Apuntes de Estadísticas

Prof. David Becerra Rojas

ESTADÍSTICA

INTRODUCCIÓN

La palabra "Estadística" ha sido frecuentemente utilizada para referirse a la información cuantitativa (o numérica). También se ha utilizado para referirse a los métodos que tratan la información. Sin embargo, cabe hacer notar que existe una diferencia entre lo que se entiende por método estadístico y dato estadístico (o información). En resumen podríamos decir que:

La Estadística es un cuerpo de conceptos y métodos empleados para recolectar e interpretar datos referentes a un área de investigación particular y para extraer conclusiones en situaciones en que la incertidumbre y la variabilidad están presentes

Para ilustrar más esta situación podemos considerar por ejemplo:

Cuando un lector tiene pocos hechos numéricos, puede utilizar la información numérica en la máxima extensión sin perder mucho tiempo o pensar demasiado en analizar los hechos. Examinemos la información:

Juan tiene 22 años y María tiene 18.

Un lector puede fácilmente interpretar la información de muchas maneras diferentes. Por ejemplo, Juan es un hombre joven de 22 años de edad, pero es cuatro años mayor que María. Sin embargo, cuando un lector tiene un gran volumen de hechos numéricos, puede encontrar que la información le es de poco valor, puesto que no puede interpretarla toda al mismo tiempo. Note la siguiente información:

> *Juan tiene 22 años, María tiene 18 años, Jaime tiene 25 años, Pedro tiene 16 años, y así sucesivamente para 1,000 estudiantes seleccionados en Swan College en octubre 1, 1966.*

Un lector tendría ciertamente dificultad en interpretar inteligentemente la distribución de edades.

El gran volumen de información numérica origina la necesidad de métodos sistemáticos, los cuales puedan ser utilizados para organizar, presentar, analizar e interpretar la información efectivamente. De esta manera pueden extraerse conclusiones válidas y tomarse decisiones razonables mediante el uso de los métodos. Los métodos estadísticos son desarrollados primeramente para llenar esta necesidad.

DATOS ESTADÍSTICOS

Información cuantitativa o numérica puede encontrarse casi dondequiera: en negocios, economía y muchas otras áreas. Por ejemplo, el precio marcado de un sombrero es mostrado en un cierto *número* de dólares, la situación de empleo en una nación es expresada en un *número* de personas, la inscripción en una universidad es registrada mediante un *número* de estudiantes, la distancia recorrida por un agente de ventas es reportada en *número* de millas y la edad de una persona es representada por un *número* de años. Sin embargo, no toda la información cuantitativa es considerada como dato estadístico. La información cuantitativa apropiada para análisis estadístico debe ser un conjunto (o conjuntos) de números que muestren *relaciones significativas*. En otras palabras, los datos estadísticos son números que pueden ser comparados, analizados e interpretados. Un número aislado que no se compara o que no muestra relación significativa con otro número no es dato estadístico.

En el ejemplo de arriba, la edad de Juan a solas no constituye dato estadístico si no hay otra disponible para comparación. Sin embargo, las edades de 1,000 estudiantes son datos estadísticos, puesto que las edades pueden ser comparadas y analizadas, y los resultados del análisis pueden ser interpretados. También las llamadas "estadísticas" de un paciente tal como son medidas por un doctor no son datos estadísticos, puesto que cada medida, tal como la estatura, no muestra relación significativa con otras medidas, tal como el número de pulsaciones por minuto o la medida de la vista del paciente. Sin embargo, la información relativa a las estaturas de todos los pacientes dentro de un cierto período de tiempo si son datos establecidos, puesto que las estaturas pueden ser comparadas, analizadas e interpretadas de acuerdo con sus relaciones.

El área de la cual los datos estadísticos son recopilados es generalmente referida como la *población o universo*. Una población puede ser *finita o infinita*. Una población finita tiene un número limitado de individuos u objetos, mientras que una población infinita tiene un número ilimitado. Por ejemplo, una clase de inglés de 25 estudiantes es una población finita. El número de estudiantes universitarios en los Estados Unidos durante el pasado, presente y futuro, es ilimitado; por lo tanto, tales estudiantes forman una población infinita.

La tarea de recopilar un conjunto completo de datos de una población finita pequeña es relativamente simple. Si deseamos obtener las edades de 25 estudiantes en la clase de inglés, podemos simplemente preguntar a cada estudiante su edad; así tenemos un conjunto completo de datos. Sin embargo, recopilar tales datos de una población finita pero grande, es algunas veces imposible o impráctico. Recopilar un conjunto completo de datos concernientes a las edades de todos los estudiantes de las escuelas de los Estados Unidos en octubre 1, 1966, por ejemplo, puede ser impráctico, aunque es posible, debido al tiempo y costo consumidos. La recopilación de datos completos de una población infinita es definitivamente imposible.

A fin de evitar la tarea imposible o impráctica, usualmente se extrae una *muestra* de elementos representativos de la población. La muestra es, entonces, utilizada para el estudio estadístico y los resultados de la muestra son usados como las bases para describir, estimar o predecir las características de la población. Supongamos que los 1.000 estudiantes presentados

anteriormente son representativos de los estudiantes de Swan College y son seleccionados del total de 30.000 estudiantes en 1966 (población). El conjunto de datos recopilados concernientes a las edades de los 1,000 estudiantes es una muestra, y un investigador puede usar estos resultados para estimar o predecir las edades de todos los estudiantes en la universidad.

MÉTODOS ESTADÍSTICOS

De acuerdo con el orden de aplicaciones en un estudio estadístico, los métodos estadísticos son divididos en cinco pasos básicos 1) recopilación, 2) organización, 3) presentación, 4) análisis y 5) interpretación.

En el ejemplo de arriba, si el encargado de los estudiantes de Swan College desea conocer el grupo de edad típico de los estudiantes en la universidad, puede primero, recopilar datos estadísticos concernientes a las edades de un grupo representativo de estudiantes de la universidad digamos 1.000 estudiantes de la población de 30.000 estudiantes. (El tamaño apropiado de una muestra se expondrá más adelante). Segundo, puede organizar las edades recopiladas clasificándolas en diferentes grupos de edad. Tercero, puede presentar los datos organizados en forma tabular, tal como la que se muestra en la tabla 1.1. Cuarto, puede analizar las edades presentadas en la tabla para obtener la información deseada. Por ejemplo, él puede encontrar que el grupo de edad típica de los estudiantes en la universidad es el grupo de edad "18 y menos de 20" puesto que es el que contiene el mayor número de estudiantes, o sea 600 estudiantes como se muestra en la tabla. Quinto, el encargado puede interpretar los resultados de su análisis de la muestra señalando que las edades típicas de todos los estudiantes en la universidad son de 18 a menos de 20 años.

Estrictamente hablando, no hay línea de división definitiva que separe los cinco pasos básicos. Algunos de los métodos pueden ser usados en más de un paso. En el ejemplo de arriba, el método de clasificar los grupos de edad usados en el paso de organización está estrechamente relacionado con los métodos empleados en el paso de análisis. Realmente, las clasificaciones de los grupos de edad son determinadas por la intención del encargado al obtener el tipo de información de los datos en el análisis. Sin embargo, la división nos da un orden lógico para estudiar los métodos estadísticos.

Tabla 1.1.

Edades de 1.000 estudiantes seleccionados en la UTFSM.

Edades	Número de estudiantes hombres	Número de estudiantes mujeres	Total
Menos de 18	120	50	170
18 y menos de 20	500	100	600
20 y menos de 22	100	80	180
22 y más	30	20	50
TOTAL	**750**	**250**	**1.000**

Fuente: Datos hipotéticos.

1. ESTADÍSTICA DESCRIPTIVA

La Estadística interviene en la investigación a través de la Experimentación.

La investigación contempla una serie de pasos que están íntimamente relacionados con los pasos mencionados anteriormente. Como son:

1.	Formulación del Problema	• Precisar conceptos a utilizar • Formulación clara de preguntas • Limitaciones del problema, etc.
2.	Diseño del Experimento	• Obtención de un máximo de información minimizando costo y tiempo • Determinar tipo de muestreo y tamaño de la muestra.
3.	Desarrollo del Experimento	• Recolección de datos.
4.	Tabulación y Descripción de Resultados (Análisis)	• Construcción de Tablas y Gráficos.
5.	Inferencia Estadística	• Conclusiones a partir de la muestra acerca de la población bajo estudio.

2. APLICACIÓN DE LA ESTADÍSTICA

2.1. AREAS DE APLICACIÓN

La estadística, prácticamente se puede utilizar en todas las actividades del ser humano, donde se presenta con mayor incidencia es en: Economía, Agronomía, Pesquería, Informática, Prevención de Riesgos, Control del Medio Ambiente, Control de Alimentos, Química Analítica, Medicina, etc., en general en todas las áreas donde se necesite realizar una investigación.

2.1.1. Malos usos de la Estadística

La estadística es una herramienta científica. Su valor depende de cómo uno la utilice como herramienta. Sin embargo, la estadística es mal utilizada muy frecuentemente en muchos lugares. El Norfolk Virginian-Pilot imprimió el siguiente fragmento de diversión en la página principal en marzo 25, 1963: "Risas de hoy--- Si el hombre está parado con su pie derecho sobre un horno encendido y su pie izquierdo en un congelador, algunos estadísticos aseverarían que, en promedio, él está confortable".

Puesto que los estudiantes a este nivel no están aún familiarizados con los métodos estadísticos, el propósito de esta sección es solamente indicar los malos usos comunes de datos estadísticos, sin incluir el uso de métodos estadísticos complicados. Un estudiante debería estar alerta en relación con estos malos usos y debería hacer un gran esfuerzo para evitarlos a fin de ser un verdadero estadístico. Las fuentes de los ejemplos en la siguiente exposición no son indicadas, puesto que puede causar dificultades.

a. Datos estadísticos inadecuados

Los datos estadísticos son usados como la materia prima para un estudio estadístico. Cuando los datos son inadecuados, la conclusión extraída del estudio de los datos se vuelve obviamente inválida. Por ejemplo, supongamos que deseamos encontrar el ingreso familiar típico del año pasado en la ciudad Y de 50,000 familias y tenemos una muestra consistente del ingreso de solamente tres familias: $ 1 millón, $ 2 millones y no ingreso. Si sumamos el ingreso de las tres familias y dividimos el total por 3, obtenemos un promedio de $ 1 millón. Entonces, extraemos una conclusión basada en la muestra de que el ingreso familiar promedio durante el año pasado en la ciudad fue de $ 1 millón. Es obvio que la conclusión es falsa, puesto que las cifras son extremas y el tamaño de la muestra es demasiado pequeño; por lo tanto la muestra no es representativa. Hay muchas otras clases de datos inadecuados. Por ejemplo, algunos datos son respuestas inexactas de una encuesta, porque las preguntas usadas en la misma son vagas o engañosas, algunos datos son toscas estimaciones porque no hay disponibles datos exactos o es demasiado costosa su obtención, y algunos datos son irrelevantes en un problema dado, porque el estudio estadístico no está bien planeado.

b. Un sesgo del usuario

Sesgo significa que un usuario dé los datos perjudicialmente de más énfasis a los hechos, los cuales son empleados para mantener su predeterminada posición u opinión. Los estadísticos son frecuentemente degradados por lemas tales como "Hay tres clases de mentiras: mentiras, mentiras reprobables y estadística", y "Las cifras no mienten, pero los mentirosos piensas".

Hay dos clases de sesgos: conscientes e inconscientes. Ambos son comunes en el análisis estadístico. Hay numerosos ejemplos de sesgos conscientes. Un anunciante frecuentemente usa la estadística para probar que su producto es muy superior al producto de su competidor. Un político prefiere usar la estadística para sostener su punto de vista. Gerentes y líderes de trabajadores pueden simultáneamente situar sus respectivas cifras estadísticas sobre la misma tabla de trato para mostrar que sus rechazos o peticiones son justificadas.

Es casi imposible que un sesgo inconsciente esté completamente ausente en un trabajo estadístico. En lo que respecta al ser humano, es difícil obtener una actitud completamente objetiva al abordar un problema, aun cuando un científico debería tener una mente abierta. Un estadístico debería estar enterado del hecho de que su interpretación de los resultados del análisis estadístico está influenciado por su propia experiencia, conocimiento y antecedentes con relación al problema dado.

c. Supuestos falsos

Es muy frecuente que un análisis estadístico contemple supuestos. Un investigador debe ser muy cuidadoso en este hecho, para evitar que éstos sean falsos.

Los supuestos falsos pueden ser originados por:

- Quien usa los datos
- Quien está tratando de confundir (con intencionalidad)
- Ignorancia
- Descuido.

2.2. TÉRMINOS COMUNES UTILIZADOS EN ESTADÍSTICA

Variable: Consideraciones que una variable son una característica o fenómeno que puede tomar distintos valores.

Dato Mediciones o cualidades que han sido recopiladas como resultado de observaciones.

Población Se considera el área de la cual son extraídos los datos. Es decir, es el conjunto de elementos o individuos que poseen una característica común y medible acerca de lo cual se desea información. Es también llamado Universo.

Muestra Es un subconjunto de la población, seleccionado de acuerdo a una regla o algún plan de muestreo.

Censo Recopilación de todos los datos (de interés para la investigación) de la población.

Estadística Es una función o fórmula que depende de los datos de la muestra (es variable).

Parámetro Característica medible de la población.

Ejemplo La universidad está interesada en determinar el ingreso de las familias de sus alumnos.
Variable: Ingreso percápita de las familias.
Dato: Ingreso percápita de la familia de un alumno específico.
Población: Las familias de todos los alumnos de la universidad.
Estadística: Ingreso percápita promedio de las familias seleccionadas en la muestra.
Parámetro: Ingreso percápita promedio de la población.

2.3. MUESTREO

Una muestra es representativa en la medida que es imagen de la población.

En general, podemos decir que el tamaño de una muestra dependerá principalmente de:

- Nivel de precisión deseado.
- Recursos disponibles.
- Tiempo involucrado en la investigación.

