se podría proceder mediante una fórmula de interpolación lineal,

$$F_x(x_k) = F_{k-1} + \frac{x_k - Li_k}{a} * f_k$$

Donde,

$Li_k = Límite\ inferior\ del\ Intervalo\ Percentilar$

$a = Amplitud\ del\ Intervalo\ Percentilar$

$f_k = Frecuencia\ Absoluta\ Simple\ del\ Intervalo\ Percentilar$

$F_{k-1} = Frec.Absoluta\ Acumulada\ del\ Intervalo\ Inmediato\ Anterior\ al\ Per$centilar

Ahora bien, si se determina

$$F(X = 112{,}5) = 9 + \frac{112{,}5 - 110}{5} * 3 = 10{,}5$$

$$F(X = 108) = 4 + \frac{108 - 105}{5} * 5 = 7$$

X	f_i
100 - 105	4
105 - 110	5
110 – 115	3
115 – 120	2

Estos valores fueron determinados previamente, son los OAP para el percentil 75 y 50.

La diferencia entre ambas frecuencias absolutas acumuladas da 10,5 – 7 = 3,5 datos.

En porcentaje, $\frac{3{,}5}{14} * 100 = 25\%$

Por lo tanto, el porcentaje de datos acumulados entre estos dos valores de variable es 25%.

Esta medida de concentración se conoce con el nombre de Rango Percentilar.

Tratamientos de los datos sin agrupar

Ahora veremos el tratamiento de los datos cuando trabajamos con datos sin agrupar.

Supongamos que contamos con los siguientes datos correspondientes a la cantidad de nuevos casos infectados con COVID-19 en 7 países de Latinoamérica al 22 de Marzo de 2020.

País	Cantidad de Infectados
Argentina	41
Brasil	74
Venezuela	7
Colombia	4
Chile	114
México	65
Bolivia	3

https://www.worldometers.info/coronavirus/

Si tuviésemos que calcular la cantidad promedio de infectados en estos países, la cantidad más frecuente, la cantidad mediana de infectados, la varianza, el desvío estándar y todas las medidas vistas, como procedemos?

Una opción es trabajar con una tabla de frecuencias de la forma siguiente y trabajar como hemos visto previamente.

X_i	f_i
3	1
4	1
7	1
41	1
65	1
74	1
114	1

Otra opción es emplear las fórmulas vistas sin considerar la tabla de frecuencias.

Por ejemplo para calcular la cantidad promedio de infectados en estos 8 países podemos emplear la siguiente fórmula,

$$\bar{X} = \sum_{i=1}^{n} \frac{\boldsymbol{X_i} * \boldsymbol{f_i}}{n} \quad = \frac{3 + 4 + 7 + \cdots + 74 + 114}{7} = 44\ infectados$$

No hay un valor de la variable que se repite con más frecuencia, por lo tanto en esta distribución de datos no hay modo.

Respecto a la mediana se procede de la forma siguiente,

1. Se ordenan los datos de forma ascendente (de menor a mayor)

2. Se ubica la posición de la Mediana, la cual es $(n + 1)/2$

$n = 7$

$x(1) = 3$
$x(2) = 4$
$x(3) = 7$
$x(4) = 41$
$x(5) = 65$
$x(6) = 74$
$x(7) = 114$

$$x\left(\frac{n+1}{2}\right) = x\left(\frac{7+1}{2}\right) = x(4) = 41$$

Con estos datos, la mediana es de 41 casos infectados. Es decir, hasta 41 infectados se acumula el 50 % de los datos y a partir de ahí el 50 % restante.

$n = 6$

$x(1) = 3$
$x(2) = 4$
$x(3) = 7$
$x(4) = 41$
$x(5) = 65$
$x(6) = 74$

$$x\left(\frac{n+1}{2}\right) = x\left(\frac{6+1}{2}\right) = x(3{,}5) = \frac{x(3)+x(4)}{2} = \frac{7+41}{2} = 24$$

Con estos datos, la mediana es de 24 casos infectados. Es decir, hasta 24 infectados se acumula el 50 % de los datos y a partir de ahí el 50 % restante.

