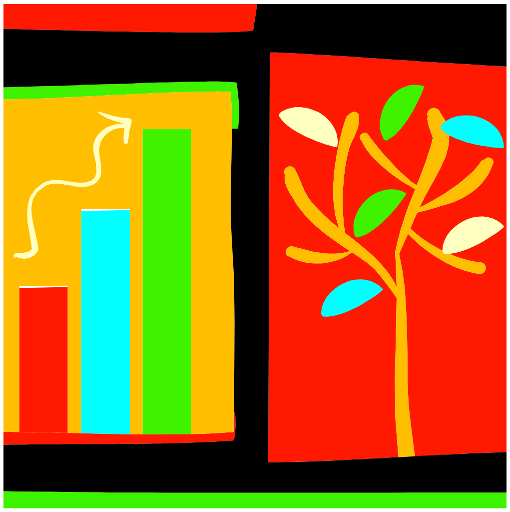
UNIVERSIDAD NACIONAL DE INGENIERÍA

UNI- NORTE - SEDE REGIONAL
Estelí, Nicaragua

Estadística Básica para Docentes de Educación Secundaria



$$\bar{x} = \frac{\sum_1^n x}{n}$$

Tercer Diplomado en Didáctica de la Matemática

12/11/2009

Luis María Dicovskiy Riobóo

Índice

Introducción

Este libro de Texto dirigido a profesores de secundaria, cubre desde séptimo hasta undécimo grado la mayoría de temas que aborda la malla de contenidos de educación secundaria del Ministerio de Educación de Nicaragua, MINED. Éste texto tiene un enfoque utilitario, práctico, respetando el principio que la Estadística debe ser una herramienta fundamental para describir procesos y tomar decisiones en el trabajo cotidiano de cualquier profesional. En el mismo se trató de romper la dicotomía entre teoría y realidad, respondiendo permanentemente a la pregunta ¿Cuándo puedo usar esta teoría? ¿Qué me permite conocer o responder la misma? Si podemos describir "la estadística" como: *"un conjunto de técnicas para describir grupos de datos y para tomar decisiones en ausencia de una información completa"*. ¡Un libro de estadística debe respetar esta definición!

Por lo anterior y respetando el principio de asequibilidad es que buena cantidad de los ejercicios fueron generados en el aula con la información que tienen los estudiantes a la mano. Creo que la estadística no puede funcionar si primero no se sabe como generar el dato, cómo organizar la información en forma de matriz y luego analizar ésta usando un programa estadístico computacional. Debo aclarar que éste texto está dirigido a docentes del área de matemáticas, pero debe ser mediado al momento de aplicar estos contenidos teóricos a jóvenes adolescentes, cada profesor debe ajustar la forma y profundidad teórica de la enseñanza según el año de académico y las características de los estudiantes.

Para hacer los ejercicios de este texto y construir gráficos digitales se sugiere utilizar el programa estadístico INFOSTAT, el cual dispone de una versión de uso libre que se puede descargar gratuitamente desde la página www.infostat.com.ar .

Recomendaciones Metodológicas

Enseñar Estadística con información construida en la clase y al mismo tiempo descifrar cómo hacer que una investigación sea el hilo conductor del curso.

En estas recomendaciones metodológicas de cómo impartir conocimientos básicos de Estadística en la educación secundaria, quisiera compartir algunos conceptos que me parecen atractivos para discutir:

- Dos propuestas a considerar.
- Una historia
- Un ejemplo de docencia a discutir.
- Una verdad relativa
- Un deseo

Dos propuestas a considerar

1- *La información que se genera en la clase puede ser la base sobre la cual se construye la teoría de una asignatura.*

2- *Si lo que se enseña en clase son herramientas operacionales, estas deben funcionar en condiciones reales.*

Una historia. La escalera.

Hace unos años tuve un encuentro con un profesor de matemáticas que impartía clases en una carrera de ingeniería, en una tarde cálida de abril él estaba explicando la teoría para resolver una derivada doble, a modo de ejemplo dio un ejercicio, el cual resolvieron los estudiantes luego de una tarde de arduo cálculo, asustado por el nivel de análisis que tenía el grupo, se me ocurrió preguntar sin ninguna malicia ¿para qué les pudiera servir dicho análisis, en cuales aspectos de la vida real podrían darle uso?

Ante mi sorpresa los alumnos no supieron que decir y el profesor muy seriamente me explicó que si alguien se subía una escalera apoyada en una pared y esta se comenzara a deslizar se podía saber por la derivada la velocidad de caída dentro de un intervalo, por ejemplo entre 3 y 2 metros antes de caer al suelo. ¡Me imaginé cayendo con la escalera y al mismo tiempo haciendo el cálculo con que me iba a dar el golpe y sinceramente no quedé convencido de la utilidad presentada! Creo que los estudiantes tampoco.

Un ejemplo de docencia a discutir.

Podemos describir "la estadística" como: *"un conjunto de técnicas para describir grupos de datos y para tomar decisiones en ausencia de una información completa"*. ¡Un curso de estadística debe respetar esta definición!

La primera pregunta que surge es ¿De dónde sacar los datos?. La enseñanza clásica diría que se debe recurrir a los ejercicios de los libros de texto donde hay ejemplos resueltos y que no tienen complicaciones extrañas. Un primera debilidad de este tipo de enseñanza es que la realidad es complicada, llena de ruidos y estos ejemplos no nos preparan para estos ruidos, sin embargo la principal debilidad es enseñar con ejemplos que llegan fuera del contexto del estudiante y este no logra apropiarse de ellos ni sentirse motivado.

La metodología que se sugiere usar, comienza por definir como sujeto de estudio al propio estudiante, cada uno de ellos serán la unidad de investigación y cada uno aportará la información de sí mismo que luego compartirá con los demás. Durante el proceso de enseñanza se sugiere recorrer los siguientes pasos:

A) Definir los "objetivos" a resolver con la información que se usará en el curso, el cual se enfoca como una investigación. Un ejemplo de objetivo podría ser: "describir las causas que inciden con el rendimiento académico del grupo". Se construyen de manera colectiva preguntas de investigación surgidas de éste tema. Por ejemplo ¿se estudia lo suficiente?, ¿la edad es relevante para prestar atención en clase?, ¿las notas son una buena medida de lo que sabe un estudiante?, etc.

B) Luego se definen conceptos básicos para poder luego construir conocimiento, por ejemplo: se explica el concepto de variable, luego se discute cuales variables pueden servir para alcanzar el objetivo definido previamente.

Variable*: es una característica observable de un objeto que varia, las variables pueden ser: a) Cualitativas ó b) Cuantitativas, que son las que analizaremos numéricamente.*

A partir de la definición se definen qué variables se recolectarán datos. Por ejemplo: Horas de estudio semanales, conformidad con la docencia recibida, edad, peso, sexo de los estudiantes, etc.

C) Se enseña a codificar y re codificar variables, se propone que junto a las variables descriptivas también halla variables actitudinales por ejemplo en escala Likert para conocer la opinión de los alumnos. Se elabora una encuesta para recolectar la información. ¡Se comienza a modelar la realidad!

D) Se llena una matriz de datos, fundamentalmente numéricos, los alumnos juegan el rol de entrevistado y entrevistador simultáneamente. Se trabaja en grupos pequeños, cada grupo genera sus datos, se entrevistan entre ellos y luego el grupo intercambian su información con los otros grupos. No es necesario que todos tengan la misma cantidad de entrevistados, solo se le pide un número mínimo, generalmente más de 30 alumnos. De esta manera al final hay varias bases de datos, cada grupo tendrá sus propios datos de una muestra. Finalmente en la matriz de datos cada fila es un alumno y cada columna es una variable.

“En este momento, a partir de la observación de la realidad se creó un modelo numérico teórico que permitirá estudiar esta realidad.”

E) Con los datos obtenidos se desarrolla el programa de estadística, se sugiere trabajar alternando breves momentos teóricos con la resolución de ejercicios. Se comienza haciendo estudios descriptivos de las variables, luego se construyen gráficos y se construyen probabilidades. Permanentemente se deben discutir los resultados observables los cuales deben generar cierta polémica. El estudiante se debe motivar con los resultados que obtiene.

Se sugiera que al inicio los ejercicios numéricos se hagan manualmente, luego con calculadora científica y finalmente con un programa de computadora, por ejemplo INFOSTAT.

Al finalizar cada curso se debe reflexionar sobre las preguntas de investigación iniciales, y que dicen los datos sobre ellas. Por ejemplo: Ser varón o mujer ó la edad inciden sobre el rendimiento académico. Entonces se discute y se trata que la información nos diga lo que puede decir. En este momento "se ha generado información que se siente viva". Simultáneamente a los resultados se debe recordar la teoría estadística que fue necesaria usar.

En resumen, difícilmente el estudiante se olvidará que es lo que hizo, en qué contexto aplicó las pruebas estadísticas y para que le sirviera la asignatura. Adicionalmente en éste proceso aprendió a recorrer un proceso de investigación con resultados que le permitirán reflexionar sobre su realidad.

Una Verdad relativa

> La realidad observable siempre es más emocionante que un ejercicio de un libro de texto.

Un Deseo

"Solo espero que disfruten enseñando estadística al mismo tiempo que sus estudiantes aprendan a mirar el mundo desde una perspectiva cuantitativa".

Capítulo I. Estadística Descriptiva

Objetivos

- Reflexionar sobre el uso de la estadística a través de situaciones de la vida cotidiana.
- Introducir a la recolección de datos a partir de un problema del entorno.
- Construir conceptos básicos de estadística desde la experiencia del estudiante.
- Ejemplificar los diferentes tipos de variables con los datos observados para construir una tabla de distribución de frecuencia, TDF.
- Realizar medidas de tendencia central, de variabilidad y diferentes tipos de Gráficos más comunes que permite una TDF.
- Valorar actitudes de orden, perseverancia, capacidades de investigación para desarrollar el gusto por la Estadística y contribuir al desarrollo del entorno social y natural.

1.1 Introducción. Tipos de Variables

La estadística, es una ciencia relativamente nueva pero con miles de años de uso empírico, María y José parten de Nazaret a Belén para ser censados por los romanos. ¡Hace 2000 años éste imperio llevaba un control estadístico de lo que poseían sus colonias para luego cobrar impuestos¡ En la actualidad los procedimientos estadísticos son de particular importancia en las ciencias biológicas y sociales para reducir y abstraer datos. Una definición que describe la estadística de manera utilitaria es la que dice que es: *"un conjunto de técnicas para describir grupos de datos y para tomar decisiones en ausencia de una información completa"*. La estadística a diferencia de la matemática no genera resultados exactos, los resultados siempre tienen asociada un grado de incertidumbre o error. La estadística trata de lograr una aproximación de la realidad, la cual es siempre mucho más compleja y rica que el modelo que podemos abstraer. Si bien esta ciencia es ideal para describir procesos cuantitativos, tiene serios problemas para explicar "el porqué" cualitativo de las cosas

En general podemos hablar de dos tipos de estadísticas, las *descriptivas* que nos permiten resumir las características de grandes grupos de individuos y *las inferenciales* que nos permite dar respuestas a preguntas (hipótesis) sobre poblaciones grandes a partir de datos de grupos pequeños o muestras.

Construcción de Variables a partir de información.

Para poder analizar datos, ya sea de forma manual o por computadora, hay que entender que trataremos a partir del estudio de la realidad observable crear un modelo numérico teórico donde se estudian variables para describirlas y analizar sus relaciones. Para hacer esto primero es necesario definir algunos términos teóricos.

Variable: es una característica observable de un objeto y que varía. Las variables se pueden clasificar de diferentes maneras, un enfoque es reconocer dos grandes grupos de variables las Cualitativas y Cuantitativas.

Variables Cualitativas, son aquellas que se ordenan en categorías debido a su carácter subjetivo y absoluto, pueden ser de dos tipos “nominales”, u “ordinales”. En las variables nominales los valores **no** pueden ser sometidos a un criterio de orden o importancia como por ejemplo “el sexo de una persona” o “el país de origen”. Las variables ordinales pueden tomar distintos valores ordenados siguiendo una escala establecida, aunque no es necesario que el intervalo entre mediciones sea uniforme, por ejemplo: leve, moderado, grave.