Además el plan de muestreo debe considerar

- La población
- Parámetros a medir.

Existe una gran cantidad de tipos de muestreo. En la práctica los más utilizados son los siguientes:

MUESTREO ALEATORIO SIMPLE:

"Es un método de selección de n unidades extraídas de N, de tal manera que cada una de las posibles muestras tiene la misma probabilidad de ser escogida".

(En la práctica, se enumeran las unidades de 1 a N, y a continuación se seleccionan n números aleatorios entre 1 y N, ya sea de tablas o de alguna urna con fichas numeradas).

Ejemplo:

Considere la producción de TV de una Compañía en un determinado turno, la cual es de N = 35 televisores. Para efectos de Control de Calidad de una de sus partes, supongamos la pantalla, se desea extraer una muestra aleatoria simple de tamaño n = 5. Si los 35 TV producidos son numerados del 1 al 35, una posible muestra podría ser 3, 5, 18, 23, 30.

¿Cuántas muestras posibles hay?

MUESTREO ESTRATIFICADO ALEATORIO:

Se usa cuando la población está agrupada en pocos estratos, cada uno de ellos son muchas entidades. Este muestreo consiste en sacar una muestra aleatoria simple de cada uno de los estratos. (Generalmente, de tamaño proporcional al estrato).

MUESTREO SISTEMÁTICO:

Se utiliza cuando las unidades de la población están de alguna manera totalmente ordenadas.

Para seleccionar una muestra de n unidades, se divide la población en "n" subpoblaciones de tamaño K = N/n y se toma al azar una unidad de la K primeras y de ahí en adelante cada K-ésima unidad, es decir, siendo n_o la primera unidad seleccionada de la sub-población (1, 2,...K).

$\{n_o, n_o + K \;\; n_o + 2K, \;\; .... \;, n_o + (n-1)\, K \}$

MUESTREO POR CONGLOMERADO

Se emplea cuando la población está dividida en grupos o conglomerados pequeños. Consiste en obtener una muestra aleatoria simple de conglomerados y luego CENSAR cada uno de éstos.

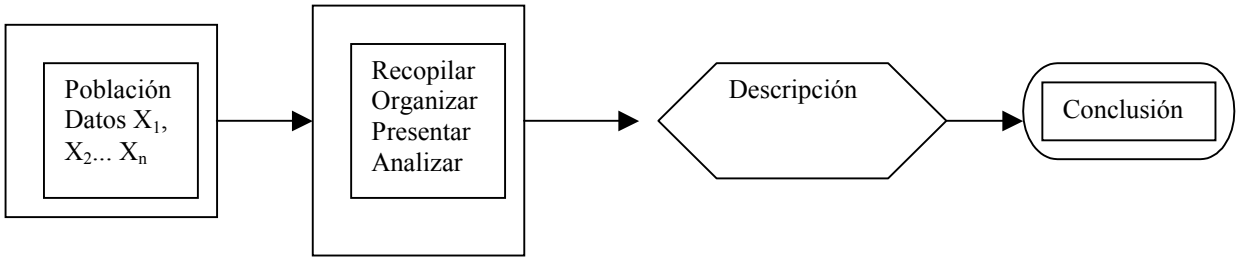
MUESTREO EN DOS ETAPAS (Bietápico)

En este caso la muestra se toma en dos pasos:

- Seleccionar una muestra de unidades primarias, y
- Seleccionar una muestra de elementos a partir de cada unidad primaria escogida.

Observación:
En la realidad es posible encontrarse con situaciones en las cuales no es posible aplicar libremente un tipo de muestreo, incluso estaremos obligados a mezclarlas en ocasiones.

En general la Estadística está encargada de llevar a cabo el siguiente esquema:



2.4. VARIABLES

2.4.1. Tipos de variables

Las variables se pueden clasificar en dos grandes grupos.

a) Variables categóricas:

Son aquellas que pueden ser representadas a través de símbolos, letras, palabras, etc. Los valores que toman se denominan categorías, y los elementos que pertenecen a estas categorías, se consideran idénticos respecto a la característica que se está midiendo. Ejemplo:
Variable: Profesión:

Valores que pueden tomar la variable:

- Programador
- Técnico en Control de Alimentos
- Técnico en Prevención de Riesgos
- Técnico en Control del Medio Ambiente
- Químico Analítico
- Técnico Mecánico
- Etc.

Las variables categóricas de dividen en dos tipos: Ordinal y Nominal.

Las Ordinales, son aquellas en que las categorías tienen un orden implícito. Admiten grados de calidad, es decir, existe una relación total entre las categorías.

Ejemplo:

Variable: Nivel de estudio de Enseñanza Básica.

Valores que toma la variable:

- Primero Básico
- Segundo Básico
- Tercero Básico
- .....
- Octavo Básico

A pesar de que esta variable admite grados de calidad, no es posible cuantificar la diferencia.

Las nominales, son aquellas donde no existe una relación de orden.

b. Variables numéricas

Son aquellas que pueden tomar valores numéricos exclusivamente (medicines). Se dividen en dos tipos. Discretas y continuas.

Discretas: son aquellas que toman sus valores en un conjunto finito o infinito numerable.

Ejemplo:
- Variable: Número de sillas por sala.
- Valores que toma la variable: 0, 1, 2, 3, ...n.

Continuas: Son aquellas que toman sus valores en un subconjunto de los números reales, es decir en un intervalo.

Ejemplo:
- Variable: Temperatura de Valparaíso en verano.
- Valores que toma la variable: entre 5 grados y 30 grados. (5º , 30º).

Observación:
En general para las variables continuas el hombre ha debido inventar una medida para poder establecer una medición de ellas:

Ejemplo: El metro, la hora.

2.5. ORGANIZACIÓN DE DATOS

Supongamos que para estudiar una variable se han definido K clases C_1, C_2, ..., C_k.

Observación: C_i puede ser un número, un intervalo o una categoría.

Algunos conceptos de interés:

Frecuencia Absoluta: (n_i): "se llama frecuencia absoluta de la clase C_i, al número de entidades que pertenecen a la clase C_i": Si el tamaño de la muestra es n entonces se cumple

$$\sum_{i=1}^{k} n_i = n$$

Frecuencia Relativa (f_i): "Se llama frecuencia relativa de la Clase C_i a la proporción de entidades, respecto al total de entidades de la muestra que pertenecen a la clase C_i". Es decir, $f_i = n_i/n$ y

$$\sum_{i=1}^{k} f_i = 1$$

Observación:

"Las dos definiciones dadas tienen sentido cuando se trabaja en cualquier escala de medida".

Los conceptos siguientes pueden definirse sólo si las clases C_i pueden ordenarse.

Frecuencia Absoluta Acumulada (N_i).

$$N_i = \sum_{i=1}^{j} n_i \quad (j = 1...k) \quad Note \quad que \quad N_1 = n_1 y N_k = n$$

Frecuencia Relativa Acumulada (F_i)

$$F_{\ i} = \sum_{i=1}^{j} f_i \quad (j = 1..k) \quad Note \quad que \quad F_k = 1$$

PRESENTACIÓN DE LA INFORMACIÓN

Tablas de Frecuencias (Modelo General)

CLASE	FRECUENCIA ABSOLUTA	FRECUENCIA RELATIVA	FRECUENCIA ABSOLUTA-ACUM.	FRECUENCIA RELATIVA ACUM.
C_1	N_1	f_1	N_1	F_1
C_2	N_2	f_2	N_2	F_2
.				
.				
C_k	N_k	f_k	$N_k = n$	$F_k = 1$
TOTAL	N	1		

OBSERVACIONES:

1. Si la variable es nominal las últimas 2 columnas carecen de sentido.

2. En el caso de variables continuas o variable de tipo discreta, cada clase C_1 puede ser representada por un valor numérico X_1 llamado MARCA DE CLASE.

3. Cada tabla de frecuencia, debe contar con un nombre en el cual se especifique la información que contiene.

Ejemplo:

Sea la variable X, definida como: Marca de bebidas gaseosas preferidas por los alumnos de la Universidad. Se consideró una muestra de tamaño n = 20, obteniéndose los siguientes resultados: Sprite 4, Fanta 5, Coca-cola 8, Bilz 0, Pepsi 3.

Luego tenemos que:

- Variable X: Marca de bebidas gaseosas preferidas por los alumnos.
- Tipo de Variable: Categórica-nominal.

La tabla de frecuencias será:

Tabla 2.1.

Bebida gaseosa preferida por una muestra de 20 alumnos de la Universidad

I	MARCA	ALUMNOS	f_i
1	Sprite	4	0.20
2	Fanta	5	0.25
3	Coca Cola	8	0.40
4	Bilz	0	0.00
5	Pepsi	3	0.15
TOTAL		**20**	**1.00**

Supongamos que la calificación de los consumidores para un nuevo producto en el mercado fue la siguiente, considerada una muestra de tamaño 40 personas: muy bueno, 8 personas, bueno 15, regular 10, malo 4 y muy malo 3.

Considerando X: calificación de los consumidores.... como una variable, categórica ordinal. La tabla de frecuencia nos queda:

Tabla 2.2.

Calificación de un nuevo producto por una muestra de 40 personas

I	CALIFICACIÓN	CONSUMIDORES	f_i	N_i	F_i
1	Muy Bueno	8	0.200	8	0.200
2	Bueno	15	0.375	23	0.575
3	Regular	10	0.250	33	0.825
4	Malo	4	0.100	37	0.925
5	Muy malo	3	0.075	40	1.000
	Total	**40**	**1.000**		

Observación:

Debido a que la variable es del tipo categórica nominal, las dos últimas columnas son presentadas y tienen sentido, no así en el ejemplo anterior.

Si en alguna ocasión la suma de las frecuencias relativas no es 1, deberá ser solamente por la aproximación de los decimales de alguna de las f_i.

Los datos siguientes representan las respuestas de 30 trabajadores de una empresa, a la consulta de: ¿Cuántos hijos tienen?

3	5	4	2	0	5	2	2	3	2
1	1	0	1	3	2	1	4	7	2
0	2	3	3	4	2	1	3	2	1

En este caso tenemos que la variable X está dada por: Número de hijos por trabajador, y es de tipo numérica discreta.

La tabla de frecuencia en este caso quedará de la siguiente manera:

Tabla 2.3.

Número de hijos por trabajador

I	CALIFICACIÓN	CONSUMIDORES	f_i	N_i	F_i
1	0	3	0.10	3	0.10
2	1	6	0.20	9	0.30
3	2	9	0.30	18	0.60
4	3	6	0.10	24	0.80
5	4	3	0.067	27	0.90
6	5	2	0.000	29	0.967
7	6	0	0.033	29	0.967
8	7	1		30	1.000
	Total	**40**	**1.000**		

Observemos que el número de valores posibles que toma la variable es de k = 8.

Supongamos que los datos siguientes, representan el número de artículos defectuosos producidos diariamente en un período de 28 días.

35	40	58	18	64	36	37
22	28	53	45	34	28	42
36	40	27	25	26	30	35
34	52	61	43	37	44	52

La variable X está definida como: Número de artículos defectuosos producidos...., y es de tipo numérica discreta.

Podemos apreciar que en este caso los valores posibles que toma la variable van de: 18 artículos a 64 artículos, es decir, existen un total de 47 valores posibles para la variable, (es decir k = 47).

En este caso podemos apreciar que una tabla como la del ejemplo 3 es impracticable puesto que la cantidad de clases, es excesiva. En estos casos (cuando $k \geq 15$), se procede agrupando los valores posibles de la variable, formando así los llamados **Intervalos de clase**.

Existen varias formas para determinar estos intervalos, puesto que es recomendable que éstos queden todos con igual amplitud. El procedimiento que propondré consta de 4 pasos.

Observación: Este procedimiento se utiliza para variables que son del tipo numérica; discretas y continuas.

DETERMINACIÓN DE INTERVALOS DE CLASES

Paso 1 Determinar el número de clases o intervalos de clases, es decir K. Cuando el valor de k no está determinado previamente una sugerencia para k está dada por la Regla de Sturgs:

$$K = 1 + 3.3 \lg (n)$$

Donde:

Lg : logaritmo decimal
N : tamaño de la muestra
K : cantidad de intervalos ($k \in N$)

Paso 2 Determinar el Rango que está dado por la diferencia entre el valor máximo (m) y el valor mínimo (m) que toma la variable en la muestra, es decir;

$$R = \left| M - m \right| + 1_u$$

1_u = indica una unidad de medida, si los datos son enteros $1_u = 1$, si los datos vienen dados por un decimal entonces $1_u = 0{,}1$, datos con dos decimales $1_u = 0.01$, etc.

Paso 3 Determinar la amplitud de los intervalos, es decir C_i, i = i, 1, 2...k, como sigue:

$$C = R / k$$

Donde
El valor de C estará dado por la unidad de medida, si no es exacto, siempre se aproxima el valor superior.

Paso 4 Determinar unidades auxiliares, como generalmente el valor de C es aproximado hacia "arriba", para que todos los intervalos de clase tengan la misma amplitud, es necesario agregar una cantidad de p unidades determinadas por:

$$P = (C * k) – R$$

Paso 5 Determinar los límites de los intervalos.

Sean L_{Ij}: "Límite inferior de la clase j"
L_{sj}: "Límite superior de la clase j"

Se definen dos tipos de intervalos: Los aparentes y los reales.

1. **Los intervalos aparentes**: utilizados principalmente para variables del tipo numérica discretas se determinan de la siguiente manera:

$L_{I1} = m : L_{s1} = L_{I1} + (C - 1_\mu)$.

$L_{I2} = L_{I1} + C, \; L_{s2} = L_{I2} + (C - 1_\mu)$.

$L_{IK} = L_{I(k-1)} + C \quad L_{sk} = L_{ik} + (C - 1_\mu)$.