¿Qué ocurre si queremos buscar un percentil distinto al 50?

¿Qué procedimiento podemos emplear?

Hay que tener en consideración lo siguiente,

Si nos piden el percentil 25, podemos calcular el orden del dato que corresponde al 25 % de los datos. La posición es $(n+1)/4$

Con esta posición se emplea una interpolación lineal entre los números ubicados esa posición.

$$P_{25} = x\left(\frac{n+1}{4}\right) = x\left(\frac{6+1}{4}\right) = x(1{,}75) = 0{,}25 * x(1) + 0{,}75 * x(2)$$

$$P_{25} = 0{,}25 * x(1) + 0{,}75 * x(2) = 0{,}25 * 3 + 0{,}75 * 4 = 3{,}75$$

Ahora bien, si fuera el percentil 75, podemos calcular el orden del dato que corresponde al 75 % de los datos. La posición es $(3n+3)/4$. Con esta posición se emplea una interpolación lineal entre los números ubicados esa posición.

$$P_{75} = x\left(\frac{3n+3}{4}\right) = x\left(\frac{3*6+3}{4}\right) = x(5{,}25) = 0{,}75 * x(5) + 0{,}25 * x(6)$$

$$P_{75} = 0{,}75 * x(5) + 0{,}25 * x(6) = 0{,}75 * 65 + 0{,}25 * 74 = 67{,}25$$

Si nos piden el decil n° i (i = 1, 2, …, 10), podemos calcular el orden del dato que corresponde al (i * 100) % de los datos.

La posición del decil i es $\frac{i*(n+1)}{10}$

Por ejemplo,

1er decil → $\frac{1*(n+1)}{10} = \frac{n+1}{10}$

5to decil → $\frac{5*(n+1)}{10} = \frac{n+1}{2}$

8to decil → $\frac{8*(n+1)}{10} = \frac{4*(n+1)}{5}$

Si nos piden el cuartil n° i (i = 1, 2, …, 4), podemos calcular el orden del dato que corresponde al (i * 100) % de los datos.

La posición del cuartil i es $\frac{i*(n+1)}{4}$

Por ejemplo,

1er cuartil → $\frac{1*(n+1)}{4} = \frac{n+1}{4}$

2do cuartil → $\frac{2*(n+1)}{4} = \frac{n+1}{2}$

3er cuartil → $\frac{3*(n+1)}{4}$

Si nos piden el quintil n° i (i = 1, 2, …, 5), podemos calcular el orden del dato que corresponde al (i * 100) % de los datos.

La posición del quintil i es $\frac{i*(n+1)}{5}$

Por ejemplo,

1er quintil → $\frac{1*(n+1)}{5} = \frac{n+1}{5}$

2do quintil → $\frac{2*(n+1)}{5}$

3er quintil → $\frac{3*(n+1)}{5}$

Luego se interpola como hemos visto.

Percentil 80 → $\frac{80*(n+1)}{100}$

$$P_{80} = x\left(\frac{80 * (n+1)}{100}\right) = x(6{,}4) = 0{,}60 * x(6) + 0{,}40 * x(7)$$

$$P_{80} = 0{,}60 * 65 + 0{,}40 * 74 = 68{,}60\ infectados$$

Percentil 43 → $\frac{43*(n+1)}{100}$

$$P_{43} = x\left(\frac{43 * (n+1)}{100}\right) = x(3{,}44) = x(3) + 0{,}44 * \big(x(4) - x(3)\big)$$

$$P_{43} = 0{,}56 * x(3) + 0{,}44 * x(4) = 0{,}56 * 7 + 0{,}44 * 41 = 21{,}96\ infectados$$

F4 | =B4+(E4-A4)*(B5-B4)

