UNIVERSIDAD NACIONAL DE INGENIERÍA

UNI- NORTE - SEDE REGIONAL Estelí, Nicaragua

Estadística Básica

Luis María Dicovskiy Riobóo

Índice

Introducción 4
Unidad I. Estadística Descriptiva 4
Objetivos 4
1.1 Introducción Unidad I 5
1.2 Análisis de datos 6
Principios a utilizar al construir una Tabla de Distribución de Frecuencias, TDF. 12
Gráficos 15
Gráficos Multivariados 19
1.3. Medidas de Tendencia Central 21
Media Aritmética 22
La Mediana 23
La Moda 24
1.4 Medidas de Dispersión o de Variabilidad 26
El Rango 27
El Desvío Estándar 28
La Varianza. 30
El Coeficiente de variación 30
1.6 Otras medidas útiles en Estadística Descriptiva. 30
La Asimetría. 31
La Curtosis. 31
1.7 Muestras y Población. 32
Muestreo Aleatorio Simple 33
Muestreo Estratificado 35
Muestreo por Conglomerados 36
Muestreo Sistemático 36
Unidad 2. Teoría Elemental de Probabilidades 38
2.1 Introducción a las Probabilidades 38
2.2 Términos Básicos. 38
Objetivos 38

2.3 Propiedades de la Probabilidad 39
Regla del producto. 40
Regla de la Suma. 40
2.4 Probabilidad condicionada 41
2.3 Teorema de Bayes 42
Regla de la probabilidad total 43
Planteo del Teorema de Bayes 44
2.4 Técnicas de conteo: Combinaciones y Permutaciones 47
Unidad 3. Variables aleatorias y sus distribuciones. 49
3.1 Distribuciones de Frecuencia, Introducción. 49
Objetivos 49
3.2 Variables aleatorias. 51
Función de densidad de probabilidad 51
Distribución acumulativa o función de distribución 53
Parámetros característicos de una función de densidad de probabilidad. 54
El Desvío Estándar y el Teorema de Chebyshev 56
3.3 Distribución Normal 57
3.4 Distribución “t” de Student. 59
3.5 La distribución X2 de Pearson. 63
3.6 La distribución “F” de Fisher. 64
3.7 La distribución Binomial. 65
3.8 Distribución de Poisson 68
Bibliografía y Documentos Consultados 70
Unidad 4. Estimación y prueba de hipótesis. 72
4.1 Estimación por Intervalos de Confianza. 72
Objetivos 72
4.2 Generalidades de las pruebas de Hipótesis 74
4.3 Prueba de hipótesis con pruebas “t” 76
El promedio de una muestra pertenece a población con promedio conocido. 76
Dos promedios tomados en una misma muestra, en momentos diferentes, son iguales. 78
Los promedios de dos muestras o grupos pertenecen a una misma población. 78

Introducción

Este texto básico de estadística está diseñando y organizado en función del contenido de la mayoría de los temas que se aborda en las asignaturas de Estadística I y Estadística II que se imparte en las carreras de Ingeniería en Sistemas, Civil, Industrial y Agroindustrial de la Universidad Nacional de Ingeniería, UNI, de Nicaragua. Sin embargo por su forma sencilla y asequible con que se trató de abordar los diferentes temas, este texto puede ser útil en cualquier otra carrera universitaria.

Este libro 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 un Ingeniero. 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?

Es por lo anterior, respetando el principio de asequibilidad es que buena cantidad de los ejercicios fueron generados en el aula con la información que tienen los estudiantes mano. Creo que la estadística no puede funcionar si primero no se sabe como generar el dato, y como organizar la información en forma de matriz de datos y que estos puedan ser analizados usando un programa estadístico computacional. 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 .

Unidad I. Estadística Descriptiva

Objetivos

- Reflexión sobre el uso de la estadística
- Introducción a la recolección de datos

- Construcción de concepto básicos
- Explicar los diferentes tipos de medidas
- Construir Distribuciones de Frecuencia.
- Realizar los tipos de Gráficos más comunes
- Construir medidas de tendencia central
- Construir medidas de variabilidad
- Utilizar las medidas en el análisis de datos
- Explicar principio básicos de muestreo

1.1 Introducción Unidad I

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 esta 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 de un cuestionario.

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 que varía, las variables pueden ser: a) Cualitativas, en dos tipos nominales o ordinales ó b) Cuantitativas, que pueden clasificarse en b1) continuas o medibles b2) discretas o contables.

El rendimiento de un lote de fríjol se mide en qq/mz y es una variable continua, se mide o pesa y el número de miembros de una familia es una variable discreta, se cuenta. Al medir una variable se tienen "datos" cada dato ocupa una celda de una matriz. En una encuesta (cuestionario) cada pregunta que se hace, 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: Definir 5 variables discretas, 5 variables continuas, 5 variables cualitativas.

Ejercicio 1.2 Clasifique las siguientes variables.

- Peso de un estudiante
- Diámetro de un casa
- Color de ojos
- Tipo de construcción
- # de vainas de frijol por planta.
- Belleza de una flor.
- Temperatura semanal.
- Largo de peces de un estanque

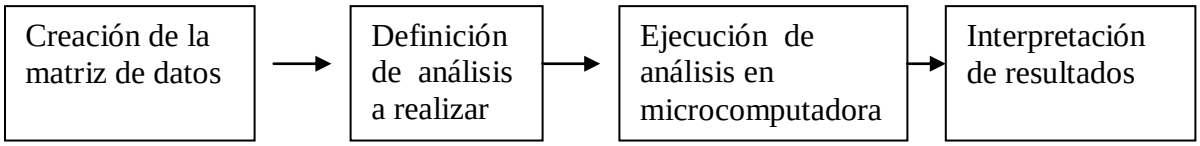
Matriz de datos: es un ordenamiento de datos en fila y columnas donde: cada fila es: un individuo, una parcela, una muestra, una unidad experimental o una encuesta determinada y cada columna: una variable. Los programas Acces, Excel, Infostat y SPSS ordenan los datos en forma de matriz.

1.2 Análisis de datos

Una vez que los datos se han codificado, transferido a una matriz y guardado en una micro computadora podemos proceder a analizarlos, proceso que generalmente se hace con un programa estadístico. En esta clase haremos a modo de ejercicio el inicio del procedimiento de forma manual, situación que difícilmente se puede hacer con

miles datos reales que puede generar una encuesta de tamaño mediano. Es por ello que el énfasis de la clase estará en la interpretación de resultados que en los procedimientos de cálculo.

El procedimiento de análisis se esquematiza en la figura siguiente:



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

Los principales análisis que pueden efectuarse son:

- Estadística descriptiva para variables tomadas individualmente
- Pruebas paramétricas.
- Pruebas no paramétricas.

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

Una primera tarea luego de construir la tabla de datos es explorar los datos buscando información atípica o anormal y corregir los datos en caso que esta 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 gráfico. Para esto los datos se agrupan en clases o categorías y para grupo se calcula la frecuencia absoluta y relativa.

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

- Las escalas Nominales, es cuando describimos algo dándole un nombre a cada categoría o clase y estas son mutuamente excluyentes. Por ejemplo la variable sexo donde “varón = 1” y “mujer = 2”.
- Las escalas Ordinales, donde hay un orden de un conjunto de objetos o eventos con respecto a 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 o un juicio de valor en una escala Likert. 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 .
- Escala de Razón Constante, tienen todas las propiedades de las clases 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 “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 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 ordenarlos 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 pueden 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 la becas que otorga una Universidad puedo codificar a estas 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 está formando 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.	ESTADO CIVIL	EDAD	ALTURA	SEXO	PESO	origen	INGRESO FAMILIAR	BECAS	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 “intervalo o 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.

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, estos no tienen 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

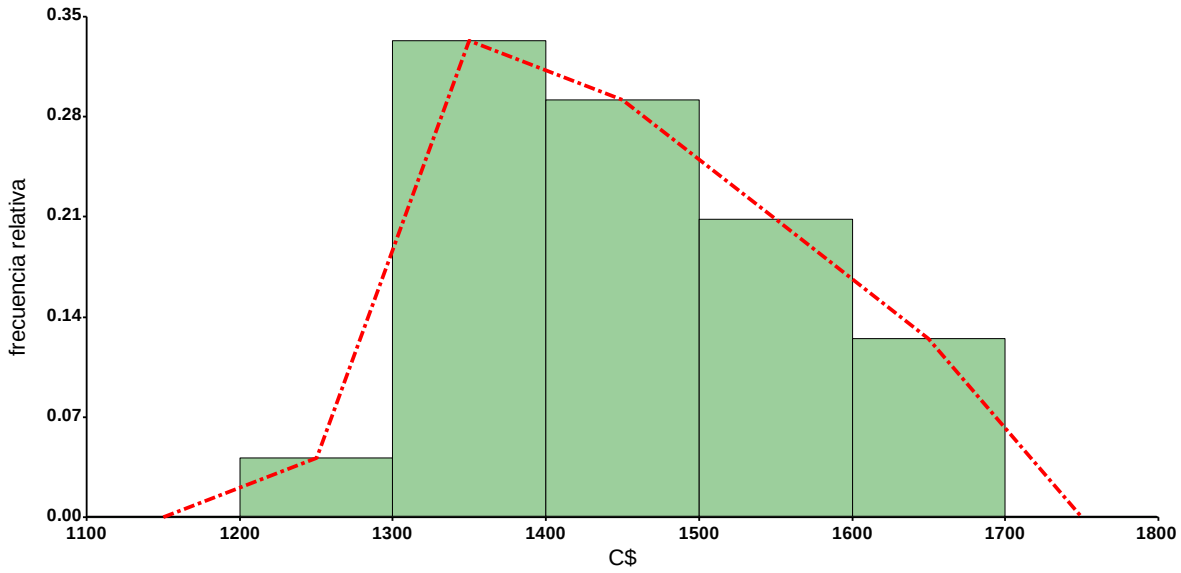
- Se calcula el Rango de los datos, valor mayor menos valor menor : 1680- 1,260 = 420 C$.
- Ancho de clase: El rango se divide en cuatro, 420/4= 105 C$, se ajusta a 100 C$ y de esta manera el número de clases queda en cinco.
- Se construye los límites inferiores y superiores de cada clase como intervalos 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"

(Observar que la información que lleva el gráfico es completa y permite explicar el contenido del mismo).

Una manera de representar una distribución de Frecuencias es:

1. Por medio de un gráfico de Barras con variables nominales.

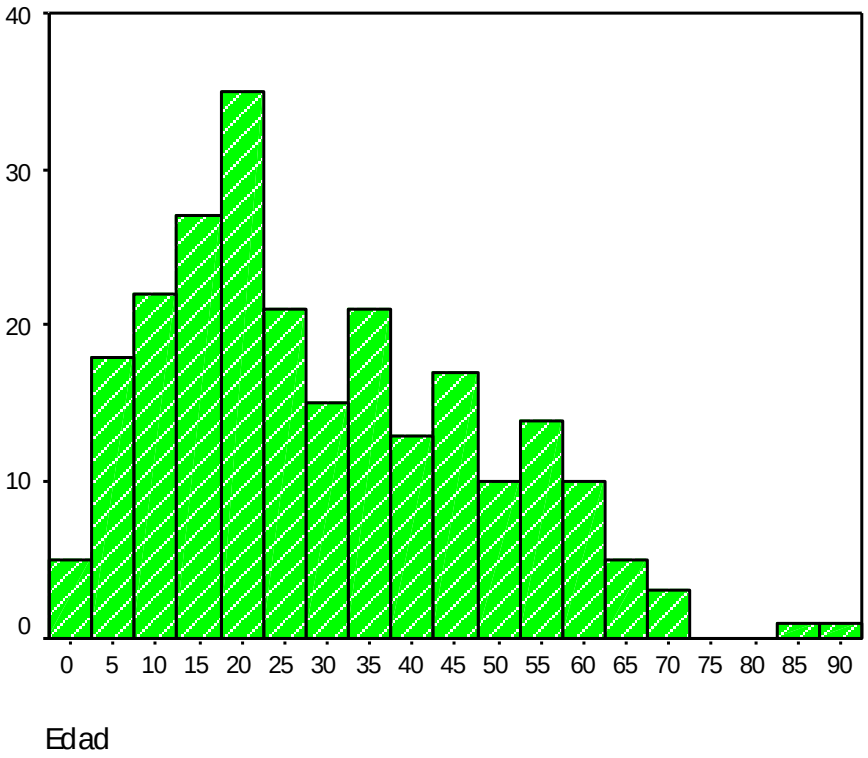
2. Con un Histograma con variables continuas.
3. Un polígono de Frecuencias cuando se quieren mostrar las frecuencias absolutas.
4. Con un gráfico de Pastel cuando se tienen porcentajes o proporciones.

