
Conceptos Básicos de Estadística para Ingenieros

Autores:

Matemático Antonio Manuel Otero Dieguéz Ph.D
Universidad Técnica Estatal de Quevedo

Ing. Héctor Raúl Reinoso Peñaherrera M.B.A
Ing. Verónica Del Consuelo Tapia Cerdá Mg.C
Ing. Edwin Homero Moreano Martínez Mg.C
Universidad Técnica de Cotopaxi

Ing. Julio César Pino Tarrago Ph.D
Universidad Estatal del Sur de Manabí - UNESUM

Ing. William Moisés Bonilla Jiménez Mg.C
Universidad de Las Fuerzas Armadas ESPE





The background features a minimalist abstract design. It consists of several nested rectangles: a large dark gray rectangle at the bottom, a medium gray rectangle above it, and a small white rectangle at the top. There are also two thin horizontal white lines extending from the left side of the medium gray rectangle towards the right.

Conceptos Básicos de Estadística para Ingenieros

Autores

Antonio Manuel Otero Dieguez, Matemático, Universidad Estatal de Odesa (I.I. Mechnikov, Ukrانيا, URSS), Doctor en Matemáticas, Universidad de Oriente. Investiga en temas: teoría asintótica de las ecuaciones diferenciales ordinarias, métodos numéricos para las Ciencias Técnicas, estudio de la convergencia de algoritmos numéricos.

Héctor Raúl Reinoso Peñaherrera, Ingeniero Mecánico, Escuela Politécnica Nacional, Diplomado en Gestión de Energías, Diplomado en Diseño y Evaluación de Proyectos Sociales, FLACSO Ecuador, Magister en Administración y Marketing, Universidad Indoamérica, Egresado de la Maestría en Diseño Mecánico, Universidad Técnica de Ambato. Investiga en temas: Aplicaciones con materiales termoplásticos, optimización de dispositivos con varios grados de libertad para niños entre 1 y 3 años, diseño de productos.

Verónica Del Consuelo Tapia Cerdá, Ingeniera en Sistemas e Informática, Universidad Regional Autónoma de los Andes, Magister en Ingeniería de Software, Universidad de las Fuerzas Armadas ESPE, Magister en Docencia Universitaria y Administración Educativa, Universidad Indoamérica. Varias publicaciones relacionadas con los Sistemas y Tecnologías de la Información, Ingeniería de Software y Gestión de Proyectos. Investiga acerca de la informática aplicada a la medicina y a la educación.

Edwin Homero Moreano Martínez, Ingeniero Electrónico, Escuela Politécnica del Ejercito, Magister en Gestión de Energías. Investiga en temas: Utilización de energías renovables en generación eléctrica con automatización de sistemas de control.

Julio Cesar Pino Tarrago, Ingeniero Mecánico, Universidad de Holguín, Doctor en Ciencias Técnica, Universidad Politécnica de Madrid. Investiga en temas: Optimización en el diseño de maquinaria agrícola.

William Moisés Bonilla Jiménez, Ingeniero Mecánico, Escuela Superior Politécnica de Chimborazo ESPPOCH, Diplomado Superior en Gestión del Aprendizaje Universitario, Magister en Gestión de Energías. Investiga en temas: Diseño de elementos de máquinas, Utilización de energías renovables en generación eléctrica para sistemas mecánicos.

Reservados todos los derechos. No se permite la reproducción total o parcial de esta obra, ni su incorporación a un sistema informático, ni su transmisión en cualquier forma o por cualquier medio (electrónico, mecánico, fotocopia, grabación u otros) sin autorización previa y por escrito de los titulares del copyright. La infracción de dichos derechos puede constituir un delito contra la propiedad intelectual.

© Copyright

Autores:

Matemático Antonio Manuel Otero Diegúez Ph.D
Universidad Técnica Estatal de Quevedo

Ing. Héctor Raúl Reinoso Peñaherrera M.B.A

Ing. Verónica Del Consuelo Tapia Cerdá Mg.C

Ing. Edwin Homero Moreano Martínez Mg.C
Universidad Técnica de Cotopaxi

Ing. Julio César Pino Tarrago Ph.D
Universidad Estatal del Sur de Manabí - UNESUM

Ing. William Moisés Bonilla Jiménez Mg.C
Universidad de Las Fuerzas Armadas ESPE

Diseño y Diagramación:

Ing. María Isabel Loján Jácome

Impresión:

IMPRESORA CHARITO

Cel.: 0995782845

Pujilí

ISBN: 978-9978-395-29-5

320 Ejemplares

PRIMERA EDICIÓN

ISBN: 978-9978-395-29-5



9 789978 395295

Prólogo

La estadística se remonta a dos tipos actividades dentro del desarrollo social, que se presentan aparentemente sin puntos comunes: los juegos al azar, y las actividades políticas.

La fundamentación matemática de los juegos al azar conllevó al desarrollo de la Teoría de las Probabilidades. Así como las necesidades que se presentan a las instituciones políticas para la descripción e interpretación de datos numéricos en los estudios sociales, económico y político de las poblaciones.

Por estadística puede ser definida la disciplina matemática que se relaciona con la recolección, procesamiento, análisis e interpretación de datos numéricos.

La estadística es presentada en dos direcciones en sus aplicaciones: la descripción de datos numéricos (estadística descriptiva) y la generalización basada en el análisis e interpretación datos numéricos (inferencia estadística).

El impacto de la estadística en la ingeniería está presente en las siguientes actividades:

- Uso eficiente de materiales para la construcción de nuevos productos.
- Uso eficiente de la fuerza de trabajo.
- Desarrollo de nuevos productos.
- Calidad de los nuevos productos.
- Mantenimiento y confiabilidad de los productos.