Variables Cuantitativas, son las que sus características están expresadas en valores numéricos, éstas asumen cualquier valor y pueden variar en cualquier cantidad, sobre una escala aritmética e infinita y pueden subdividirse en dos tipos “continuas o medibles” y “discretas o contables”.

Las variables continuas pueden adquirir cualquier valor dentro de un intervalo especificado de valores, permite siempre que se encuentre un valor nuevo entre dos valores previos. El rendimiento de un lote de fríjol se mide en qq/mz es una variable continua, se mide o pesa.

Las variables discretas presentan interrupciones en la escala de valores que puede tomar. Estas separaciones o interrupciones indican la ausencia de valores entre los distintos valores específicos que la variable pueda asumir por número de miembros de una familia es una variable discreta, se cuenta y entre dos personas no hay un valor intermedio, no existe 1.5 personas . Los atributos, en control de calidad, son variables discretas.

Las variables generan "datos", con ellos se hace la estadística y cada uno de éstos ocupa una celda de una matriz o base de datos. La Matriz de datos es un ordenamiento de datos en fila y columnas donde cada fila es un individuo, una parcela, una muestra, una unidad experimental o una encuesta determinada y cada columna: una variable. Los programas Access, Excel, Infostat y SPSS ordenan los datos en forma de matriz. Por ejemplo en una encuesta (cuestionario) cada pregunta que se tiene, genera al menos, una variable generalmente discreta. Hay casos donde una pregunta puede generar muchas variables de tipo dicotómico, SI- NO, que se suele codificar como 1= SI y 0= NO.

Ejercicio 1.1: Construya variables relacionadas con su entorno, 5 nominales, 5 ordinales, 5 continuas y 5 ordinales.

Ejercicio 1.2 Clasifique las siguientes variables.

- Peso de un estudiante.
- Diámetro de una casa.
- Color de ojos.
- Tipo de techo.
- Vida útil de un monitor
- # de ladrillos de una pared.
- Belleza de una flor.
- Temperatura semanal.
- Largo de peces de un estanque.
- Diámetro de un tornillo

1.2 Análisis de datos, TDF y Gráficos

> ***"A partir de la realidad observable se debe crear un modelo numérico teórico para intentar estudiar ésta realidad"***

Una vez que los datos se han codificado, transferidos a una matriz y guardado en una computadora podemos proceder a analizarlos, proceso que se hace con un programa estadístico como SPSS o INFOSTAT, de forma manual solo se pueden manejar pocos datos y variables es por ello que el énfasis de este libro está más en la interpretación de resultados que en los procedimientos de cálculo.

El procedimiento de análisis sugerido se esquematiza en la figura siguiente:

Creación de la matriz de datos → Definición de análisis a realizar → Ejecución de análisis en computadora → Interpretación de resultados

En general el investigador debe buscar de primero cómo describir sus datos y posteriormente efectuar el análisis estadístico para relacionar las variables generadas. Los tipos de análisis son variados y cada método tiene su razón de ser un propósito específico, "la estadística no es un fin en sí misma, sino una herramienta para analizar datos".

Los principales análisis que pueden efectuarse son:

- Estadística descriptiva de las variables.
- Pruebas de hipótesis para la toma de decisiones.

> ***"la estadística está ligada a la toma, organización, presentación y análisis de un grupo de datos".***

Una primera tarea luego de construir una tabla o matriz de datos, es explorarlos buscando información atípica o anormal y corregir los casos que la información atípica se deba a una mala digitación o error en la recolección de datos.

Lo siguiente para observar el comportamiento de los datos es realizar una “distribución frecuencias” en forma de tabla y gráficos. Para esto, los datos se agrupan en clases o categorías y para grupo se calcula las frecuencias absolutas y relativas.

En este momento es importante poder definir el tipo de escala de medición usada, sucesión de medidas que permite organizar datos o para agrupar los datos, en este sentido se pueden reconocer diferentes escalas:

- Las Escalas Nominales, son discontinuas y se usan cuando describimos algo dándole un nombre a cada categoría o clase y estas son mutuamente excluyentes. A cada categoría se le adjudica un valor numérico. Por ejemplo la variable sexo donde “varón = 1” y “mujer = 2”.
- Las Escalas Ordinales, son discontinuas y se usan donde hay un orden jerárquico de un conjunto de objetos o eventos con respecto a algún atributo específico, por ejemplo ordenar los ingresos en tres niveles: “alto =1”, “medio = 2” y “bajo = 3”.
- Las Escalas de Intervalos Iguales, estas pueden ser sumadas, restadas multiplicadas y divididas sin afectar las distancias relativas entre las calificaciones. Por ejemplo las medidas de temperatura en Grados C^0, las calificaciones de un examen en una escala de 1 a 100. En esta escala el “0” es arbitrario y no necesariamente representa ausencia, también nos dice que un valor de 30 puntos de un examen de español no necesariamente representa la mitad de conocimiento de un valor de 60 puntos.
- Las Escala de Razón Constante, tienen todas las propiedades de las Escalas de intervalos más un cero absoluto, por ejemplo las medidas de tiempo, peso y distancia, el valor “0” representa ausencia del valor.

Un caso especial de escala ordinal es la escala de Likert, esta escala es muy usada en las ciencias sociales y se usa para medir actitudes, “Una actitud es una predisposición aprendida par responder consistentemente de una manera favorable o desfavorable ante un objeto de sus símbolos”. Así las personas tenemos actitudes hacia muy diversos objetos o símbolos, por ejemplo: actitudes hacia la política económica, un profesor, la ley, nosotros, etc. Las actitudes están relacionadas con el comportamiento que mantenemos. Estas mediciones de actitudes deben interpretarse como "síntomas" y no como hechos. Esta escala es bipolar porque mide tanto el grado positivo como negativo de cada enunciado y consiste en un conjunto de ítem presentado en forma de afirmaciones o juicios ante los cuales se pide reacción a los sujetos en estudio en una escala de 5 puntos, cada punto tiene un valor numérico. Un ejemplo de cómo calificar con afirmaciones positivas es ¿Le gusta cómo se imparte la clase de estadística?:

1- Muy en desacuerdo, 2- En desacuerdo, 3- Ni de acuerdo, ni en desacuerdo, 4- De acuerdo, 5-Muy de acuerdo.

Estar de acuerdo con la idea presentada significa un puntaje mayor.

Ejercicio 1.3: entre los participantes de la clases tomar datos de 15 variables al menos por ejemplo: Edad, Sexo, Procedencia, etc. Y luego ordénelos en forma de matriz de datos, recodifique la información cualitativa en numérica.

Organización de una matriz de información a partir de un cuestionario.

Una encuesta impersonal con preguntas cerradas es una manera de recolectar mucha información rápidamente que luego se puede codificarla fácilmente, la debilidad de este instrumento es que no siempre la gente responde adecuadamente y que las respuestas generadas se limitan a las opciones previamente definidas y la experiencia nos dice que la realidad es mucho más rica que lo que creemos ocurre a priori. Para los que trabajan con entrevistas hay que saber que también la información que se genera de las entrevistas puede luego tabularse numéricamente de la misma manera que una encuesta.

Encuestas o Cuestionarios: Al diseñar una encuesta esta debe ayudar a responder a las preguntas que genera la hipótesis del trabajo, un error común es hacer una encuesta primero y luego que se han recolectado los datos, se solicita a un estadístico que no ayude a analizar la información, "la lógica es al revés" se debe pensar como se analizará la información desde el mismo momento que se diseña la encuesta. Se sugiera que las variables cualitativas (ej. nombres) se deben recodificar al momento del llenado de la base de datos creando variables numéricas discretas, por ej. Si quiero clasificar las becas que otorga una Universidad puedo codificarlas de la siguiente manera: Beca interna =1, Beca externa =2 y No beca =0.

Si las opciones que genera una variable discreta permite hacer combinaciones de las respuestas se sugiere crear muchas variables dicotómicas del tipo Si o No (1,0). Veamos un ejemplo: Si se pregunta: que prácticas de en los cultivos realiza un campesino, estas pueden ser varias y combinadas como: Insecticidas Botánicos, Trampas amarillas, Barreras vivas, Semilla resistente etc. En este caso lo que se hace es generar un variable del tipo 0-1 para cada opción de práctica de cultivo, generando muchas variables en una sola pregunta.

Para crear una base de datos hay que recordar que se está obteniendo una matriz de datos donde en la primera fila se tiene el nombre abreviado de la variable y en el resto de las filas los datos para cada encuesta o individuo en estudio. Las variables cualitativas se deben recodificar, veamos el siguiente ejemplo hipotético de 8 encuestas:

Encuesta	**Sexo**	**Edad**	**Ingresos semanales C$**	**Comunidad**	**Labor realizada**
1	1	31	1,394	2	3
2	1	35	1,311	4	2
3	1	43	1,300	2	3
4	1	28	1,304	3	1
5	2	45	1,310	1	3
6	2	36	1,443	2	2
7	2	21	1,536	2	3
8	2	32	1,823	1	3

Esta matriz se codifica así: la variable “Sexo”: 1= varón, 2 = mujer. Para la variable “comunidad” hay 4 tipos diferentes donde: 1= Estelí, 2= Condega, 3= Pueblo Nuevo y

4= Limay y para "Labor realizado": 1= en otra finca, 2= en la cuidad y 3= en la propia finca.

De esta manera se transforma en datos numéricos una información descriptiva, estos números permiten luego hacer estadística.

Ejercicio 1.4: Intente codificar numéricamente las respuestas que se generan a partir de la encuesta de caracterización socioeconómica, que a continuación se detalla, discuta las posibles respuestas, diga si las preguntas están bien formuladas, sugiera si alguna de ellas está de más y que preguntas propone para completar la información.

Hoja de Encuesta

Número de ficha___________

Fecha: ____

Primer Apellido____

Segundo Apellido____

Nombres:____

Año____________

Dirección: ____

Estado Civil: ____________

Número de personas que habitan la vivienda__________________________

Nivel de estudio de ellos____

Edad de cada una de ellos____

Profesión: ____

Ejercicio 1.5:

- Defina variables para caracterizar a los estudiantes del curso con el objetivo de determinar posibles causas que tengan influencia en el rendimiento académico del grupo.
- Cree una base de datos de al menos 25 individuos. Ver ejemplo.

Ejemplo de una matriz de datos generados con datos de estudiantes.

Códigos: Estado Civil: 1 Soltero, 2 Casado; Origen: 1 Estelí, 2 No Estelí; Sexo: 1 Varón, 2 Mujer; Becas: 1 Si 2 No; Opinión: 1 Negativa 5 Positiva

GENERACION DE DATOS

NOMBRE	**NOTAS Prom.**	**EST ADO CIVIL**	**EDAD**	**ALTU RA**	**SEXO**	**PESO**	**origen**	**INGRE SO FAMI LIAR**	**BE CA S**	**Opinión**
Abel	74	2	25	1.75	1	140	2	1	0	3
Adely	70	2	18	1.55	2	110	1	1	0	3
Alexis	80	2	24	1.85	1	150	1	1	1	2
Aracely	70	2	20	1.54	2	117	1	1	1	4
Candelario	78	1	24	1.65	1	150	2	1	0	5
Carlos	85	2	19	1.8	1	150	1	2	0	5
Cesar	70	2	19	1.7	1	140	2	1	0	5
Cleotilde	75	1	20	1.5	2	112	1	1	1	1
Danny T	70	2	18	1.7	1	160	1	1	0	4
Danny	85	2	18	1.67	1	120	2	1	0	4
David N	77	2	18	1.63	1	135	1	1	0	2
Deice	75	2	20	1.52	2	110	1	1	1	3
Edwin	80	1	18	1.75	1	110	1	1	0	3
Ronal	80	2	21	1.73	1	160	2	1	0	3
Sara	80	2	17	1.6	2	114	2	1	0	2
Sayda	78	2	18	1.5	2	128	2	1	0	5
Seyla	75	2	20	1.7	2	120	1	1	1	5
Tania	90	2	19	1.65	2	130	2	1	0	4
Uriel	70	2	22	1.65	1	140	2	1	0	2
Yilmar	78	2	18	1.8	1	174	2	2	0	4

Principios a utilizar al construir una Tabla de Distribución de Frecuencias, TDF.

Aunque esta tabla sirve para resumir información de variables discretas ó continuas, de manera particular la TDF permite transformar una variable continua, a una variable discreta definida por el número de intervalos y su frecuencia. Esta transformación permite construir gráficos de histogramas o polígonos. Con Variables continuas como (peso, altura, producción / superficie, etc.) el recorrido de la variable se parte en intervalos semiabiertos, las clases.