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.

Este gráfico univariado se acompaña de estadística descriptiva como promedios, 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, obtenida como salida de un análisis con SPSS. Este Gráfico es creado con frecuencias y porcentajes, permite resaltar segmentos de clases determinadas.

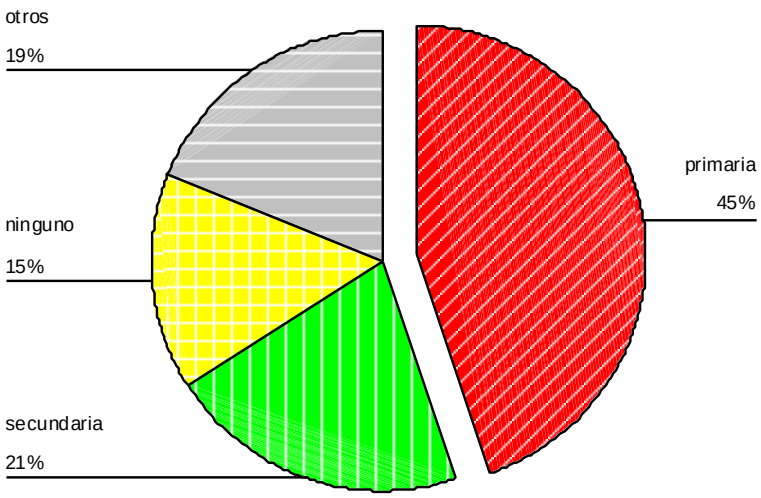
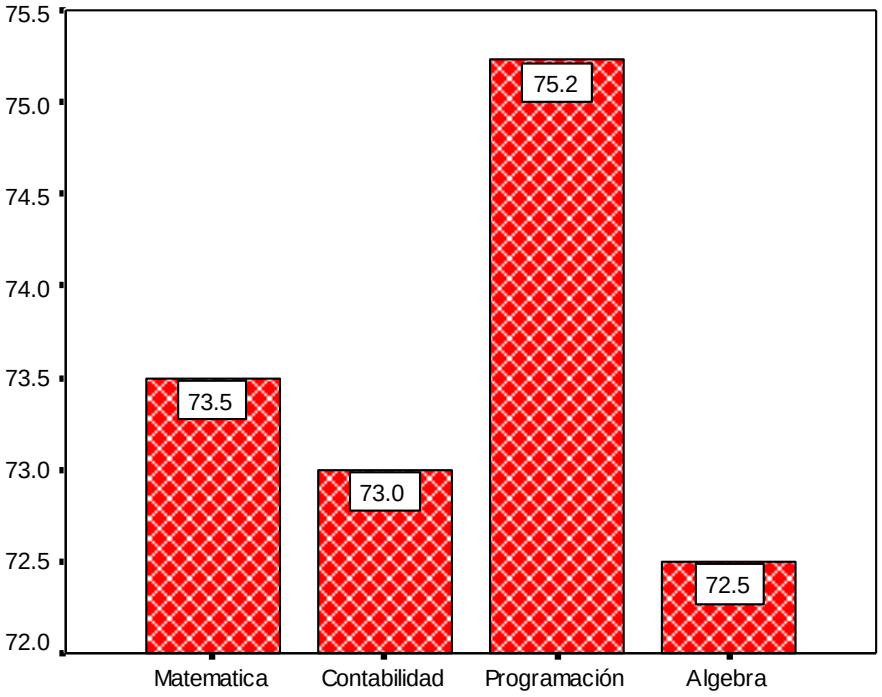


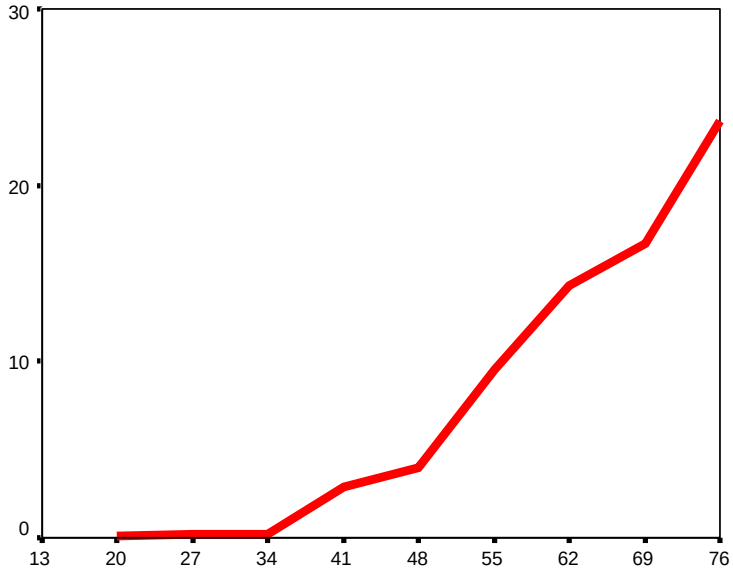
Gráfico de pastel o sectores.

"Gráfico de Barras bivariado". Ejemplo de las notas de tres asignaturas presentadas en forma de barras. Este resume el promedio de notas obtenido por asignatura. Entre barra y barra hay un espacio. El gráfico observado a continuación se construyó con una variable nominale, asignatura y una variable continua, nota.



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

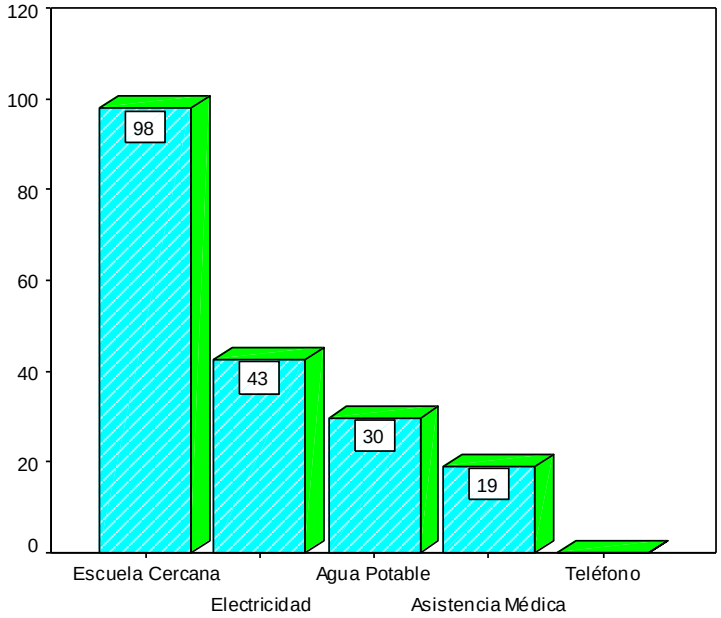
El Polígono es una línea quebrada que se construye uniendo los puntos medios en la parte superior de cada barra, marca de clase de un histograma



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

Gráficos Multivariados

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

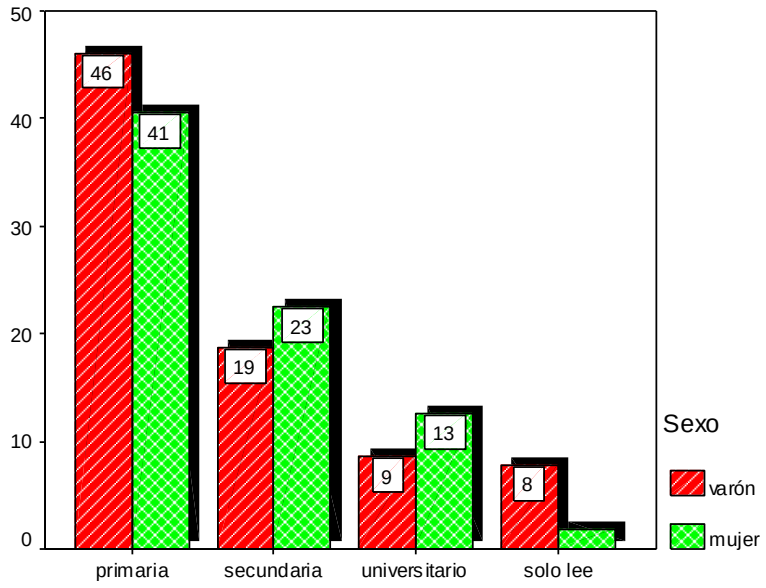


Gráfico Bivariado De Barras Apiladas

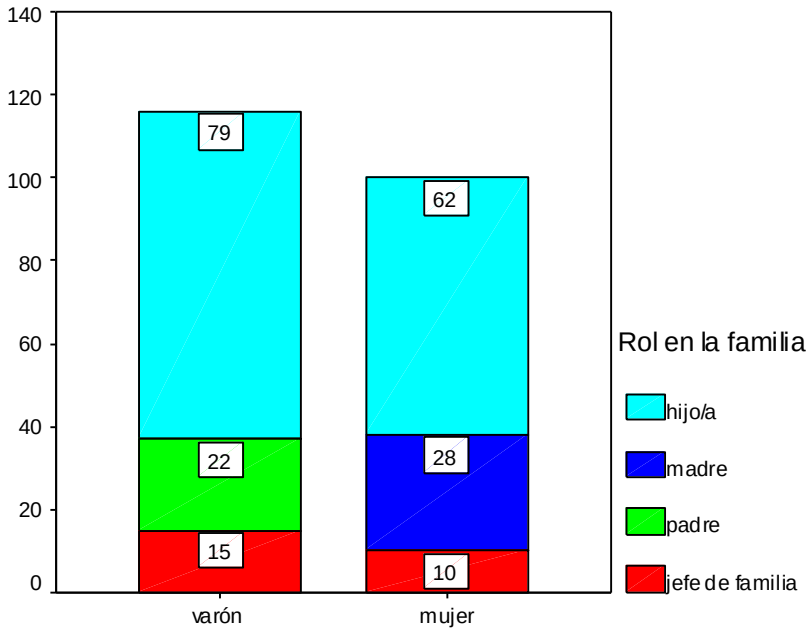
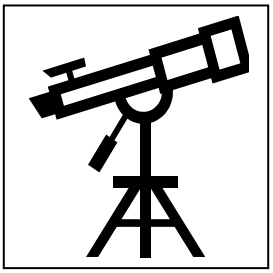


Gráfico bivariado que reduce el número de barras y por lo tanto se simplifica el diseño. Se puede construir con frecuencias o porcentajes



"Un Gráfico es una manera de ver rápidamente lo que nos dicen los datos"

"A partir de la realidad observable debo crear un modelo numérico teórico para intentar estudiar esa realidad"

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

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

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".

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 el promedio común o "media aritmética", pero también son de importancia, según las circunstancias y el tipo de variables la "moda" y la "mediana".