Siendo $L_{sk} = M + p$

2. **Los intervalos reales**: utilizados para variables numéricos discretos se obtienen de la siguiente manera:

$L_{I1} = m - (1/2)_\mu : L_{s1} = L_{I1} + C$

$L_{I2} = L_{I1}$, $\quad L_{s2} = L_{I2} + C$

$L_{IK} = L_{s(k-1),} \quad L_{sk} = L_{ik} + C$

Siendo $L_{sk} = (M + p + 1_u/2)$

Paso 6 Se determina un representante de los intervalos de clase, el cuál recibe el nombre de **Marca de Clase**, y está dado por el punto medio de cada intervalo, es decir:

$$X_i = \frac{Límite \;\; \text{int}erior \;\; + \;\; Límite \;\; \text{sup}erior}{2}$$

Luego la tabla de frecuencias se forma de la siguiente manera:

I	**TOTAL**	$\mathbf{n_i}$	$\mathbf{f_i}$	$\mathbf{N_i}$	$\mathbf{F_i}$	$\mathbf{X_2}$
1	C_1	n_1	f_1	N_1	F_1	X_2
2	C_2	n_2	f_2	N_2	F_2	X_2
.						
.						
K	C_k	n_k	f_k	$N_k = n$	$F_k = 1$	X_k
		n	1			

En el caso del ejemplo anterior la tabla se formará de la siguiente manera:

Paso 1 $K = 1 + 3.3 \lg (28) = 5.8 \approx 6$

Paso 2 $R = \left| 64 - 18 \right| + 1$

Paso 3 $C = (47) \div 6 = 7.8 \approx 8$

Observación: Si el valor de $c = 7.1 \approx 8$

Paso 4 $P = (8 \cdot 6) - (46 + 1) = 1 \Rightarrow L_{sk} = M + 1 = 6.5$

Paso 5 Determinación de los intervalos

Paso 6 Determinar Marcas de Clase

Tabla de Frecuencia

X_A	X_R	n_i	f_i	N_i	F_i	X_i
18-25	17.5-25.5	3	0.11	3	0,.11	21.5
26-33	25.5-33.5	5	0.18	8	0.29	29.5
34-41	33.5-41.5	10	0.3	18	0.64	37.5
42-49	41.5-49.5	4	0.14	22	0.78	45.5
50-57	49.5-57.5	3	0.11	25	0.89	53.5
58-65	57.5-65.5	3	0.11	28	1.00	61.5
Total		28	1.0			

X_A = Intervalos aparentes.

X_R = Intervalos reales.

En general no es necesario que una tabla de frecuencias incluya ambos intervalos.

Representación gráfica

Actualmente, se reúne con mucha frecuencia al lenguaje visual, se puede apreciar esto en la prensa, televisión, computación, etc., esto deja de manifiesto lo importante que es el comunicarse de esta forma con otras personas. Su importancia radica principalmente en el hecho que esta comunicación es masiva, es decir, muchas personas pueden acceder a la información a través de este medio. La Estadística utiliza también esta técnica de comunicación de tal manera de poder transmitir información a un gran número de personas, las cuales no necesitan tener conocimiento de Estadística.

A continuación se presentan algunos tipos de gráficos utilizados con más frecuencia.

Tabla A.

Notas de los 30 alumnos de un curso.

Notas	n_i	f_i	N_i	F_i
1	2	0.067	2	0.067
2	5	0.167	7	0.234
3	6	0.200	13	0.434
4	7	0.233	20	0.667
5	5	0.167	25	0.834
6	3	0.100	28	0.934
7	2	0.067	30	1.001
Total	**30**	**1.001 ≈ 1.0**		

Tabla B

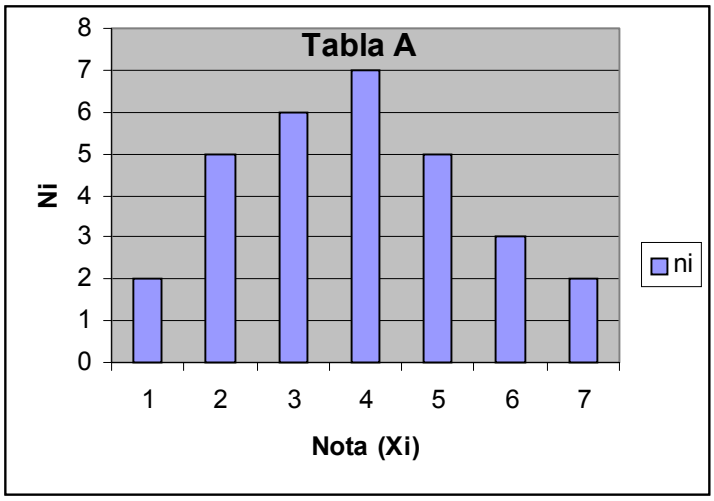
Distancia en metros recorridos por 50 ejemplares de una variedad de caracoles en un día

CLASE	**X**	n_i	f_i	N_i	F_i	X_i
1	7.1 – 7.7	2	0.04	2	0.04	7.4
2	7.7 - 8.3	2	0.04	4	0.08	8.0
3	8.3 – 8.9	6	0.12	10	0.20	8.6
4	8.9 – 9.5	14	0.28	24	0.48	9.2
5	9.5 – 10.1	12	0.24	36	0.72	9.8
6	10.1 - 10.7	10	0.20	46	0.98	10.4
7	10.7 - 11.3	4	0.08	50	1.00	11.0
TOTAL		**50**	**1.00**			

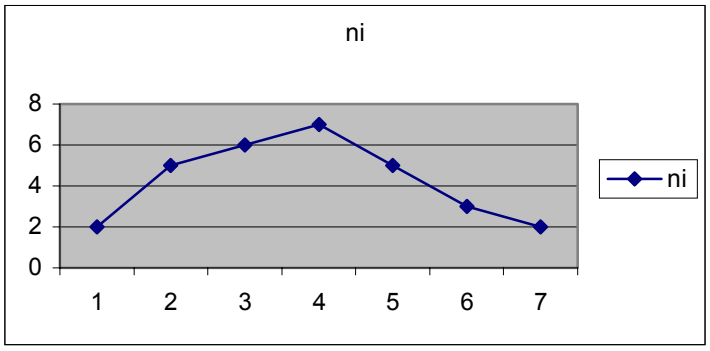
REPRESENTACIÓN GRÁFICA

Generalmente se usa la representación cartesiana en el plano, con un par de ejes coordenadas para representar el par (punto, m frecuencia) el cual puede estar acompañado por barras o unidos por una poligonal para destacar más las características de la distribución.

NOTAS DE ALUMNOS (Tabla A)



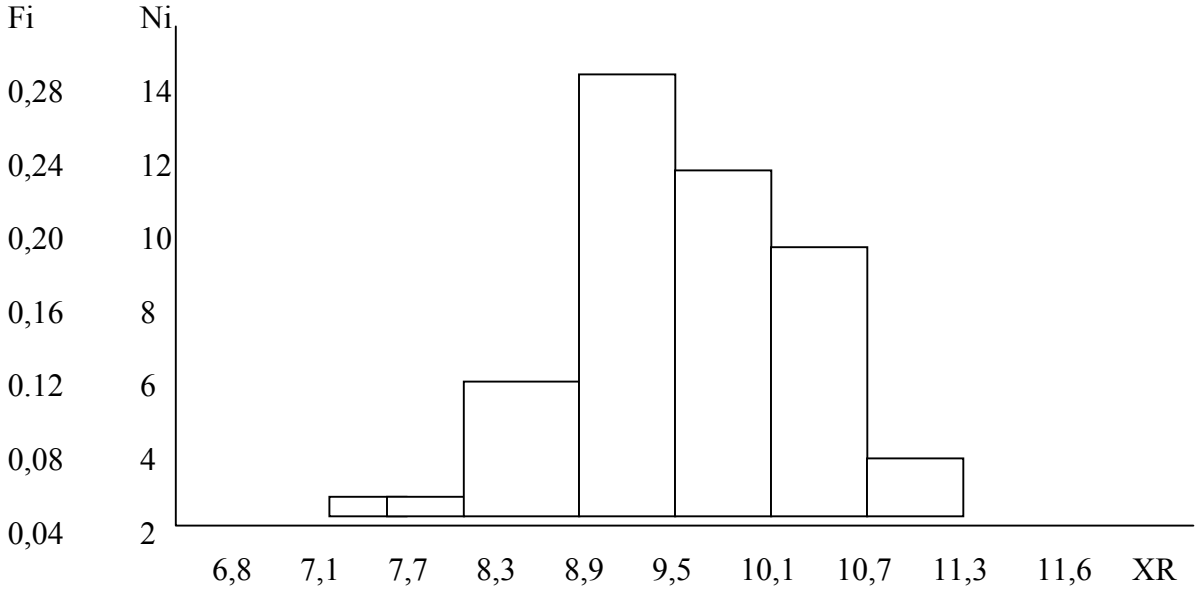
Notas de Alumnos (Tabla A)



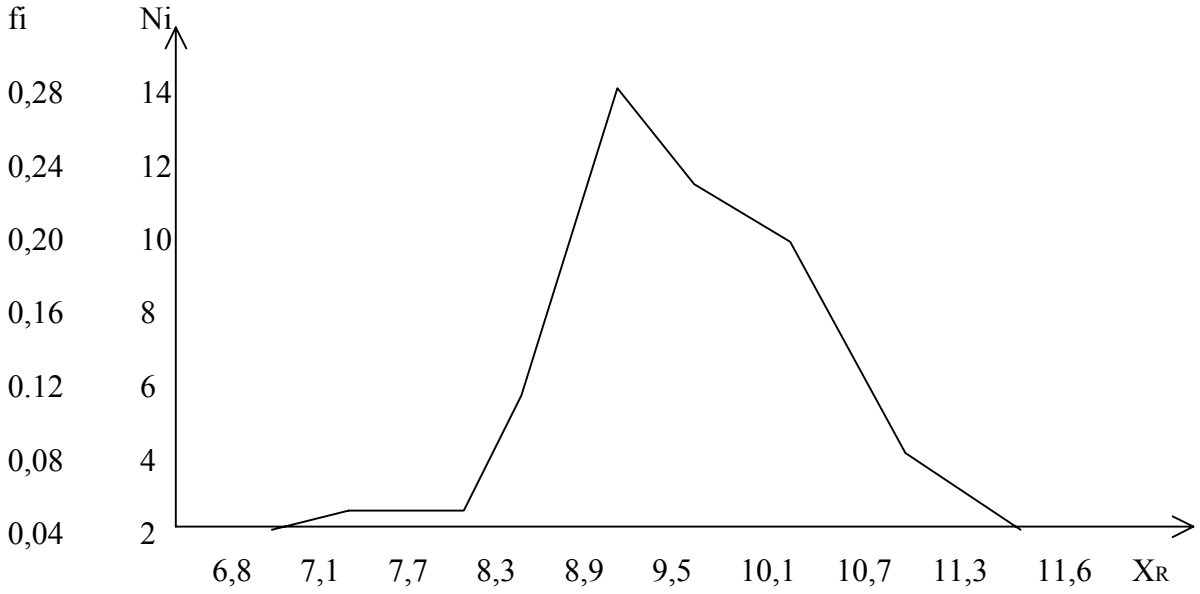
En el caso de tener intervalos el ancho de la barra se toma de la longitud de clase y la marca de clase para dibujar la poligonal.

