2.1 Análisis conjunto de variables

Hasta ahora se han analizado las variables de una población independientemente unas de otras, en la práctica muchas de estas variables se relacionan apareciendo el concepto de distribución conjunta de variables. Para simplificar su estudio solo se considerarán relaciones entre dos variables (X, Y)[1] de una población. X e Y pueden ser variables cualitativas o cuantitativas.

Interesa medir para las dos variables la distribución conjunta de frecuencias absolutas y relativas, las distribuciones marginales, las distribuciones condicionales, indicadores estadísticos y la relación de dependencia entre ellas.

Definición 2.1: Distribución conjunta de frecuencias de dos variables $(\boldsymbol{X, Y})$

Se denominará Distribución conjunta de frecuencias de dos variables (X, Y) a la tabla que representa los valores que toma la variable bidimensional (x_i, y_j) con $i = 1,2,3, \dots, k$ y $j = 1,2,3, \dots, l$ junto con las frecuencias absolutas y relativas de éstos en la población.

Definición 2.2: Frecuencia Absoluta: $\boldsymbol{n_{ij}}$

En una tabla de distribución conjunta de frecuencias de dos variables (X, Y) en una población de tamaño n, se llamara Frecuencia Absoluta de (x_i, y_j), que se denotara como n_{ij}, al número de elementos de la población para los que X toma el valor x_i y Y toma el valor y_j

Definición 2.3: Frecuencia Relativa: $\boldsymbol{f_{ij}}$

En una tabla de distribución conjunta de frecuencias de dos variables (X, Y) en una población de tamaño n, se llamara Frecuencia Relativa de (x_i, y_j), que se denotara como f_{ij}, a la proporción de elementos de la población para los que X toma el valor x_i y Y toma el valor y_j y se expresa como:

$$f_{ij} = \frac{n_{ij}}{n}$$

Cada una de las variables (X, Y) se puede analizar separadamente a través de lo que se denomina distribución marginal.

Definición 2.4: Distribución marginal de X: $\boldsymbol{n_{i*}}$

Se denomina Frecuencia Marginal Absoluta de x_i al valor:

$$n_{i*} = n_{i1} + n_{i2} + n_{i3} + \cdots + n_{il} = \sum_{j=1}^{l} n_{ij}$$

Definición 2.5: Distribución marginal de Y: $\boldsymbol{n_{*j}}$

Se denomina Frecuencia Marginal Absoluta de y_j al valor:

$$n_{*j} = n_{1j} + n_{2j} + n_{3j} + \cdots + n_{kj} = \sum_{i=1}^{k} n_{ij}$$

Las frecuencias marginales relativas se obtienen mediante las fórmulas:

$$f_{i*} = \frac{n_{i*}}{n} \quad \text{y} \quad f_{*j} = \frac{n_{*j}}{n}$$

Definición 2.6: Frecuencia condicional de X: $\boldsymbol{f(y|x)}$

[1] Se usaran letras mayúsculas para representar los nombres de las variables y se dejan las minúsculas para indicar los valores que toman dichas variables

La frecuencia condicional de X se define mediante la fórmula:

$$f(Y = j|X = i) = \frac{f(i,j)}{f_{i*}}$$

Definición 2.7: Frecuencia condicional de Y: $f(x|y)$

La frecuencia condicional de Y se define mediante la fórmula:

$$f(X = i|Y = j) = \frac{f(i,j)}{f_{*j}}$$

Ejercicio: El Administrador de un local de comidas rápidas en un reconocido centro comercial de la ciudad de Cali ha registrado el número de clientes de acuerdo al promedio del consumo: X (en miles de pesos) y al número de productos comprados: Y (en unidades)

No de productos: Y / Promedio del Consumo: X	**1**	**2**	**3**	**4**	**5**
$ 5000	10	5	1	0	0
$ 10000	15	20	5	1	0
$ 15000	20	35	12	4	1
$ 20000	10	30	12	7	2

Resuelva:

1. Interprete el valor de 35 registrado en la tabla.
2. Obtenga la Frecuencia Relativa: f_{ij}
3. Obtenga e Interprete $f(x = 5000, y = 2)$
4. Obtenga la frecuencia marginal de X: f_{i*}
5. Obtenga la frecuencia marginal de Y: f_{*j}
6. Dado que un cliente consume en promedio $15000, ¿Qué porcentaje de clientes compra exactamente tres productos? Respuesta: 0.17