Lo primero para construir una TDF es definir el “número de clases” ó intervalos a crear y el “ancho” de cada intervalo. Para que los gráficos permitan visualizar tendencias de la variable en estudios, el número de clases se recomienda que no sean menor de 5 ni mayor de 20. Al ancho de clase se calcula dividiendo el Rango (valor mayor – valor menor), con un valor que debe variar entre 5 y 20. Hay que utilizar más clases cuando se tiene más datos disponibles, si el número de clases es muy grande es posible tener muchas clases vacías, si es demasiado pequeño podrían quedar ocultas características importantes de los datos al agruparlos. Se tendría que determinar el número de clases a partir de la cantidad de datos presente y de su uniformidad, en general con menos de treinta datos se usa una TDF con 5 clases, para tener un criterio sobre el número de clases en función del número de datos ver la tabla siguiente .

Tabla para determinar el número de clases de una TDF

Número datos	Número de clases
30-50	5-7
51-100	6-10
101-250	7-12
+250	10-20

El valor central de una clase se llama “marca de clase”, este valor se usa para construir los gráficos de polígonos de frecuencia. Veamos un ejemplo de cómo se construye una Tabla de Distribución de Frecuencias. Es importante resaltar que con las variables nominales no se construyen intervalos, límites ó marcas de clase, esto no tiene sentido con este tipo de variable.

Ejemplo con Datos de ingresos de 24 familias. Variable: Ingresos semanales en C$ por familia, n = 24 datos.

1,450	1,443	1,536	1,394	1,623	1,650
1,480	1,355	1,350	1,430	1,520	1,550
1,425	1,360	1,430	1,450	1,680	1,540
1,304	1,260	1,328	1,304	1,360	1,600

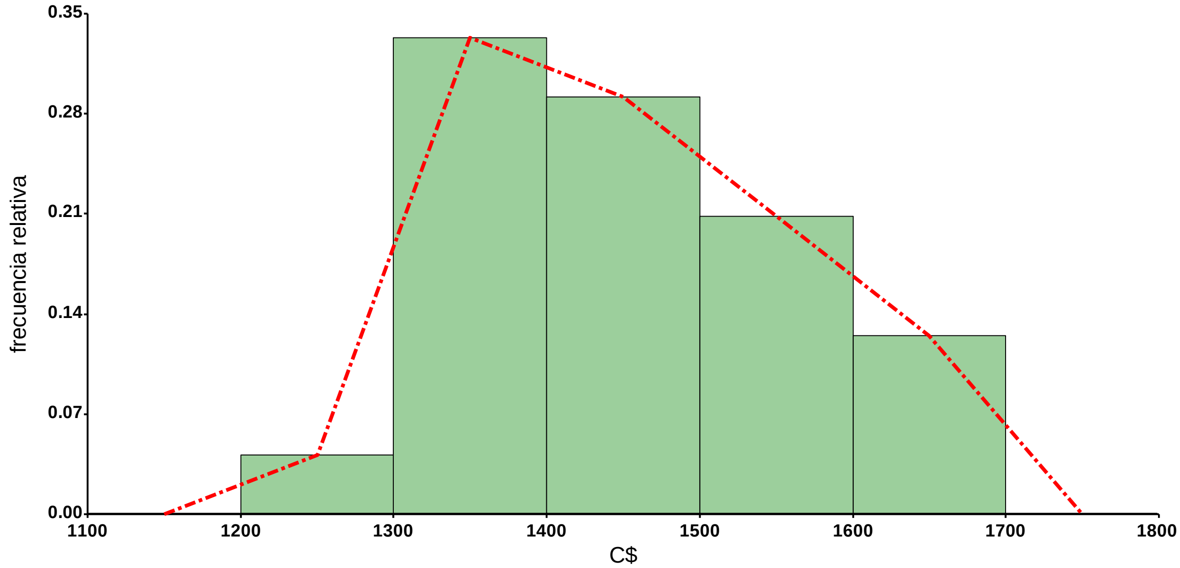
Secuencia de actividades

- Se calcula el Rango de los datos, valor mayor menos valor menor: 1680- 1,260 = 420 C$.
- Ancho de clase: El rango se divide en cuatro, 420/4= 105 C$, se ajusta a 100 C$ y de esta manera el número de clases queda en cinco.
- Se construye los límites inferiores y superiores de cada clase como intervalos semiabiertos,
- Luego se cuentan las frecuencias por clase, esto es la Frecuencia Absoluta
- Se calcula la Frecuencia Relativa (Frecuencia Absoluta / n)
- Se hace Frecuencia Acumulada. que es la suma de las frecuencias absolutas. También se pueden hacer las frecuencias expresadas en porcentajes.

Tabla de Distribución de frecuencias, TDF.

Clase	Límite Inferior Igual a	Lim. Superior Menor a	Marca de clase	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia Acumulada
1	1,200	<1,300	1,250	1	0.04	1
2	1,300	<1,400	1,350	8	0.33	9
3	1,400	<1,500	1,450	7	0.29	16
4	1,500	<1,600	1,550	4	0.17	20
5	1,600	<1,700	1,650	4	0.17	24
			Total	**24**	**1.00**	

Ejemplo de gráfico construido con estos datos



"Histograma y Polígono de Frecuencias Relativas de Ingresos semanales de 24 familias del Barrio Virginia Quintero, Estelí. 2008"

Se puede observar que la información que lleva el gráfico es completa, incluye todos los datos y permite explicar el contenido del mismo por ejemplo: la barra de mayor altura contiene la moda y al no ser un gráfico simétrico concluyo que la media y mediana son diferentes y que los datos son sesgados hay un agrupamiento de frecuencias a la izquierda del centro.

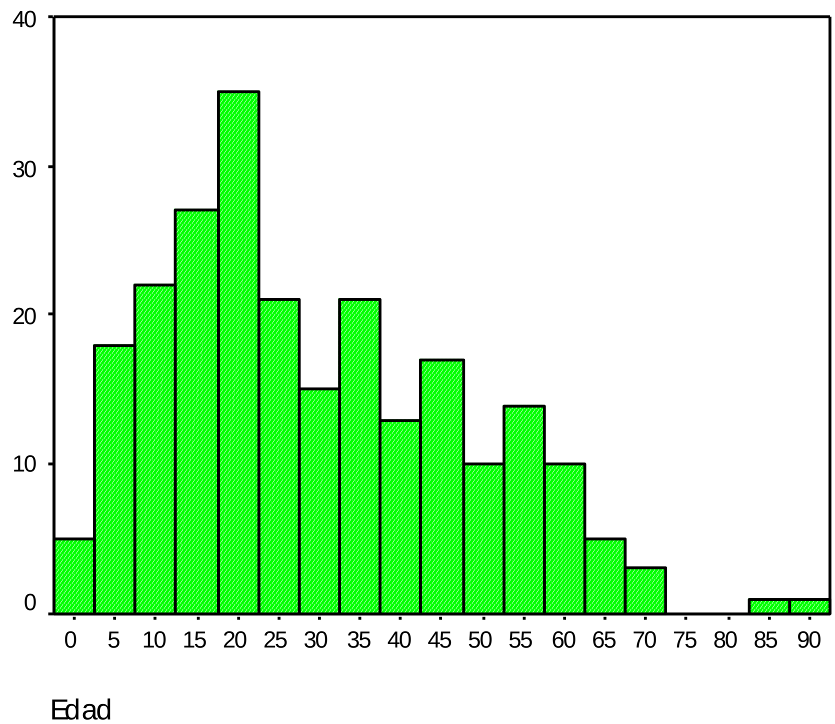
Una manera de representar una distribución de Frecuencias es:

1. Por medio de un gráfico de Barras con variables nominales.
2. Con un Histograma con variables continuas.
3. Un polígono de Frecuencias cuando se quieren mostrar las frecuencias absolutas.
4. Con un gráfico de Pastel cuando se tienen porcentajes o proporciones.

Ejercicio 1.6 Realizar una tabla de frecuencias con una variable discreta (contable) y una variable continua (medible) de la matriz generada con los datos obtenidos en clase.
Ejercicio 1.7. Realizar un gráfico de barras y un gráfico de Pastel a partir de los datos recolectados.

Gráficos. Los gráficos nos permiten presentar la información que san los datos de manera resumida y gráfica, fácil de entender. Los gráficos pueden ser univariados, bivariados y multivariados, según el número de variables involucradas.

Gráficos univariados, Ejemplo de edad de una muestra de personas, datos presentados en forma de Histograma de frecuencias. En este gráfico las barras se encuentran unidas, no habiendo espacio entre las barras. Para su construcción primero se tiene que hacer una tabla de distribución de frecuencias, TDF, donde se precisen los límites reales de frecuencia, que se usan para construir las barras. El centro de cada barra es la "marca de clase", esta medida se usa para construir polígonos.



Histograma de Frecuencias absolutas, de la edad, de una muestra de personas de una comunidad rural del Departamento de Estelí. 2008.

Para describir los datos éste gráfico univariado se acompaña de estadística descriptiva como medias, medianas, desvíos estándares e intervalos de confianza.

“Gráfico de Pastel o Sectores” Ejemplo del nivel de educación, de una muestra de 598 personas de origen rural. Este Gráfico creado con frecuencias y porcentajes, permite resaltar segmentos de clases determinadas.

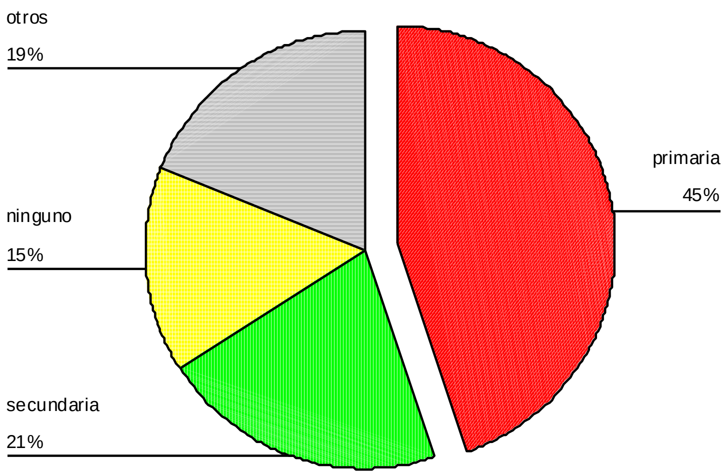
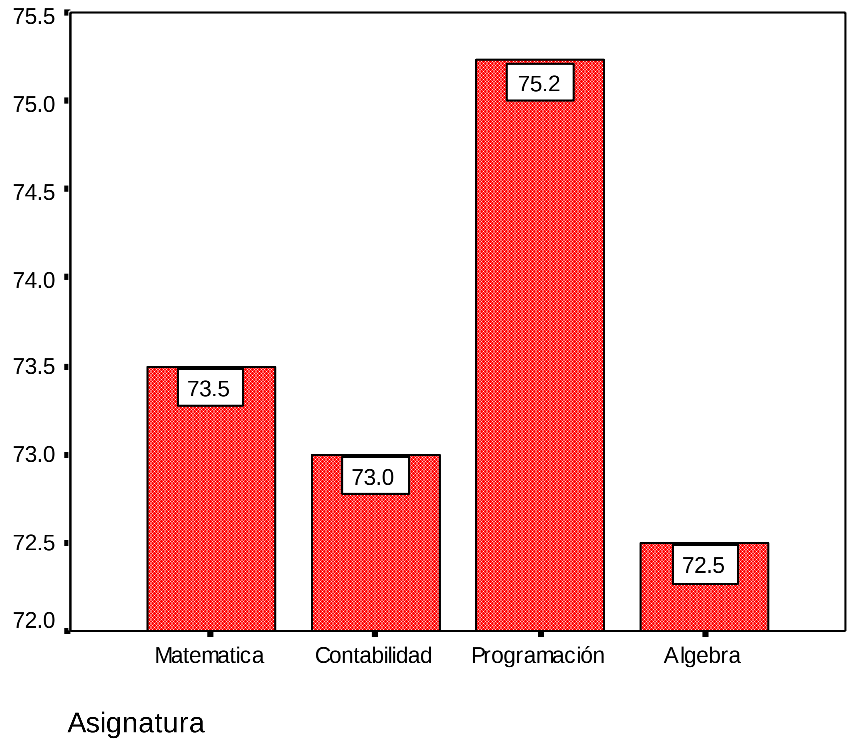


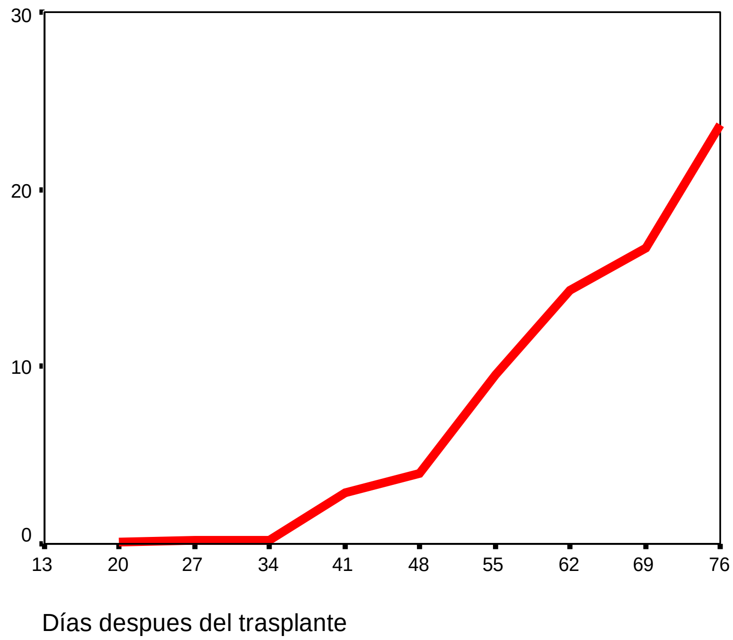
Gráfico de pastel o sectores que representa el nivel educativo alcanzado en un barrio de Estelí.