Lo que muestra la necesidad e importancia de su presencia dentro de los currículos de las carreras de ingeniería.

Sin embargo, la Estadística presenta ciertas dificultades de aprendizaje por parte de los estudiantes de titulaciones técnicas, ya que su aspecto cuantificador produce un cierto desasosiego e inseguridad en ellos. En este sentido, debemos concienciarlos de la peculiaridad de que el aprendizaje de nuestras asignaturas trasciende más allá del ámbito académico y que basta, por ejemplo leer un periódico, para encontrar la necesidad del estudio de los conceptos y herramientas estudiadas en las clases. En muchas ocasiones, mal utilizadas por los medios de información y políticos.

Este libro pretende ser un complemento didáctico en el estudio de la Teoría Básica Estadística, evitando la alta abstracción y el formalismo de las teorías matemáticas, presentando las herramientas y métodos estadísticos con un enfoque algorítmico, lo que pensamos facilita el empleo de las técnicas estadísticas por parte de los ingenieros.

En cada capítulo se presentan e ilustran los contenidos a través de ejemplos, se resuelven problemas concretos y se proponen diferentes actividades a realizar por los alumnos.

El objetivo nuestro es que este texto sirva de ayuda complementaria a todos aquellos estudiantes que se enfrentan por primera vez a la resolución de problemas que requieren: recolección, procesamiento, análisis e interpretación de datos numéricos.

Para los ya familiarizados con la literatura sobre temas estadísticos puede llamar la atención que la obra no presenta en el índice los temas relacionados con la Teoría de la Probabilidad. Una introducción y presentación de elementos de la Teoría de la Probabilidad está presente en el Anexo G. Donde se introduce y fundamenta la teoría de Fiabilidad (Confiabilidad o Sobrevivencia). Esto puede no ser usado en el curso básico de estadística, pero puede ser útil para los estudiantes en el desarrollo de proyectos de investigación.

Esperamos que esta primera edición contribuya a mejorar las experiencias del aprendizaje sistemático de la Estadística en las Ciencias de las Ingenierías. Agradeceremos todos los aportes que puedan hacernos para, a su vez, mejorar este instrumento didáctico.

MsC. Mirian Susana Pallasco Venegas

ÍNDICE

CAPÍTULO I: INTRODUCCIÓN AL ANÁLISIS DE DATOS

1.1 Las variables. Medición y clasificación.	11
1.2 Tipos de datos que analiza la estadística en la investigación.	12
1.3 Tablas de frecuencias.	17
1.4 Representación gráfica de los datos.	22
1.5 Medidas descriptivas.	26
1.5.1 Medidas de posición.	26
1.5.2 Diagramas de caja.	30
1.5.3 Medidas de dispersión.	30
1.6 Distribuciones de frecuencias bivariadas.	32

CAPÍTULO II: MUESTREO Y ESTIMACIÓN

2.1 Tipos de muestreo.	41
2.2 La tabla de números aleatorios.	44
2.3 Tamaño de muestra.	45
2.4 El Muestreo Aleatorio Simple (M.A.S.)	47
2.4.1 El muestreo sistemático:	48
2.4.2 El muestreo estratificado	49
2.4.3 El muestreo por conglomerado.	50
2.5 Estimación.	51
2.5.1 Estimación puntual	52
2.5.2 Distribuciones muestrales.	53
2.5.3 Distribución muestral de \bar{X} para σ^2 conocida	54
2.5.4 Distribución muestral de \bar{X} para σ^2 desconocida.	54
2.5.5 Distribución muestral de la varianza (s^2)	55
2.5.6 Error Máximo permisible	56
2.6 Estimación por intervalos	59
2.6.1 Intervalo de μ con σ^2 conocida.	60
2.6.2 Intervalo de confianza para la varianza poblacional	61

CAPÍTULO III: PRUEBA DE HIPÓTESIS

3.1 Prueba de hipótesis para la media de una distribución normal con varianza conocida.	70
3.2 Prueba de hipótesis para la media de una distribución normal con varianza desconocida.	71
3.3 Dócimas de hipótesis para la varianza de una distribución normal.	77
3.4 Pruebas X^2 de bondad de ajuste. Pruebas de Kolmogorov-Smirnov de bondad de ajuste.	86
3.5 Pruebas de Kolmogorov - Smirnov para la bondad de ajuste.	92
3.5.1 Pruebas de Kolmogorov - Smirnov para una muestra.	92
3.5.2 Dócima de Kolmogorov - Smirnov para la comparación de dos poblaciones	95
3.6 Distribuciones empíricas de frecuencia	97
Referencias Bibliográficas.	106
Anexos	109

CAPÍTULO I: INTRODUCCIÓN AL ANÁLISIS DE DATOS

1.1 Las variables. Medición y clasificación.

La Estadística Matemática es la rama de la Matemática Aplicada que se dedica al análisis de datos. Existen varias razones por las que el conocimiento de esta ciencia es fundamental para los que desarrollan cualquier investigación, entre ellas tenemos :

- Comprender la literatura profesional. Muchos libros y artículos de revista presentan informes experimentales en forma de resúmenes estadísticos o presentan teorías y argumentos utilizando conceptos estadísticos.

- La formación de un profesional exige que se diseñen y lleven a cabo experimentos. El diseño de un experimento es inseparable del tratamiento estadístico de los resultados y de una buena interpretación. Si el diseño de un experimento es defectuoso, ninguna manipulación estadística puede conducir a la extracción de inferencias válidas.