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 : " $\bar{x} = x_1 + x_2 + x_3 + .. x_n / n$ " 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:

$$\bar{x} = \sum_1^n x\, f / n$$

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 o ordinal.

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

Ejemplo de cálculo de una media o promedio.

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
Robero	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 pocas 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. Además se suele preferir en la construcción de pruebas de hipótesis, en la estadística inferencial. Se usa normalmente con datos de intervalo y de razón constante y cuando las distribuciones tiene forma simétrica.

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 del promedio. Por ejemplo si tengo 9 personas con 0 ingresos y uno sola que tiene ingresos de 10 unidades, el promedio 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.

No necesariamente una escala de medida nos debe decir que tipo de medida de tendencia central debemos usar, pero si nos ayuda a determinar cual es la más apropiada.

Un aspecto interesante entre las tres medidas es su comportamiento referente a la simetría que toma una distribución. Cuando las distribuciones son simétricas, sin sesgo, caso de la distribución Normal que tiene forma de campana, "la media, la mediana y la moda coinciden". Si la distribución es asimétrica con sesgo positivo, 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 es negativo, entonces "la media es menor que la mediana y esta 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 $\bar{x}_g = \sqrt[4]{(4)(5)(4)(6)} = 4{,}68$

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 $\bar{x}_c = \sqrt[2]{\frac{4^2 + 5^2 + 4^2 + 6^2}{4}} = 4.81$

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.

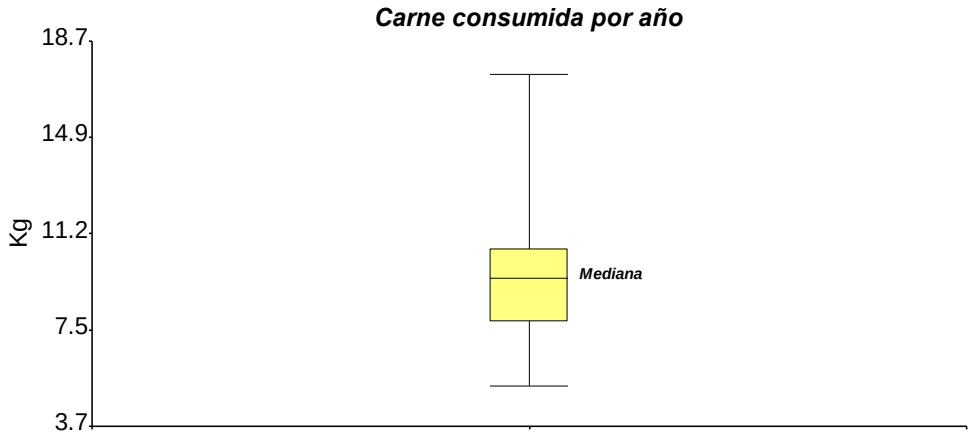


Diagrama de caja

Los bordes de la caja amarilla son el Q_1 y el Q_3.

Deciles, 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.

Percentiles, 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. Así como las medidas de tendencia central son valores 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.

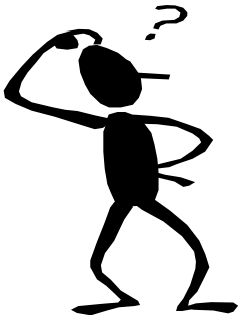
El Rango, Recorrido o Amplitud de un conjunto de mediciones, “es la diferencia entre el valor mayor y el valor menor”, indica el número necesario y mínimo de unidades, en la

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 el promedio de desviación de los valores con respecto a al 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 en promedio y 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 de “S” por suma de cuadrados

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, pero esta descomposición es la más usada en estadística inferencial, por ejemplo en la construcción de tablas de Análisis de Variancia, ANDEVA.

Ejemplo de cálculo de Desvío Estándar “S”

Con el ejemplo de las notas de matemáticas haremos cálculo de “S”

“S”= $\sqrt{((55-76.1)^2+(62-76.1)^2+(67-76.1)^2+(68-76.1)^2+(69-76.1)^2+(79-76.1)^2+(88-76.1)^2+(89-76.1)^2+(92-76.1)^2+(92-76.1)^2)/9}$

= 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, por ejemplo "el Análisis de Varianza, ANDEVA" que se basa en la descomposición y relación de las varianzas de las causas de variación de los datos. 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, $CV = \left({}^{S}/_{\bar{X}}\right) 100$. Este coeficiente permite comparar la variabilidad de diferentes muestras de una población ó la variabilidad entre variables diferentes. En general un CV menor al 10 %, dice que los datos tienen poca variabilidad, que es lo mismo que decir que los valores observados son en general, cercanos al valor medio.

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 en el caso que la distribución sea normal. La moda, mediana y el rango pueden completar la información sobre la distribución 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.*

1.6 Otra

s 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 es positiva, el promedio es mayor que la mediana, quiere decir que hay valores agrupados hacia la izquierda de la curva, por debajo del valor de la media. Cuando es negativa, 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.

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

$$Sesgo = \frac{3(\overline{X} - Me)}{S}$$

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).

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

Definición:
Las medidas calculadas a partir de la población, Ej. "μ" y "σ" se llaman PARÁMETROS

Las medidas calculadas a partir de las muestras, Ej. "$\bar{x}$" "S" se llaman ESTADÍSTICOS

Ejercicio 1.8:

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

1.7 Muestras y Población.

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.

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) .

Tenemos varias formas de calcular el "n", número de de elementos de la muestra.

Cálculo de "n" por ecuación: Cuando la fracción n / N a priori se determina que será mayor que 0.1, un método para determinar "n" de manera aproximada es el siguiente:

$$n = \frac{N * p * q}{(N-1) * D + p * q}$$

Donde:

- Los valores "p" y "q" cumplen que "p + q = 1" y generalmente se acepta que "p = q = 0.5".
- "D" es un valor que se vincula al error de estimación prefijado donde "$D = B^2/4$"
- "B" es el error de estimación que se debe fijar y generalmente fluctúa entre 0.01 y 0.10 .
- "p x q" es la variancia de una distribución binomial de una pregunta dicotómica, con tiene 2 posibles respuestas.

Si bien este modelo es bastante teórico es un método muy usado para aproximar un valor de "n" entrevistados, cuando se realiza investigación social.

Cálculo de "n" Gráficamente: Se sabe que a más grande la muestra mejor ésta estima el promedio de la población, sin embargo hay un momento que el promedio 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 promedios se puede observar gráficamente con un gráfico de promedios. El primer promedio 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 los promedios 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 los promedios consecutivos, con un dato, dos datos, tres datos… hasta 15 datos. Se observa que a partir de 10 datos, el promedio 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

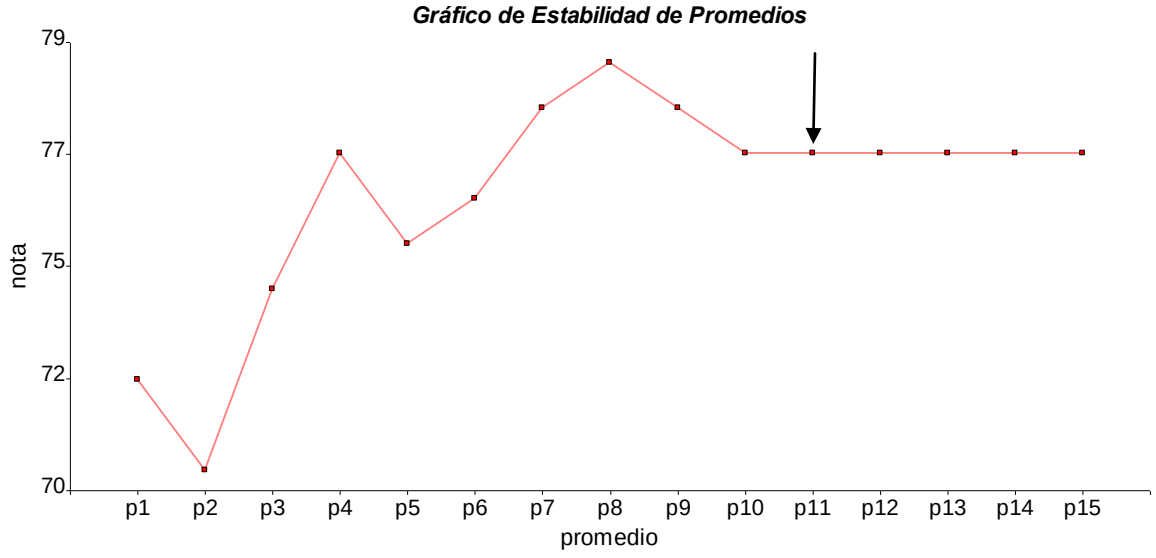


Gráfico de Estabilidad de Promedios

Muestreo Estratificado

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. 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.

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.

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 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 de la muestra. En general se hará de la manera **"$n_i = *N_i/N$".**

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 son 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.

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 subdividir la población y asegurar mayor representatividad de la muestra”.

Tarea:

- 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 comunidades semejantes en cuanto a su nivel de vida y forma de producir la tierra.

Bibliografía consultada:

Mendenhall . Estadística para administradores. México: Iberoamericana

Sampieri, R; Fénández, C y Baptista, P (2003). Metodología de la Investigación. México: Mc Graw Hill.

Spiegel, M y Stephens, L. (2002). Estadística. México: Mc Graw Hill.

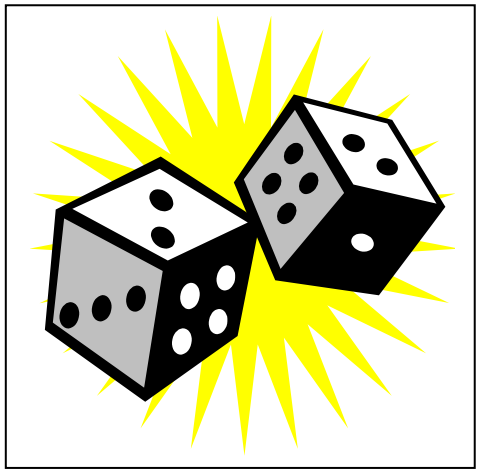
Young y Veldaman . Introducción a la Estadística aplicado a las ciencias de la conducta. México: Trillas.

Unidad 2. Teoría Elemental de Probabilidades

Objetivos

- Definir conceptos básicos de probabilidad
- Explicar las Reglas de Adición y Multiplicación..
- Valorar la importancia de la probabilidad para la selección de una muestra
- Introducir la probabilidad condicional y Bayesiana

2.1 Introducción a las Probabilidades



Con esta teoría se estudia fenómenos naturales con el fin de descubrir regularidades en la ocurrencia de los mismos. Sus fundamentos aunque parezca extraño se basó en un inicio en el estudio de los juegos al azar (dados, cartas, ruletas, etc), así comenzó esta ciencia en la Francia Monárquica. Sus aplicaciones hoy día abundan desde ramas de las ciencias como por ejemplo en la genética mendeliana, genética de poblaciones, análisis de experimentos, predicciones del tiempo, predicción de ataque de plagas, etc. En nuestra vida diaria aplicamos inconscientemente probabilidades cuando compramos un billete de lotería o llevamos un paraguas cuando vemos el cielo nublado.

2.2 Términos Básicos.

Experimento: Es el proceso que permite obtener una o varias observaciones.

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: 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)**.

Por ejemplo si en una comunidad hay 64 campesinos que siembran frijol de forma manual y 16 con bueyes. En este caso hay 2 eventos: Siembra manual y Siembra con bueyes y existe la P(bu) y la P(ma) asociados a la frecuencia de ocurrencia de cada evento. La probabilidad que al elegir una parcela al azar esta fue sembrada de forma manual P(ma) es de 16/80 = 0.20 ó 20 % .