Ejercicio: Una agencia de empleo profesional evalúa la relación entre la experiencia laboral (en años) (***X***) y el número de títulos académicos (pregrado, postgrados) (***Y***) en 500 solicitudes. Se tiene la siguiente información:

		Y			
		1	2	3	4 o más
	1	30	5	0	0
	2	20	15	10	5
X	3	15	15	30	25
	4	10	25	30	50
	5 o más	5	60	90	60

a. Obtenga la frecuencia relativa: $\boldsymbol{f_{ij}}$
b. Indique e interprete el valor más probable de la distribución conjunta $\boldsymbol{f_{ij}}$
c. Determine las frecuencias marginales de X e Y.

2.2 **Taller**

1. Un estudio de mercadeo en cierta ciudad ha registrado para una muestra de 1000 negocios establecidos legalmente en la Cámara de Comercio, la cantidad de negocios según El Número de Empleados (***X***) y las Ventas Diarias en millones de pesos (***Y***). Se tiene la siguiente información:

		Y (Ventas diarias en millones de pesos		
		Menos o igual a 10	(10;20]	(20;50]
X	[1;10]	293	98	78
No. de	(10;20]	122	73	68
Empleados	(20;50]	98	54	44
	Más de 50	48	24	0

   a. Interprete el valor de 293
   b. Obtenga la frecuencia relativa: $\boldsymbol{f_{ij}}$
   c. Obtenga las distribuciones marginales de X e Y.
   d. Obtenga los valores promedio de X e Y. De una interpretación a estos resultados
   e. Obtenga la desviación estándar de X e Y. De una interpretación a estos resultados
   f. Dado que un negocio tiene más de 50 empleados, ¿cuál es el porcentaje de negocios con ventas superiores a 10 millones de pesos?

2. ASONALCO VALLE construirá un nuevo centro recreacional en el sur de la ciudad de Cali. Se debe decidir sobre la manera de diseñar el centro recreacional sobre la base del tipo de actividades que ofrecerá el centro a los clientes. Una encuesta reciente de 300 posibles clientes mostró los resultados relacionados a continuación:

Tipo de actividad de interés	No. respuestas
Deportivas: natación, futbol, tenis	63
Lúdicas: baile, caminatas, etc.	135
No les interesa	78
No responden	24

a. Diseñe una gráfica de barras para representar los resultados de la encuesta
b. Trace una gráfica de circular para los resultados de la encuesta
c. Si usted está preparando un informe para el gerente de ASONALCO, ¿Cuál es la mejor gráfica para representar los resultados?. ¿Cómo se llama esta gráfica?. **Ayuda**: investigue la gráfica de Pareto

3. Se reportan a continuación los ingresos por semana (en millones de pesos) de una compañía comercial durante un largo periodo.

55,8	35,0	37,0	91,3	30,2	42,3	76,8	60,6	76,0	40,5
45,9	39,1	55,5	56,0	44,6	71,7	42,7	45,3	47,2	36,4
83,2	40,0	51,7	36,7	45,8	47,3	94,6	56,3	30,0	68,2
75,3	71,4	65,2	52,6	58,2	48,0	33,3	36,6	39,8	35,3
49,2	77,1	59,1	49,5	28,3	45,9	41,3	29,4	87,1	66,3

Construya para la variable ingresos las siguientes herramientas:

a. La tabla de frecuencias. Ayuda: use siete intervalos de igual amplitud.
b. El histograma. Comente sobre la forma de distribución de los ingresos.
c. El diagrama de tallo y hojas. Compare con el histograma anterior y comente sobre la utilidad de esta herramienta.
d. El diagrama de caja y extensión.
e. ¿Cuáles de las tres herramientas representa mejor el comportamiento de los datos? Concluya sobre el uso apropiado de cada una de las herramientas utilizadas.

4. Un supervisor de control de calidad recoge continuamente muestras de una máquina empacadora de botellas de agua para determinar si el proceso de llenado está cumpliendo con los estándares establecidos (volumen promedio de 250 mililitros y un coeficiente de variación menor del 5%. Después de analizar la última muestra tomada, el supervisor decide parar el proceso de llenado, resuelva:

a. Calcule los indicadores de tendencia central
b. Determine el grado de homogeneidad del volumen en la muestra tomada
c. Pruebe que la muestra no tiene datos atípicos. Ayuda: use los estándares del promedio y la desviación estándar.
d. ¿tiene razón el supervisor en parar el proceso? Justifique.