- La formación en Estadística es a su vez formación en método estadístico. La inferencia estadística es inferencia científica, lo que es a la vez inferencia inductiva, es decir, la extracción de afirmaciones generales a partir del

estudio de casos particulares. Estos términos son, a efectos prácticos, y en un cierto nivel de generalidad, sinónimos. La estadística intenta hacer una inducción rigurosa.

Así, pues, comprender la literatura científica, diseñar y llevar a cabo experimentos, y comprender las reglas del método científico como medio de formación intelectual son tres razones básicas por las cuales es conveniente estudiar la Estadística (Kim and Mueller 1978). Kerlinger (1975) define la Estadística del siguiente modo: «La teoría y el método de analizar datos cuantitativos obtenidos de muestras de observaciones para estudiar y comparar fuentes de variancia de fenómenos, ayudar a tomar decisiones sobre aceptar o rechazar relaciones hipotéticas entre los fenómenos y ayudar a hacer inferencias fidedignas de observaciones empírica»

De dicha definición se sugieren cuatro metas para la Estadística:

- A través de los estadísticos (índices de muestras) y de parámetros (índices de poblaciones) se pueden reducir grandes cantidades de datos a formas manejables y comprensibles.
- Ayudar en el estudio de poblaciones y muestras.
- Ayudar en la toma de decisiones.
- Ayudar a obtener inferencias fidedignas de datos de observaciones.

1.2 Tipos de datos que analiza la estadística en la investigación.

En las definiciones de Estadística, dada aquí y en otras, de un modo o de otro se ha hecho alusión a datos: a datos cuantitativos, a observaciones, a recogida de información, a recogida de datos, ¿Qué datos? La respuesta no es otra que ésta: la información recogida mediante un proceso de medida.

La medición consiste en la estimación del grado en que una cualidad es poseída, siendo expresada esa estimación numéricamente. En su sentido más amplio, dice Kerlinger (1975) medición es la asignación de numerales a objetos o acontecimientos.

En las investigaciones, la información sobre las variables se obtiene mediante dos procesos fundamentales: La clasificación, que es no cuantitativa, y la medición, que sí lo es. Hay dos tipos de variables, discretas y continuas: La base de esta distinción estriba en si solamente se puede clasificar o medir la variable por unidades enteras

(discretas) o si también puede haber unidades fraccionales (continuas).

Algunos expertos consideran que la clasificación no es medición. Dice Kerlinger (1975) que no existe la llamada variable «cuantitativa», puesto que siempre podemos asignar unos y ceros a variables categóricas, que son así susceptibles de cuantificación. Cuando los números o símbolos asignados a objetos no tienen significado numérico más allá de la presencia o ausencia de la propiedad o atributo que se mide, la medición se llama nominal.

Una variable que se expresa mediante medición nominal es, por supuesto, lo que se ha llamado categórica. Si, por el contrario, trabajamos con variables que tienen aspecto cuantitativo, entonces podemos utilizar el proceso de medición. En él intentamos obtener cierta estimación cuantitativa de la variable, es decir, de la cantidad de la variable que tiene cada uno de los sujetos. En este caso se puede aspirar a tres niveles de medición, que son, desde el más débil al más fuerte, el ordinal, el de intervalo y el de razón.

Antes de seguir adelante mencionando las tres escalas de medición más usuales, se centrará la atención en el análisis de algunas variables en orden a examinar su naturaleza y las formas como pueden ser presentadas. Variable es una característica que tiene más de un valor. Se contrapone a constante. Hay dos tipos de variables cuantitativas, continuas (las que se pueden expresar por unidades enteras y fraccionales) y discretas (aquéllas que solamente se pueden clasificar o medir por unidades enteras). Cuando se establecen categorías para cada valor de la variable, refiriéndose a características que no se pueden cuantificar pueden ser dicotómicas: solo pueden tomar dos valores (Krathwohl, 1998). Ejemplo: Variable sexo, o polítómicas: pueden darse más de dos valores en la característica medida. Ejemplo: Variable procedencia social Tabla I.

Examinemos estas variables: Peso, rendimiento académico, actitud, edad y sexo. Cuando se miden estas variables, es distinta la forma como son tratadas porque tienen significados distintos. Véase el cuadro para expresar lo que se quiere decir. Se alude a la naturaleza de la variable: si es continua o discreta; a su manifestación: si es cuantitativa o cualitativa; y las posibles escalas en que se puede expresar: razón, intervalo, ordinal o nominal.

Tabla 1. Algunas variables sociales.

Variable	Manifestación	Naturaleza	Escala de medición
Peso	Cuantitativa	Continua	Razón
Rendimiento	Cuantitativa	Continua	Intervalo, ordinal
Número de hijos	Cuantitativa	Discreta	Razón
Sexo	Cualitativa	Discreta	Nominal

Examinemos el peso. Es una variable de naturaleza continua se es más o menos pesado con un grado continuo de diferenciación. Lo peculiar de esta variable, a diferencia de las otras, es que tiene un cero absoluto. Es decir, cuando medimos a alguien que pesa cero (0) kg estamos diciendo que no existe, y si decimos de alguien que pesa 30 kg estamos diciendo que es el doble de otro que pesa 15 y el triple de quien pesa 10. Podemos, pues, afirmar, al dividir $30:15 = 2$ y $30:10 = 3$, etc., que una persona es el doble, triple, de pesado que otra. Ésta es la cualidad de algunas variables que permiten que se realice la operación de dividir. Por eso se dice que dichas variables se pueden expresar en una escala de **razón o cociente** porque es conocida la proporción de un valor de la escala a cualquier otro.