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

Se debe considerar que la P(A) se hace más real en la medida el número de elementos de Ω se hace más grande

2.3 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.

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)

Por ejemplo si la Probabilidad de obtener cara al arrojar una moneda es 0.5, P(cara) = 0.5, la probabilidad que al arrojar dos veces la moneda salgan dos caras, ya que ambos eventos son independientes, uno no influye sobre otro, la P(cara, cara) es de “0.5 x .05 = 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

Tarea: estimar la probabilidad que al elegir por sorteo dos estudiantes del grupo, ambos sean varones. Determinar también cuales eventos forman **“Ω”**es este caso.

Si los sucesos son independientes: P(A) x P (B) es igual p(A $\cap$ B)

Otro enfoque de mirar independencia es, dos eventos *A* y *B* son independientes si y sólo si: $P(A/B) = P(A)$ y $P(B/A) = P(B)$ o, que es lo mismo: $P(A \cap B) = P(A) \times P(B)$

Regla de la Suma.

Para que dos eventos “A” y ”B” se puedan sumar directamente, estos deben se 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 no 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**

2.4 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. El proceso de realizar la historia de un caso, explorar y realizar pruebas complementarias ilustra 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).

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

Ejemplo:

Se conoce que a los estudiantes de la UNI 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 de la universidad, 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:

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

Ejercicio 2.1

Si se tiene una escuela de 200 alumnos distribuidos en tres aulas y por sexo: mujer M, y varón, V; 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?

Ejercicio 2.2

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?

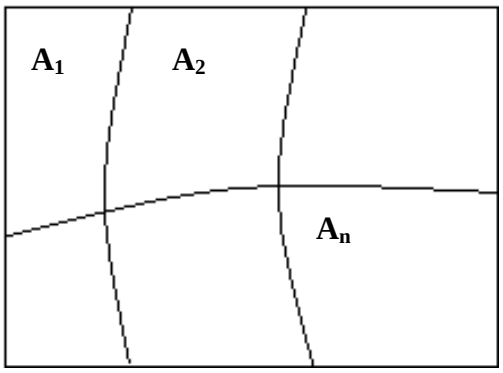
2.3 Teorema de Bayes

Regla de la probabilidad total

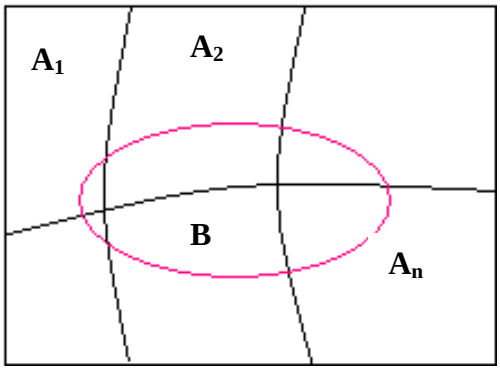
Se llama partición a un conjunto de sucesos A_i tales que.

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

Es decir, son un conjunto de sucesos mutuamente excluyentes y que cubren todo el espacio muestral.



Regla de la probabilidad total: Si un conjunto de sucesos A_i forman una partición del espacio muestral y $p(A_i) \neq 0\ \forall\ A_i$, para cualquier otro suceso B se cumple



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

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

Siendo P(B) la *probabilidad total*, la cual se puede interpretar como un promedio ponderado de los diferentes $\boldsymbol{P(B\backslash A_i)P(A_i)}$

Planteo del Teorema de Bayes

Si los sucesos A_i son una partición y B un suceso tal que p (B) $\neq 0$ y para i= 1,2,..n, como lo visto en la teoría de Probabilidad Total

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

Ejercicio resuelto:

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)

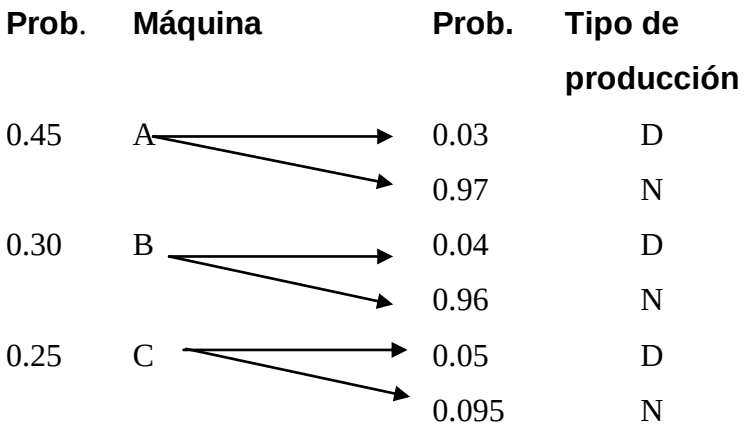
b. Tomamos, al azar, una pieza y resulta ser defectuosa; calcula la probabilidad de haber sido producida por la máquina *B*.
c. ¿Qué máquina tiene la mayor probabilidad de haber producido la citada pieza defectuosa?

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.



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 2.2 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?

2.4 Técnicas de conteo: Combinaciones y Permutaciones

Las técnicas de conteo son aquellas que son usadas para enumerar eventos difíciles de cuantificar.

Combinaciones:

La expresión **"$C_{m,n}$"** representa las combinaciones de "m" elementos, formando subgrupos de "n" elementos. Para calcular el número de combinaciones se aplica la siguiente fórmula:

$$C_{m.n} = \frac{m!}{n!(m-n)!}$$

El término **" n ! "** se denomina "factorial de n" y es la multiplicación de todos los números que van desde "n" hasta 1.

Por ejemplo: 4 ! = 4 * 3 * 2 * 1 = 24

Ejemplo: $C_{10,4}$ son las combinaciones de 10 elementos agrupándolos en subgrupos de 4 elementos sin importar el orden:

$$C_{10,4} = \frac{10!}{4!(10-4)!} = \frac{10.9.8.7.6.5.4.3.2.1}{(4.3.2.1)(6.5.4.3.2.1)} = 210$$

Es decir, podríamos formar 210 subgrupos diferentes de 4 elementos, a partir de los 10 elementos.

Por ejemplo: Si tomamos el conjunto A={a,b,c,d}, ¿cuántos subconjuntos de 2 elementos cada uno se pueden obtener?

Haciéndolos se obtienen: {a,b}, {a,c}, {a,d}, {b,c}, {b,d}, {c,d}. Son seis los subconjuntos.

Permutaciones:

La expresión **"$P_{m,n}$"** representa las variaciones de "m" elementos, formando subgrupos de "n" elementos. En este caso, un subgrupo se diferenciará del resto, bien por los elementos que lo forman, o bien por el orden de dichos elementos. Para calcular el número de permutaciones se aplica la siguiente fórmula:

$$P_{m.n} = \frac{m!}{(m-n)!}$$

Ejemplo: $P_{(10,4)}$ son las permutaciones de 10 elementos agrupándolos en subgrupos de 4 elementos:

$$P_{10,4} = \frac{10!}{(10-4)!} = \frac{10.9.8.7.6.5.4.3.2.1}{6.5.4.3.2.1} = 5{,}040$$

Es decir, podríamos formar 5.040 subgrupos diferentes de 4 elementos, a partir de los 10 elementos.

Ejemplo: Sea *A*= letras {*a*,*b*,*c*,*d*}, ¿cuántos "conjuntos" de dos letras se pueden obtener?

Lo que se pide es formar permutaciones u ordenaciones de 2 letras, cuando el total de letras es 4. En este caso *n*=2 y *m* =4. Las "palabras" de 2 letras formadas son: *aa*, *ab*, *ac*, *ad*, *ba*, *bb*, *bc*, *bd*, *ca*, *cb*, *cc*, *cd*, *da*, *db*, *dc*, *dd*. En total son 16.

Bibliografía y Documentos Consultados

Abraira V. PROBABILIDAD.. Centro de Estudios Ramón Areces. Madrid. 1996.

Dicovskiy, L. Módulo Nro 1. Bioestadística. Folleto de clase. UCATSE.

Sampieri R, Collado C y Lucio P. 2004.Metodología de la Investigación. Tercera∩ Edición edit. Mc Graw Hill.

Vermeer I.1996. Estadística, Curso Básico. EAGE. 104 p.

Unidad 3. Variables aleatorias y sus distribuciones.

Objetivos

- Aplicar el concepto de variable aleatoria
- Explicar las distribuciones más usadas en la ingeniería
- Identificar que modelos se usan con variables discretas y continuas

3.1 Distribuciones de Frecuencia, Introducción.

"Los modelos estadísticos son un puente entre la muestra observada y la población desconocida."

Hasta esta unidad nos hemos ocupado de descripciones de muestras usando tablas, gráficos y medidas como la media y la varianza. Pero generalmente nuestro interés va más allá que una simple descripción, suele haber interés en tratar de generalizar los resultados de la muestra hacia el grupo total, es decir la Población.

Para generalizar podemos usar modelos estadísticos teóricos diseñados por estadísticos famosos como Gauss, Fisher, Gosset y otros.

Hoy en día los modelos estadísticos teóricos son frecuentemente utilizados para observar y comprender fenómenos naturales que implican el estudio de variables o características de poblaciones naturales. El instrumento conceptual que permitirá esta generalización es un modelo de la población, es decir una representación simbólica de su comportamiento. Los modelos estadísticos van a actuar de puente entre lo observado, la muestra y lo desconocido, la población.

Las distribuciones de probabilidad están relacionadas con las distribuciones de frecuencias. Una distribución de frecuencias teórica es una distribución de probabilidades que describe la forma en que se *espera* que varíen los resultados. Debido a que estas distribuciones tratan sobre expectativas de que algo suceda, resultan ser modelos útiles para hacer inferencias y para tomar decisiones en condiciones de incertidumbre.

Las distribuciones de probabilidad son idealizaciones de los polígonos de frecuencias. En el caso de una variable estadística continua consideramos el histograma de frecuencias relativas, y se puede comprobar que al aumentar el número de datos y el número de clases el histograma tiende a estabilizarse llegando a convertirse su perfil en la gráfica de una función.

Una distribución de frecuencias es un listado de las frecuencias observadas de todos los resultados de un experimento que se presentaron realmente cuando se efectuó el experimento, mientras que una distribución de probabilidad es un listado de las probabilidades de todos los posibles resultados que podrían obtenerse si el experimento se lleva a cabo.

Las distribuciones de probabilidad pueden basarse en consideraciones teóricas o en una estimación subjetiva de la posibilidad. Se pueden basar también en la experiencia.

Las distribuciones de probabilidad se clasifican como *continuas* y *discretas*. En la distribución de probabilidad discreta la variable aleatoria, la que toma los posibles resultados del experimento, sólo toma un número limitado de valores, por ejemplo que un ladrillo tomado “sea defectuoso” o “no”. En una distribución de probabilidad continua, la variable que se está considerando puede tomar cualquier valor dentro de un intervalo dado, por ejemplo “los ladrillos de una población que pesen entre 1,5-1,6 Kg”. Las distribuciones continuas son una forma conveniente de presentar distribuciones discretas que tienen muchos resultados posibles, todos muy cercanos entre sí.

3.2 Variables aleatorias.

Una variable es aleatoria si toma los valores de los resultados de un experimento aleatorio. Esta variable puede ser discreta o continua. De manera general se puede decir que si el experimento toma un número finito de valores o un número infinito pero numerable, que se puede contar, tenemos una variable aleatoria discreta. En el otro extremo, si el experimento puede tomar cualquier valor dentro de un intervalo dado, entonces se trata de una variable aleatoria continua, generalmente son aquellas variables que se miden ó se pesan. Las variables aleatorias definidas sobre espacios muestrales discretos se llaman variables aleatorias discretas y las definidas sobre espacios muestrales continuos se llaman continuas.