Datos del volumen empacado por botella (en mililitros) de la última muestra tomada:

248, 249, 250, 251, 252, 253, 254, 255, 252, 250, 251, 252, 253, 254, 249, 250, 251, 252, 253, 254, 252, 251, 252, 253, 251, 252, 253, 251, 252, 253, 252, 256.

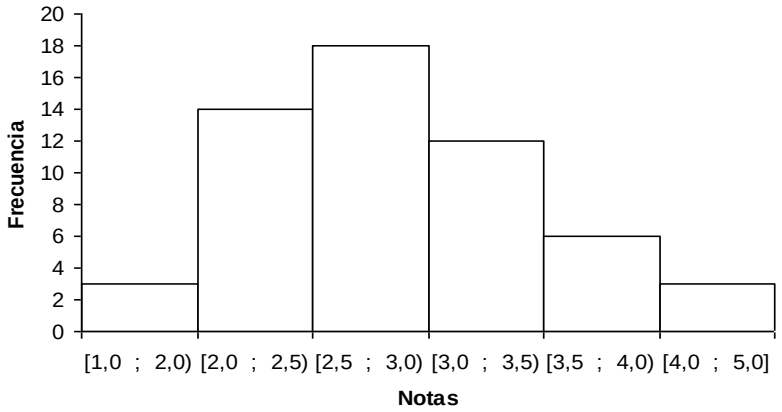
5. A continuación se indican la Tasa de desempleo (promedio enero-diciembre de cada año) y el Número de desocupados en Colombia desde el año 2001.

a. Calcule e interprete el promedio de la Tasa de desempleo en Colombia
b. Calcule e interprete la desviación estándar de la Tasa de desempleo en Colombia.
c. Haga un análisis descriptivo indicando tendencias, forma y variación de la Tasa de desempleo.
d. Utilice una herramienta estadística para indicar un pronóstico para la Tasa de desempleo en Colombia en el año 2015. Indique el valor y justifique.

AÑO	NUMERO DE DESOCUPADOS (EN MILES)	TASA DE DESEMPLEO
2001	2.782	15,0
2002	2.927	15,5
2003	2.724	14,0
2004	2.632	13,6
2005	2.280	11,8
2006	2.311	12,0
2007	2.152	11,2
2008	2.214	11,3
2009	2.515	12,0
2010	2.564	11,8
2011	2.426	10,8
2012	2.394	10,4
2013	2.243	9,6
2014	2.151	9,1

Fuente: El Dane

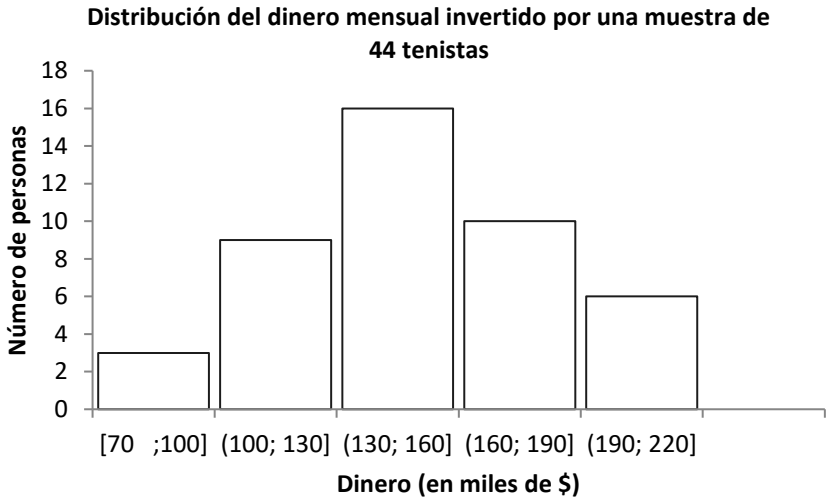
6. El siguiente histograma muestra los resultados en el primer parcial de un curso de Estadística



a. ¿Cuántos estudiantes presentaron el examen?
b. ¿Qué porcentaje de estudiantes ganaron el examen?
c. ¿Cuántos estudiantes presentaron un resultado inferior a 2.8?
d. Calcule el promedio y la desviación estándar de las notas