El rendimiento académico es una variable en la que no existe un cero absoluto: De una persona que obtiene un 0 en un examen no se puede decir que no sabe nada, al igual que se decía que no existía una persona que tenía de peso cero. Y tampoco se puede afirmar que quien ha sacado un 10 sabe el doble de otro que ha sacado un 5. La escala de mayor nivel en la que es susceptible de ser expresado el rendimiento, es la de intervalo: Un alumno que ha obtenido un 9, tiene dos puntos más que otro que ha sacado en una prueba un 7, y éste dos puntos más que otro que obtuvo un 5. Pero, al igual que con el peso, podemos definir el rendimiento, en términos de orden éste rinde más que este otro (escala ordinal).

El número de hijos, aun siendo cuantitativa no es continua sino discreta, ya que tenemos 2, tres o no tenemos hijos pero no tenemos 2.3 hijos.

La medición nominal se caracteriza por atribuir números o símbolos a las diferentes categorías o clases en que se ha dividido un conjunto de tal forma que el mismo número o letra indique la pertenencia al mismo grupo o categoría (profesión: ladrero, 2 auxiliar, 3 agricultor, 4 empresario). El número no tiene ningún valor operativo, simplemente señala la pertenencia o no a ese grupo o categoría previamente establecida. No

se trabaja directamente con los números como tales, sino con sus frecuencias, es decir, el número de veces que se presenta un hecho o fenómeno en el grupo objeto de investigación, y en cada una de las categorías definidas con anterioridad. Como aplicaciones de este tipo de datos nominales, se encuentran la moda, la frecuencia, el coeficiente de correlación (C) o de Contingencia, la prueba de χ^2 Cuadrado con sus diferentes modalidades. En similar situación se encuentran las variables medidas en escala ordinal.

Cuando la medición es de intervalos las estadísticas que se pueden calcular son la media, la desviación típica, la correlación de Pearson y, en general, todas aquellas pruebas de resolución de contraste de hipótesis englobadas en la denominación genérica de pruebas paramétricas.

Cuando la medición es de razón están justificadas todas las operaciones matemáticas de suma, resta, multiplicación y división, además de poder determinar lo que es el doble, el triple, la mitad. En el campo de la estadística tenemos la media geométrica y el coeficiente de variación, que requieren de la existencia del punto 0 de la escala (Krathwohl, 1998).

Resumen

Medir es cuantificar y por tanto necesitamos establecer ciertas escalas para poder llevar a cabo la medición. Emplearemos 4 escalas de medición o cuantificación: nominal, ordinal, de intervalo y de razón o proporción (ver Tabla 2).

Tabla 2. Escalas utilizadas en las mediciones.

ESCALAS	DEFINICIÓN	EJEMPLO
Nominal	Datos Categóricos	Colores, Sexo, Estado Civil, nacionalidad
Ordinal	Datos ordenados por rangos con orden creciente o decreciente (rango)	Altos/Bajos Pesados/Ligeros Interesados/Desinteresados Nivel de Escolaridad
Intervalo	Intervalos iguales siendo el cero arbitrario	Tiempo, Test
Razón	Intervalos iguales, el cero se define como ausencia de la característica	Temperatura, Peso, Longitud

Escala nominal: Cuando se define una relación de equivalencia entre los elementos de la población, esto es, se establece un número determinado de clases o categorías tales que cada elemento pertenezca a una y solo una clase. Se establecen atributos o valores dados por cualidades y no hay relación matemática entre los elementos.

Se emplea sólo una escala nominal para distinguir a las unidades de análisis de una muestra (dividen a las unidades de análisis según sean iguales o no respecto a una característica).

Ejemplo (de una escala dicotómica): la variable sexo, tiene dos posibilidades de encasillamiento para las unidades de análisis: masculino y femenino. En muchos casos suele emplearse el siguiente código de transformación, Masculino: 1 y Femenino: 2.

Cuando la variable se especifica a nivel nominal, los únicos análisis matemáticos permitidos son aquellos a base de porcentajes, o frecuencias por categorías.

Escala ordinal: es una escala nominal entre cuyas clases (puntajes) está definido un orden de modo que cualesquiera que sean dos de ellas una será mayor o superior, en algún sentido que la otra.

Escala de intervalo: es una escala ordinal en la que se ha definido una distancia, una unidad de medida entre sus clases o puntajes, de modo que para un par de puntajes x y z cualesquiera tales que $x < z$ se puede expresar la cantidad de unidades, de igual medida, en que z es mayor a x .

Llamaremos longitud de un intervalo a la distancia entre dos clases. En este caso se tiene que la proporción o razón entre las longitudes de dos intervalos cualesquiera permanece invariable ante toda transformación de la escala de intervalo, o sea, ante toda transformación del tipo $y = ax + b$.

En las escala de intervalo se le atribuyen valores numéricos a las unidades de análisis. La mayoría de las variables cuantitativas en Ciencias Sociales suelen ser medidas en escalas de intervalo.

Ejemplos: el rendimiento académico, la escala de temperatura medida en grados centígrados, etc.

Escala de razón: es una escala de intervalo que posee un cero absoluto.

El cero absoluto se considera como la ausencia total de cualidad medida, y por

tanto es el valor que no puede ser rebasado en la parte inferior. Muchas variables cuantitativas de tipo físico se miden en escalas de razón como la edad, el peso, la longitud, la temperatura en grados Kelvin, etc.

Es muy importante saber distinguir el tipo de variable a utilizar, pues los procedimientos estadísticos están asociados a los tipos de variables y se usa uno u otro en dependencia de ello.