Se puede pensar en una variable aleatoria como un valor o una magnitud que cambia de una presentación a otra, sin seguir una secuencia predecible. Los valores de una variable aleatoria son los valores numéricos correspondientes a cada posible resultado de un experimento aleatorio.

Una variable aleatoria asocia un número o más generalmente una característica a todo resultado posible del experimento. Por ejemplo, si consideramos el experimento que consiste en realizar mediciones de la concentración de un producto en una solución, nos interesa la variable aleatoria X= “valor medido de la concentración de azúcar en una salsa.” Otro ejemplo de variable aleatoria asociada a un proceso de fabricación, al experimento de escoger un elemento producido, y considerar la variable aleatoria X= “duración de vida hasta el fallo”.

La *distribución de probabilidad* de una variable aleatoria proporciona una probabilidad para cada valor posible, y estas probabilidades en su totalidad deben sumar uno.

Función de densidad de probabilidad: función que mide concentración probabilidad alrededor de los valores de una variable aleatoria. A cada valor de una variable aleatoria discreta o a un intervalo de una variable aleatoria continua, le corresponde una probabilidad asociada.

Ejemplo: Va a nacer tres bebes. Representamos “varón” por *v* y “niña” por ñ.

S = {vvv, *vvñ, vñv, ñvv, vññ, ñvñ, ññv, ññ*ñ}

La probabilidad de cada suceso elemental es 1/8. Por ejemplo p (vvv)=1/8, ya que la probabilidad de nacer un varón en un nacimiento es 1/2 según la definición clásica y los nacimientos son independientes, p(vvv)= $(½)^3$.

Definimos la variable aleatoria. X: número varones nacidos, la cual puede tomar los valores {0, 1, 2, 3}. Se buscan todos sucesos de la muestra que dan lugar a cada valor de la variable y a ese valor se le asigna la probabilidad del suceso correspondiente.

x	Sucesos	p_x
0	{ññ*ñ*}	1/8
1	{vññ, ñvñ, ññv}	3/8
2	{vvñ, vñv, ñvv}	3/8
3	{vvv}	1/8

A esta función se le denomina *función densidad de probabilidad (fdp)*, que "funciona" de distinta manera en las variables discreta que en las continuas. En el caso de las variables discretas, como en el ejemplo, es una función que para cada valor de la variable hay una probabilidad.

Sin embargo para las variables continuas la probabilidad de que una variable tome **cualquier** valor concreto es 0, por lo tanto la *fdp* sólo permite calcular la probabilidad para un intervalo del tipo (a<X<b), mediante el área bajo la curva de la *fdp.*

Para las variables aleatorias de interés hay tablas, y programas de computacionales, donde buscar esos valores.

Distribución acumulativa o función de distribución

Función que acumula probabilidades asociadas a una variable aleatoria. Su notación es $F(x) = p(X \leq x)$. Para el ejemplo anterior, F (X) es:

X	f(x)	F(x)
0	1/8	1/8
1	3/8	4/8
2	3/8	7/8
3	1/8	8/8

En variables continuas $F(X) = P(X \leq a) = \int_{-\alpha}^{a} f(x)dx$

La probabilidad de que la variable esté dentro de un intervalo [a - b] se calcula:

$P (a \leq x \leq b) = F(b) - F(a)$

La probabilidad de que la variable continua tome un valor particular se puede expresar como: $F(c) - F(c) = 0$. Esto explica la idea de que para el caso de una variable aleatoria continua no tiene sentido trabajar con la probabilidad de un valor particular, ya que esta vale 0. Por ejemplo de una población de personas, la probabilidad que una persona

pese exactamente 1.75000… cm es de cero. Sin embargo hay posibilidad para un intervalo dado de altura como 1,70-1.80 cm.

Parámetros característicos de una función de densidad de probabilidad.

Valor esperado o *esperanza matemática o media*

$\mu_X = E(x) = \sum xf(x)$ *Caso discreto*

$\mu_X = E(x) = \int_{-\infty}^{\infty} xf(x)dx$ *Caso continuo*

Si X es una variable aleatoria cualquier función de ella, h(x), es también una variable aleatoria, en consecuencia también se define este parámetro para una función de variable aleatoria.

$\mu_X = E[h(x)] = \sum h(x) f(x)$ *Caso discreto*

$\mu_X = E[h(x)] = \int_{-\infty}^{\infty} h(x) f(x)dx$ *Caso continuo*

Ejemplo: Se tira un dado. Se define como variable aleatoria el número que sale ¿Cuál es su media?

La variable X puede tomar los valores 1, 2, ..., 6 y para todos ellos f(x) = 1/6. En consecuencia la media es

$$\mu_X = \sum_{x=1}^{6} xf(x) = 1\frac{1}{6} + 2\frac{1}{6} + ... + 6\frac{1}{6} = 3.5$$

Obsérvese que es un número que la variable aleatoria no puede alcanzar.

Si se define ahora una nueva función sobre X: el premio: si sale 1 ó 2 se gana 100 C$, si sale 3 se gana 500 y si sale 4, 5 ó 6 no se gana nada

X	h(x)
1	100
2	100
3	500
4	0
5	0
6	0

¿Cuál es el valor medio de esta función?

$$\mu_X = \sum_{x=1}^{6} h(x) f(x) = 100\frac{1}{6} + 100\frac{1}{6} + 500\frac{1}{6} + 0 + 0 + 0 = 116.6$$

¿Qué significa? es el valor medio a la larga: si se juega un número grande de veces la ganancia final es como si en cada jugada se hubiera ganado 116,6 C$. Si la apuesta costara menos de eso el juego sería ventajoso para el jugador, si costara más, para la banca. Se debe considera que el juego sería justo si la apuesta costara exactamente 116.6 C$.

Ejercicio 3.1: En los casino el juego de ruleta mesa tiene 38 números, esto incluye el número 0 y doble 00. Si usted apuesta a una moneda a un número y gana, el casino le paga 36 monedas. ¿Este es un juego justo? Justificar la respuesta.

Varianza:

Se define como:

$$\sigma_X^2 = E(x - \mu_X)^2$$

Aunque para el cálculo se suele usar esta otra fórmula equivalente:

$$\sigma_X^2 = E(x^2) - \mu_X{}^2$$

¿Qué mide la varianza? Mide la dispersión de la variable alrededor de la media.

Ejemplo de cálculo de varianza:

Si ocurren tres nacimientos de bebes, la esperanza y la varianza de la variable aleatoria X "varones nacidos" es:

E (X) = 0* 1/8 + 1* 3/8 + 2* 3/8 + 3 * 1/8 = 3/2= 1,5

$\sigma_X^2 = E(x^2) - \mu_X{}^2$ = 0^2 * 1/8 + 1^2 * 3/8 + 2^2 * 3/8 + 3^2 * 1/8 – $(3/2)^2$ = ¾

$$\sigma_x = \sqrt{\sigma_x^2} = \sqrt{3/4} = 0.866$$

El Desvío Estándar y el Teorema de Chebyshev

Es conocida en el área de la probabilidad y estadística, la desigualdad de Chebyshev, matemático Ruso del siglo XIX, que dice que la probabilidad de que una variable aleatoria esté distanciada de su media en más de "*a"* veces la desviación estándar, es menor o igual que"$1/a^2$". Si E(x) es la media (o la esperanza matemática) y σ es la desviación estándar, entonces podemos redefinir la relación como:

$$P(|x - E(x)| \geq a\sigma) \leq \frac{1}{a^2}$$

Tomando en cuenta el teorema de Chebyshev se puede construir las siguientes reglas sobre el uso del desvío estándar:

Según el teorema de Chebyshev, y sin importar el tipo de distribución de los datos, se cumple que:

- *El intervalo $\bar{x} \pm$ 2 "S" contendrá al menos ¾ de los datos.*
- *El intervalo $\bar{x} \pm$ 3 "S" contendrá al menos 8/9 de los datos.*

3.3 Distribución Normal

La distribución Normal es un modelo teórico para variables aleatorias y continuas y representa la distribución de frecuencias de una población de valores.

La *curva normal* es una campana simétrica cuya forma y posición depende de dos parámetros

- μ, media poblacional, que se localiza en el centro de la del eje horizontal.
- σ, desviación estándar que determina el ancho de la curva.

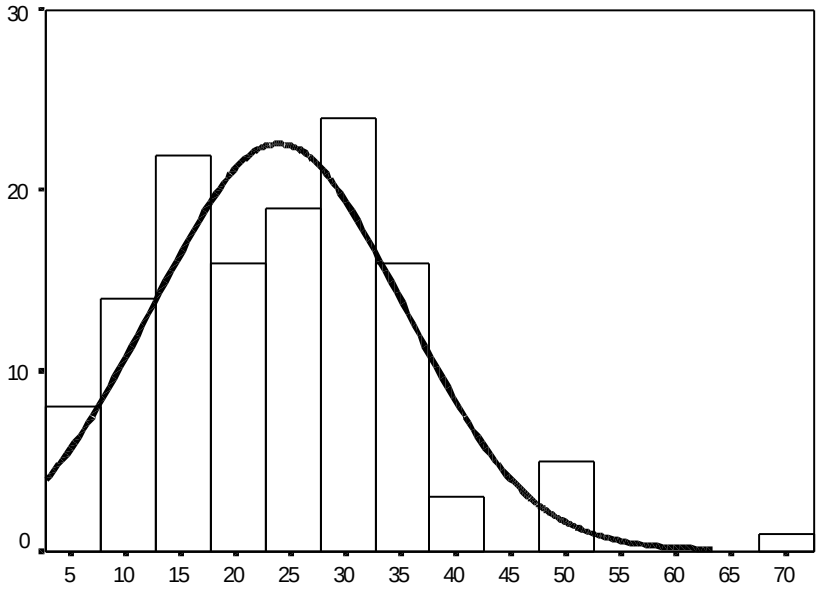
Para una variable "x" con media μ y desviación estándar σ, que está normalmente distribuida, escribimos: "x" es N (μ, σ).

La función de densidad de la distribución normal es:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Ejemplo de una distribución de frecuencias de Mg. de Aflotoxinas (toxinas) en maíz y la curva Normal teórica que genera el programa SPSS.

Histograma de frecuencias y curva teórica Normal



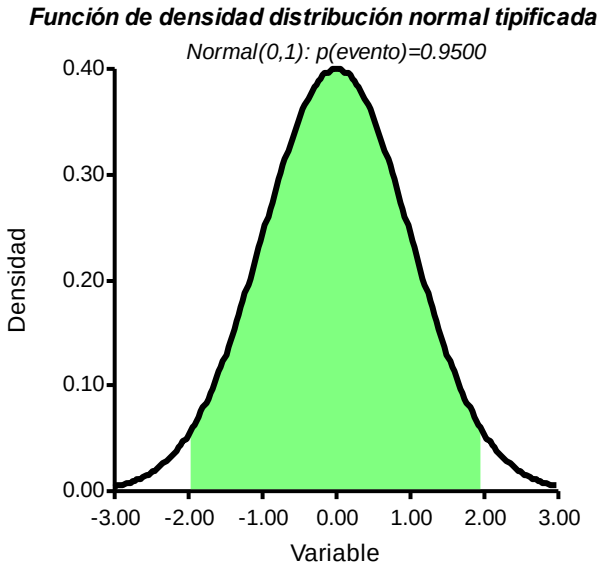
Si un Distribución de datos tiene aproximadamente el perfil o forma de campana se cumple que:

- El intervalo $\mu \pm$ "σ" contendrá aproximadamente el 68 % de los datos.
- El intervalo $\mu \pm 2$ "σ" contendrá aproximadamente el 95 % de los datos.
- El intervalo $\mu \pm 3$ "σ" contendrá aproximadamente casi la totalidad de los datos.