TABLA B

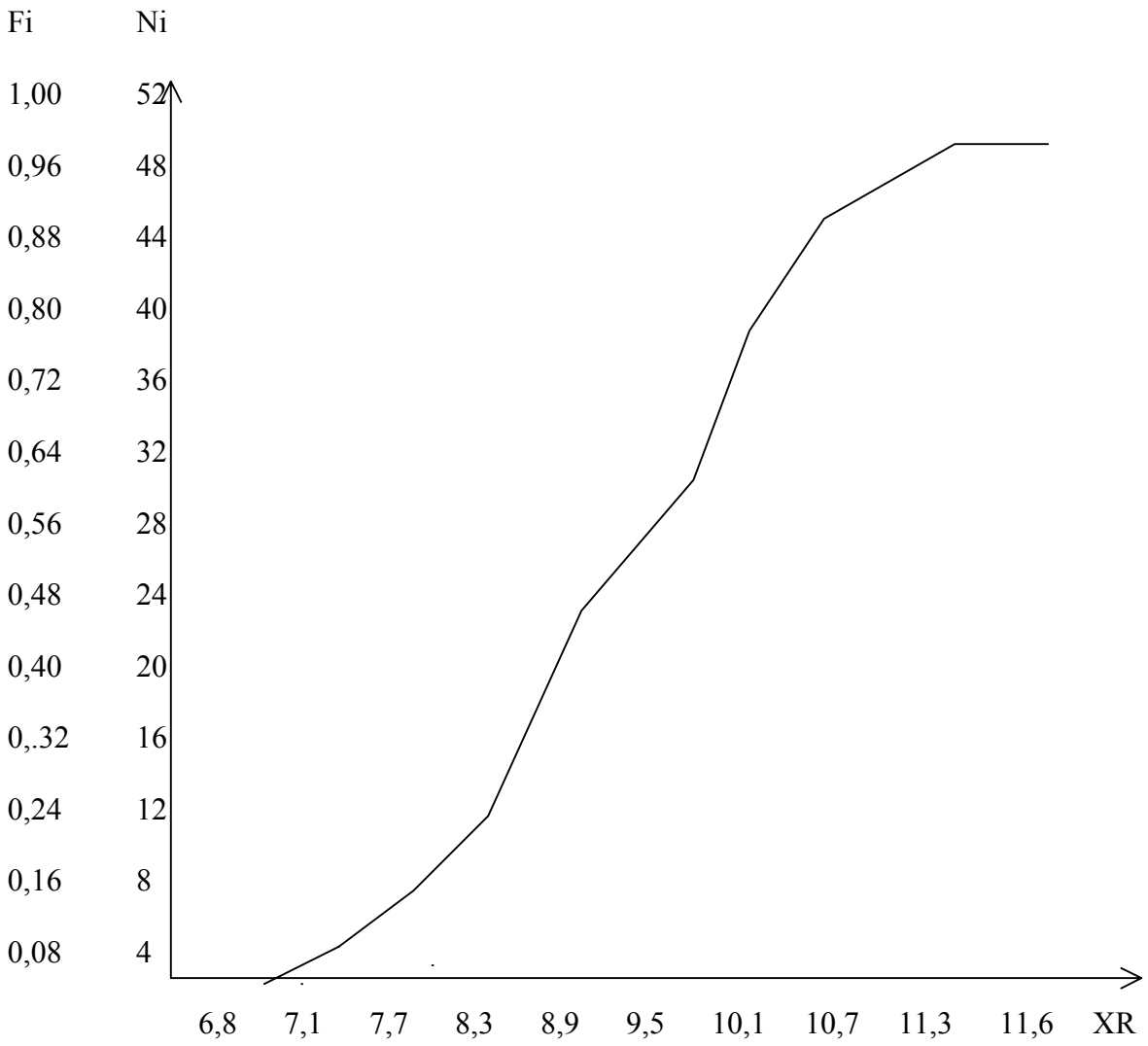
Histograma



Poligono de Frecuencia



Frecuencias acumuladas y **Ojiva**



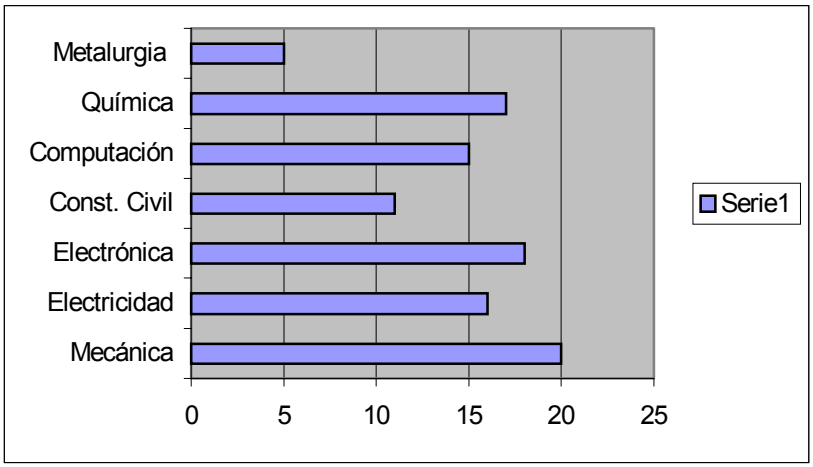
Otros ejemplos de representaciones gráficas son:

1 El número de estudiantes en el primer curso de estadística de la UTFSM por áreas principales sexo desde 1965 a1972, es:

2

AREA PRINCIPAL	HOMBRES	MUJERES	TOTAL DE ESTUDIANTES
Mecánica	12	8	20
Electricidad	6	10	16
Electrónica	15	2	17
Construcción Civil	7	3	10
Computación	8	6	14
Química	10	5	15
Metalurgia	2	2	4
TOTAL	**60**	**36**	**96**

Represente esta tabla según el gráfico de barras horizontales:



2. El número de aparatos de televisión producidos por la I.R.T. durante los años 1970, 1971, 1972 y 1973 ha sido el siguiente:

AÑOS	NÚMERO DE APARATOS
1970	4.500
1971	6.000
1972	8.500
1973	6.500

1970 

1971 

1972 

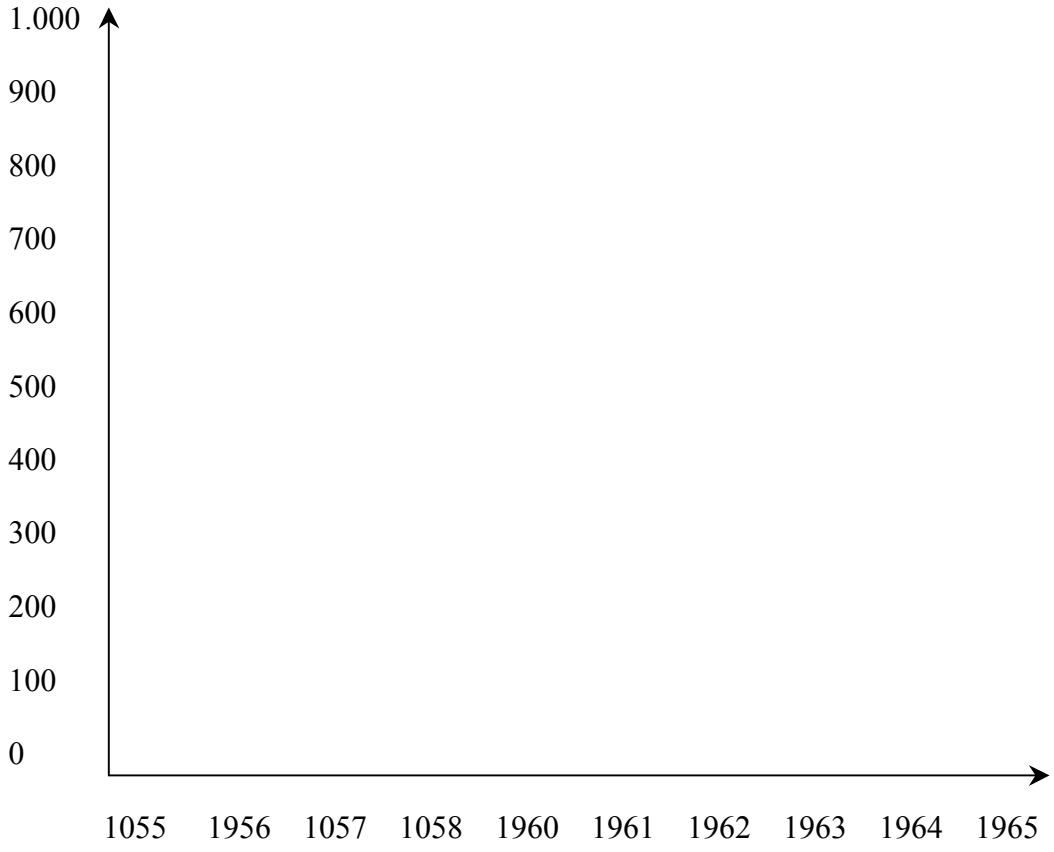
1973 

 = 1.000 unidades.

Pictograma

3. Datos de producción anual Loren Manufacturing Company por plantas de 1955 a 1965, son:

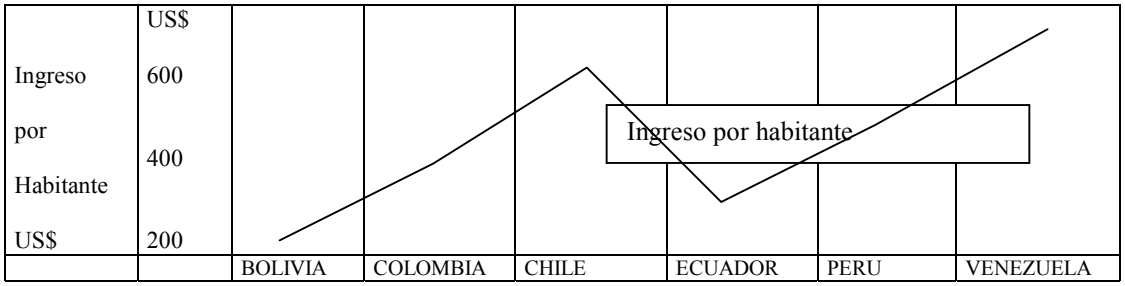
AÑOS	(MILES DE UNIDADES				
	PLANTA A	PLANTA B	SUBTOTAL	PLANTA C	TOTAL
1955	150	190	340	160	500
1956	170	230	400	170	570
1957	200	150	350	200	550
1958	240	210	450	150	600
1959	200	280	480	220	700
1960	250	300	550	100	650
1961	270	230	500	200	700
1962	300	220	520	260	780
1963	280	320	600	200	800
1964	350	280	630	270	900
1965	400	250	650	150	800

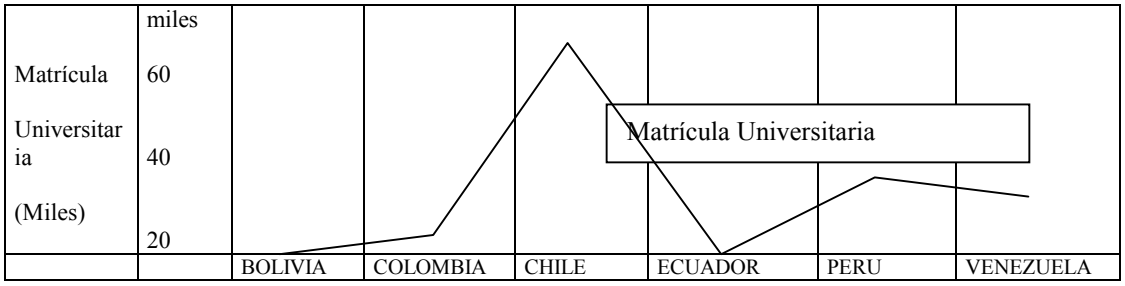


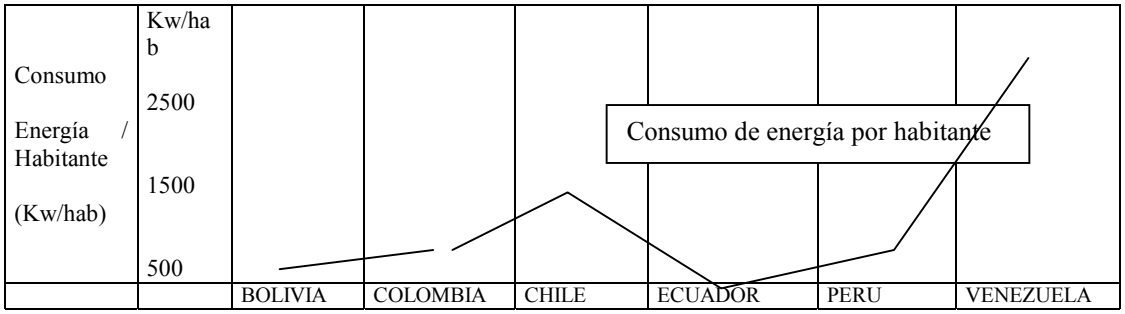
Construya el gráfico

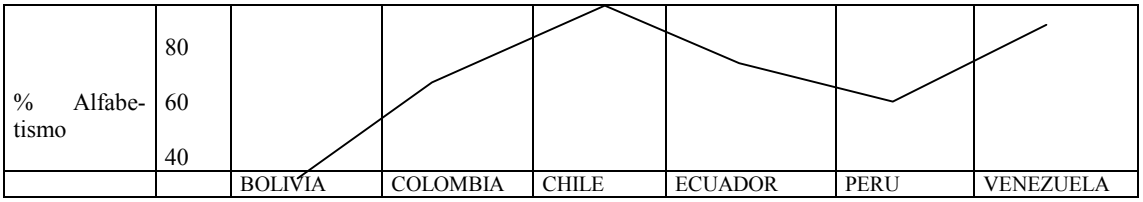
4. Algunos indicadores estadísticos de algunos países del pacto Andino son (CEPAL, 1972).

	BOLIVIA	**COLOMBIA**	**CHILE**	**ECUADOR**	**PERU**	**VENEZUELA**
Ingreso por habitante (US$)	200	359	645	300	420	627
Matrícula Universitaria. (miles)	12	20	70	10	40	34
Consumo energía/habitante (Kw/hab)	183	506	1020	102	537	2620
% Alfabetismo	39,8	72,0	88,8	72,0	67,0	85,0





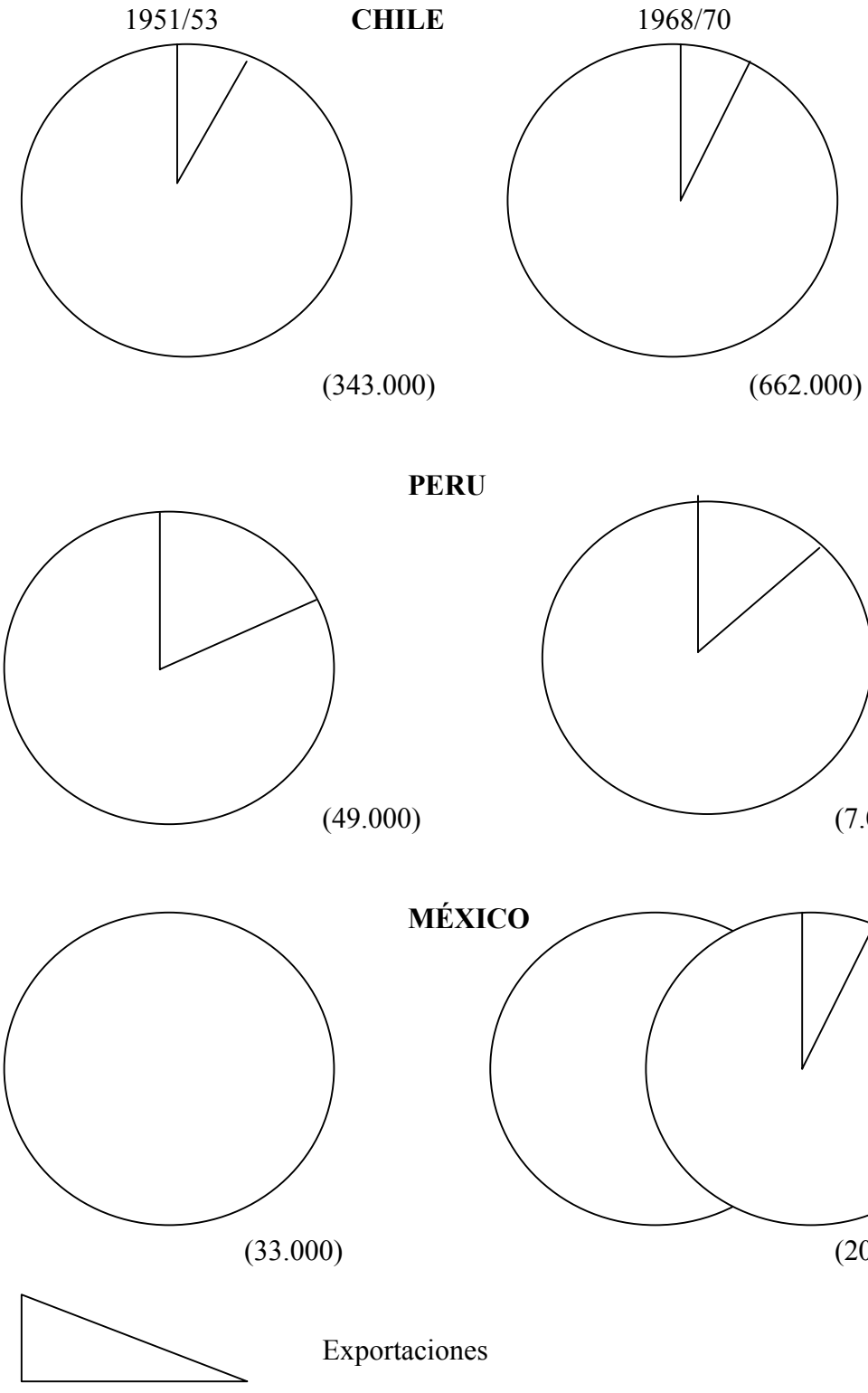




5. La producción de cobre de Chile, Perú y México y las exportaciones de estos países durante los períodos 1951/53 y 1968/70 fueron los siguientes:

	PRODUCCIÓN EN MINAS (miles de toneladas cortas)	
	1951/53	1968/70
Chile	384	675
Perú	62	63
México	33	204
Mundial	2.721	5.816
Porcentaje del total mundial		
Chile	14.1	11.6
Perú	2.3	1.1
México	1.2	3.5
Exportaciones		
Chile	343	662
Perú	49	7
México	33	207

Gráfico de Sectores o Círculos



Achure donde corresponda

4. MEDIDAS DE TENDENCIA CENTRAL Y DE DISPERSIÓN

4.1. La idea es resumir los datos en un solo valor, un valor que represente a todo un conjunto de datos, este tiene que ser un número hacia el cual tienen tendencia a concentrarse los datos, o sea, que es un valor central o de posición central a cuyo alrededor se distribuyen todos los datos del conjunto. Los más comunes son: la mediana, moda, media o promedio, media geométrica, etc.

Estas y otras medidas nos sirven para resumir la información presentada en cuadros y poder relacionar y comparar entre sí, de una manera sencilla, un conjunto de distribuciones de frecuencias.

Una vez determinadas las medidas de tendencia central de una distribución nos interesa determinar cómo se reparten (dispersan, desvían) los datos a uno y otro lado de la medida central. O sea, es necesario cuantificar la representatividad de la medida de tendencia para poder caracterizar la distribución. Si la dispersión es pequeña indica gran uniformidad y la información tiende a concentrarse en torno a la medida central, por el contrario una gran dispersión indica que los datos están alejados de ella.

Las salidas de dispersión más usuales son: desviación media, desviación típica o estándar, rango, etc.

4.2. Hay que hacer notar que toda variable puede clasificarse en uno de los niveles de medición que se darán en orden creciente en cuanto a la riqueza de la información y de acuerdo a ese nivel de calidad se darán sus medidas de tendencia central y de dispersión.

a. Variable nominal:

La variable induce en la población una subdivisión y la información se puede clasificar en clases, donde cada clase está completamente definida y diferenciada de las demás.

La recopilación se reduce a contar el número de individuos de la muestra que pertenecen a cada clase.

Ejemplo
Variable = color de ojos
Clases = negro, café, verde, azul, etc.

b. Variable Ordinal

La variable admite grados de calidad u ordenamiento, esto significa que existe una relación de orden entre las clases.

Ejemplo:

Variable = rendimiento académico
Clases = 7, 6, 5, 4, 3...

c. Variables numéricas de rango grande

La información obtenida en este caso es de tipo cuantitativo o numérico y es posible agruparla en intervalos.

Ejemplo:

Variable = estatura
Clases = [1,20 ; 1,59]
[1,50 ; 1,80]
[1,80 ; 2,10]

4.3. La medida de tendencia central que se utiliza en el nivel nominal es la moda o clase modal (se anota o por Mo).

4.3.1.a. Modal:

Definición: La clase modal es aquella clase cuya frecuencia es mayor que la frecuencia de todas las demás clases (O sea $f_M > f_i$ para todas las clases).

Hay variables que pueden ser unimodales, bimodales, trimodales, etc.

b. Una medida de dispersión de la clase modal es la TASA o RAZÓN DE VARIACIÓN, ella nos entrega la proporción que NO está contenida en la clase modal.

Ella se define como $V = 1 - f_M$

La moda es altamente significativa si $V \simeq 0$ y no es significativa si $V \simeq 1$.

Ejemplo:
En una muestra de 50 fumadores clasificándolos según sus preferencias se obtuvo:

Clase-Marca	η	f
C_1 – H	6	0,12
C_2 – B	30	0,60
C_3 - V	10	0,20
C_4 L	2	0,04
C_5 - W	2	0,04

Clase modal : C_2 ; $f_M = \frac{30}{50} = 0,60$

$V = 1 - 0,6 = 0,4$

La clase modal es representativa.

4.3.2. En el nivel ordinal se definen los fractiles o cuantiles (ellos dividen o fraccionan la muestra en partes más o menos iguales) destacándose los cuartiles (dividen en 4 partes),

4.3.2.1. Cuartil

La clase cuartil de orden k es la primera clase cuya frecuencia relativa acumulada es mayor o igual a k/4 (o sea, la clase que tiene $F_k \geq k/4$).

4.3.2.2. Decil

La clase decil de orden k es la primera clase cuya frecuencia relativa acumulada es mayor o igual a k/10.

4.3.2.3. Percentil

La clase percentil de orden k es la primera clase cuya frecuencia relativa acumulada es mayor o igual a k/100.

4.3.2.4. Mediana

a. La medida tendencia central característica del nivel ordinal es la MEDIANA (se abrevia M_d).

La clase mediana es la primera clase cuya frecuencia relativa acumulada es mayor o igual a ½ (es el cuartil de orden 2 o el decil de orden 5 o el percentil de orden 50).

Observación: La mediana divide a la muestra en dos mitades aproximadamente.

b. Una medida de dispersión respecto de la mediana es:

$$D = \frac{rango\ \ clase\ \ tercer\ \ cuartil - rango\ \ clase\ \ primer\ \ cuartil}{número\ \ total\ \ de\ \ clases - 1}$$

donde el rango de la clase es su número de orden.

Esta medida de dispersión indica el grado de concentración en torno a la clase mediana. Si D ≅ 1 la muestra está muy disgregada.

Ejemplo 4.1.:

Una muestra de 50 estudiantes clasificados en cinco clases ordenadas según su rendimiento.

Clases	n	f	N	F
C_1 – (3)	10	0,20	10	0,20
C_2 – (4)	7	0,14	17	0,34
C_3 – (5)	14	0,28	31	0,62
C_4 - (6)	6	0,12	37	0,74
C_5 – (7)	13	0,26	50	1,00

Clase Modal: C_3

Tasa de variación: V = -0,28 = 0,72 baja representatividad.

Clase mediana: C_3 (0,62 > 0,5) (casualmente coincidió con la clase modal)

CLASES	C_1	C_2	C_3	C_4	C_5
RANGO	1	2	3	4	5

CLASE CUARTIL ORDEN 1 C_1 (9,34 > 0,25)

CLASE CUARTIL ORDEN 3 C_5 (1 > 0,75)

$D = \frac{5-2}{5-1} = 0,75$ *(indica alta dispersión en torno a la mediana)*

4.3.3. Media

4.3.3.1 La medida de tendencia central más utilizada en el nivel intervalar es la MEDIA o promedio (se designa por X)

Donde $X = \frac{1}{n}\sum_{1=1}^{i=k} n_i M_{ci} = \sum_{i=1}^{i=k} f_i M_{ci}$

en que:

$\bar{X}$ = media
n = número total de datos
n_i = frecuencia absoluta de la clase i
f_i = frecuencia relativa de la clase i
k = número de clases
M_{ci} = marca de clase de la clase i

Nota: para este nivel se definen la moda y la mediana como:

4.3.3.2. Moda

$$Mo = L + C \cdot \left[\frac{d_1}{d_1 + d_2} \right]$$

donde:

L = límite inferior real de la clase modal
I = amplitud del intervalo
d_1 = $n_M - n_{M-1}$
n_M = $n_M - n_{M-1}$
n_M = frecuencia absoluta de la clase modal
n_{M-1} = frecuencia absoluta anterior a la clase modal
n_{M+1} = frecuencia absoluta posterior a la clase modal

4.3.3.3. Mediana

$$M_d = L + C \cdot \left[\frac{n/2 - N_{d-1}}{n_d} \right]$$

donde:

L = limite inferior real de la clase mediana
I = amplitud del intervalo
n = número total de datos
n_d = frecuencia absoluta de la clase mediana
N_{d-1} = frecuencia absoluta acumulada hasta la clase anterior a la mediana.

4.3.3.4. Medidas de dispersión

a. Rango

El rango es un suplemento de la media definida como:

R = dato mayor – dato menor.

Ejemplo:

El promedio semanal de una fábrica A es 40 (unidades) con un rango de 15 a 60 (unidades); y el de la fábrica B es 40 (unidades) con un rango de 30 a 50 (ud). Por lo tanto B es más representativo de los dos.

El rango es una medida de dispersión muy pobre puesto que puede ser afectada por un dato no usual muy pequeño o muy grande. Una medida de dispersión que no se ve afectada por los valores extremos es la desviación cuartílica.

b. Desviación cuartílica (se abrevia DQ o RSQ)

$$Q = ½ \ (Q_3 + Q_1)$$

donde:

$$Q_q = Li + (q\ n /4 \ - \ N_{d-1}) / n_d$$ en que:

L_i = límite inferior real de la clase del cuartil de orden i (1 ó 3)
n = número total de datos
n_{ci} = frecuencia absoluta del cuartil de orden i.
N_{ci-1}= frecuencia absoluta acumulada hasta la clase anterior a la del cuartil de orden i.

La desviación cuartílica no se ve afectada por los valores extremos como el rango, pero aunque es mejor que el rango, ella no está basada en cada valor incluido en una distribución dada.

c. Rango percentil (se abrevia RP)

$$RP = P_{90} - P_{10}$$

donde:

$$P_p = Li + (pn/100 - N_{p-1}) \times c \div nd$$

en que:

L_i = límite inferior real de orden i (10 o 90)
I = amplitud del intervalo
n = número total de datos
N_{p-1} = frecuencia absoluta acumulada hasta la clase anterior a la del percentil de orden i.
n_p = frecuencia absoluta de la clase del percentil de orden i.

d. Desviación media (se abrevia M.D.)

La desviación media es una medida de dispersión que está basada en todos los datos y mide la dispersión alrededor de una medida central (que puede ser x, M_o, M_d).

$$M.D. = \sum_{i=1}^{i=k} f_i |X_i - x| = \frac{1}{n} \sum_{i=1}^{i=k} n_i |X_i - x|$$

donde:

k = número de clases
n = número total de datos
f_i = frecuencia relativa de la clase i
n_i = frecuencia absoluta de la clase i
X_i = marca de clase de la clase i

e. Desviación estándar (se designa por σ)

$$\sigma^2 = \frac{1}{n} \left[\sum_{i-1}^{i-k} n_1 (X_i)^2 \right] - \overline{x^2}$$

Una vez obtenido σ^2 (que se llama varianza o variancia) se puede obtener S sin dificultad.

f. Dispersión relativa

Sirve para comparar dos conjuntos de datos.

$$V = \frac{\sigma}{x} \quad coeficiente \quad de \quad \text{var}iación$$

$$V_{MD} = \frac{MD}{M_d} \quad coeficiente \quad de \quad la \quad desviación \quad media$$

$$V_{DQ} = \frac{D.Q.}{M_d} \quad coeficiente \quad de \quad desviacon \quad cuartíica$$

En nuestro ejemplo 3.2. dado en la tabla del cuadro resumen, se tiene:

X_i	n_i	$X_i\, n_i$	$X_i^2 \cdot n_i$
74	2	148	10.952
80	2	160	12.800
86	6	516	44.376
92	14	1.288	118.496
98	12	1.176	115.248
104	10	1.040	108.160
110	4	440	48.400
	n = 50	**4.768**	**458.4342**

Por lo tanto:

a. Media

$$\bar{x} = \frac{4.768}{50} = 95{,}36$$

b. Desviación estándar

$$\sigma^2 = \frac{458.432}{50} - 95{,}36$$

$\sigma^2 = 9168.64 - 9093.53$

$\sigma^2 = 75.11$

$\sigma = 8.67$

c. Clase modal: C_4

$$M_o = 89 + 6 \cdot \frac{8}{8+2} = 93{,}8$$

V = 1 – 0,28 = 0,72 (revela que la moda no es significativa)

d. Clase mediana: C_5

$$M_d = 95 + \frac{\frac{50}{2} - 24}{12} \cdot 6 = 95{,}5$$

($F_i = 0{,}5 > 0{,}48$)

$$D = \frac{6-4}{7-1} = 0{,}33$$

$$MD = \frac{353{,}32}{50} = 7{,}066$$

$$MD = \frac{7{,}066}{95{,}5} = 0{,}074$$

d. Clase cuartil de orden 3

C_6 ($F_i = 0{,}92 > 0{,}75$)

$$Q_3 = C_3 = 101 + 6 \frac{3\,50/4 - 36}{10} = 101{,}90$$

Clase cuartil de orden 1:

C_4 ($F_i = 0{,}48 > 0{,}25$)

$$Q_1 = C_1 = 89 + 6 \frac{50/4 - 10}{14} = 89{,}07$$

DQ = ½ (101,90 – 90,07) = 5,91

$$DQ = \frac{5{,}91}{95{,}5} = 0{,}062$$

e. Clase percentil de orden 90

C_6 ($F_i = 0,9 > 0,90$)

$$P_{90} = 101 + 6 \frac{\frac{90\,50}{100} - 36}{10} = 106,4 \qquad P_{10} = 84$$

$$RP = 22,4 \qquad RP = \frac{22,4}{95,5} = 0,235$$

4.4. MEDIDAS DE FORMA: SESGO Y CURTOSIS

La asimetría (sesgo) de una distribución está referida a un eje que pasa por su media. El coeficiente se basa en el hecho de que cuanto mayor sea la asimetría, mayor será la diferencia entre la media y la mediana.