Autoevaluación:

Teniendo en cuenta su experiencia profesional o por necesidades del trabajo que desempeña, defina 5 variables, exprese su nivel de medición y clasifíquelas.

1.3 Tablas de frecuencias.

Existen dos enfoques en el análisis de datos, que más que excluyentes consideramos como complementarios: el enfoque descriptivo, y el enfoque inferencial.

La Estadística Descriptiva es la parte de la Estadística que opera con estadísticos usados sólo con fines descriptivos de muestras de las que derivan y no para describir una población o universo relacionado. Uno de los propósitos es resumir y describir de forma clara y conveniente las características de uno o más de un conjunto de datos.

La Estadística Descriptiva Univariada trata de describir una distribución de datos que provienen de la medición de una variable en una muestra. ¿Cómo se presentan o se pueden presentar los datos que provienen de una medición de una variable en una muestra?

Básicamente son tres las formas como los datos se presentan para el análisis en una investigación:

- Como puntuaciones directas,
- Como puntuaciones directas agrupadas en frecuencias,
- Como puntuaciones directas agrupadas en intervalos de frecuencias.

En Estadística, los datos que no han recibido ningún procesamiento y que el investigador los tiene, tal y como han resultado de su proceso de recolección, se denominan datos primarios.

Una vez que los datos primarios han sido recopilados el investigador debe

proceder a analizarlos. Para la presentación de la "información" recopilada se pueden utilizar tablas y gráficos estadísticos (Marascuilo and Serlin 1988)

Ejemplo: Los siguientes datos constituyen las mediciones de cuatro variables, realizadas a una muestra aleatoria de 40 estudiantes, donde: X es la calificación en determinada asignatura (en puntos), Y es el número de asignaturas en las que desarrollaron las habilidades en el uso de la computación (en cantidad), Z: es la valoración del material docente (en Excelente, Muy Bien, Bien, Regular y Mal) y W: es el interés profesional (en sí o no) (ver tabla 3).

Tabla 3. Muestra de datos del ejemplo

Estudiante	Calificación	Habilidades	Valoración	Interés
1	84	5	R	sí
2	72	5	B	sí
3	70	2	R	no
4	72	3	M	no
5	85	4	R	sí
6	84	4	R	sí
7	74	3	M	sí
8	77	3	M	sí
9	77	1	B	no
10	77	2	R	no
11	79	3	B	no
12	68	1	B	sí
13	79	2	R	sí
14	82	4	M	no
15	76	3	M	sí
16	78	3	B	sí
17	86	5	R	no
18	88	5	B	sí
19	80	4	R	no
20	81	4	M	no
21	66	3	M	no
22	75	4	M	no
23	67	3	M	sí
24	84	4	R	sí
25	77	3	R	sí
26	75	2	M	no
27	82	5	B	sí

28	67	1	R	no
29	71	2	R	no
30	88	4	B	sí
31	78	3	R	sí
32	76	3	M	sí
33	74	3	M	sí
34	87	5	B	sí
35	70	3	R	no
36	69	2	R	no
37	73	3	R	no
38	86	5	B	sí
39	73	3	R	sí
40	80	4	B	sí

Sería muy difícil, utilizando estos datos, tal y como aquí se muestran, responder las siguientes interrogantes: ¿cuántos de estos estudiantes tienen interés profesional y qué por ciento ellos representan del total?, ¿cuántos tienen notas, en la asignatura entre 66 y 70 puntos y qué por ciento ellos representan con respecto al volumen de esa muestra?, ¿cuántos de los estudiantes tienen una calificación superior a 85 puntos y han desarrollado las habilidades en el uso de la computación?, etc.

Las tablas y los gráficos, que son formas complementarias de presentación de los datos primarios, nos ayudarán a responder, con cierta facilidad, las anteriores preguntas y otras muchas. Estudiaremos primeramente las tablas.

Una tabla estadística (o simplemente, una tabla) es una disposición, arreglo o agrupamiento de los datos primarios, de modo tal, que el "investigador" pueda encontrar "regularidades esenciales" presentes en esos datos.

Una forma de organizar los datos en tablas, consiste en escribir ordenadamente todos los valores posibles, registrando al lado de cada uno el número de veces que ha aparecido. A esta organización se llama distribución de frecuencias.

Distribuciones de frecuencias univariadas y sus elementos.

La tabulación de los datos primarios de una sola variable, bien sea en una tabla simple o en una de agrupación, recibe el nombre de distribución de frecuencias univariada o distribución empírica univariada (Rivas et. al 1991).

Elementos de una distribución de frecuencia univariada.

Clases o intervalos: es el "arreglo" que se utiliza para distribuir los datos de la variable que se tabula. Se denota por k.

Si la variable es discreta se utilizan las clases y se tendrán tantas clases como valores tenga la variable. Se pueden nombrar categorías.

Si la variable es continua se utilizan los intervalos y para formarlos se tienen en cuenta un grupo de pasos que veremos a continuación.

A los extremos de cada intervalo se le denominan límites del intervalo: particularmente el menor de esos extremos, situado en la parte izquierda de la clase, se le llama límite inferior, (Li), y al otro, límite superior, (LS), ubicado al lado derecho de la clase (Quivy and Campenhoudt 2000)

A la diferencia entre el LS y el Li de la clase k se le denomina Amplitud del intervalo k y se denota Cj.

Punto medio o marcas de clases: Es la semisuma de los límites del intervalo. Se denota por Xi.

Frecuencia absoluta: Se denomina frecuencia absoluta al número de veces que aparece repetido un dato (ni).