“Gráfico de Barras bivariado”. Ejemplo de las notas de tres asignaturas presentadas en forma de barras. Este resume la media de notas obtenido por asignatura. Cómo la variable en estudio, Asignatura, no es continúa, entre barra y barra hay un espacio. El gráfico observado a continuación se construyó con una variable nominal, asignatura y una variable continua, nota.



"Polígono de Frecuencias" Este polígono se construye con los valores medio de cada clase, Marca de clase y las frecuencias por clase. En el ejemplo se grafica muestra en el tiempo el desarrollo de una enfermedad, tizón temprano, en el follaje de las platas de tomate.

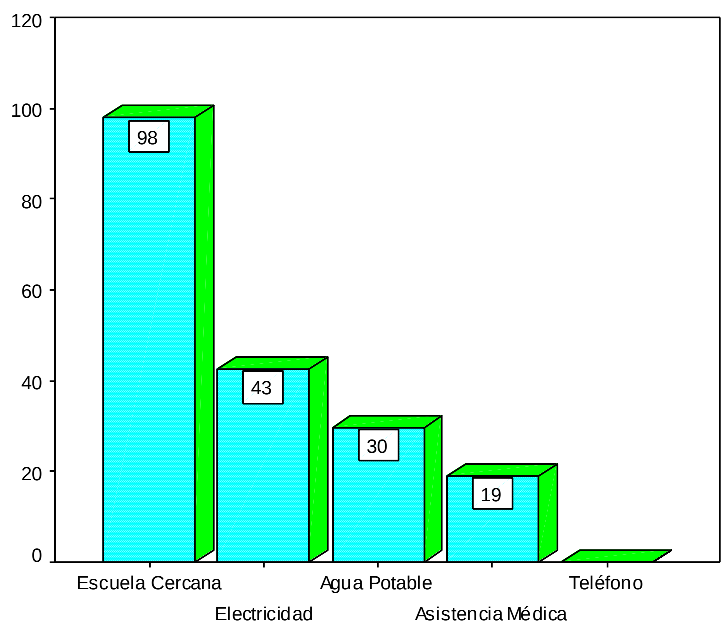
El Polígono es una línea quebrada que se construye uniendo los puntos medios en la parte superior de cada barra, marca de clase de un histograma



Polígono de frecuencias acumuladas en porcentaje, del desarrollo de una enfermedad fungosa, en plantas de tomate.

Gráficos Multivariados. Son gráficos que incorporan 2 o más variables.

Gráfico de Barras que incorpora 4 variables dicotómicas (si- no)



Este tipo de gráfico permite resumir de manera muy eficiente la información de hasta 6 o 7 variables. Es ideal para usar con escalas de opinión como la escala Likert o variables dicotómica, SI y NO.

Gráfico De Barras, Bivariado en Cluster o Agrupamientos

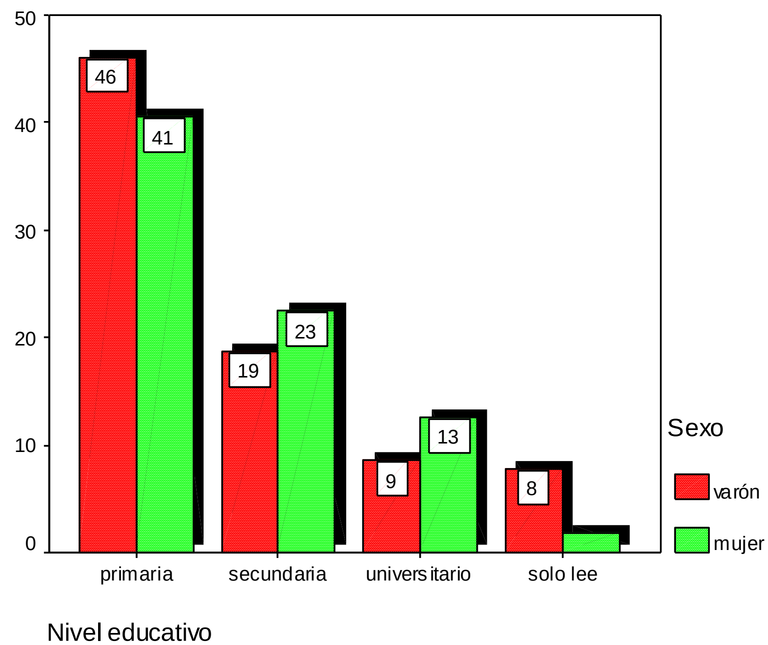


Gráfico bivariado, se puede acompañar de una tabla cruzada de frecuencias y porcentajes con una prueba estadística X^2 de independencia.

Gráfico Bivariado De Barras Apiladas

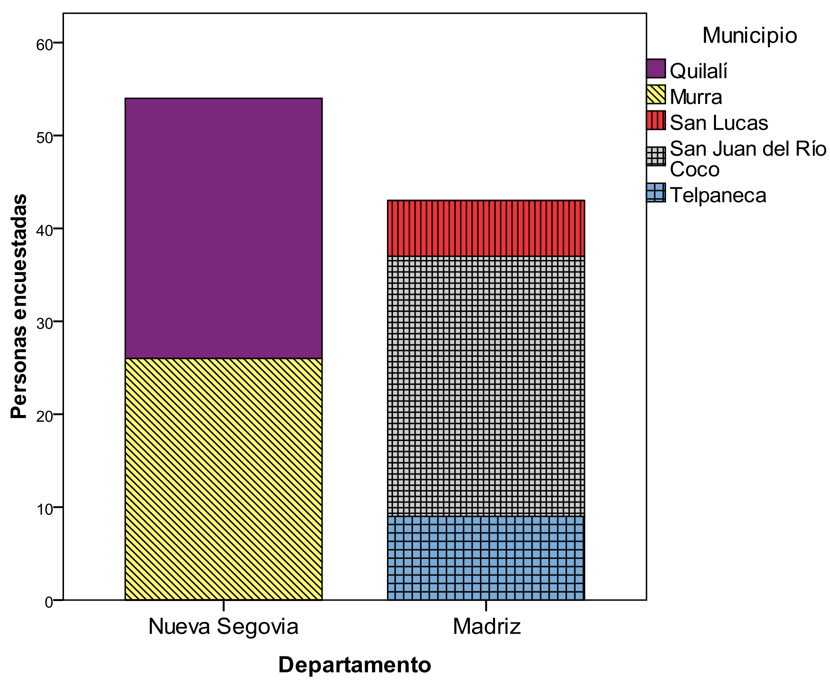
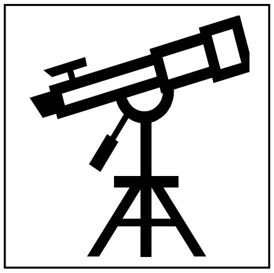


Gráfico bivariado que reduce el número de barras y por lo tanto se simplifica el diseño. Se puede construir con frecuencias o porcentajes



"Un Gráfico permite ver rápidamente lo que dicen los datos"

1.3. Medidas de Tendencia Central

Al forjarnos una imagen mental de la distribución de frecuencias de un conjunto de mediciones, una de las primeras apreciaciones descriptivas de interés es una medida de tendencia central, es decir, una que localiza el centro de la distribución.

Una de las medidas de tendencia central más común y útil es la media común o "media aritmética", pero también son de importancia, según las circunstancias y el tipo de variables la "moda" y la "mediana". Otras medidas de tendencia central menos usadas son la "media geométrica" y la "media cuadrática".

La sumatoria, un concepto básico introductorio:

En matemática, el símbolo Griego "Σ" en mayúscula se utiliza para indicar sumatoria de datos donde:

$$\sum_1^n x_i = \mathbf{x_1} + \mathbf{x_2} + \mathbf{x_3} + \mathbf{x_4} + \mathbf{\ldots\ldots} + \mathbf{x_n}$$

Siendo "x" un valor de una medición de la variable en estudio e "i" un índice que varía de "1 a n ".El número de datos de la muestra se identifica con la letra "n".

Media Aritmética

La media aritmética o simplemente media de un conjunto de mediciones es la medida de tendencia central más usada y conocida. Esta medida se simboliza como $\bar{x}$ (x con raya) cuando representa la media muestral y como μ (letra griega minúscula) para representar la media poblacional. "$\bar{x}$" o "μ" es la suma de todos los valores de la muestra o población divididos por el número de casos. En el caso de la media muestral esta es igual a: "**$\bar{x} = (x_1 + x_2 + x_3 + \ldots + x_n) / n$**" donde "n" es el número de datos de la muestra y "x" el valor numérico del dato. La fórmula simplificada de la media es:

"$\bar{x} = (\sum_{1}^{n} x_i / n)$", donde $\sum$ representa la letra griega sigma que en matemáticas es el símbolo de sumatoria de datos, el subíndice "i" es un valor que varía desde "1" a "n".

Cuando se tienen datos agrupados en una distribución de frecuencias se obtiene el punto medio de cada intervalo y se determina media de la siguiente manera:

$$\bar{x} = \sum_{1}^{n} x\,f \Big/ n$$

Donde "f" es la frecuencia de la clase y "x" el punto medio de cada intervalo.

Una debilidad de la media aritmética es que es sensible a valores extremos de la distribución y que carece de sentido para variables medidas con un nivel nominal u ordinal.

Media Aritmética	$\boldsymbol{\bar{x} = \frac{\sum_{1}^{n} x}{n}}$

Ejemplo de cálculo de una media.

Si tengo la nota de un examen de matemáticas de 10 estudiantes en una escala de 1 a 100 donde:

Estudiante	**"Variable Nota = x_i"**	**Valor de x_i**
Luis	X_1	62
Alberto	X_2	68
Juan	X_3	92
Pedro	X_4	88
Roberto	X_5	55
María	X_6	79
Raquel	X_7	89

Luisa	X_8	92
Rosa	X_9	67
Diana	X_{10}	69
	$\sum_1^{10} x_i =$	761.

En este caso "i" varia de 1 a 10.

Media de notas de los estudiantes = $\sum_1^{10} x_i /10 =$ 761/10 = **76.1**

La Mediana

La segunda medida de tendencia central es la mediana. La mediana "m" de un conjunto de mediciones "x_1, x_2, x_3,...., x_n" es el valor de "x" que se encuentra en el punto medio o centro cuando se ordenan los valores de menor a mayor.

Si las mediciones de un conjunto de datos se ordenan de menor a mayor valor y "n" es impar, la mediana corresponderá a la medición con el orden "(n + 1) / 2". Si el número de mediciones es par, n = par, la mediana se escoge como el valor de "x" a la mitad de las dos mediciones centrales, es decir como el valor central entre la medición con rango "n/2" y la que tiene rango "(n/2) + 1".