7. Una cadena de tiendas deportivas que satisface las necesidades de los tenistas de la región, planea llevar a cabo un estudio sobre la cantidad de dinero mensual que un tenista invierte en este deporte. Sobre esta base, se desea analizar la posibilidad de ofrecer equipo gratis (bolas, encordados, etc.) para los tenistas que inviertan más. Una muestra de 44 tenistas tomada el mes pasado revelo las inversiones (en miles de pesos) indicadas en la gráfica adjunta. Resuelva:

a. Organice los datos en una tabla de distribución de frecuencias
b. ¿Cuál es la tendencia en los gastos de los tenistas? Use un indicador estadístico para marca esta tendencia.
c. ¿Qué tan variable son las inversiones que hacen estos tenistas? Justifique con un indicador estadístico.
d. Si solo hay presupuesto para el 2.5% de los tenistas que más inviertan, ¿Cuál sería el valor mínimo (valor aproximado) a considerar para ofrecer el equipo gratis? Justifique su respuesta.
e. Dada la promoción de equipo gratis, Alonso, un gran aficionado al tenis invierto $ 245,000 en su última compra, ¿se puede considerar este valor como un dato atípico?



8. Un supervisor de control de calidad recoge continuamente muestras de dos máquinas llenadoras de latas para una libra de café (16 onzas) En un procedimiento rutinario se recogieron las dos siguientes muestras:

Maquina 1:	16.0	16.0	15.9	15.9	15.8	15.7	15.6	16.4	16.3	16.0	16.0	16.0
15.9	15.9	16.1	16.1	16.2	16.1	16.1	16.1	16.2	16.2	16.2	16.2	
16.3	16.3	16.1	16.0	16.0	15.8	15.8						
Maquina 2:	16.0	16.0	16.0	16.1	16.1	16.1	15.9	15.9	15.9	15.8	15.8	15.8
15.8	15.7	15.7	15.6	15.7	15.8	15.8	15.9	15.9	16.0	16.0	16.2	
16.2	16.4	16.5	16.2	16.4	16.1	15.9	16.8					

Resuelva las siguientes preguntas:

a. En el proceso de empacado se identifican las siguientes variables:
Peso empacado (en onzas)
Numero de máquinas (una o dos)
Estado del proceso (estable, no estable)

Clasifique cada variable según sean Cuantitativas o Cualitativas y según la escala de medición (nominal, ordinal, de intervalo o razón)

b. Analice individualmente a cada máquina, ¿Cuál de ellas parece presentar un comportamiento anormal (no estable)? Justifique su respuesta con un procedimiento estadístico

c. Si el proceso de pesado es estable, la compañía ha establecido que el 95.5% de las latas deben ser empacadas para exportación, ¿Cuál es este intervalo de peso?

d. En la maquina 1 se encontró una lata con peso 15.2 onzas, ¿este valor se puede considerar atípico?

9. Responda Falso o verdadero

a. La desviación estándar del conjunto de valores 2, 2, 2, 2 y 2 es 2. ( )
b. En una distribución normal (en forma de campana), el rango de los datos es aproximadamente igual a seis desviaciones estándar. ( )
c. Dos conjuntos de datos de n_1 y n_2 elementos tienen medianas $\tilde{X}_1$ y $\tilde{X}_2$ respectivamente. La mediana del total de datos $(n_1 + n_2)$ es $(\tilde{X}_1 + \tilde{X}_2)$. ( )
d. Dos conjuntos de datos de n_1 y n_2 elementos tienen medias $\bar{x}_1$ y $\bar{x}_2$ respectivamente. La media del total de datos $(n_1 + n_2)$ es $\frac{(\bar{x}_1+\bar{x}_2)}{2}$. ( )

10. Una muestra tiene media 40, varianza 16 y distribución simétrica acampanada. Esto significa que: Responda falso o verdadero justificando la respuesta.

a. Aproximadamente el 99.5 de los datos está entre 38 y 42
b. Entre 38 y 48 hay más del 75% de los datos
c. El valor aproximado para el percentil 84 es 44
d. El valor aproximado hasta el que se acumula el 98% de los datos es 52.

11. Una muestra aleatoria de 7 días de operaciones produjo los siguientes valores de los datos (precio, cantidad):

Precio por litro, en miles de pesos, de pintura (X)	Cantidad Vendida, en litros (Y)
10	100
8	120
5	200
4	200
10	90
7	110
6	150

Describa los datos numéricamente: calcule: media, mediana, moda, desviación estándar, varianza, rango y C.V. para cada variable, ¿Cuál de las variables presenta un mejor comportamiento? Justifique su respuesta en términos estadísticos.