El aplanamiento (curtosis) se refiere al valor máximo de la curva de la distribución en comparación con la curva normal.

a. Sesgo o coeficiente de simetría.

Se define como:

$$\gamma_1 = \frac{m_3}{s^3} \quad donde \quad m_3 = \frac{1}{n} \sum n_i \quad (X_i - \bar{x})^3$$

el cual no siempre es fácil de calcular. Por lo tanto se prefiere definirlo como:

$$\gamma = \frac{\bar{x} - Mo}{s}$$

Si $\Upsilon_1 = 0$ la distribución es simétrica con respecto a la media (esto también se visualiza cuando $M_d = M_o = \bar{X}$)

Si $\Upsilon_1 < 0$ la distribución tiende a concentrarse en valores bajos de la variable (asimetría positiva) (cuando $M_d < M_o < \bar{X}$).

b. Curtosis

La cual está definida por:

$$\gamma_2 = \frac{m_4}{s^4} - 3$$

Si $\Upsilon_2 = 0$ da una curva más puntiaguda que la normal, mientras $\Upsilon_2 < 0$ da una curva achatada.

Problemas propuestos

1. En una gran empresa metalúrgica se computa al azar el número de inasistencia a las labores de sus trabajadores eligiendo al azar una tarjeta de asistencia diaria por cada una de las 51 semanas del año, y así se obtiene la siguiente serie de inasistencias.

120	110	119	121	107	94
118	116	108	114	103	104
119	113	114	116	110	116
115	116	109	118	116	102
116	118	113	109	113	105
117	112	110	120	101	116
122	98	118	104	116	
114	115	106	116	108	

Represente estos datos mediante:

- Un histograma de frecuencia
- Polígono de frecuencias absolutas y relativas
- Frecuencias acumuladas absolutas y relativas

Calcule

- $\bar{X}$, σ, M_{o2}, M_d
- Sesgo

2. La siguiente tabla muestra los diámetros en pulgadas de una muestra de 60 cojinetes de bolas fabricadas por una empresa metalúrgica.

   a. Construir una distribución de frecuencia de los diámetros utilizando intervalos de clases adecuadas.

MEDIDAS					
0,738	0,729	0,743	0,740	0,736	0,741
0,728	0,737	0,736	0,735	0,724	0,733
0,745	0,736	0,742	0,740	0,728	0,738
0,733	0,730	0,732	0,730	0,739	0,734
0,735	0,732	0,735	0,727	0,734	0,732
0,732	0,737	0,731	0,746	0,735	0,735
0,735	0,735	0,733	0,726	0,736	0,732
0,742	0,729	0,739	0,739	0,730	0,735
0,725	0,731	0,741	0,734	0,737	0,744
0,738	0,736	0,734	0,727	0,735	0,740

- Represente y calcule además lo solicitado en problema anterior.

- Mediante uso de una ojiva, ¿Qué porcentaje está comprendido entre 0,728 y 0,733?

3. Idem para; supongamos que, la siguiente tabla muestra el número de trabajadores agrícolas y no agrícolas en la primera región en los años 1850 y 1950, expresados en miles:

AÑO	AGRÍCOLAS	NO AGRÍCOLAS
1850	4.9	2.8
1860	6,2	4,3
1870	6,9	6,1
1880	8,6	8,8
1890	9,9	13,4
1900	10,9	18,2
1910	11,6	25,8
1920	11,4	31,0
1930	10,5	38,4
1940	8,8	42,9
1950	6,8	52,2

4. La tabla siguiente muestra la cantidad en milímetros de agua caída en Santiago durante los años 1957- 1967.

$$fij = \frac{n_g}{n} \quad es\ la\ frecuencia\ relativa\ de\ la\ \text{mod}alid\ A_i\ B_j$$

$$\Rightarrow \sum_{i=1}^{r}\sum_{i=1}^{s} fij = 1$$

La información acerca de las frecuencias ya sea absoluta o relativa, se pueden resumir en un cuadro denominado "Tabla de Contingencia".

X | Y	B_1	B_2	---	B_j	---	B_s	TOTAL
A_1	n_{11}	n_{12}	---	n_{Ij}	---	n_{1s}	$n_{1.}$
A_2	n_{21}	n_{22}	---	n_{Ij}	---	n_{2s}	$n_{2.}$
||	|	|		|		|	
A_i	n_{i1}	n_{i2}	---	n_{ij}	---	n_{is}	$n_{i.}$
||	|	|		|		|	
A_r	n_{r1}	n_{r2}	---	n_{rj}	---	n_{rs}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$		$n_{.j}$		$n_{.s}$	n

Definamos para y = 1..., r.

$n_{i.} = \sum_{j=1}^{s} n_{ij}$ (suma de los valores de la fila i-ésima de la tabla de contingencia de frecuencias).

n_i corresponde al número de elementos de la muestra que pertenecen a la clase A_i según X independiente de la modalidad B_j a la que estén asociados.

AÑO	AGUA CAÍDA
1957	309
1958	336
1959	320
1960	194
1961	261
1962	227
1963	446
1964	187
1965	414
1966	364
1967	173

5. ESTADÍSTICA DESCRIPTIVA BIVARIADA

Nos corresponde tratar ahora el problema de analizar simultáneamente dos "variables estadísticas" de una población para lo cual la "censamos" o tomamos una muestra de ella estudiando sobre la base de ésta ambos caracteres.

Sean X, Y los caracteres a estudiar, y supongamos que hemos obtenido una muestra de tamaño n de la población.

Dividamos la muestra en r clases A_i según X y en S clases B_j según Y. Llamamos n_{ij} al número de elementos de la muestra que pertenecen simultáneamente a la clase A_i según X y a la clase B_j según Y. Podemos luego considerar una clase o modalidad A_iB_j formada por los elementos de la muestra que pertenecen simultáneamente a A_i según X y a B_j según Y. Se observa que hay r·s modalidades A_iB_j.

n_{ij} : Llamamos a la frecuencia absoluta de la modalidad A_iB_j.

$n_{.j} = \sum_{j=1}^{s} n_y$ (suma de los valores de la columna j-ésima de la tabla de contingencia de frecuencias).
n_{ij} corresponde al número de elementos de la muestra que pertenecen a la clase B_j según Y independientemente de la modalidad A_i a la que estén asociados.

5.1. DISTRIBUCIÓN DE FRECUENCIAS MARGINALES

i) De la variable estadística X

$$f_i = \frac{n_{i.}}{n} \quad j = I, \quad ..... \quad s$$

(conjunto de variables relativas a las clases A_i considerándolas independiente de las B_j).

ii) De la variable Y

$$f_i = \frac{n_{j.}}{n} \quad j = I, \quad ..... \quad s$$

(conjunto de variables relativas a las clases B_j considerándolas independiente de las A_i).

5.2. DISTRIBUCIÓN CONDICIONAL

La distribución condicional consiste en estudiar las frecuencias asociadas a las clases de una variable cuando nos restringimos a los elementos de una clase dada según la otra variable, esto es estudiar el comportamiento de una variable dado un valor fijo de la otra.

5.3. DISTRIBUCIÓN CONDICIONAL DE X DADO Y (X/Y)

$$f_{i/j} = \frac{f_{ij}}{f_{.j}} = \frac{n_{ij}}{n_{.j}} \quad i = 1, \quad ..... \quad r$$

El conjunto $\{f_{1/j}, f_{2/j}, ...f_{r/j}\}$ constituye la distribución condicional del carácter X dada la clase B_j de Y (es decir, la distribución de frecuencias según X cuando tomamos sólo los elementos pertenecientes a la clase B_j según Y).

Análogamente podemos definir distribución condicional de Y dado X (Y(C))

$$f_{j/i} = \frac{f_{ij}}{f_{.i.}} = \frac{n_{ij}}{n_{i.}} \quad j = 1, \quad ..... \quad s$$

5.4. INDEPENDENCIA DE VARIABLES

Decimos que una "variable estadística" Y es independiente de X si las frecuencias condicionales de Y/X son todas iguales, es decir, no depende de la clase X condicionante.

Y es independiente de X $\Rightarrow f_{j/1} = f_{j/2} = ... = f_{j=r}$
$V_j = 1, ...s$

y esto es $\frac{n_{1j}}{n_1} = \frac{n_{2j}}{n_2} = ... = \frac{n_{rj}}{n_r} = \frac{n_{1j} + n_{2j} + ... + n_{rj}}{n_1 + n_2 + ... + n_r} = = \frac{n_{ij}}{n} = f_i$

Entonces Y es independiente de X $\Leftrightarrow$ las frecuencias condicionales de Y/X son iguales a la frecuencia relativa marginal correspondiente, cualquiera que sea la clase de X condicionante y para toda clase de Y.

De manera análoga se define "X independiente de Y".

OBSERVACIÓN:

Y es independiente de x $\Leftrightarrow$ la frecuencia conjunta es igual al producto de las frecuencias marginales.

$f_{ij} = (f_i)\ (f_j) = f_{i\bullet} \times f_{\bullet j}$

DEFINICIÓN:

X e Y no son independientes entre sí, se dice que existe relación o ligazón entre ellos. De modo que el conocimiento de una de las variables presente alguna información respecto a la otra.

Nuestro objetivo es medir de alguna forma porcentual esta relación existente y poder además describir de que forma (lineal, exponencial, potencial, etc.) existe.

5.5. ASOCIACIÓN EN EL NIVEL INTERVALAR

Frecuentemente nos hallamos ante tablas donde se ha recogido datos sobre dos variables intervalares. Nos interesa estudiar la asociación que entre ellas pudiera existir. A manera de motivación consideremos el ejemplo siguiente:

Ejemplo:

Se toma una muestra de 5 individuos y encuestamos

X = sueldo Y = cargas familiares

X	Y	x_i- x	y_i-y	(x_i-x)(y_i-y)
10	1	-14	-1,4	19.6
8	1	-16	-1,4	22.4
22	3	-2	0.6	-1.2
30	3	6	0.6	3.6
50	4	26	1.6	41.6
Media 24	2.4			17.2

La idea que hay tras esta tabla es la siguiente:

Si X e Y están asociadas de modo "favorable" (es decir, se comportan igual: aumenta X $\Rightarrow$ aumenta Y; disminuye X $\Rightarrow$ disminuye Y) entonces las columnas X_i-x e y_i-y deberían tener los mismos signos ya que situarse a la izquierda de la media de x implicaría estar también a la izquierda de y y viceversa.

Por el contrario si X e Y estuvieran asociadas "repulsivamente" (variarán en direcciones opuestas) entonces los signos de las columnas $x_i - x$ e $y_i - y$ serían contrario.

Nuestra intención es construir una medida de asociación que tenga la propiedad de ser positiva si X e Y juegan favorablemente y de ser negativa en caso contrario. Ello se podría conseguir promediando la columna de productos.

$(x_i\text{-}x)(y_i\text{-}y)$

Tal medida de asociación se llama covarianza entre X e Y.

$$\text{cov}(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

en que n es la cantidad de individuos.

En nuestro ejemplo n = 5 y cov(x,y) = 17,2 que al ser positiva muestra que X e Y están asociadas favorablemente.

La covarianza, sin embargo, no permite tener una noción del "grado de asociación" ya que puede variar entre - ∞ y + ∞ y no hay modo de saber si es "grande" o "chica".

Para poder cuantificar el nivel de asociación se utiliza el llamado coeficiente de correlación.

$$r = \frac{\mathrm{cov}(x,y)}{s(x)\cdot s(y)}, \quad en \quad que \quad s(x) \quad \sqrt{V(x)} \quad y \quad s(y) \quad = \quad \sqrt{V(y)}$$

Nota: (1) $|r| \leq 1 \quad \Leftrightarrow \quad -1 \leq r \leq 1$

en que se tiene:

Correlación positiva = asociación favorable
Correlación negativa = asociación impulsiva
Se entiende que r es de fácil interpretación.

Correlación						
Negativa				**Positiva**		
Alta	Media	Baja	Nula	Baja	Media	Alta
0,1	0,5	0,3	0	0,3	0,5	1

En síntesis, el ejemplo nos permite establecer que:

1. Cuando se estudian dos variables intervalares pretendemos medir de alguna forma la asociación (llamada correlación) que existe entre ellas.

2. Una forma de investigar la dependencia es ver cómo se comportan

3. ambas características en torno a sus respectivas medias, lo cual se hace mediante la "covarianza".

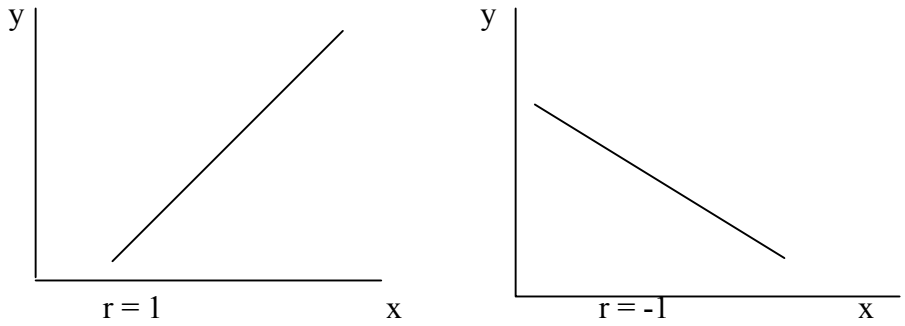
   La covarianza permite decidir el tipo de asociación (favorable o repulsiva) sobre la base de su signo pero no permite cuantificar el grado de asociación).