Reglas para calcular la mediana

- Ordenar las mediciones de menor a mayor
- Si "n" es impar, la mediana "m" es la medición con rango "(n + 1) / 2"
- Si "n" es par, la mediana "m" es el valor de "x" que se encuentra a la mitad entre la medición con rango "n / 2" y la medición con rango "(n /2)+1".

datos de menor a mayor:

Estudiante	**"Datos ordenados"**	**Valor de x_i**
Roberto	1	55
Luis	2	62
Rosa	3	67
Alberto	4	68
Diana	5	69
María	6	79
Pedro	7	88
Raquel	8	89

Juan	9	92
Luisa	10	92

Como “n” es par, la mediana es igual a la mitad entre la medición con rango “n / 2” y la medición con rango “(n/2) +1”, donde n / 2 = 5 y (n /2) +1)= 6.

El dato 5 vale 69 y el dato 6=79, entonces “la mediana” es igual a 69 + 79 / 2= **74**

En este ejemplo la mediana es semejante a la media.

La Moda

La moda es la medida de tendencia central más fácil de calcular y también es la más sujeta a fluctuaciones cuando cambian unos pocos valores de la distribución. Por esta razón la moda se suele usar para una evaluación rápida de la tendencia central. La moda se define como “el valor más frecuente de una distribución”. En una tabla de frecuencias, la frecuencia mayor es la que contiene a la moda. Esta medida se usa más y tiene más sentido cuando se describen datos nominales, de hecho es la única medida de tendencia central que funciona con este tipo de escala.

La moda es el valor más frecuente y funciona bien con escalas nominales

Comparaciones entre las diferentes medidas.

Las tres medidas de tendencia central, la media, mediana y moda, no son igualmente útiles para obtener una medida de tendencia central. Por el contrario, cada una de estas medidas tiene características que hacen que su empleo sea una ventaja en ciertas condiciones y en otras no.

La media es la medida de tendencia central, generalmente más usada y tiene la característica que incorpora todos los datos de la variable en su cálculo por lo tanto su valor suele ser más estable.

La mediana suele ser la medida preferida cuando se emplea una escala ordinal, estas son las situaciones donde el valor asignado a cada caso no tiene otro significado más que el indicar el orden entre los casos. Por ejemplo saber en una clase cuales alumnos están dentro del 50% con mejores notas y cuales dentro del 50% con peores notas. También se suele preferir la mediana cuando unos pocos valores extremos distorsionan el valor de la media. Por ejemplo si tengo 9 personas con 0 ingresos y uno sola que tiene ingresos de 10 unidades, la media me puede dar a entender que la mayoría recibe 1 unidad, cuando esto no es real.

La moda en ciertas condiciones puede ser la más apropiada, por ejemplo cuando se quiere información rápida y cuando la precisión no sea un factor especialmente importante. En ciertos casos solo esta medida tiene sentido por ejemplo en un equipo de fútbol llevo la estadística por jugador (escala ordinal) de la cantidad de pases que realiza por juego, esto para detectar quien es el que mejor distribuyendo la pelota, en este caso la media y la mediana no tendrían significado, solo la moda.

Un aspecto interesante entre las tres medidas es su comportamiento referente a la simetría que toma una distribución. Cuando las distribuciones son simétricas, <u>sin sesgo</u>, caso de la distribución Normal que tiene forma de campana, "la media, la mediana y la moda coinciden". Si la distribución es asimétrica con <u>sesgo positivo</u>, hay más datos hacia la izquierda de la media, entonces "la media es mayor que la mediana y esta mayor que la moda". Si ocurre lo contrario, el sesgo <u>es negativo</u>, entonces "la media es menor que la mediana y ésta menor que la moda".

Otras medidas de tendencia central.

La Media Geométrica.

La media geométrica se define como $\bar{x}_g = \sqrt[n]{x_1 x_2 x_3 .. x_n}$, por ejemplo la media geométrica de los valores "4, 5, 4, 6" es $\bar{x}_g = \sqrt[4]{(4)(5)(4)(6)} = 4.68$

Una ventaja de su uso es que considera todos los valores de la distribución y es menos sensible que la media aritmética a los valores extremos, sin embargo es de cálculo complicado y si un valor vale 0 se anula.

La Media Cuadrática.

Se construye a partir de suma de los cuadrados de un conjunto de valores. Su forma de cálculo es $\bar{x}_c = \sqrt[2]{\frac{x_1^2 + x_2^2 + x_3^2 + ... + x_n^2}{n}}$, si tomamos los valores anteriores la media cuadrática tiene el siguiente valor $\bar{x}_c = \sqrt[2]{\frac{4^2 + 5^2 + 4^2 + 6^2}{4}} = 4.81$

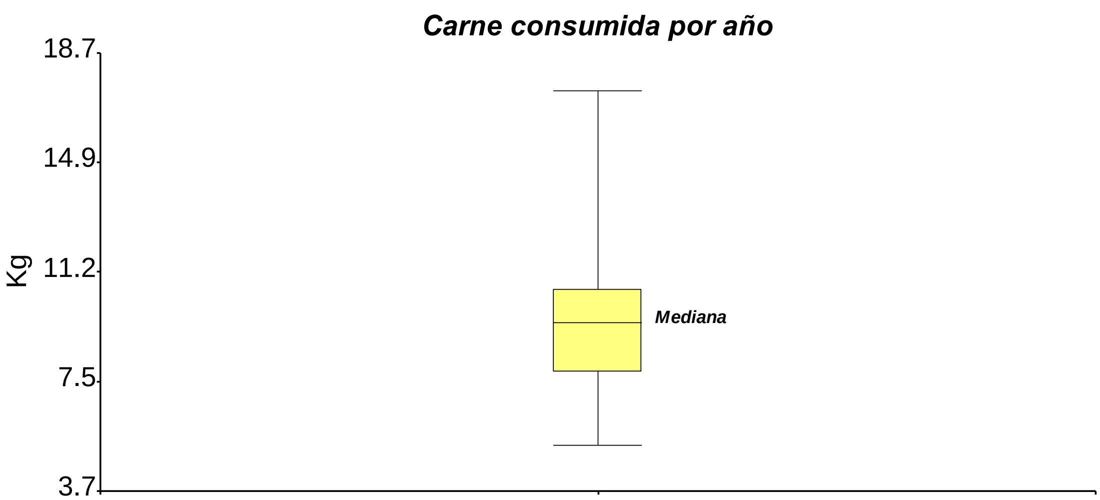
Se utiliza cuando se quiere evitar los efectos de los signos. Ésta media solo puede tomar valores positivos.

Cuartiles, Deciles y Percentiles.

Cuartiles: si a un conjunto de datos se ordena de mayor a menor, el valor central es la mediana, este valor divide el grupo, en dos subgrupos cada uno con el 50 % de los datos. Si a cada subgrupo ordenado se le marca el valor central, tenemos así tres valores seleccionados que llamaremos Cuartiles, Q_1, Q_2 y Q_3. Estos valores dividen al conjunto de datos en cuatro grupos con igual número de términos, cada cuartil contiene el 25% de los datos. La mediana es el cuartil dos, Q_2. Con los Cuartiles se construye un gráfico especial, "el diagrama de caja", este permite visualizar la variabilidad de los datos por Cuartil.

En el diagrama de caja, el centro de la caja es el Q_2, la mediana, los bordes de la caja son el Q_1 y el Q_3. En los extremos del diagrama se trazan dos rayas horizontales que representan los valores máximo y mínimo de la distribución y que no se consideran anómalos. Para hallar los valores de las rayas se multiplica la amplitud inter cuartil (Q_3 - Q_1) por 1,5 y el resultado se suma a Q_3 y se resta a Q_1.Por último, por encima y por debajo de las rayas se representan de forma individual los valores extremos y anómalos de la distribución.

Diagrama de caja, variable: cantidad de carne consumida por año.



<u>Deciles</u>, si el conjunto de valores, ordenados de de mayor a menor, se dividen en diez partes iguales, los valores que dividen los datos se llaman deciles y son nueve, D_1, D_2,..D_9.

<u>Percentiles</u>, si se tiene un conjunto de datos muy numerosos y a este se lo divide en 100 partes iguales, cada valor que divide los datos se llama percentil, P_1, P_2, P_3…P_{99}.

1.4 Medidas de Dispersión o de Variabilidad

Las medidas de variabilidad indican la dispersión de los datos en la escala de medición y son tan importantes como las medidas de tendencia central y así como éstas son valores puntuales en una distribución, las medidas de dispersión son “intervalos”, distancias o un número de unidades en la escala de medición. Este tipo de medida se complementa con las medidas de centralidad y ambas permiten describir a la mayoría de las distribuciones. Los tipos de medidas de Dispersión más comunes son: “el Rango”, “el Desvío Estándar” y la “Varianza”.

El Rango.

El Rango, Recorrido o Amplitud de un conjunto de mediciones, “es la diferencia entre el valor mayor y el valor menor”, indica el número necesario y mínimo de unidades, en la escala de medición, para incluir los valores mínimo y máximo. Es la medida de dispersión más fácil de calcular, pero también es la menos estable al estar fuertemente influenciada por valores extremos atípicos.

Cuanto más grande es el rango, mayor será la dispersión de los datos de una distribución. Es adecuada para medir la variación de pequeños conjuntos de datos.

El Desvío Estándar.

El Desvío Estándar es la medida de dispersión más ampliamente usada y es la más estable ya que depende de todos los valores de la distribución. Es la media de desviación de los valores con respecto a la media, aunque una definición completa sería: “la raíz cuadrada de la suma de las desviaciones alrededor de la media, elevadas al cuadrado y divididas entre el número de casos menos uno” en el caso de “S”.

Desvío Estándar “S”: la raíz cuadrada de la suma de las desviaciones alrededor de la media, elevadas al cuadrado y divididas entre el número de casos menos uno.

Cuando se trabaja con muestras el desvío estándar se simboliza con una “S” y con la letra sigma minúscula “σ” cuando se usan datos de una población. Su fórmula de cálculo tradicional es:

$$\sigma = \sqrt{\sum_{1}^{N} \frac{(x_i - \mu)^2}{N}}$$

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

Donde i es cualquier valor de “uno” a “n o N”, y “n” es el número total de datos de la muestra y “N” de la población.

El desvío estándar, “S” o “σ”, se interpreta como cuanto se desvía de la media un conjunto de valores. Este valor se grafico como un intervalo. Esta medida solo se utiliza con variables continuas u ordinales.

Cálculo del desvió estándar “S” por suma de cuadrados, para datos no agrupados.

El desvió estándar se puede expresa también de la siguiente manera:

$$S = \sqrt{\frac{\sum_1^n x^2 - \frac{(\sum_1^n x)^2}{n}}{n-1}}$$

Esta forma de resolución es equivalente a la forma de cálculo tradicional, es de más fácil resolución cuando se tiene calculadoras de mano que hacen sumas de cuadrados.

Cálculo del desvió estándar “S” para datos agrupados

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

Donde “x_i” es la marca de clase “i”, “k” en el número de clases y “n” en número total de datos.

Ejemplo de cálculo de Desvío Estándar "S", datos no agrupados

Con el ejemplo de las notas de matemáticas haremos cálculo de "S"

"S"= $$\sqrt{\begin{array}{l}((55-76.1)^2+(62-76.1)^2+(67-76.1)^2+(68-76.1)^2+(69-76.1)^2+(79-76.1)^2+ \\ (88-76.1)^2+(89-76.1)^2+(92-76.1)^2+(92-76.1)^2)/9\end{array}}$$

= 13.6

Se sugiere hacer estos cálculos usando una calculadora científica en función estadística.

La Varianza.

La varianza es el desvío estándar elevado al cuadrado y se simboliza con "S^2" cuando es muestral, o "σ^2 cuando es poblacional. Este es una medida que se usa en muchas pruebas de Hipótesis estadísticas inferenciales, por ejemplo "el Análisis de Varianza, ANDEVA". Pero para fines descriptivos se prefiere usar el desvío estándar en vez de la varianza, que suele ser un valor mayor y difícil de interpretar.