	A	B	C	D	E	F
2	SR. No	Digit				
3	1	23				
4	2	24				
5	3	27				
6	4	30				
7	5	32				
8	6	32				
9	7	32				
10	8	33				
11	9	36				
12	10	36				
13	11	42				
14	12	45				
15	13	51				
16	14	54				
17	15	55				
18	16	55				
19	17	56				
20	18	57				
21	19	60				
22	20	62				
23	21	63				
24	22	72				
25	23	77				
26						
27	n	23				

Decile	Data position	Value
D1	2.4	25.2
D2	4.8	31.6
D3	7.2	32.2
D4	9.6	36.0
D5	12	45.0
D6	14.4	52
D7	16.8	55.8
D8	19.2	60.4
D9	21.6	68.4

Para calcular las medidas de variabilidad se procede de igual forma,

$$S_x^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n} = \sum_{i=1}^{n}\frac{X_i^2}{n} - \bar{X}^2 \qquad S_x = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n} = \sum_{i=1}^{n}\frac{X_i^2}{n} - \bar{X}^2}$$

$$CV_X = \frac{\boldsymbol{S_x}}{\bar{X}}$$

En nuestro ejemplo estas medidas nos dan,

$$S_x^2 = \frac{3^2 + 4^2 + 7^2 + 41^2 + 65^2 + 74^2 + 114^2}{8} - 44^2 = 3056{,}5 \; - 1936 = 1120{,}5$$

$$S_x^2 = 1120{,}5 \; infectados^2$$

$$S_x \cong 33{,}47 \; infectados$$

$$CV_X \cong 0{,}76$$

Momentos Absolutos y Centrados

De una distribución de datos

Los momentos absolutos y centrados se utilizan en estadística descriptiva como en la teoría de la probabilidad para calcular medidas de forma.

Un momento absoluto de orden k representa el promedio de la potencia k-ésima de los valores observados de la variable en estudio. En términos formales,

$$m_k = \sum_{i=1}^{n} \frac{X_i^k * \boldsymbol{f_i}}{n}$$

Por lo tanto, si hacemos variar k, es decir darle distintos valores obtendremos distintos promedios que serán de utilidad para el cálculo de ciertas medidas estadísticas.

Por ejemplo,

$$m_o = \sum_{i=1}^{n} \frac{X_i^0 * \boldsymbol{f_i}}{n} = \sum_{i=1}^{n} \frac{\boldsymbol{f_i}}{n} = 1$$

Si k = 1, obtenemos el promedio aritmético simple de los valores observados. En términos formales,

$$m_1 = \sum_{i=1}^{n} \frac{X_i^1 * \boldsymbol{f_i}}{n} = \bar{X}$$

Si k = 2, obtenemos el promedio aritmético simple del cuadrado de los valores observados. En términos formales,

$$m_2 = \sum_{i=1}^{n} \frac{X_i^2 * \boldsymbol{f_i}}{n}$$

Ahora bien, hemos visto que la varianza se calcula en función de estos dos momentos. Recordamos por un momento su fórmula de cálculo,

$$S_x^2 = \sum_{i=1}^{n} \frac{X_i^2 f_i}{n} - \bar{X}^2 = m_2 - m_1^2 \qquad S_x = \sqrt{m_2 - m_1^2}$$

Cuando k = 3 y k =4, obtenemos el promedio aritmético simple del cubo y cuádruple de los valores observados. En términos formales,

$$m_3 = \sum_{i=1}^{n} \frac{X_i^3 * \boldsymbol{f_i}}{n}$$

$$m_4 = \sum_{i=1}^{n} \frac{X_i^4 * \boldsymbol{f_i}}{n}$$

Estos dos momentos absolutos son importantes para calcular las medidas de forma asimetría y kurtosis, tal como se verán en unas slides más adelante.