Un tipo de distribución Normal especial es la distribución Normal Tipificada (0,1), simbolizada con la letra "z". Esta distribución se usa mucho para resolver pruebas de hipótesis ya que cualquier dato "x_i" de una variable normal (μ, σ) se puede convertir en dato "z_i" de una variable normal tipificada con la siguiente transformación

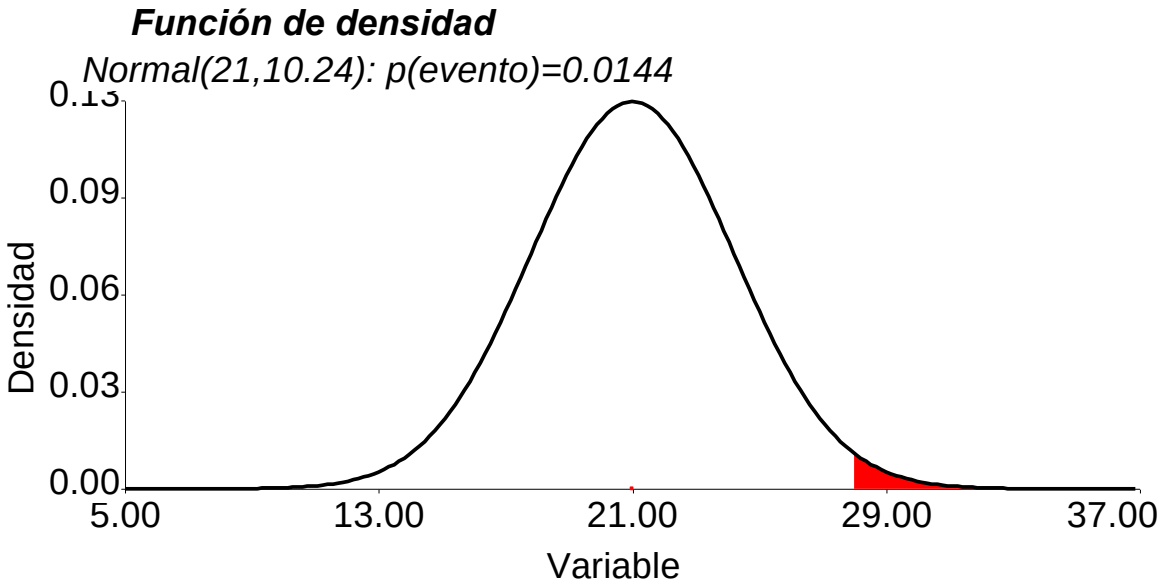
$$z_i = \frac{x_i - \mu}{s}$$

Luego con una tabla normal tipificada es fácil determinar probabilidades por intervalos para diferentes valores de la variable "x". Esta distribución funciona relativamente bien para hacer probabilidades cuando se tiene más de 30 datos, y estos tienen una distribución en forma de campana. A continuación se observa un gráfico de una distribución normal tipificada (0,1) donde está sombreado un intervalo de ± 1.96 desvió estándar.



Ejercicio 3.2. Si el promedio de edad de los alumnos de la universidad es de 21 años, con un desvío estándar de 3.2 años. ¿Cuál es la probabilidad que un estudiante tenga más de 28 años?

$Z_{28} = {28 - 21}/{3.2} = 2.1875$, se debe buscar la P($z_i \geq$ 2,1875) en una tabla normal tipificada que resulta como 0.5 - 0.4854 (el valor de tabla) = 0.14 . Este problema se puede resolver gráficamente usando el programa INFOSTAT, con el módulo aplicaciones didácticas.



El área sombreada es la respuesta, que un estudiante tenga más de 28 años y tiene una probabilidad de 0,014.

Ejercicio 3.3 La estatura de una población tiene una distribución normal con media de 170 cm y una desviación estándar de 7.60
¿Cuál es la probabilidad que una persona seleccionada de este grupo tenga una estatura entre 165 y 180?

Ejercicio 3.4 Una población de niños en edad escolar tiene una media de 11.5 años y un desvío estándar de 3 años. ¿Cuál es la probabilidad de que un niño sea entre 8.5 y 14.5 años, más de 10, y menos de 12?

Ejercicio 3.5 El promedio de notas de un grupo de estudiantes es 70 y el desvío estándar es 10. ¿Cuál es la probabilidad que un estudiante obtenga más de 80 ptos.? ¿Cuál es la proporción de aplazados esperados (P<60 ptos.)?

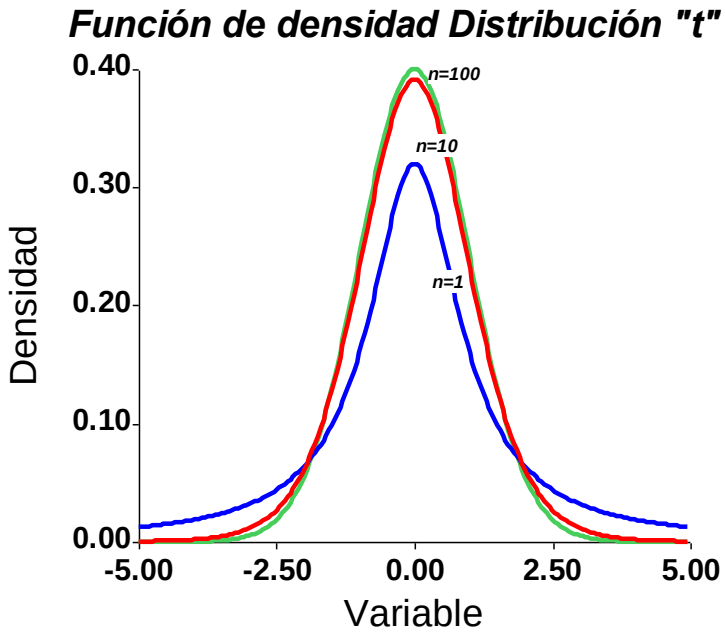
Ejercicio 3.6 Se producen quesos con un diámetro es 35cm y se acepta una varianza de 0.1 cm2. Si por problemas de envase se rechaza productos con diámetros menores a 34.5cm y mayores a 35.5 ¿Cuál es la probabilidad de rechazo de la producción por problemas de envase?

3.4 Distribución "t" de Student.

La curva Normal y Normal Tipificada son modelos teóricos adecuados para describir muchas poblaciones, basándose en dos parámetros μ y σ. Sin embargo por lo general, trabajamos con muestras, con μ y σ desconocidos, lo que da alguna inseguridad sobre el modelo empleado al desconocerse estos parámetros. Un investigador, Gosset (seudónimo Student) estudio este problema y llegó a la conclusión que la distribución Normal no funciona bien con muestras pequeñas, de tamaño menor a 30 datos, y encontró una distribución que supera este problema. Luego esta distribución se llamaría "t" de Student en honor a Gosset.

Esta distribución es simétrica, con forma de campana y su promedio vale 0. Cuando hay pocos datos la campana es más aplanada que una campana Normal, con de 30 datos la distribución "t" es casi igual que la distribución Normal Tipificada (0,1).

Esta Distribución se usa mucho para construir de intervalos de confianza de μ y realizar pruebas de hipótesis de uno y dos promedios con variables continuas.

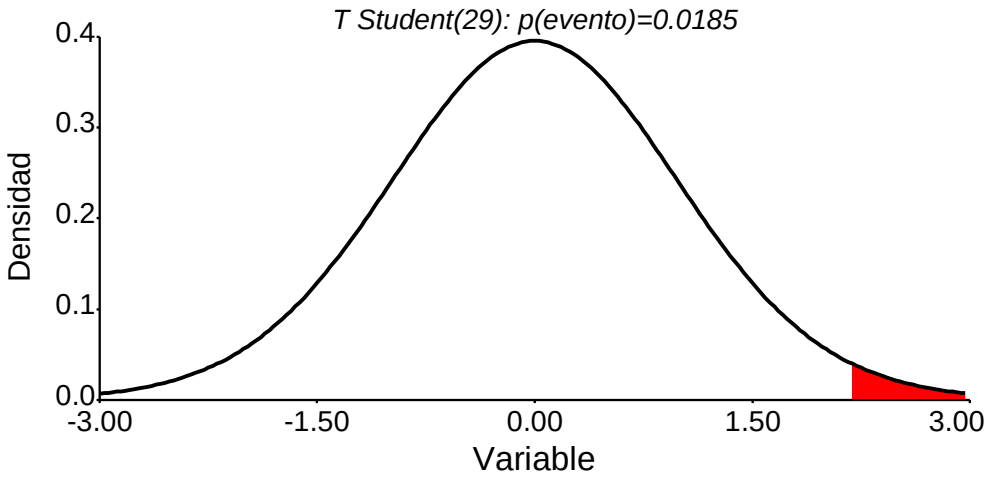


Se observa que a más datos, la campana es más alta, con valores menos dispersos y semejante a una curva Normal.

Ejercicio 3.7 Si consideramos los datos del ejercicio 3.2 de una distribución Normal, donde ahora $\bar{x}$ = 21 años y S = 3.2 años, pero obtenidos de una muestra de 30 alumnos y no de la población de estudiantes, la probabilidad que un estudiante tenga más de 28 años calculada con el módulo aplicaciones didácticas de INFOSTAT es de 0.018, levemente mayor que el valor 0.014 calculado con la distribución normal tipificada Se puede a continuación observar el gráfico de densidad de una “t” con 29 grados de libertad, n-1, donde se muestra la probabilidad de un valor de “t” mayor a 2.1875, el valor Z_{28} antes calculado.

Si la muestra aumenta a 300 estudiantes el valor de “P” es de 0.14, igual al obtenido de la normal, lo que demuestra con muchos grados de libertad la distribución “t” es casi igual a una distribución Normal 0,1.

Gráfico de Función de Densidad de una Distribución "t" con 29 gl



3.5 La distribución X2 de Pearson.

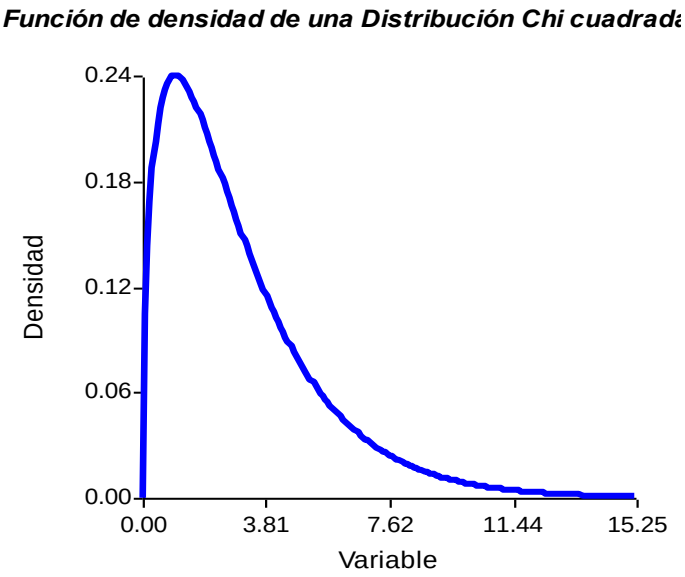
La distribución X^2 se genera a partir de "n" variables aleatorias independientes normales con media "0" y varianza "1" . Si realizamos la siguiente operación:

$$X_n^2 = z_1^2 + ...z_n^2$$

Es decir elevamos los "n" valores generados al cuadrado y los sumamos. Si aplicamos este procedimiento muchas veces, obtendremos la distribución de una variable que solo depende del número de sumandos. Esta distribución se denomina X^2 con "n" grados de libertad. Esta distribución no posee valore negativos por ser creada a partir de suma de valores cuadrados.