El Coeficiente de variación

El coeficiente de variación, CV, es un cociente entre el desvío estándar y la media de los datos, expresado en porcentaje, $\mathrm{CV} = \left(S/_{\bar{X}} \right) 100$. Este coeficiente permite comparar la variabilidad de diferentes muestras en una misma variable ó la variabilidad existente entre variables diferentes. Una investigación experimental en el campo agropecuario que tenga un CV menor al 10 %, muestra que en el experimento hubo un muy buen control del error experimental entre las diferentes repeticiones, sin embargo en procesos productivos industriales éste valor de variabilidad en una variables de salida, sería muy alto, en general se aceptan valores muy pequeños, inferiores al 1%.

Interpretación de las medidas de tendencia central y de la variabilidad.

Cabe destacar que al describir nuestros datos, debemos interpretar nuestros datos de tendencia central y de variabilidad en conjunto y no de manera separada. Con la media y el desvío estándar se pueden construir intervalos donde supongo están la mayoría de los datos. La moda, mediana y el rango pueden completar la información sobre los datos y así tener una buena idea de lo que sucede con la variable en estudio.

En una variable continua:

- *La media, la mediana y la moda son puntos en una recta.*
- *El desvío estándar y el rango son intervalos.*

1.5 Otras medidas útiles en Estadística Descriptiva.

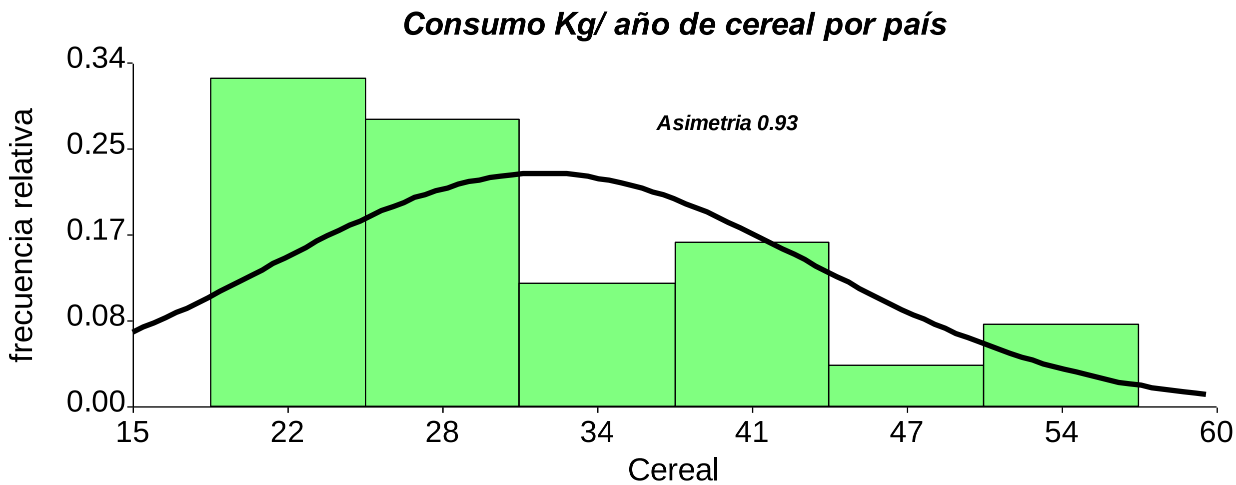
Cuando los polígonos de frecuencia de una variable se presentan en forma de curva hay dos medidas esenciales para describir estas curvas: “La Asimetría” y la “Curtosis”.

La Asimetría o Sesgo.

La Asimetría es una medida necesaria para conocer cuánto se parece nuestra distribución a la distribución teórica de una “curva normal”, curva con forma de campana, y constituye un indicador del lado de la curva donde se agrupan las frecuencias. Esta medida se construye con el valor medio, la mediana y el desvió estándar. Si el valor del sesgo es cero (asimetría = 0), la curva de distribución es simétrica, en este caso coinciden los valores de la media, la mediana y la moda. Cuando el sesgo es positivo, la media es mayor que la mediana, quiere decir que hay valores agrupados hacia la izquierda de la curva y la cola de la distribución es más larga a la derecha. Cuando el sesgo es negativo, la media es menor a la mediana, significa que los valores tienden a agruparse hacia la derecha de la curva, por encima de la media y la cola de la distribución es más larga a la izquierda.

Su forma de cálculo original es: $\boldsymbol{Sesgo} = \frac{\boldsymbol{3(\bar{x} - Moda)}}{\boldsymbol{s}}$ pero como aproximadamente se cumple que “Media – Moda = 3 (Media - Mediana)”, se usa la siguiente forma de cálculo práctico del sesgo:

$$Sesgo = \frac{3(\bar{x} - Me)}{S}$$



Histograma de consumo de cereal en Kg/ año por habitante de diferentes países. En este gráfico se observa una asimetría o sesgo positivo de 0.93, hay un agrupamiento de datos a la izquierda de la curva de distribución normal, curva en color negro.

La Curtosis.

La curtosis es una medida que indica o mide lo plano o puntiaguda que es una curva de distribución. Cuando esta es cero, curtosis = 0, significa que se trata de una curva Normal. Si es positiva, quiere decir que la curva o distribución o polígono es más puntiaguda o levantada que la curva normal (curva leptocúrtica). Si es negativa quiere decir que es más plana (curva mesocúrtica).

$$\textbf{Curtosis} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4 / n}{S^4}$$

Ejercicio 1.8:

Tomando como fuente de datos las variables continuas recolectadas a partir de los datos que generen los estudiantes en clase deben construir :

- medidas de tendencia central: medias, modas, medianas.

- medidas de dispersión: desviación estándar y rango.
- distribución de frecuencias.
- espacios: $\bar{X} \pm 2$ "S" y determinar cuantos datos entran en este intervalo.
- gráficos de barras, histogramas y gráficos de pastel.

Capítulo II. Muestras y Población.

Objetivos

- Explicar principios básicos de muestreo con ejemplos cotidianos.
- Diferenciar las diferentes formas de realizar muestreos que permitan estudiar el contexto social y productivo.
- Aprender a calcular de forma ordenada el tamaño de una muestra con variables construidas en el aula.

Llamaremos **población** a un conjunto homogéneo de elementos en el que se estudia una característica dada. El censo es la forma de estudio de todos los elementos de una población. Frecuentemente no es posible estudiar toda la población ya que suele ser económicamente inviable o llevar tanto tiempo que es impracticable.

Como generalmente no se puede estudiar la población, se selecciona un conjunto representativo de elementos de esta, que llamaremos **muestra**. Cuando la muestra está bien escogida podemos obtener información de la población similar a la de un censo, pero con mayor rapidez y menor costo.

La clave de un procedimiento de muestreo es garantizar que la muestra sea representativa de la población. Por lo tanto cualquier información al respecto de las diferencias entre sus elementos debe tenerse en cuenta para seleccionar la muestra, esto origina diferentes tipos de muestreo, los cuales se describen a continuación.

2.1 Muestreo Aleatorio Simple

Es la manera más sencilla de hacer muestreo. Decimos que una muestra es aleatoria cuando:

- Cada elemento de la población tiene la misma probabilidad de ser elegido.

- La población es idéntica en todas las extracciones de muestreo. Esta característica es irrelevante si el tamaño de la población (N) es grande en relación al tamaño de la muestra (n).

El muestreo aleatorio simple debe utilizarse cuando los elementos de la población son homogéneo respecto a las características a estudiar, es decir a priori no conocemos que elementos de la población tendrán valores altos de ella. El primer problema al aplicar esta forma de muestreo, es calcular el "n", número de de elementos de la muestra.

Cálculo de "n" Gráficamente: Se sabe que a más grande la muestra mejor ésta estima la media de la población, sin embargo hay un momento que la media que se calcula a partir de la muestra casi no cambia, aunque ésta aumente de tamaño, en ese momento el tamaño de la muestra comienza a ser óptimo.

Esta estabilidad de medias se puede observar gráficamente con un gráfico de medias. La primera media de este gráfico se hace con un dato de la población, el segundo con dos datos, el tercero con tres datos y así sucesivamente, hasta que en el gráfico las medias casi no fluctúen entre muestra y muestra. A continuación se muestra un ejemplo de 15 datos de notas que obtuvieron 15 estudiantes en la asignatura de Física. En la fila tercera se calcularon las medias consecutivos, con un dato, dos datos, tres datos… hasta 15 datos. Se observa que a partir de 10 datos, la media se estabiliza en el valor 75, el valor de "n", tamaño de muestra para esta variable estaría entre 11 y 12 datos.

72	68	82	88	65	79	89	92	67	69	75	79	71	78	75
$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	$\bar{x}_4$	$\bar{x}_5$	$\bar{x}_6$	$\bar{x}_7$	$\bar{x}_8$	$\bar{x}_9$	$\bar{x}_{10}$	$\bar{x}_{11}$	$\bar{x}_{12}$	$\bar{x}_{13}$	$\bar{x}_{14}$	$\bar{x}_{15}$
72	70	74	77	75	76	78	79	78	77	77	77	77	77	77

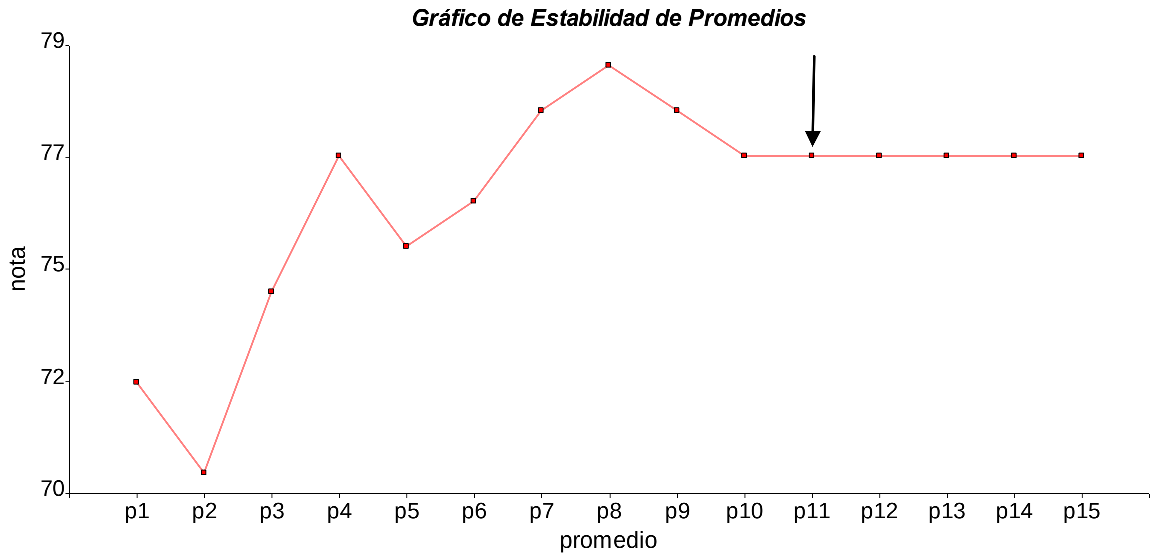


Gráfico de Estabilidad de Medias

Cálculo empírico para lotes y atributos. Si en un proceso industrial se tienen lotes, volumen de producción por tiempo ó por cantidad de materia prima, para realizar un muestreo del proceso productivo por atributos, por ejemplo artículos sanos ó defectuosos, se puede seguir el siguiente criterio, el cual es una adaptación resumida del método desarrollado por el ejército de EEUU en su norma Military Standar 414.

Tabla sobre el número de piezas a muestrear

Tamaño del Lote	**% de piezas de la muestra**
60-300	10
301-1000	5
1001-5000	2
+ 5000	1

2.2 Muestreo Estratificado

Se denomina muestra estratificada aquél en que los elementos de la población se dividen en clases o estratos. La muestra se toma asignando un número o cuota de miembros a cada estrato y escogiendo los elementos por muestreo aleatorio simple dentro del estrato.

Cuando dispongamos de información sobre la población conviene tenerla en cuenta al seleccionar la muestra. Un ejemplo clásico son las encuestas de opinión, donde los elementos (personas) son heterogéneas en algunas variables como: sexo, edad, profesión, etc. Interesa en estos casos que la muestra tenga una composición análoga a la población, lo que se consigue mediante una muestra estratificada. En concreto si existen “k” estratos de tamaño $N_1...N_k$ y tales que “$N = N_1 + N_2 +....+ N_k$” se tomará una muestra “n” que garantice una presencia adecuada de cada estrato “n_i”.