12. Para el rector de una universidad, los puntos obtenidos por los aspirantes en las pruebas de admisión constituyen una variable aleatoria con polígono de frecuencias relativas que sugiere una distribución simétrica y en forma de campana. A su juicio la proporción de estudiantes que obtienen más de 400 puntos es 0.025 y además la proporción de estudiantes que obtienen más de 370 puntos es 0.16. Cuáles son la media y la desviación estándar en esta prueba.

13. La Sociedad Colombiana de Economistas – SCE, es una Asociación Sin Ánimo de Lucro –ONG, de carácter gremial y académico, que agrupa y organiza a los profesionales de la carrera de economía en Colombia. En su última reunión, 25 de sus miembros recibieron el encargo de predecir el crecimiento porcentual que experimentará el *Índice de Precios de Consumo (IPC)* en el próximo año. Sus predicciones fueron:

    2.6, 3.4, 3.9, 4.0, 4.5, 5.2, 5.5, 3.4, 3.9, 4.0, 4.5, 5.0, 3.2, 3.8, 4.1, 4.6, 3.8, 4.1, 4.6, 3.7, 4.1, 4.7, 4.2, 4.4, 4.3

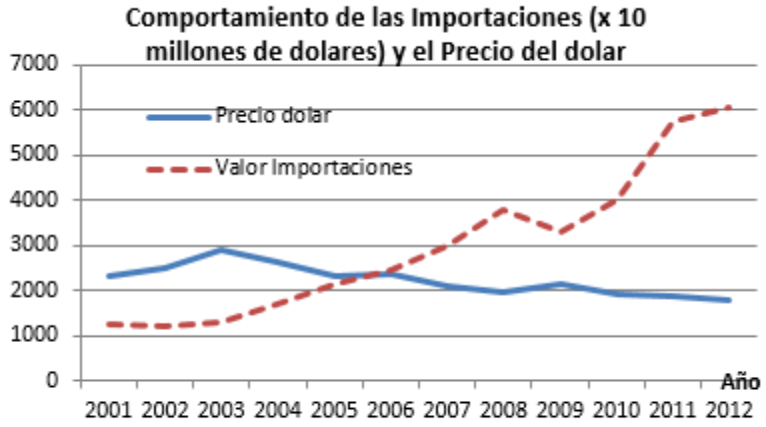
a. Identifique la variable de los datos presentados. Indique su nombre y su escala de medición
b. Organice los datos en un diagrama de tallo y hojas y concluya sobre la forma que toman los datos.
c. Calcule las medidas de tendencia central y especifique cuál de ellas marca mejor la tendencia de los datos.
d. Calcule la varianza de los datos e intérprete su valor.
e. Con base en los datos suministrados y su análisis estadístico haga una predicción del crecimiento porcentual que experimentará el *Índice de Precios de Consumo (IPC)* en el próximo año. Justifique su respuesta.

14. Un procedimiento de control de calidad en producto terminado en una empresa de producción de consiste en revisar una muestra de 50 unidades por día y contar el número de defectos (*D*) resultantes. Para un día determinado se tienen los siguientes datos:

Número de defectos (D)	0	1	2	3
Número de unidades	12	15	17	6

a. Calcule el número promedio de defectos (D) en la muestra indicada.
b. Si el costo de producción (C) de una unidad está relacionado con el número de unidades defectuosas producidos (D) mediante la expresión $C = \$200 + 4.5 * D$ ¿Cuál es el costo promedio de la producción de una muestra de 50 unidades?
c. ¿Qué porcentaje de la muestra no presento defectos?

15. ¿Cómo el precio del dólar a ha afectado el comportamiento de las Importaciones Colombianas? Para tener claridad sobre esta problemática realice un análisis de descriptivo con la siguiente información: **Precio promedio del dólar** por año y Las **Importaciones**, en millones de dólares por año en Colombia.



Año	Precio dólar	Importaciones
2001	2299,77	1282,1
2002	2507,96	1269,5
2003	2877,5	1388,2
2004	2626,22	1676,4
2005	2320,77	2120,4
2006	2357,98	2616,2
2007	2078,35	3289,7
2008	1966,26	3966,9
2009	2156,29	3289,8
2010	1897,89	4068,3
2011	1848,17	5467,5
2012	1798,23	5415,2