Este tipo de distribución se usa en pruebas de hipótesis sobre:

- Distribuciones, por ejemplo para verificar si una distribución observada se comporta como una distribución Normal.
- Independencia, para verificar si dos variables discretas son independientes o no.



3.6 La distribución "F" de Fisher.

La distribución "F" de Fisher surge del cociente de dos distribuciones X^2 independientes, con "n" y "m" grados de libertad respectivamente. Un valor "F" se define matemáticamente de la siguiente manera:

$$F_{n,m} = \frac{\frac{X_n{}^2}{n}}{\frac{X_m{}^2}{m}}$$

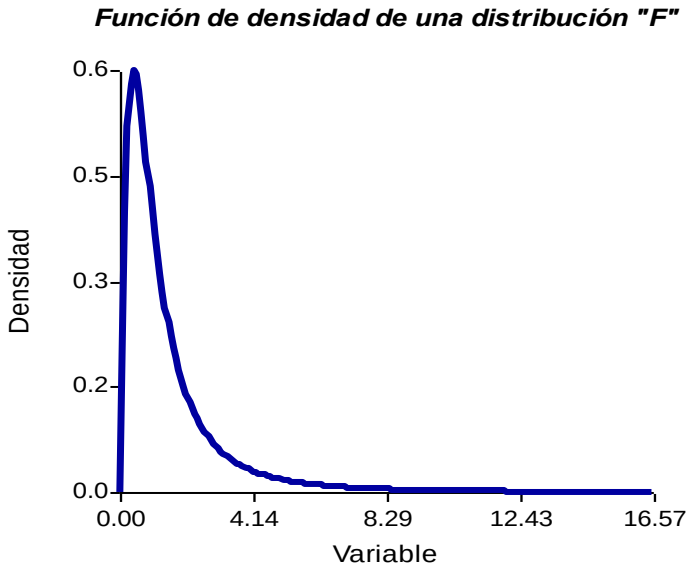
La distribución de "F" es asimétrica y comienza del valor "0", no posee valores negativos, al igual que la distribución X^2.

Este tipo de distribución se usa mucho con pruebas de hipótesis de promedios , Análisis de Variancia, donde:

- Hipótesis nula, los promedios de los tratamientos pertenecen a un mismo promedio poblacional

$$H_0 : x_1, x_2 .... x_n \in \mu$$

- Hipótesis alternativa, al menos un promedio de los tratamientos evaluados no pertenecen al mismo promedio poblacional



3. 7 La distribución Binomial.

Se usa con variables discretas, es decir cuyos valores son contables. Este modelo se aplica a poblaciones finitas de las que tomamos elementos al azar con reemplazamiento y también a poblaciones conceptualmente infinitas, como son piezas que generara una máquina, siempre que el proceso generador sea estable (proporción de pieza defectuosas constante a largo plazo) y sin memoria (el resultado en cada momento es independiente de lo previamente ocurrido).

Un experimento Binomial tiene las siguientes características:

- Las observaciones se clasifican en dos categorías, por ejemplo A = aceptable y D = defectuoso.
- La proporción de elementos A y D en la población es constante y no se modifica, siendo en este caso “p” la probabilidad de defectuosos y “q” la probabilidad de aceptables.

- Las observaciones son independientes, es decir que la probabilidad de elemento defectuoso es siempre la misma y no se modifica por cualquier combinación de elementos defectuosos o aceptables observados.

Ejemplos de este proceso son:

- Observar cinco de cerdos hembras de una camada de 12 lechones recién nacidos,
- Ganar 4 veces apostando a docena en diez tiradas sucesivas de una ruleta
- La aparición de 10 plantas planta enferma en 100 plantas de cultivo.
- Tener 5 ladrillos defectuosos en un lote de 500 ladrillos.

Generalizando, la variable binomial posee siempre 2 eventos, por ejemplo “A” y “B”. Se define como “r”:

r = número de elementos del evento “A” al observar “n” experimentos

Conociendo que:

- “p” es la probabilidad de ocurrencia del evento A
- “q” es la probabilidad de ocurrencia del evento B

Por lo tanto la probabilidad de encontrar “r” elementos que cumplen el evento “A” luego de “n” repeticiones del experimento, se define como P (r):

$$P(r) = \binom{n}{r} p^r q^{n-r} \qquad \text{siendo } r = 0, 1, ..., n$$

Siendo $\binom{n}{r}$ las posibles combinaciones de ocurrencia de “r” en “n” experimentos y esto se resuelve de la siguiente manera:

$$\binom{n}{r} = n!/r!(n-r)!$$

Estos problemas se pueden resolver directamente o con una tabla de probabilidades binomiales.

Una distribución binomial **B(n,p)** se parece a una normal tanto más cuanto mayor es el producto n * p (o n * q si q<p, siendo q=1-p). Cuando “n * p” y “n * q” superan el valor 5, la aproximación es casi perfecta.

En estas condiciones:

B(n,p) se aproxima a un distribución normal, $N(np, \sqrt{npq})$

Veamos un ejemplo donde se usa esta distribución, ¿Cual es la probabilidad de nacer 5 varones en 12 nacimientos? Este problema se puede resolver con un diagrama de árbol de probabilidades, pero se hace muy complicado. Por distribución Binomial se resuelve el problema de la siguiente manera.

Si sabemos que:

- “A” evento varón
- “B” evento no varón, es decir mujer.
- “p” probabilidad de varón = 0.5
- “q” probabilidad de mujer = 0.5
- “n” son 12 nacimientos totales
- “r” son 5 nacimientos de varones

Por lo tanto:

$$P(\text{5 varones}) = \binom{12}{5} 0.5^5 0.5^{12-5}$$

Donde $\binom{12}{5} = 12!/5!(12-5)! = 792$

P (5 varones) = $792(0.5^5)0.5^7$ = 792 (0.03125) (0.0078125) = **0.19**

Ejercicio 3.8 El Ministerio del Trabajo reporta que 20% de la fuerza de trabajo en un pueblo está desempleada. De una muestra de 14 trabajadores, calcule las siguientes probabilidades con la fórmula de la distribución binomial (n=14, p=0.2): Resuelva:

1. Tres están desempleados: P(x=3)=.250
2. Al menos un trabajador está desempleado:

   P(x $\geq$ 1) = 1 - P(x=0) =1 - .044 = .956
3. A lo más dos trabajadores están desempleados:

   P(x $\leq$ 2)=.044 +.154 +.250 =.448

Ejercicio 3.9. Si el 20% de las piezas producidas por una máquina son defectuosas, ¿cuál es la probabilidad de que entre cuatro piezas elegidas al azar, a lo sumo 2 sean defectuosas?

Ejercicio 3.10. Si de seis a siete de la tarde se admite que un número de teléfono de cada cinco está comunicando, ¿cuál es la probabilidad de que, cuando se marquen 10 números de teléfono elegidos al azar, sólo comuniquen dos?

3.8 Distribución de Poisson

En teoría de probabiliddes y estadística, la distribución de Poisson es una distribución de probabilidad discreta. Expresa la probabilidad de un número de eventos ocurriendo en un tiempo fijo si estos eventos ocurren con una tasa media conocida, y son independientes del tiempo desde el último evento. La distribución fue descubierta por Siméon Poisson en 1781–1840.

Una característica de la distribución de probabilidades binomial es que se hace cada vez más sesgada a la derecha conforme la probabilidad de éxitos disminuye. La forma límite de la distribución binomial donde la probabilidad de éxito es muy pequeña y n es

grande se llama distribución de probabilidades de Poisson. La distribución de Poisson se puede describir matemáticamente por la fórmula:

$$P(x) = \frac{\mu^x e^{-\mu}}{x!}$$

Donde *u* es la media aritmética del número de ocurrencias en un intervalo específico de tiempo, *e* es la constante 2.71828 y "x" es el número de ocurrencias.

El número medio de éxitos "*u"* se puede determinar en situaciones binomiales por "n p", donde "n" es el número de ensayos y p la probabilidad de éxito. La varianza de la distribución de Poisson también es igual a "n p".

En general utilizaremos la distribución de Poisson como una aproximación de experimentos binomiales donde el número de pruebas es muy alto, pero la probabilidad de éxito muy baja. En la práctica esta distribución se utiliza para calcular la probabilidad de un número específico de eventos, durante un período o espacio particular. El tiempo o la cantidad de espacio suele ser la variable aleatoria.

Ejemplo: un Hospital se especializa en el cuidado de lesiones menores. En las horas de la tarde de 6-10 PM el número medio de llegadas es 4.0 personas por hora. ¿Cuál es la probabilidad de 4 llegadas en una hora?
$P(4) = (4^4)\ (e^{-4}) / 4! = 0.1954$.

Ejercicio 3.11. La producción de computadoras trae asociada una probabilidad de defecto del 1.5%, si se toma un lote o muestra de 100 computadoras, obtener la probabilidad de que existan 4 computadoras con defectos.

Ejercicio 3.12. Se calcula que en la ciudad el 20% de las personas tienen afición a mirar TV de noche, si tomamos una muestra de 150 personas al azar ¿ Calcular la probabilidad de que 10 de ellos tengan el hábito de mirar TV de noche.

Ejercicio 3.13 El 6% de los registros contables de una empresa presentan algún problema, si un auditor toma una muestra de 50 registros ¿ Calcular probabilidad de que existan 5 registros con problemas ?

Ejercicio 3.14 En promedio, cada una de las 18 gallinas de un gallinero pone 0.5 huevos al día. Si se recogen los huevos cada 8 horas.
¿Cuál es el número medio de huevos que se recogen en cada visita? ¿Con qué probabilidad encontraremos x huevos para x = 0,1, 2, 3? ¿y la probabilidad de que $x \geq 4$?

Bibliografía y Documentos Consultados

Cebrían, M. 2001. Distribuciones continuas. Ministerio de Educación y ciencia. España. http://descartes.cnice.mecd.es/Bach_HCS_2/distribuciones_probabilidad/dis_continuas.htm

Dicovskiy, L.1998. Módulo Nro 3. Bioestadística. EAGE. Folleto de clase.

CYTA. Guía de Estadísticas. Distribución de Poisson http://www.cyta.com.ar/biblioteca/bddoc/bdlibros/guia_estadistica/index.htm.

Hospital Universitario Ramón y Cajal. Material docente de la Unidad de Bioestadística Clínica. Madrid. http://www.hrc.es/bioest/M_docente.html#tema2

Kessler, M. 2005. Apuntes de Métodos estadísticos de la Ingeniería http://filemon.upct.es/~mathieu/metodos/teoria/pdftema3.pdf

Peña D. Estadística 1, modelos y métodos, Fundamentos. 551 p.

Vermeer I..1996. Estadística, Curso Básico. EAGE. 104 p.

Zadú, I. Metodología. Variable Aleatoria.

http://www.southlink.com.ar/vap/VARIABLE%20ALEATORIA.htm

Unidad 4. Estimación y prueba de hipótesis.

Objetivos

- Desarrollar el concepto de estimación de parámetros
- Explicar que es una prueba de hipótesis
- Diferenciar grupos de una población utilizando pruebas de Student
- Diferenciar grupos de una población usando pruebas de varianzas
- Realizar pruebas de independencia chi cuadrado

Se debe poder hacer conclusiones generales para toda la población, a partir del estudio de las muestras.

4.1 Estimación por Intervalos de Confianza.

En estadística se llama estimación al conjunto de técnicas que permiten dar un valor aproximado de un parámetro (Ej.: μ, σ) de una población a partir de estadísticos, generados por los datos (Ej: $\bar{x}$, S, n). Un estimador puntual de un parámetro es un valor que puede ser considerado representativo de este y se obtiene a partir de alguna función de la muestra, por Ej. $\bar{x}$, promedio muestral, estima puntualmente a μ, el promedio poblacional.