Una forma sencilla para dividir el tamaño total de la muestra “n” entre los estratos de “n_i” es por el Método de Asignación Proporcional, el cual toma en cuenta el tamaño relativo del estrato de la población, por ejemplo si en la población hay un 55 % de mujeres y un 45 % de hombres, mantendremos esta proporción en la muestra. En general se hará de la manera **“$n_i= n\ N_i/N$”.**

2.3 Muestreo por Conglomerados

Existen situaciones donde ni el muestreo aleatorio simple ni el estratificado son aplicables, ya que no disponemos de una lista con el número de elementos de la población ni en los posibles estratos. En estos casos típicamente los elementos de la población se encuentran de manera natural agrupados en conglomerados, cuyo número es conocido, por ejemplo la población rural se distribuye en comunidades y los habitantes de un barrio en manzanas. Si suponemos que cada uno de estos habitantes es parte de un conglomerado que pertenece a una población total de conglomerados semejantes para una variable dada, podemos seleccionar algunos conglomerados al azar y dentro de ellos analizar a todos sus elementos o una muestra aleatoria simple. Este método se conoce como muestreo por conglomerados y tiene la ventaja de simplificar la recogida de la información muestral, no es necesario visitar todos los conglomerados para recolectar una muestra. El inconveniente obvio es que si los conglomerados son heterogéneos entre sí, cómo se analizan solo algunos de ellos la muestra final puede ser no representativa de la población, algo así sucede si estudio a fondo una comunidad en lo referente a un opinión dada y supongo que los resultados son representativos de un conjunto de comunidades, pero si esta comunidad estudiada tiene opiniones distintas del resto, los resultados no serán representativos de la población, por ejemplo las comunidades más ricas suelen tener opinión diferente a las más pobres respecto a la ayuda social que da el estado

En resumen las ideas de estratificación y de conglomerados son opuestas, la estratificación funciona tanto mejor cuanto mayor sean las diferencias entre los estratos y más homogéneas sean estos internamente. Los conglomerados funcionan si hay poca diferencia entre ellos y son muy heterogéneos internamente, que incluyan toda la variabilidad de la población en el conglomerado.

2.4 Muestreo Sistemático

Cuando los elementos de la población están en una lista o un censo, se puede utilizar el muestreo sistemático. Supongamos que tenemos una población de tamaño “N” y se desea una muestra de tamaño “n” y sea “K” un valor entero más próximo a la relación “n/N”. La muestra sistemática se toma eligiendo al azar, con números aleatorios, un elemento entre los primeros “K” elementos y se denomina “n_1”. El muestreo se realiza seleccionando los elementos “(n_1 + K); (n_1 + 2 K), etc.” a intervalos fijos de “K” hasta completar la muestra. Si el orden de los elementos en la lista es al azar, este procedimiento es equivalente al muestreo aleatorio simple, aunque resulta más fácil de llevar a cabo sin errores.

Si el orden de los elementos es tal que los más próximos tienden a ser más semejantes que los alejados, el muestreo sistemático tiende a ser más preciso que el aleatorio simple al cubrir más homogéneamente toda la población.

El muestreo sistemático puede utilizarse conjuntamente con el estratificado para seleccionar le muestra dentro de cada estrato.

La regla general que se aplica a los procedimientos de muestreo es que: "cualquier información previa debe utilizarse para asegurar mayor representatividad de la muestra".

Ejercicio 2.1:

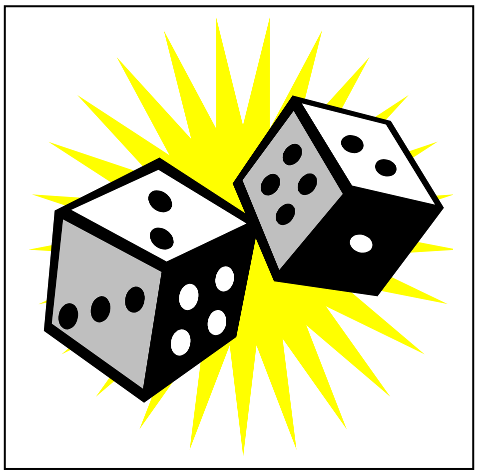
- Suponga que quiere conocer la opinión de una comunidad donde hay 50 personas adultas, N = 50. ¿Cuál es la es tamaño de "n" mínimo a calcular?
- ¿Cuál sería el valor de "n" con una ciudad de 50,000 habitantes?
- Discuta que método de muestreo usaría si quiere estudiar la opinión de la gente de 12 barrios semejantes en cuanto a su nivel de vida y forma de de generar sus ingresos.

Capítulo III. Teoría Elemental de Probabilidades

Objetivos

- Definir conceptos básicos de probabilidad a partir de situaciones cotidianas.
- Explicar las Reglas de Adición y Multiplicación de probabilidades en la resolución de problemas observables.
- A través del trabajo en equipo, valorar la importancia de utilizar probabilidad condicional al describir situaciones de nuestro entorno.

3.1 Introducción a las Probabilidades



Con esta teoría se estudian fenómenos naturales con el fin de descubrir regularidades en la ocurrencia de los mismos. Esta ciencia comenzó a desarrollarse en la Francia Monárquica cuando los aristócratas se preocuparon en el estudio de los juegos de azar, dados, cartas, ruletas, etc. Sin embargo, hoy día, sus aplicaciones abundan en las diferentes ciencias, por ejemplo su teoría se usa en el diseño de modelos de mejoramiento genético, análisis de experimentos, predicciones del tiempo, predicción de vida útil de un equipo, etc. En nuestra vida diaria aplicamos inconscientemente probabilidades cuando compramos un billete de lotería o llevamos un paraguas cuando vemos el cielo nublado.

3.2 Términos Básicos.

Experimento aleatorio: Es el proceso que permite obtener una o varias observaciones, de los cuales no se puede predecir de antemano su resultado.

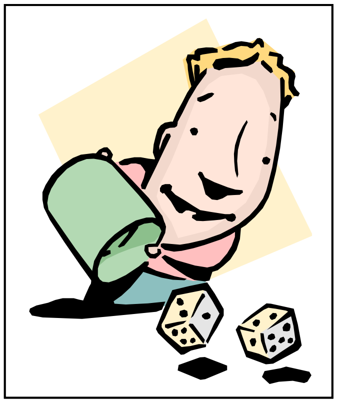
Espacio Muestral “S” ó “Ω”: Todos los posibles resultados de un experimento.

Evento “A”: Algún resultado del experimento que nos interesa.

Ejemplo: Experimento: tirar un dado.

Espacio muestral **"Ω"**= (1, 2, 3, 4, 5, 6)

Evento "A" = sale 3.



Probabilidades, definición Clásica:

Si la probabilidad de un evento "A" se define como la frecuencia relativa de "A" en el espacio muestral **"Ω"**y se denota como P(A).

P(A) = # casos favorables A / # casos Totales de Ω

Es la definición más antigua y se atribuye al matemático francés Pierre Laplace (1749-1827); también se conoce con el nombre de probabilidad a priori pues, para calcularla, es necesario conocer, antes de realizar el experimento aleatorio, el espacio muestral y el número de resultados o sucesos elementales que entran a formar parte del suceso.
La aplicación de la definición clásica de probabilidad puede presentar dificultades de aplicación cuando el espacio muestral es infinito o cuando los posibles resultados de un experimento no tienen iguales probabilidades. Ej: En un proceso de fabricación de artículos puede haber algunos defectuosas y si queremos determinar la probabilidad de que uno que sea defectuoso, no podemos utilizar la definición clásica pues necesitaríamos conocer previamente el resultado del proceso de fabricación. Para resolver estos casos, se hace una extensión de la definición de probabilidad, de manera que se pueda aplicar con menos restricciones, llegando así a la definición frecuentista de probabilidad.

Probabilidades, definición frecuencial:

$$P(A) = \lim_{n \to \infty} \frac{n(A)}{n}$$

La definición frecuencial consiste en definir la probabilidad de la proporción o frecuencia relativa del suceso como el límite cuando “n” tiende al infinito. Es imposible llegar a este límite, ya que no podemos repetir el experimento un número infinito de veces, pero si podemos repetirlo muchas veces y observar como las frecuencias relativas tienden a estabilizarse.

Esta definición frecuentista de la probabilidad se llama también probabilidad a posteriori ya que sólo podemos dar la probabilidad de un suceso después de repetir y observar un gran número de veces el experimento aleatorio correspondiente.

Por ejemplo si en una región hay 640 campesinos que siembran frijol de forma manual y 160 con bueyes. En este caso hay 2 eventos: Siembra manual y Siembra con bueyes y existen las probabilidades, P (bueyes) y la P (manual) asociados a la frecuencia de ocurrencia de cada evento. La probabilidad que al elegir una parcela al azar esta fue sembrada con bueyes, P (bueyes) es de 160/800 = 0.20 ó 20 %.

Ley de los Grandes Números.

En un experimento aleatorio, en la medida que aumenta “n”, la frecuencia relativa de un suceso tiende a estabilizarse y se aproxima cada vez más un número fijo que es su probabilidad teórica.

3.3 Propiedades de la Probabilidad

- $0 \leq P(A) \leq 1$
- El evento A es más probable que B $\Leftrightarrow P(A) \geq P(B)$
- Un Evento cierto, que seguramente ocurre, tiene probabilidad 1.
- Un Evento imposible, que nunca ocurrirá, tiene probabilidad 0.
- Tiene dos reglas básicas que la estructuran: la regla del producto y la regla de la suma.

Regla del producto.

Si dos evento "A" y "B" son independientes si "A" no influye de ninguna manera en "B" y viceversa. La probabilidad que los eventos independientes "A" y "B" ocurran al mismo tiempo es P(A y B) = P (AB) = P(A) x P (B)= $P(A \cap B)$, A intersección B.

Por ejemplo si la Probabilidad de un nacimiento de un niño es 0.5, P (niño) = 0.5, la probabilidad que dos mujeres su primer parto tengan hijos varones es un evento independientes, uno no influye sobre otro, la P (niño, niño) es de "0.5^2 = 0.25"
Una paradoja es que una persona que compra todas las semanas la lotería, para un sorteo dado, tiene la misma probabilidad de sacar el premio mayor que una persona que compró un número por primera vez.

Ejercicio 3.1: estimar la probabilidad que al elegir por sorteo dos estudiantes del grupo, ambos sean varones. Determinar también cuales eventos forman "Ω"es este caso.

Regla de la Suma.

Para que dos eventos "A" y "B" se puedan sumar directamente, estos deben ser incompatibles, es decir ellos no pueden ocurrir al mismo tiempo $P(A \cap B) = 0.$

La probabilidad que ocurra "A" ó "B" para eventos incompatibles "A" y "B" es P(A ó B) = $P(A) + P(B) = P(A \cup B)$.

Si los eventos <u>no</u> son incompatibles $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Esta sería la regla general de la suma de probabilidades.

En el ejemplo de arrojar dos veces una moneda al aire, la probabilidad que salga una vez cara y el otro sol sin importar el orden, es la probabilidad de los eventos "cara, sol" y "sol, cara". Debido a que son cuatro los eventos posibles "Ω"= cara –cara, sol – cara, cara – sol y sol-sol y cada uno con igual probabilidad, cada uno de esto eventos tiene una P = 0.25, de ocurrencia. Por lo tanto la ocurrencia de "cara-sol" más "sol – cara" es de “P (c, s) + P (s, c)”), que en valore de probabilidades es de **P (0.25) + P (0.25) = 0.5**

Ejercicio 3.2. En la matricula de primer año de la escuela secundaria, 150 estudiantes son originarios del departamento de Estelí, 60 estudiantes del departamento de Nueva Segovia y 100 estudiantes del resto del país. ¿Cuál es la probabilidad que un estudiante tomado al azar no sea del departamento de Estelí?

3.4 Probabilidad condicionada

Como la probabilidad está ligada a nuestra ignorancia sobre los resultados de la experiencia, el hecho de que ocurra un suceso, puede cambiar la probabilidad de los demás que ocurran luego. El proceso de realizar la historia de un caso, explorar y realizar pruebas complementarias ilustra este principio, cuando más se conoce de lo que ocurrió, mejor puedo predecir el futuro, la “probabilidad condicionada” se nutre de este principio.