Frecuencia absoluta acumulada: Se denomina frecuencia absoluta acumulada correspondiente a un dato, a la suma de la frecuencia de este dato y la de los datos anteriores (Ni).

Frecuencia relativa: Se denomina al cociente de las frecuencias absolutas por el número de datos (fi).

Frecuencia relativa acumulada: Se denomina frecuencia relativa acumulada correspondiente a un dato, a la suma de la frecuencia relativa del dato y la de los datos anteriores a él (Fi).

Propiedades:

- La suma de las frecuencias absolutas coincide con el número de datos y son siempre números no negativos.
- Las frecuencias relativas y las frecuencias relativas acumuladas son siempre números fraccionarios no mayores que 1 y su suma es aproximadamente igual a 1.

Ejemplo la tabla de frecuencias para la variable habilidades (tabla 4)

Tabla 4. Tabla de frecuencias para ejemplo

Xi	ni	fi	Ni	Fi
1	3	0,075	3	0,075
2	6	0,15	9	0,225
3	15	0,375	24	0,6
4	9	0,225	33	0,825
5	7	0,175	40	1

De aquí podemos conocer que hay 15 estudiantes que tienen 3 habilidades, que representan un 37,5 del total de estudiantes. Que existen 24 estudiantes que tienen a lo sumo tres habilidades. Hay un 40% de los estudiantes con más de tres habilidades.

Cuando queremos formar una tabla por intervalos procedemos así:

Mínimo = 66,0; Máximo = 88,0; Rango = 22,0

En este caso vamos a formar 5 intervalos por lo que la amplitud se divide entre el número de intervalos de donde obtenemos un valor de 4,4 este valor se aproxima para que sea mas fácil trabajar con él, por lo que el rango de la tabla sería de $5 * 4 = 25$. Lo que origina una diferencia entre ambos rangos de 3 unidades, las cuales repartimos entre el valor mínimo y máximo de los datos quedando Mínimo = 65 y Máximo 90. Este proceso lo realiza cualquier paquete de programa en segundos.

Tabla 5. Tabla de frecuencias por intervalos.

Intervalos	Xi	ni	fi	Ni	Fi
65-70	67,5	7	0,175	7	0,175
70-75	72,5	9	0,225	16	0,4
75-80	77,5	12	0,3	28	0,7
80-85	82,5	7	0,175	35	0,875
85-90	87,5	5	0,125	40	1

Hay 12 estudiantes que sus calificaciones se encuentran entre 75 y 80 puntos, estos representan el 30 % de los analizados. La nota máxima del 70 % de los

estudiantes con más bajos resultados es 80 puntos. Hay 12 estudiantes con resultados superiores a los 80 puntos.

Autoevaluación:

1. Las horas de trabajo en la elaboración de un plan de desintoxicación de 40 psicólogos de una clínica especializada están registradas en la tabla siguiente:

61	65	75	87	74	62	95	78
96	78	89	61	75	95	60	79
79	62	67	97	74	85	76	65
86	67	73	81	72	63	76	75
76	85	63	68	83	71	53	85

Determine:

- El mayor tiempo de trabajo.
 - El menor tiempo de trabajo.
 - Construya una tabla de frecuencias de 5 intervalos
 - El tiempo de trabajo del programador que está en el primer cuartil.
 - ¿Cuántos psicólogos trabajaron por encima del tiempo de trabajo promedio?
 - Haga el histograma y el polígono de frecuencias de esa distribución.
2. En una investigación sobre el número de niños agresivos detectados diariamente en 20 aulas de ua escuela se obtienen los siguientes resultados.

4 5 6 3 7 4 8 3 5 9 3 6 8 7 5 3 6 5 5

- Haga una distribución de frecuencias por puntos.
- Determine las medidas descriptivas.
- Diga a que por ciento de las aulas se le detectaron más de 5 niños agresivos.
- Represente gráficamente la información.

1.4 Representación gráfica de los datos.

Como ya se había planteado, las dos ayudas gráficas que más se utilizan en los informes de investigación son las tablas y las gráficas.

Cuando es necesario presentar datos las ayudas gráficas pueden facilitar la comunicación de la información a su audiencia en una forma más rápida.

Además de hacer el informe más fácil de leer y de entender, las ayudas gráficas mejoran su apariencia física: El gráfico tiene la ventaja de que permite apreciar más rápidamente el comportamiento de los datos

Las representaciones gráficas que puede utilizar para la visualización de los datos son muy variadas, desde gráficos de líneas, de pastel, de barra hasta gráficos en tres dimensiones (Peña and Romo 1997).

Diagramas de barras: nombre que recibe el diagrama utilizado para representar gráficamente distribuciones discretas de frecuencias no agrupadas. Se llama así porque las frecuencias de cada categoría de la distribución se hacen figurar por trazos o columnas de longitud proporcional, separados unos de otros. Existen tres principales clases de gráficos de barras:

Barra simple: se emplean para graficar hechos únicos

Barras múltiples: es muy recomendable para comparar una serie estadística con otra, para ello emplea barras simples de distinto color o trámado en un mismo plano cartesiano, una al lado de la otra

Barras compuestas: en este método de graficación las barras de la segunda serie se colocan encima de las barras de la primera serie en forma respectiva (Ibáñez 1993).