La estimación por intervalos consiste en la obtención de un intervalo dentro del cual estará el valor del parámetro estimado, con una cierta probabilidad. Un uso de la distribución Normal y de la "t" de Student es la creación de Intervalos de confianza, estimación por intervalos, de los promedios poblacionales, μ.

El promedio poblacional, μ, se estima por un intervalo calculado a partir de "S" y $\bar{X}$ de muestras.

El intervalo de confianza de μ con un 95 de confianza, $IC_{95\%}$, es el más usado y para muestras de más de 30 datos se calcula como :

$$IC_{95\%} = \bar{x} \pm 1.96 \ (s/\sqrt{n})$$

Para menos de 30 datos se usa:

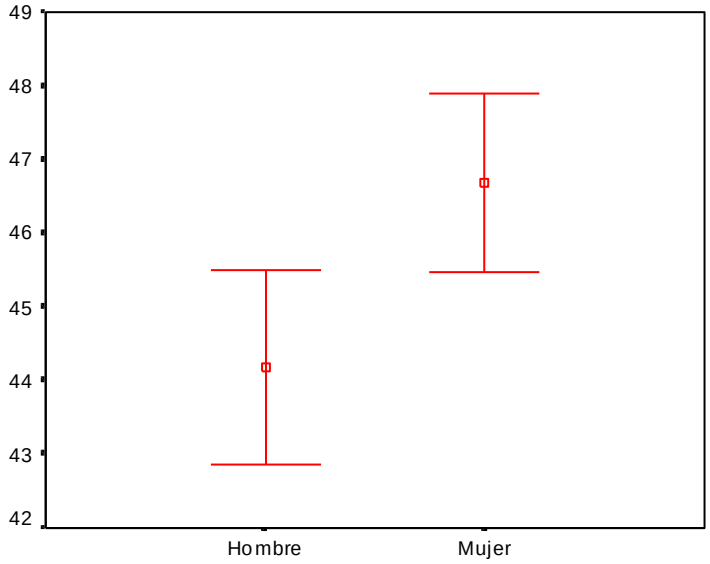
$$IC_{95\%} = \bar{x} \pm \text{“}t_{95}\text{”} \ (s/\sqrt{n-1}),$$

Donde “t” es el valor dado por la distribución “t” de Student con “n-1” Grados de Libertad, para un 95 % se busca el valor del “t” 0.975, ya que esta es una prueba de dos colas.

El $IC_{95\%}$ nos dice que con un 95 % de confiabilidad en este intervalo encuentro el promedio de la población, el cual desconozco. Para esto necesito conocer de la muestra los siguientes estadísticos: $\bar{x}$ **, S** y **“n”**.

El gráfico de $IC_{95\%}$ se usa cuando se cruza una variable discreta que genera grupos, con una variable continua. En este gráfico se observan los promedios de cada grupo con sus intervalos de confianza al 95 %, estos en forma de dos rayas. Veamos un ejemplo de este tipo.

Gráfico de Promedios e Intervalos de Confianza de μ, "$t_{95\%}$", desagregada por sexo, de la Edad de una población adulta.



En este tipo de gráfico es interesante observar si los intervalos de confianza de los diferentes promedios tienen valores superpuestos, ya que si es así, al hacer una prueba de hipótesis lo más probable que la respuesta sea de hipótesis nula, es decir los "promedios superpuestos" pertenecen a un mismo promedio poblacional.

4.2 Generalidades de las pruebas de Hipótesis

Una *hipótesis estadística* es una asunción relativa a una o varias poblaciones, que puede ser cierta o no. Las hipótesis estadísticas se pueden contrastar con la información extraída de las muestras y tanto si se aceptan como si se rechazan se puede cometer un error.

La hipótesis formulada con intención de rechazarla se llama hipótesis nula y se representa por H_0. Rechazar H_0 implica aceptar una hipótesis alternativa (H_1).

La situación se puede esquematizar:

	H_0 cierta	**H_0 falsa** **H_1 cierta**
H_0 rechazada	Error tipo I (α)	Decisión correcta (*)
H_0 no rechazada	Decisión correcta	Error tipo II (β)

(*) Decisión correcta que se busca

α = p (rechazar H_0 siendo H_0 cierta)

β = p (aceptar H_0 siendo H_0 falsa)

Potencia =1-β = p (rechazar H_0 siendo H_0 falsa)

Detalles a tener en cuenta

1 α y β están inversamente relacionadas.

2 Sólo pueden disminuirse las dos, aumentando n.

Los pasos necesarios para realizar un contraste relativo a un parámetro θ son:

1. Establecer la hipótesis nula en términos de igualdad

H_0: $\theta = \theta_0$

2. Establecer la hipótesis alternativa, que puede hacerse de tres maneras, dependiendo del interés del investigador

H_0: $\theta \neq \theta_0$ ó $\theta > \theta_0$ ó $\theta < \theta_0$

En el primer caso se habla de contraste *bilateral* o de *dos colas*, utilizada generalmente con la distribución Normal o la "t" de Student . En los otros dos casos se tiene un contraste *lateral derecho* en el segundo caso, o *izquierdo* en el tercer caso, ambos son pruebas de *una cola*. Con la distribución X^2 y la "F", por ser distribuciones asimétricas positivas los contrastes son unilaterales por el lado derecho.

3. Elegir un *nivel de significación*: nivel crítico para α, generalmente del 5 %.

4. Elegir un *estadístico de contraste y verificar hipótesis*: Este estadístico de contraste es un estadístico cuya distribución es conocida en H_0, que esté relacionado con θ y permite establecer, en base a dicha distribución, la *región crítica*: región en la que el estadístico calculado tiene una probabilidad menor que α.

Obsérvese que, de esta manera, se está más seguro cuando se rechaza una hipótesis que cuando no. Por eso se fija como H_0 lo que se quiere rechazar. Cuando no se rechaza, no se ha demostrado nada, simplemente no se ha podido rechazar. Por otro lado, la decisión se toma en base a la distribución de la muestra en H_0.

5. Calcular el estadístico para una muestra aleatoria y compararlo con la región crítica. Si el estadístico tiene en valor absoluto, un valor menor al valor tabular de la distribución conocida correspondiente al α se acepta H_0.

Esto es equivalentemente calcular el "valor p" del estadístico (probabilidad de obtener ese valor, u otro más alejado de la H_0, si H_0 fuera cierta) y compararlo con valor de α. Este valor “p” es el valor de una integral y generalmente lo calculan los programas estadísticos como el Infostat o SPSS. Si el valor de α es el 5 % la regla de decisión para aceptar o rechazar una hipótesis en pruebas unilaterales es la siguiente:

- Si el valor “p” calculado es > a 0.05 ocurre H_0
- Si el valor “p” calculado es ≤ a 0.05 ocurre H_1

4.3 Prueba de hipótesis con pruebas “t”

El promedio de una muestra pertenece a población con promedio conocido.

Esta es una prueba que permite contrastar si una muestra de una variable difiere significativamente de un promedio poblacional dado o no. Generalmente este promedio es histórico.

La hipótesis nula es H_0: $\mu = \bar{x}$

El estadístico de contraste es el valor “t” calculado:

$$t_C = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

Ejemplo: Históricamente la edad de los alumnos que entran a primer año de la Universidad es de 18 años. Se quiere saber si para el año que viene la edad de ingreso será la misma a la histórica, para estudiar esto se tomó una muestra de 36 estudiantes del último año de secundaria y se calculó la edad de ingreso a la universidad. En función de los datos observados surge la hipótesis de que la edad de los estudiantes es mayor que 18 años. La muestra de 36 sujetos dio los siguientes datos:

$\bar{x}$ = 18.5 S=3.6

Se trata de un contraste sobre medias. La hipótesis nula (lo que queremos rechazar) es:

$H_{0:}\ \mu = 18$

La hipótesis alternativa

$H_{1:}\ \mu > 18$

Este un contraste lateral derecho.

Fijamos "a priori" el nivel de significación en 0,05 y la región crítica $T>t_{\alpha}$ Si el contraste hubiera sido lateral izquierdo, la región crítica sería $T<t_{1-\alpha}$ y si hubiera sido bilateral $T<t_{1-\alpha/2}$ o $T>t_{\alpha/2}$. En este ejemplo $t_{(35)0,05}=1,69$.

Calculamos el valor de t_c en la muestra

$$"t_c" = \frac{18.5 - 18}{\frac{3.6}{\sqrt{36-1}}} = 0.82$$

No está en la región crítica (no es mayor que 1,69), por tanto no rechazamos H_0, concluimos que la edad histórica de ingreso se mantiene.

Dos promedios tomados en una misma muestra, en momentos diferentes, son iguales.

Esta es una prueba "t" para muestras relacionadas, donde pretendemos contrastar las medias de una misma muestra que se ha medido dos veces en los mismos sujetos, por ejemplo si en grupo de estudiantes queremos comparar el resultado del primer examen parcial de una asignatura con el del segundo parcial, se pretende saber si estos promedios difieren o no.

El estadístico de contraste es

$$t_c = \frac{\bar{d}}{S_d / \sqrt{n}}$$

Donde $\bar{d}$ es el promedio de las diferencias de los datos repetidos, S_d es la desviación estándar de las diferencias. "n" es el número de pares (diferencias).

Los promedios de dos muestras o grupos pertenecen a una misma población.

Esta es una prueba de hipótesis muy usada cuando se tienen dos grupos y se quiere saber si estos tienen un mismo promedio poblacional.

La hipótesis nula es $H_{0:}\ \mu_1 = \mu_2$, la hipótesis alternativa es H_0: $\mu_1 \neq \mu_2$

Hay diferentes tipos de prueba "t", pero suponiendo varianzas iguales, el estadístico a calcular se hace:

$$"t_c" = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{S_1^2}{n_1 - 1} + \frac{S_2^2}{n_2 - 1}}}$$

Ejemplo

En un ensayo para evaluar la vida útil de dos productos. La variable medida es el tiempo de vida útil en años: producto "T", n = 35; $\bar{x}$ = 3,7 años de vida y s^2 =13,9; producto "P" n = 40; $\bar{x}$ = 15,1 años y s^2 = 12,8. ¿El producto "P" tiene igual vida útil que el producto "T"? Se trata de un contraste sobre diferencias de medias

Como no conocemos como son las varianzas entre sí, el modelo nos obliga a verificar si la varianzas son iguales, si fueran distintas es otra la prueba "t" a realizar. Para ello se debe plantear primero un contraste de prueba de hipótesis de variancias. Si las variancias son iguales se sigue con la prueba "t" que se presenta, sino se debe hacer otra variante de prueba "t" de más difícil cálculo.

Hipótesis de Variancias

$$H_0 : \sigma_T^2 = \sigma_P^2 \qquad H_1 : \sigma_T^2 \neq \sigma_P^2$$

El estadístico es de contraste es una prueba "F"= $S_P^2 / S_T^2 = 13.9 / 12.8 = 1.09$, como el valor "F" de tabla es 1.74, en consecuencia aceptamos la H_0 y concluimos que las varianzas son iguales. Luego se hace la prueba de hipótesis de promedios con el estadístico antes detallado.

$$"t" = \frac{15.1 - 3.7}{\sqrt{\frac{13.9_P}{35-1} + \frac{12.8_T}{40-1}}} = 13.28$$

Se concluye que se rechaza la H_0, ya que el valor "t" calculado es mayor que el valor de tabla con $n_1 + n_2 - 2$, 35 + 40 -2 = 73 grados de libertad. Con estos grados de libertad y con un alfa del 5% bilateral (2.5 % de rechazo en cada extremo) el valor "t" es de 2.0, valor menor que 13.28. Concluimos que los promedios de años de vida útil de los dos productos son distintos.