A continuación, se exhibe la forma de cálculo de los momentos centrados,

Un promedio centrado de orden k es el promedio aritmético simple de la potencia k-ésima de los desvíos de la variable respecto a su valor promedio. En términos formales,

$$mc_k = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^k * f_i}{n}$$

Por ejemplo, si k = 0 dicho momento central vale 1. Es decir,

$$mc_0 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^0 * f_i}{n} = 1$$

Por ejemplo, si k = 1 dicho momento central vale 0. Es decir,

$$mc_1 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^1 * f_i}{n} = 0$$

Ahora bien, si k = 2 se obtiene la varianza,

$$mc_2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2 * f_i}{n} = S_x^2$$

Hemos visto que la varianza se desagrega en una relación con los momentos absolutos. Por lo tanto, si se desarrolla el Binomio que figura en el denominador de la expresión anterior se obtiene lo siguiente,

$$mc_2 = S_x^2 = m_2 - m_1^2$$

Con los momentos centrados siguientes ocurre algo similar. Por ejemplo, si k = 3 se puede expresar en función de los momentos absolutos,

$$mc_3 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^3 * f_i}{n} = m_3 - 3m_2m_1^2 + 2m_1^3$$

Si k = 4 se puede expresar en función de los momentos absolutos de la forma siguiente,

$$mc_4 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^4 * f_i}{n} = m_4 - 4m_3m_1 + 6m_2m_1^2 - 3m_1^4$$

Estos momentos centrados serán de utilidad para calcular las medidas de forma, conocidas como coeficiente de asimetría y de curtosis.

Ahora bien, si quisiéramos calcularlos con nuestro de variable discreta debemos usar más columnas en la tabla de frecuencias.

X_i	f_i	$X_i * f_i$	$X_i^2 * f_i$	$X_i^3 * f_i$	$X_i^4 * f_i$
0	4	0	0	0	0
1	5	5	5	5	5
2	3	6	12	24	48
3	2	6	18	54	162

$$m_1 = \frac{0 + 5 + 6 + 6}{14} \cong 1{,}21$$

$$m_2 = \frac{0 + 5 + 12 + 18}{14} = 2{,}5$$

$$m_3 = \frac{0 + 5 + 24 + 54}{14} \cong 5{,}93$$

$$m_4 = \frac{0 + 5 + 48 + 162}{14} \cong 15{,}36$$

$$mc_2 = 2{,}5 - 1{,}21^2 \cong 1{,}03$$

$$mc_3 = 5{,}93 - 3 * 2{,}5 * 1{,}21^2 + 3 * 1{,}21^3 \cong 9{,}78$$

$$mc_4 = 15{,}36 - 4 * 5{,}93 * 1{,}21 + 6 * 2{,}5 * 1{,}21^2 - 3 * 1{,}21^4 \cong 2{,}19$$

Medidas de Forma

Asimetría y Curtosis

La curtosis de una variable estadística es una característica de forma de su distribución de frecuencias.

Según su concepción clásica, una curtosis grande implica una mayor concentración de valores de la variable tanto muy cerca de la media de la distribución (pico) como muy lejos de ella (colas), al tiempo que existe una relativamente menor frecuencia de valores intermedios.

Esto explica una forma de la distribución de frecuencias con colas más gruesas, con un centro más apuntado y una menor proporción de valores intermedios entre el pico y colas.

Una mayor curtosis no implica una mayor varianza, ni viceversa.

Un coeficiente de apuntamiento o de curtosis es el cuarto momento con respecto a la media estandarizado que se define como,

$$K[X] = \frac{mc_4}{S_x^4} - 3$$

Este ratio se encuentra estandarizado al dividirse por la potencia cuarta del desvío estándar y será:

- **Leptocúrtica**, cuando $K[X] > 0$, significa que es más apuntada y con colas más gruesas que la Distribución de Probabilidad Normal (la veremos más adelante!!!!)

- **Platicúrtica**, cuando $K[X] < 0$, significa que es menos apuntada y con colas menos gruesas que la Distribución Normal.

- **Mesocúrtica**, cuando $K[X] = 0$ y es cuando se corresponde con una Normal.

Las medidas de asimetría son indicadores que permiten establecer el grado de simetría (o asimetría) que presenta una distribución de probabilidad de una variable aleatoria sin tener que hacer su representación gráfica.

Como eje de simetría consideramos una recta paralela al eje de ordenadas que pasa por la media de la distribución.