El diagrama de barras proporciona información comparativa principalmente y este es su uso principal, este diagrama también muestra la información referente a las frecuencias

Tabla 6. Distribución de temperatura por ciudad

CIUDAD	TEMPERATURA
A	12
B	18
C	24

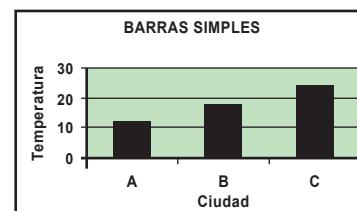


Figura 1. Gráfico que representa la distribución de temperatura por ciudad

Tabla 7. Distribución de las ganancias de las tiendas por meses

Tienda	Enero	Febrero	Marzo	Abril	Mayo	Junio
A	800	600	700	900	1100	1000
B	700	500	600	1000	900	1200

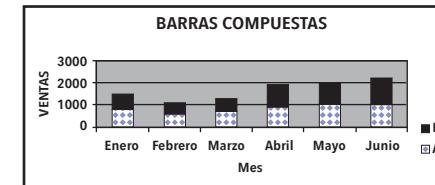


Figura 2. Gráfico que representa la distribución de las ganancias de las tiendas por meses

Histogramas: Se emplea para ilustrar muestras agrupadas en intervalos. Está formado por rectángulos unidos a otros, cuyos vértices de la base coinciden con los límites de los intervalos y el centro de cada intervalo es la marca de clase, que representamos en el eje de las abscisas. La altura de cada rectángulo es proporcional a la frecuencia del intervalo respectivo. Esta proporcionalidad se aplica por medio de la siguiente fórmula.

$$Ar = \frac{fi}{l}$$

Donde:

Ar = Altura del rectángulo

fi = frecuencia relativa

l = longitud de base

El histograma se usa para representar variables cuantitativas continuas que han sido agrupadas en intervalos de clase, la desventaja que presenta que no funciona para variables discretas, de lo contrario es una forma útil y práctica de mostrar los datos estadísticos.

Ejemplo: La representación gráfica de la tabla de frecuencias por intervalos (Tabla 5) se representa en la figura 3.

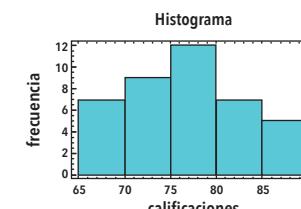


Figura 3. Histograma que representa la tabla de frecuencias por intervalo de la tabla 5.

Gráficos de sectores: es un gráfico que se basa en una proporcionalidad entre la frecuencia y el ángulo central de una circunferencia, de tal manera que a la frecuencia total le corresponde el ángulo central de 360º. Para construir se aplica la siguiente fórmula:

$$X = \frac{\text{frecuencia relativa} * 360^\circ}{S * \text{frecuencia relativa}}$$

Este se usa cuando se trabaja con datos que tienen grandes frecuencias, y los valores de la variable son pocos, la ventaja que tiene este diagrama es que es fácil de hacer y es entendible fácilmente, la desventaja que posee es que cuando los valores de la variable son muchos es casi imposible o mejor dicho no informa mucho este diagrama y no es productivo, proporciona principalmente información acerca de las frecuencias de los datos de una manera entendible y sencilla (Solanas et al 2002).

Ejemplo: Representar mediante un gráfico de sectores la frecuencia con que aparece cada una de las cinco vocales en el presente párrafo:

Tabla 8. Frecuencias de la cinco vocales en el párrafo anterior

Vocal	a	e	i	o	u	
Frecuencia	13	20	4	6	3	S 46

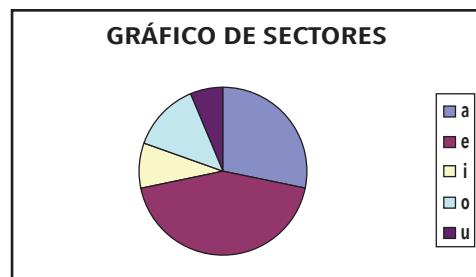


Figura 4. Gráfico de pastel que representa la distribución de las vocales de la tabla 8

Este gráfico es muy recomendable cuando hay que dividir el pastel en pocas partes o cuando hay varias partes pero una de ellas es muy superior a las demás

Autoevaluación:

- Obtenga, usando cualquier herramienta Informática, el histograma y un gráfico de pastel para las variables analizadas.

1.5 Medidas descriptivas.

Existen diversas situaciones en las que más que tener una presentación de los datos en una tabla de frecuencias o distribución empírica. Se necesitan "valores representativos" de estos a los que se les da el nombre de medidas descriptivas. Estas medidas ayudan a encontrar regularidades entre los datos que ellas describen. Las medidas descriptivas se pueden calcular para una variable de modo individual o para describir la "relación" existente entre dos o más variables, en cuyo caso se denominan medidas descriptivas de asociación. Las medidas descriptivas para una variable, de acuerdo con la "información" que proporcionan al investigador, se clasifican en medidas de posición, de dispersión, de deformación y de apuntamiento (Glass and Stanley 1980).

1.5.1 Medidas de posición.

Son medidas descriptivas que tienden a ubicarse hacia el centro de los datos de la muestra. Los valores que asumen estas medidas están incluidos entre el menor y el mayor de los datos medidos en la muestra. Esto no significa que una medida de este tipo ocupará exacta y necesariamente el centro de los datos, ni que el valor que ella toma tiene que coincidir con algún valor de los que han sido recolectados. A las medidas descriptivas de posición también se les denominan medidas de tendencia central o promedios.

Entre estas medidas tenemos: la media, la moda, la mediana, la media aritmética, los percentiles (entre ellos fundamentalmente los cuartiles).

La media.

La media aritmética o simplemente la media es la más importante medida de tendencia central. Ella representa un valor alrededor del cual oscilan los valores de la variable observada, constituyendo el centro de gravedad de la distribución. Se denota \bar{X} .

Ella solo tiene validez práctica cuando se le aplica a variables que estén medidas en escala métrica (intervalo y razón).