4. Se puede demostrar que una forma más fácil para el cálculo de r es:

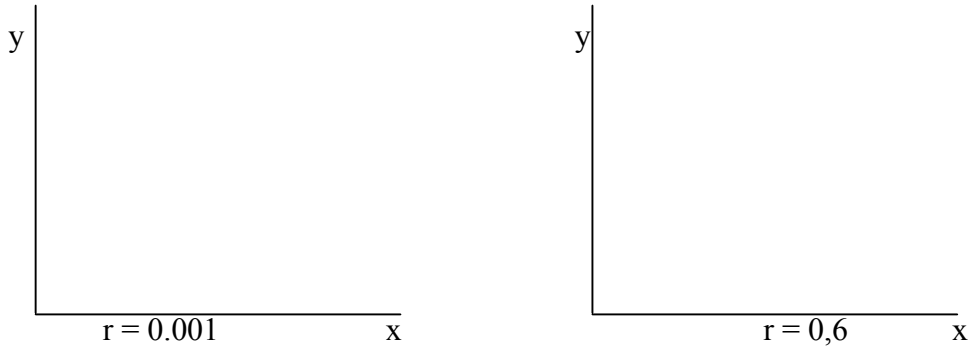
$$R = \frac{\overline{XY} - \bar{X}\,\bar{Y}}{S_x \times S_y}$$

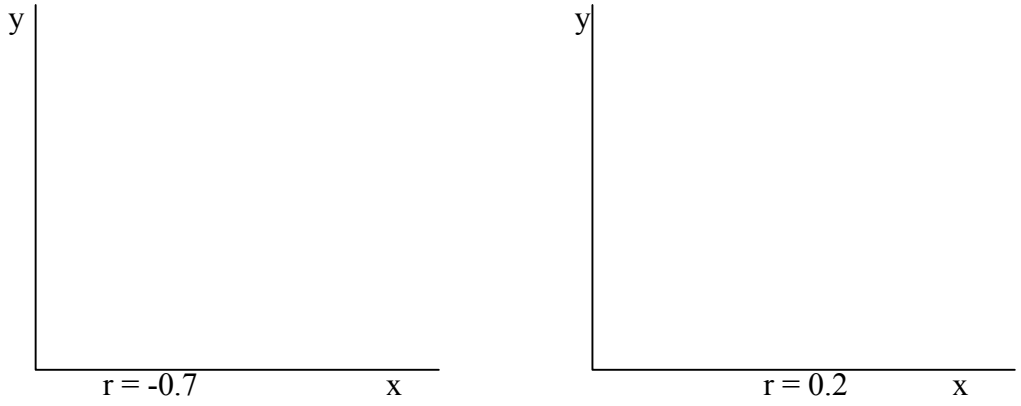
   donde $\overline{XY}$ es el promedio de los productos $x_i y_i$.

Si graficamos las observaciones (x_i, y_i) tendremos situaciones como las siguientes, reflejadas en el valor de r:



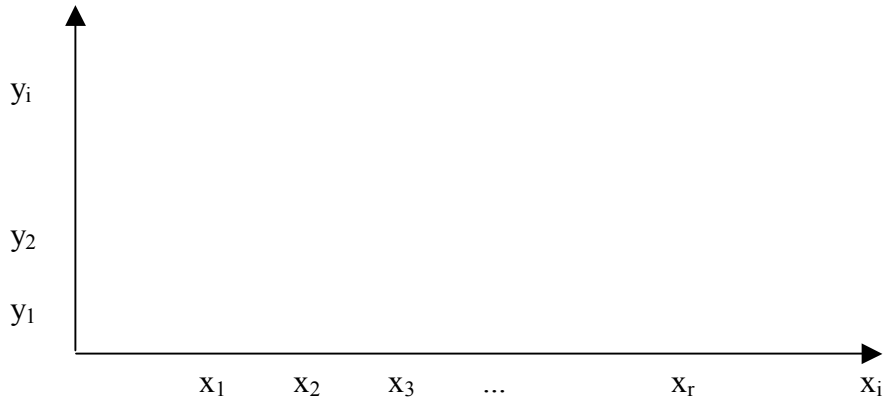
Grafique los puntos.





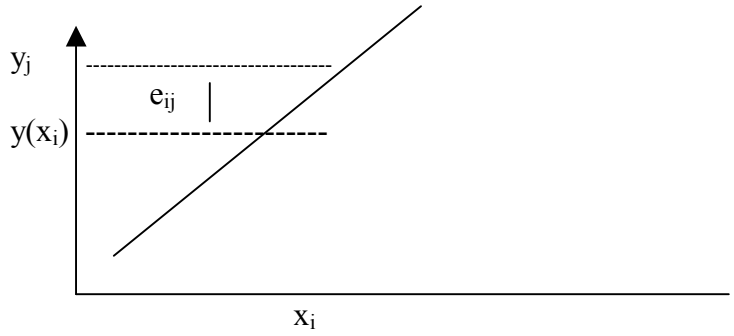
6. CURVAS DE REGRESIÓN

La curva de regresión de Y en X es el gráfico representativo de las medias condicionales y, en función de x_i.



En general nuestra intención es construir una función y = y(x) que "represente" en la mejor forma posible la relación entre x_i e y_j. Esta función permitiría "predecir" (aunque con cierto error) el valor que tomaría la variable Y dado un valor X no encuestado. El criterio para decidir la función se basa en dos fundamentos:

1º Ajustarse a la forma de la nube de puntos (x_i, y_j) = 1, ... r, j = 1, ..., s.
2º Minimizar la media de los cuadrados de los errores.



$y_i = y(x_i) + e_{ij}$

y se determina con la condición

Minimizar $\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij} \quad e_{ij}^2$

Ejemplo

6.1. REGRESIÓN LINEAL

La nube de puntos presenta forma alineada. Suponemos y(x) = ax+b. Entonces:

$y_j = ax_i + b + e_{ij}$

$e^2_{ij} = (y_j - ax_i - b)^2$

Ahora minimizamos:

$$A = \sum_{i=1}^{r} \sum_{j=1}^{s} f_{ij} \; e_{ij}^2 = \sum_{i=j}^{r} \sum_{j=1}^{s} (y_j - ax_i - b)^2$$

Para ello imponemos:

$$\frac{dA}{da} = 0 \quad y \quad \frac{dA}{db} = 0$$

$$\frac{dA}{da} = -2\sum_{i=1}^{r} \sum_{j=1}^{s} f_{ij}(y_j - ax_i - b)x_i = 0$$

$$\frac{dA}{db} = -2\sum_{i=1}^{r} \sum_{j=1}^{s} f_{ij}(y_j - ax_i - b) = 0$$

Resolviendo este sistema de ecuaciones (2 ecuaciones y 2 incógnitas) llamado "Sistema de Ecuaciones Normales" se encuentra

$$a = \frac{\sum\sum f_{ij} x_i y_j - \overline{xy}}{V(x)}$$

$$b = y - ax$$

Se puede demostrar que la ecuación antes obtenida:

y = ax + b se puede poner como

$$a = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$b = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n\sum x^2 - (\sum x)^2}$$

Nota:

Si se puede redefinir los valores de X tales que $\Sigma x = 0$, la expresión se reduce a:

$$a = \frac{\sum xy}{\sum x^2} \qquad b = \bar{y}$$

6.2. OTRAS CURVAS DE REGRESIÓN

1. Si $\gamma(x) = ax^2+bx+c$. Entonces debemos minimizar $A = \Sigma(yj-a\ x_i^2 - bx_i - c)^2$ y sus ecuaciones normales se dan de:

$$\frac{\delta A}{\delta a} = 0 \quad \frac{\delta A}{\delta b} = 0 \quad \frac{\delta A}{\delta c} = 0$$

2. En el caso de tener una curva del tipo exponencial

   $y = a\ e^{bx}$

   podemos tomar logaritmos y queda $1ny = Ina + bx$

   por lo que basta hacer un ajuste lineal ente x y ln y.

3. Parecido ocurre para ajustes del tipo $y = a\ x^b$

   ya que se convierte en $1n\ y = 1n\ a + b\ 1n\ x$

   o sea, debemos hacer un ajuste lineal entre 1nx y 1ny.

   Ejemplo:

   Se da la siguiente tabla de alturas en pulgadas de 12 padres e hijos:

Altura X del padre (pulg)	65	63	67	64	68	62	70	66	68	67	69	71
Altura Y del hijo (pulg)	68	66	68	65	69	66	68	65	71	67	68	70

Por lo tanto, se tiene:

X	Y	X^2	XY	Y^2
65	68	4225	4420	4624
63	66	3969	4158	4356
67	68	4489	4556	4624
64	65	4096	4160	4225
68	69	4624	4692	4761
62	66	3844	4092	4356
70	68	4900	4760	4624
66	65	4356	4290	4225
68	71	4624	4828	5041
67	67	4489	4489	4489
69	68	4761	4692	4624
71	70	5041	4970	4900
$\Sigma X = 800$	$\Sigma Y = 811$	$\Sigma X^2 = 53,418$	$\Sigma Xy = 54,107$	$\Sigma Y^2 = 54,849$

Así $\bar{x} = \frac{800}{12} = 66,67$

$\bar{y} = \frac{811}{12} = 67,58$

$\bar{x} = \frac{53418}{12} - 66,67^2 = 4451,5 - 4444,89$

$= 6,61$

Así s(x) = 2,57

$Vy = \frac{54849}{12} - 67,58^2 = 4578,75 - 4567,06$

$= 11,69$

Así $\sigma(y) = 3,42$

y $\overline{xy} = \frac{54107}{12} = 4508.92$

Por lo tanto:

$$r = \frac{4508{,}92 - 66{,}67 \cdot 67{,}58}{2{,}57 \cdot 3{,}42}$$

$$r = \frac{4508{,}92 - 4505{,}56}{8{,}79} = 0{,}38$$

Si se hace un ajuste lineal

y = ax + b

$$a = \frac{12 \cdot 54.107 - 800\,811}{12 \cdot 53418 - 800^2}$$

$$= \frac{64284 - 648800}{641016 - 64000} = \frac{484}{1016} = 0{,}476$$

$$b = \frac{811 \cdot 53418 - 800\,54107}{1016}$$

$$= \frac{43.321.998 - 43.285.600}{1016} = 35.825$$

Por lo tanto y = 0,476x + 35.825 es la recta ajustada.

Por ejemplo para x = 70 y = 69.145

EJERCICIOS PROPUESTOS

1. El número de bacterias en un cultivo por unidad de tiempo está dado por la tabla

Nº	3	7	21	62	180	500
t	1	2	3	4	5	6

Ajuste una curva no lineal . Justifique la bondad del ajuste.

2. En la siguiente tabla se muestra la potencia (en kg) de un tractor en 1ª en relación con su velocidad (km/hora)

Vol.	1.4	1.8	2.3	3.0	4.0
Pot	7400	7500	7600	7500	7200

Ajuste una recta a estos datos. Calcule r.
Estime la potencia a 5 km/hora.

3. Las notas de Matemática M y Física F de un curso están dadas por:

M/F	2	3	4	5	6
3	1	1	0	0	0
4	2	3	1	1	0
5	0	1	2	3	1
6	0	0	1	2	1

Calcule las medias condicionales de F para los distintos valores de M.
Ajuste una recta a estos datos.
Calcule R.

CAPÍTULO II

TEORÍA DE PROBABILIDADES

Antes de estudiar teoría de probabilidades que nos dará el soporte teórico para el resto del curso, debemos analizar algunos métodos de conteo, que nos permitan conocer el número de maneras que un suceso pueda ocurrir.

2.1. COMBINATORIA

Las reglas básicas para estos métodos de enumeración son:

Regla del y

Si se tiene dos sucesos independientes A y B, A puede ocurrir de "m" maneras y B de "n" maneras, entonces A y B puede ocurrir de m x n maneras.

Regla del o

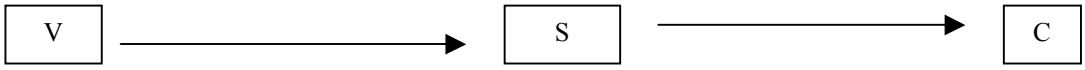
Si se tienen dos sucesos A y B excluyentes o disjuntos. Entonces A ó B puede ocurrir de m + n maneras.

Ejemplo 1

Se quiere viajar de Santiago a Concepción y lo podemos hacer en bus o avión. Si en avión podemos escoger entre 3 aerolíneas y en bus entre 6 empresas entonces pueden viajar de 6 + 3 maneras.

Ejemplo 2

Se quiere viajar de Valparaíso y Santiago y de ahí a Concepción. Si podemos viajar de Valparaíso a Santiago de 4 maneras y de Santiago a Concepción de 3 maneras. Entonces por cada forma escogida en el primer tramo tengo 3 para continuar, como son 4 estas alternativas para el primero, da 4 x 3 = 12 formas de viaje completo.



Ejemplo 3

Cuántas placas patentes se pueden emitir que puedan tres letras y tres dígitos?

25	25	25	10	10	10

Respuesta: $25^3 \cdot (10^3 - 1)$, ya que no existen placas con 000.

Ejemplo 4

¿Cuántas apuestas diferentes se pueden hacer a la polla gol?

1. 13 simples.
   Como cada partido tiene 3 posibilidades, en total se tiene 3^{13} formas.
2. Con 1 doble
   Además de marcar una cruz por partido, se debe escoger uno de los 13 y ahí se disponen de 2 casilleros libres.

Respuesta: $3^{13} \times 2 \times 13$

Ejemplo 5

Se dispone de una bandera blanca, un a azul y una roja. ¿cuántas señales se pueden hacer izando banderas en un mástil?

Si se iza solo una : 3
Si se izan dos : 3 x 2 = 6
Si se izan las tres : 3 x 2 x 1 = 6
Como se pueden izar 1 ó 2 ó 3 da 15 señales.