Si una distribución es simétrica, existe el mismo número de valores a la derecha que a la izquierda de la media, por tanto, el mismo número de desviaciones con signo positivo que con signo negativo.

Decimos que hay asimetría positiva (o a la derecha) si la "cola" a la derecha de la media es más larga que la de la izquierda, es decir, si hay valores más separados de la media a la derecha.

Diremos que hay asimetría negativa (o a la izquierda) si la "cola" a la izquierda de la media es más larga que la de la derecha, es decir, si hay valores más separados de la media a la izquierda.

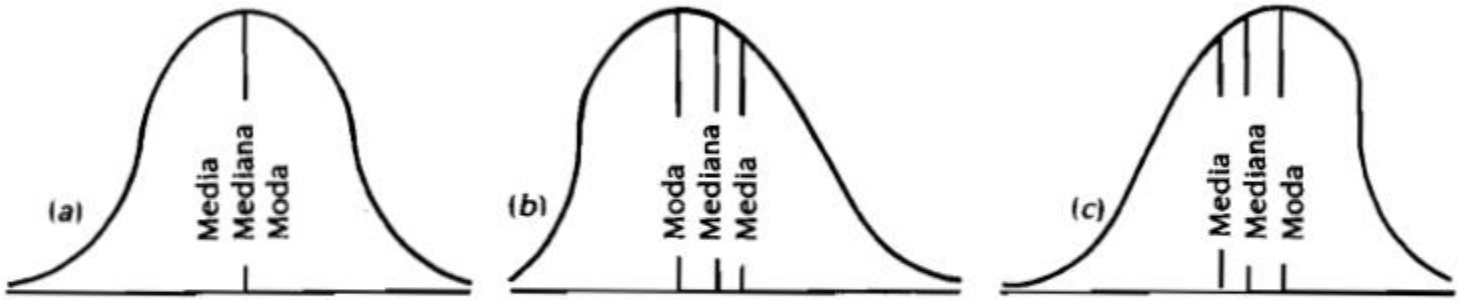


Figura 3.1.a Figura 3.1.b Figura 3.1.c

Figura 3.1 *La posición de la media, mediana y moda en distribuciones simétricas y sesgadas.*

El coeficiente de asimetría de Fisher se define como,

$$AS[X] = \frac{mc_3}{S_x^3}$$

Si $\mathrm{AS}[X] > 0$ la distribución es asimétrica positiva (o a la derecha)

Si $\mathrm{AS}[X] < 0$ la distribución es asimétrica negativa (o a la izquierda)

Si la distribución es simétrica entonces sabemos que $\mathrm{AS}[X] = 0$.

El reciproco no es cierto: es un error común asegurar que si $\mathrm{AS}[X] = 0$ entonces la distribución es simétrica (lo cual es falso).

Con los resultados de nuestro ejemplo para variable discreta, se determinará el coeficiente de asimetría y de curtosis,

$$\mathrm{AS}[X] = \frac{9{,}78}{1{,}03^3} > 0$$

La distribución de los datos tiene asimetría positiva, es decir se acumula mayor cantidad de datos del lado izquierdo de la distribución de los datos.

Respecto al coeficiente de curtosis esta arroja el siguiente resultado,

$$K[X] = \frac{2{,}19}{1{,}03^4} - 3 \cong -1{,}05$$

La distribución de los datos es platicúrtica, es decir los datos se encuentran aplanados.

Propiedades de la Media Aritmética y Varianza

Media Aritmética:

1. El valor promedio de una constante es la constante misma. Si $X = k$ entonces $\overline{X} = k$

2. El valor promedio de una constante multiplicada por una variable es la constante multiplicada por el promedio de la variable. Si $Y = kX$ entonces $\overline{Y} = k\overline{X}$

3. El valor promedio de una transformación afín es si $Y = a + bX$ entonces $\overline{Y} = a + b\overline{X}$

4. El promedio de una suma de variables es la suma del promedio de las variables

$$\text{Si } W_i = X_i + Y_i \; entonces \; \overline{W} = \overline{X} + \overline{Y}$$