La probabilidad de que ocurra el suceso A si ha ocurrido el suceso B, “P (A\B)”, se denomina probabilidad condicionada y se define.

$$P(A \backslash B) = \frac{P(A \cap B)}{P(B)} \quad \text{Si p (B)} \neq 0$$

La condición que P (B) > 0, esto es necesario para una buena definición de probabilidad condicional. Es de notar que si A y B son sucesos independientes, la P (A\B) es igual a la P(A), es otro enfoque de mirar independencia. Cómo regla general se enuncia que:

Dos eventos A y B son independientes si y sólo si: P (A\B) = P (A) y P (B\A) = P (B)

que es lo mismo: $P(A \cap B)$ = P (A) x P (B)

De lo anterior se deduce que: $P(A \cap B) = P(A \backslash B)\ P(B)$

Ejemplo:

Se conoce que los estudiantes tienen las siguientes preferencias en el consumo de gaseosas:

Consumo de Gaseosas por semana	**Varones**	**Mujeres**	**Total**
No consume	30	10	40
1-5 veces	50	25	75
Más de 5 veces	20	15	35
Total	100	50	150

Si de un grupo de jóvenes del bar del colegio, se selecciona al azar un estudiante varón ¿Cuál es la probabilidad que ese que ese joven halla consumido más de 5 gaseosas por semana? En este problema ya no es necesarios conocer el número total de estudiantes, porque al seleccionar a un individuo del sexo masculino, los individuos del sexo femenino no son tomados en cuenta. Entonces se puede definir la probabilidad deseada como ¿Qué probabilidad existe de que un individuo beba más de 5 gaseosas a la semana dado que el individuo seleccionado sea varón? Esta es una probabilidad condicional y se resuelve de la siguiente manera:

$P(C_{+5} \backslash S_v) = \dfrac{P(C_{+5} \cap S_v)}{P(S_v)}$ = (20/150) / (100/150) = 20/100= 0.2, donde “C” es por consumo y “S” por sexo.

Ejercicio 3.3 Si se tiene una escuela de 200 alumnos distribuidos en tres aulas: A, B y C. Por sexo: mujer, y varón; como sigue:

Aula/ Sexo	**Varón**	**Mujer**
A	20	20
B	30	30
C	56	44
Total	**106**	**94**

¿Cuál es la probabilidad que un estudiante, sin importar el sexo, sea del aula B?

¿Cuál es la probabilidad que un estudiante sea del aula A, si el estudiante es mujer?

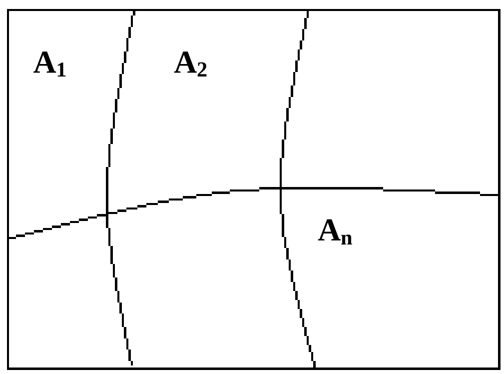
3.5 Uso de la Probabilidad condicional en el Teorema de Bayes

3.5.1 Regla de la probabilidad total

Si se tiene una partición de sucesos A_i que son un conjunto de sucesos mutuamente excluyentes y que cubren todo el espacio muestral.

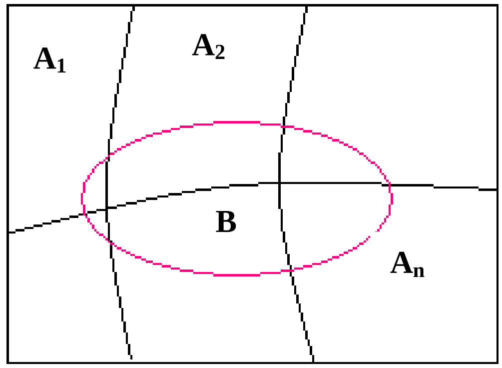
.

$$\boxed{A_1 \cup A_2 \cup \ldots \cup A_n = \Omega \text{ y } A_i \cap A_j = \varnothing \ \forall\, i \neq j}$$



Y si el conjunto de sucesos A_i que forman una partición del espacio muestral y sucede que $p(A_i) \neq 0 \ \forall A_i$. Entonces si ocurre un suceso B dentro del mismo espacio muestral y se cumple que:

$$P(B) = P(B \cap A_1) + P(B \cap A_2) + .. + P(B \cap A_n)$$



Entonces se llamara a ***P (B)*** cómo "*probabilidad total"*, la cual se puede interpretar como una media ponderada de los diferentes $\boldsymbol{P(B\backslash A_i)}$.

P (B) también se puede expresar cómo la sumatoria de las probabilidades condicionadas por la probabilidad del evento *A* correspondiente.

$$\boldsymbol{P(B) = P(B\backslash A_1)P(A_1) + P(B\backslash A_2)P(A_2) + .. + P(B\backslash A_n)P(A_n) = \sum_{i=1}^{n} P(B\backslash A_i)P(A_i)}$$

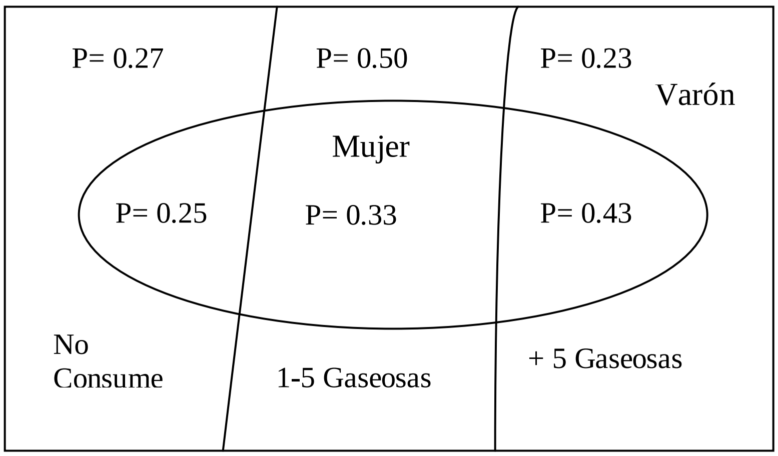
3.5.2 Planteo del Teorema de Bayes

El teorema de Bayes, enunciado por Thomas Bayes y publicada por primera vez en 1763, parte de una situación en la que ocurran una serie de sucesos A_i que son una partición completa de un espacio muestral Ω y donde P (A_i) ≠ 0. Pero también dentro del mismo espacio muestral existe un suceso *B*, tal que *P (B)* ≠ 0, y que las probabilidades de ocurrencia de B son distintas según el suceso A_i que haya ocurrido, tal como se explica en la regla de la probabilidad total.

Conociendo que ha ocurrido el suceso *B,* la fórmula del teorema de Bayes nos indica como modifica esta información las probabilidades de los sucesos A_i . Se resalta que al disponer información de *B* se cambian las probabilidades de A_i. El teorema se presenta algebraicamente de la siguiente manera:

$$P(A_i \backslash B) = \frac{P(B \backslash A_i)P(A_i)}{\sum_{i=1}^{n} P(B \backslash A_i)P(A_i)}$$

Ejemplo con los datos de preferencias de consumo de gaseosas de los estudiantes de la UNI se puede construir el siguiente diagrama de Bayes:



Resolviendo por Bayes, la probabilidad que una mujer no consuma gaseosas es:

$$P(NoC \backslash M) = \frac{P(M \backslash NoC)P(NoC)}{P(M \backslash NoC)P(NoC) + P(M \backslash 1-5)P(1-5) + P(M \backslash +5)P(+5)}$$

$$P(NoC \backslash M) = \frac{0.27\ (0.25)}{0.27\ (0.25) + 0.50\ (0.33) + 0.23\ (0.43)} = 0.20$$

Ejercicio resuelto usando el teorema de Bayes:

Tres máquinas, *A, B* y *C*, producen el 45%, 30% y 25%, respectivamente, del total de las piezas producidas en una fábrica. Los porcentajes de producción defectuosa de estas máquinas son del 3%, 4% y 5%.

a. Seleccionamos una pieza al azar; calcula la probabilidad de que sea defectuosa. (probabilidad Total)

b. Tomamos, al azar, una pieza y resulta ser defectuosa; calcula la probabilidad de haber sido producida por la máquina *B*.
c. ¿Qué máquina tiene la mayor probabilidad de haber producido la citada pieza defectuosa?

Sea *D*= "la pieza es defectuosa" y *N*= "la pieza no es defectuosa". La información del problema puede expresarse en el diagrama de árbol adjunto.

a. Para calcular la probabilidad de que la pieza elegida sea defectuosa, *P(D)*, por la propiedad de la probabilidad total,

$$P\ (Total) = P(D) = P(A) \cdot P(D \backslash A) + P(B) \cdot P(D \backslash B) + P(C) \cdot P(D \backslash C) =$$
$$= 0.45 \times 0.03 + 0.30 \times 0.04 + 0.25 \times 0.05 = 0.038$$

Resolución por diagrama de árbol. Un diagrama de árbol es una representación gráfica de un experimento que consta de pasos, donde cada uno de los pasos tiene un número finito de maneras de ser llevado a cabo.

Prob.	**Máquina**	**Prob.**	**Tipo de producción**
0.45	A	0.03	D
		0.97	N
0.30	B	0.04	D
		0.96	N
0.25	C	0.05	D
		0.095	N

b. Debemos calcular *P(B\D)*. Por el teorema de Bayes,

$$P(B/D)=\frac{P(B).P(D/B)}{P(A).P(D/A)+P(B).P(D/B)+P(C).P(D/C)}$$

$$=\frac{(0.30)(0.04)}{(0.45)(0.03)+(0.3)(0.04)+(0.25)(0.05)}=\frac{12}{38}=0.316$$

c. Calculamos *P(A\D)* y *P(C\D)*, comparándolas con el valor de *P(B\D)* ya calculado. Aplicando el teorema de Bayes, obtenemos:

$$P(A/D)=\frac{(0.45)(0.03)}{(0.45)(0.03)+(0.3)(0.04)+(0.25)(0.05)}=\frac{135}{380}=0.355$$

$$P(C/D)=\frac{(0.25)(0.05)}{(0.45)(0.03)+(0.3)(0.04)+(0.25)(0.05)}=\frac{125}{380}=0.329$$

La máquina con mayor probabilidad de haber producido la pieza defectuosa es *A*

Ejercicio 3.4 El reporte meteorológico ha anunciado tres posibilidades para el día de mañana: que llueva: probabilidad del 50%, que salga el sol: probabilidad del 30% y que esté nublado: probabilidad del 20%.

Según estos posibles estados meteorológicos y datos históricos de comportamiento vehicular, la posibilidad de que ocurra un accidente es la siguiente: si llueve: probabilidad de accidente del 20%, si sale el sol: probabilidad de accidente del 10% y si está nublado: probabilidad de accidente del 5%.

Si se sabe que ocurrió un accidente,

¿Cuál es la probabilidad de que haya llovido?

¿Cuál es la probabilidad de que haya salido el sol?

¿Cuál es la probabilidad de que haya estado nublado?

Ejercicio 3.5 Cierto artículo es manufacturado por tres fábricas: F1, F2 y F3. Se sabe que la primera produce el doble de artículos que la segunda y que ésta (F2) y la tercera producen el mismo número de artículos (durante un período de tiempo especificado, el mismo para las tres). Se sabe también que el 1.5% de los artículos producidos por las dos primeras fábricas es defectuoso, mientras que en la tercera los es el 3.5%.

Se colocan juntos todos los artículos producidos por las tres fábricas y se escoge uno al azar.

¿Cuál es la Probabilidad de que un artículo sea Defectuoso?

¿Cuál Fábrica tiene la mayor probabilidad de haber producido el artículo Defectuosos?

Ejercicio 3.6

En un aula hay 6 estudiantes realizando un examen, dos son mujeres y cuatro son varones. ¿Cuál es la probabilidad que finalice una mujer de segunda dado que el primero en finalizar fue un hombre?

Si la solución es:

$$P(M \backslash V) = \frac{P(M \cap V)}{P(V)} = \frac{8/30}{4/6} = \frac{2}{5}$$

¿Explicar cómo se construyeron los valores 8/30 y 4/6?