Como se puede apreciar la naturaleza de los problemas planteados es variada, por lo tanto, trataremos de clasificarlos de acuerdo a dos pautas:

1. Se puede o no repetir
2. Importa o no el orden en que se encuentran.

2.1.1. Arreglos

Se llama arreglos o variaciones de k en n si disponemos de n objetos y escogemos k de ellos importando el orden.

Como el 1º lo escogemos de n maneras, el 2º de (n-1), el 3º de (n-2), etc., nos da que los arreglos sin repetición son:

$A^n_k = n(n-1)\,(n-2)...(n-k+1)$
Si se define : $1 \cdot 2 \cdot 3 \cdot 4 \ldots \cdot n|$ (n factorial)

Entonces la fórmula anterior se puede escribir

$$A_k^n = \frac{n!}{(n-k)!}$$

Ejemplo 6

Si se dispone de 8 libros. ¿De cuantas maneras puedo escoger 3 para ponerlos en un estante?

$$A_3^8 = \frac{8!}{5!} = 876 = 336 \quad maneras$$

Si se permite repetición, entonces en cada oportunidad puedo escoger de n maneras, así

$$\overline{A_k^n} = n^k$$

Ejemplo 7

Si dispongo de banderas rojas, blancas y azules. ¿Cuántas señales puedo hacer al izar 3 banderas?

$$A_3^3 = 3^3 = 27$$

2.1.2. Permutaciones

Si el número k coincide con n, o sea, se toman todos, se convierte en ¿cuántas maneras se pueden arreglar u objetos? esto recibe el nombre de permutaciones de n objetos.

$P_n = n!$

Ejemplo 8

¿De cuántas maneras se pueden ordenas 4 libros?

$P_4 = 4! = 1 \cdot 2 \cdot 3 \cdot 4 = 24$ maneras.

Ejemplo 9

Si se dispone de 4 libros, pero se quiere que dos queden juntos.
Tomemos esos 2 como uno sólo y da $P_3 = 3\text{-}1 = 6$, pero se pueden poner de dos formas AB o BA, así.

Respuesta = $6 \cdot 2 = 12$ maneras.

Si se toma la diferencia $24 - 12 = 12$ da el número de maneras que están separados.

2.1.3. Combinaciones

Si se dispone de n objetos y se quiere escoger k de ellos, sin importar el orden se llama combinaciones de k entre n.

Se puede calcular tomando como si importase el orden A_k^n y dividiendo por el número de veces que cada uno está repetido k!, así

$$C_k^n = \frac{n!}{(n-k)!\cdot k!} = \binom{n}{k}$$

Ejemplo 10

Se dispone de 10 personas. ¿Cuántos tríos de personas diferentes se pueden formar?

$$C_3^{10} = \frac{10!}{7!3!} = \frac{10\,9\,8}{1\,2\,3} = 120$$

Ejemplo 11

Existe un grupo de 4 hombres y 3 mujeres. ¿Cuántos tríos se pueden formar de modo que haya al menos una mujer?

con 1 mujer $C_1^3 C_2^4 = 3 \cdot \frac{4\,3}{1\,2} = 18$

Con 2 mujeres $C_2^3 C_1^4 = 3 \cdot 4 = 12$

Con 3 mujeres $C_3^3 = 1$

Total = 18 + 12 + 1 = 31.

2.2. TEORÍA DE PROBABILIDADES

Si se quiere estudiar una característica X de una población, X recibe el nombre de variable aleatoria y todos los posibles valores que ella puede tomar se llama espacio muestral M. Un subconjunto de M recibe el nombre de suceso.

Ejemplo:.

Se lanza un dado

X = número que resulta del dado.

M = {1, 2, 3, 4, 5, 6}

Un suceso S podría ser "ser par" o sea S = {2,4,6,}

Nos interesa definir la probabilidad que un suceso dado S suceda, o sea definir a cada suceso un número que esté dado en porcentaje, es decir, entre 0 y 1, lo haremos de la siguiente manera:

Definición:

La probabilidad p, será una función ñp : {sucesos} $\rightarrow$ R.

tales que:

1) $0 \leq p(s) \leq 1$
2) $p(m) = 1$
3) Si $A \cap B = \varnothing$, entonces $p(A \cup B) = p(A) + p(B)$

Consecuencias:

a) $p(\varnothing) = 0$
b) $p(A^c) = 1 - p(A)$
donde A^c es el complemento de A
c) $P(A \cup B) = p(A) + p(B) - p(A \cap B)$

Esta definición tendrá distintas formas en su aplicación dependiendo de cómo es el espacio muestral M que tiene la variable aleatoria.

2.2.1. Espacio Muestral Finito

Supongamos de M = $\{x_1, x_2 ... , x_n\}$ donde los x_i son los valores que puede tomar x, entonces se tiene que:

$p(x_{1)} + p(x_2) + ... + p(x_n) = p(M) = 1$

Si todos ellos tienen igual probabilidad p, entonces p + p + p .-..+ p = np = 1, lo que da $p = \frac{1}{n}$

Ejemplo 1:

¿Cuál es la probabilidad de obtener un as con un dado?

$p = \frac{1}{6}$

Si se toma un suceso S = $\{x_1 , x_2 , ... x_k\}$, entonces $p(s) = p(x_{1)} + p(x_2) + ... * p(x_k) =$ p + p ... + $p = \frac{k}{n}$, lo que nos lleva a la siguiente definición

Definición

$$p(S) = \frac{N \ de \ casos \ favorables}{N^o \ de \ casos \ totales}$$

Ejemplo 2:

Sacar suma 7 con dos dados:

Cada dado puede caer de 6 maneas, por lo tanto los dos dados pueden caer de 36 maneras, mientras que 7 se puede obtener como: 16, 25, 34, 43, 52, 61, o sea

$$p(7) = \frac{6}{36} = \frac{1}{6}$$

Ejemplo 3:

Probabilidad de al sacar 2 cartas sean dos ases.

$$p(AA) = \frac{4}{52} \bullet \frac{3}{51}$$

Nota: Sea A: se extrae un A_s

Ejemplo 4

Si se dispone de máquinas nuevas y usadas, eléctricas y manuales en una oficina dadas por la tabla:

	N	U	
E	20	10	30
M	4.	15	19
	24	25	49

¿Cuál es la probabilidad que sea?

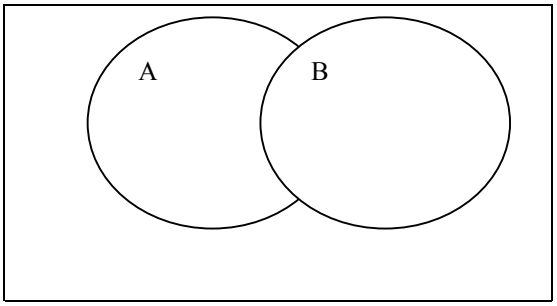
a) Eléctrica : $p(E) = \frac{30}{49}$

b) Nueva : $p(NE) = \frac{24}{49}$

c) Nueva y Eléctrica : $p(N \text{ y } E) = \frac{20}{49}$

2.2.2. Probabilidad Condicional

Definiremos la probabilidad condicional como la probabilidad de que ocurra un suceso A si se sabe que ha ocurrido un suceso B.



Definición: $P(A/B) = \dfrac{p(A \cap B)}{p(B)}$

Ejemplo 5

En el caso de las máquinas del ejemplo 4. Calcule ¿Cuál es la probabilidad que sea nueva una máquina si se sabe que es eléctrica?

$$p\,(Nueva/Eléctrica = \frac{p(nueva \quad y \quad eléctrica}{p(eléctrica)}$$

$$= \quad \frac{20/49}{30/49} \quad d = \quad \frac{20}{30} \quad = \quad \frac{2}{3}$$

Diremos que dos sucesos son independientes $\Leftrightarrow p(A \cap B) = p(A)\,p(B)$

En este caso:

$$p(A/B) \quad = \quad \frac{p(A \cap B)}{p(B)} \quad = \quad \frac{p(A)p(B)}{p(B)} \quad = \quad p(A)$$

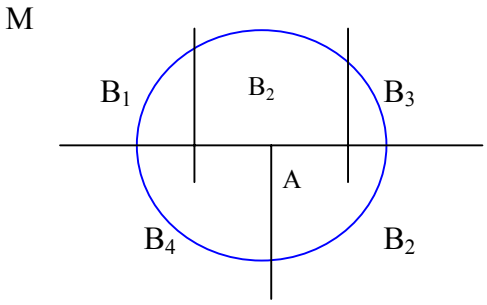
2.2.3. Espacio Muestral Particionado

Entenderemos que el espacio muestral está particionando si:

$M = B_1 \cup B_2 \ ... \cup \ B_6$

con $B_i \cap B_j = \varnothing$

Se quiere expresar la probabilidad de un suceso cualquiera A en ese espacio muestral.



$p(A) = p(A \cap B_1) + p(A \cap B_2) + ... + p(A \cap B_n)$ usando la fórmula anterior.

$$\boxed{p(A) = p(B_1)\,p(A/B_1) + p(B_2)\ p(A/B_2) + ... + p(B_n)\,p(A/B_n)}$$

Ejemplo 6

Una fábrica tiene tres plantas de producción, una en Arica que produce el 30% de la producción y muestra un 5% de productos defectuosos. Santiago que produce el 50% con 3% de defectuosos y Valparaíso con el resto y un 2% de defectuosos. Si se toma un artículo producido al azar ¿Cuál es la probabilidad que sea defectuoso?

B_1 = Arica B_2 = Santiago B_3 = Valparaíso

Def: defectuoso

A = Ser defectuoso

p(def) = p(Arica · p(def)/Arica + p(Santiago)· p(def/Santiago) + p(Valparaíso) · p/def/Valparaíso)

= 0,3 · 0,05 + 0,5 3 0,03 + 0,2 · 0,02

= 0,015 + 0,015 + 0,004 = 0,034

2.2.4. Probabilidad de Causas. Bayes

La pregunta es ahora al revés, si se sabe que algo pasó ¿Cuál es la probabilidad de alguna causa?

O sea:

$$p(B_i / A) = \frac{p(B_i \cap A)}{p(A)}$$

$$= \frac{p(A / B_i d)\, p(B_i)}{p(A)}$$

Ejemplo 7

En el ejemplo 6 anterior si un artículo fue defectuoso ¿Cuál es la probabilidad de que haya sido producido en Arica?

$$p(A / Def) = \frac{p(Def / Arica) \cdot p(Arica)}{p(Def)}$$

$$= \frac{0{,}05\,0{,}3}{0{,}034} = \frac{0{,}015}{0{,}034} = \frac{15}{34}$$

EJERCICIOS

1. ¿Cuál es la probabilidad de obtener una figura y un as, al sacar dos cartas de un naipe? (21 real).

2. Un dispositivo eléctrico tiene probabilidades 0,1 de fallas. ¿Cuál es la probabilidad de que falle un sistema que tiene 2 dispositivos
   a) En serie
   b) En paralelo?

3. ¿Cuál es la probabilidad de ganar el loto, si se trata de acertar 6 números de 36? ¿del Kino, si se trata de acertar 15 de 25?

4. Un agricultor planta 3 tipos de manzanas. El 40% son del tipo A y produce un 85% de exportación, 30% son del tipo B y el 90% son de exportación, el resto es del tipo C y el 80% es de exportación. ¿De su producción que porcentaje es de3 exportación? Si una manzana es de exportación ¿Cuál es la probabilidad que sea del tipo B?

5. En una ciudad se venden 1000 diarios A, 2000 B y 5000 C. Si de los lectores de A el 25% fuma, el 50% de los B y el 10% de los de C. Si suponemos que cada persona lee un solo diario

   a) ¿Qué porcentaje de los lectores fuman?
   b) Si una persona no fuma ¿Cuál es la probabilidad que lea B?
   c) De los lectores de B el 25% bebe alcohol y el 15% bebe y fuma ¿Cuál es la probabilidad que un lector de B no beba ni fume?

2.3. VARIABLES ALEATORIAS

Dependiendo del problema a estudiar definimos una variable aleatoria X y su correspondiente espacio muestral M. Si M es un conjunto finito o numerable se dice que la variable y el espacio muestral es <u>discreto</u>, si M es un intervalo de números reales se dice que son continuos.

En el caso que la variable sea discreta llamaremos función de probabilidad a:

$p : M \rightarrow R$

tales que:

a) $(X) > 0 \qquad \forall x \in M$

b) $\sum_{x \in M} p(x) = 1$

Definición

Si X es continua, llamaremos función de densidad a la función

$F : M \rightarrow R$

tales que

a) $f(x) \geq 0 \qquad \forall x \in M$

b) $\int_M f(x)dx = 1$

En este caso

$P(a \leq x \leq b) = \int_a^b f(x)dx$

2.3.1. Funciones de Probabilidad más Usuales

a) Geométricas

Un experimento tiene probabilidad p de éxito, se debe repetir el experimento hasta tener éxito.

X = Nº de experiencias necesarias hasta tener éxito.

M = {1,2,3,43...}

$p(x = k) = (1-p)^{k-1} \cdot p$

b) Binomial

La probabilidad de tener éxito en una experiencia es p. Si realizamos n experiencias independientes, ¿Cuál es la probabilidad de tener un Nº dado de éxitos?

X = Nº de éxitos entre los n

M = {0,1,2... n}

$p(x = k) = \binom{n}{k} p^k (1-p)^{n-k}$

c) Hipergeométrica

Se disponen de N objetos, de ellos r son de una cierta clase A y el resto (N-r) no lo son. Si tomamos n de ellos ¿Cuál es la probabilidad que k sean del tipo A?

X = Nº de objetos del tipo A

M = {0,1,2... n}

$$p(x=k) = \frac{\binom{r}{k}\binom{N-r}{n-k}}{\binom{N}{n}}$$

Observación:

A veces ocurre que por ser los números grandes es muy difícil calcular estos coeficientes binomiales y como para estos valores no son muy diferentes las probabilidades calculadas por binomial e hipergeométrica se aproxima la 2ª por la primera.