En algunos casos cada uno de los números de la sucesión $x_1, x_2, x_3, \ldots, x_n$ tiene una importancia relativa (peso) respecto de los demás elementos de la sucesión. Cuando esto sucede, la media está dada por

$$\bar{x}_p = \frac{w_1x_1 + w_2x_2 + w_3x_3 + \quad + w_kx_k}{w_1 + w_2 + w_3 + \quad + w_k} \qquad (3\text{-}10)$$

en donde $x_1, x_2, x_3, \ldots, x_k$ son los datos; y $w_1, w_2, w_3, \ldots, w_k$ son los pesos respectivos.

Varianza:

1. La varianza de una constante es cero

   Si $X = k$ entonces $S_k^2 = 0$

2. La varianza de una constante por una variable es la constante al cuadrado por la varianza de la variable

   Si $Y = kX$ entonces $S_Y^2 = k^2 S_X^2$

3. La varianza de la transformación afín

   Si $Y = a + bX$ entonces $S_Y^2 = b^2 S_X^2$

4. La varianza de una suma de variables

   Si $W_i = X_i + Y_i \; entonces \; S_W^2 = S_X^2 + S_Y^2 \pm 2 \; Cov(X, Y)$

Varianza:

4. La varianza de una suma de variables

Si $W_i = X_i + Y_i \; entonces \; S_W^2 = S_X^2 + S_Y^2 \pm 2\, Cov(X,Y)$

Si $X \perp Y \Rightarrow Cov(X,Y) = 0 \;\; (\Rightarrow \rho(X;Y) = 0) \Rightarrow S_W^2 = S_X^2 + S_Y^2$

$$Cov(X,Y) = \frac{\frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n}}{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

$$\rho(X;Y) = \frac{Cov(X,Y)}{S_X * S_Y}$$

Independencia ⇒ Incorrelación

Varianza:

4. La varianza de una suma de variables

Si $W_i = aX_i + bY_i \; entonces \; S_W^2 = a^2 S_X^2 + b^2 S_Y^2 \pm 2\, a\, b\, Cov(X,Y)$

$$S_W^2 = a^2 S_X^2 + b^2 S_Y^2 \pm 2\, a\, b\, \rho(X;Y) S_X * S_Y$$

$$\rho(X;Y) = \frac{Cov(X,Y)}{S_X * S_Y}$$

Resumen de Fórmulas

$$\bar{X} = \sum_{i=1}^{n} \frac{\boldsymbol{X_i} * \boldsymbol{f_i}}{n}$$

$$S_x^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 f_i}{n} = \sum_{i=1}^{n} \frac{X_i^2 f_i}{n} - \bar{X}^2$$

$$S_x = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 f_i}{n} = \sum_{i=1}^{n} \frac{X_i^2 f_i}{n} - \bar{X}^2}$$

$$\text{IQR} = P_{75} - P_{25}$$

$$Mo_X = Li_m + \frac{d_1}{d_1 + d_2} * a$$

$$\text{IDR} = P_{90} - P_{10}$$

$$Me_X = Li_m + \frac{\frac{n}{2} - F_{m-1}}{f_m} * a$$

$$CV_X = \frac{\boldsymbol{S_x}}{\bar{X}}$$

$$P_k = Li_k + \frac{\frac{k*n}{100} - F_{k-1}}{f_k} * a$$

$$F_x(x_k) = F_{k-1} + \frac{x_k - Li_k}{a} * f_k$$

Próxima Clase

Nos introduciremos en la teoría de la Probabilidad y se darán las definiciones más importantes y los axiomas y teoremas que sustentan dicha teoría.

Se espera que repasen todo lo visto y que puedan ejercitar los ejercicios sugeridos en la página web y en el grupo de la materia.

Ante cualquier consulta, no duden en avisarnos mediante el foro.


TO BE
CONTINUED...

Preguntas, Sugerencias y Comentarios

RDelRosso-ext@austral.edu.ar

¡Muchas Gracias!