Para un conjunto de n datos primarios $x_1, x_2, x_3, \dots, x_n$, la media se calcula: Suma de todos los datos de la muestra dividida por el volumen de esta.

Donde: x_i representa a cada dato o valor de la variable, el signo Σ significa la suma de todos los datos de la muestra y n es el tamaño de esta.

Si los datos están previamente organizados en una tabla o distribución empírica:

$\bar{X} = \frac{1}{n} \left(\sum_{i=1}^k n_i Y_i \right)$. Donde Y_i representa para datos discretos los diferentes valores de la variable y para datos continuos el punto medio o marca de clase.

A esta medida es común llamarle, simplemente, media. También, se le suele decir promedio, aunque este último nombre se puede prestar a confusión, ya que sabemos que la media no es el único promedio que existe.

La media aritmética para cada muestra siempre existe, es única, puede o no coincidir con uno o más datos de esa muestra y no depende del tamaño de esta. Para su cálculo no requiere que los datos sean ordenados, ni tabulados y puede o no ser igual a la moda. Además, está "afectada" por cada elemento de la muestra, y principalmente, por los "valores extremos", es decir, por aquellos datos que se alejan mucho de los demás (Amón 1980). Quizás sea esta la gran deficiencia o limitación de esta medida, ello hace que, en ocasiones, la media no sea una "buena representación" de los datos de la muestra.

Propiedades de la media aritmética:

1. Si en una muestra todos los datos son iguales (constantes), entonces la media aritmética de esa muestra es esa misma constante.

2. La suma de las desviaciones o diferencias de cada dato de la muestra con respecto a su media aritmética, siempre es cero.

3. Si una muestra de tamaño n se subdivide en k submuestras, mutuamente excluyentes y exhaustivas, de volúmenes n_1, n_2, \dots, n_k ($n = n_1 + n_2 + \dots + n_k$), entonces la media de la muestra de extensión n es igual a cada n_i por su respectiva media dividido entre n .

En símbolos: $\bar{X} = \frac{(n_1 \bar{X}_1 + n_2 \bar{X}_2 + \dots + n_k \bar{X}_k)}{n}$

Ejemplo:

$$\bar{X} = \frac{1}{n} \left(\sum_{i=1}^n x_i \right)$$

Si $n = 40$, se ha subdividido en 4 submuestras con $n_1 = 8, n_2 = 10, n_3 = 12$ y $n_4 = 10$ y en cada submuestra se obtiene la media resultando:

$\bar{X}_1 = 3,5, \bar{X}_2 = 3,8, \bar{X}_3 = 3,1, \bar{X}_4 = 4,0$,

$$\bar{X} = \frac{(8)3,5 + (10)3,8 + (12)3,1 + (10)4,0}{40} = \frac{143,2}{40} = 3,58$$

Muy atentos a esta propiedad pues existe una tendencia a promediar promedios incorrecta.

La moda.

En una muestra de tamaño n , la moda, si existe, es el dato o los datos, que tienen mayor frecuencia absoluta. Se denota M_o .

De lo anterior se infiere que en una muestra para que haya moda, tiene que existir por lo menos un dato que se repita una cantidad de veces mayor que la que aparecen los demás. Por tanto, en una muestra la moda puede, o no existir, y si existe puede ser única o no. Así, si la moda es única, la muestra se dice que es unimodal, si existen dos modas es bimodal. La moda se puede calcular para cualquier escala de medición de la variable que se estudia.

La mediana.

La mediana de una muestra de volumen n está dada por el valor que supera a no más de la mitad de los datos y a la vez es superado por la mitad de los datos, estos datos han sido ordenados ascendente o descendente (es el valor (único) que ocupa el propio centro de dichos datos). Se denota M_e .

Es necesario tener en cuenta si la muestra que se estudia tiene una cantidad impar o par de datos. Si los datos están sin agrupar y n es impar, la mediana ocupa la posición $(n+1)/2$ de los datos; en cambio si n es par, entonces la mediana se encuentra entre los datos que ocupan las posiciones $n/2$ y $(n/2)+1$.

Cuando los datos están agrupados para localizar el intervalo que contiene a la mediana, se obtiene $n/2$, luego en las frecuencias absolutas acumuladas (N_i) se busca el primer valor que lo supere, el intervalo al que pertenece ese valor es el intervalo mediano.

Se aplica a niveles de medición ordinal, por intervalos y de razón..

La mediana para cada muestra siempre existe, es única, puede o no coincidir con uno o más datos de esa muestra, y no depende del tamaño de esta. Para su cálculo requiere que los datos estén ordenados; puede o no ser igual a la moda. Además, no está "afectada" por cada elemento de la muestra, y mucho menos, por los "valores extremos" de esta.

En el momento de realizar la interpretación de la mediana se deberá tener mucho cuidado, ya que en ocasiones esta coincide con algunos de los datos primarios y en otras no.

Cuartiles y percentiles.

los cuartiles dividen a los datos de la muestra en cuatro partes, por lo tanto, existen tres cuartiles que denotaremos por C_1 , C_2 y C_3 . El primer cuartil C_1 , es el valor que supera a no más de la cuarta parte de los datos, y a la vez, es superado por no más de las tres cuartas partes de esos datos; el segundo cuartil C_2 es igual a la mediana y el tercer cuartil C_3 , es el valor que supera a no más de las tres cuartas partes de los datos, y a la vez, es superado por no más de la cuarta parte de los datos. De igual forma se definen los quintiles, deciles y centiles, los cuales son valores de X que dividen a la distribución en cinco, diez y cien partes iguales respectivamente.

Un percentil es un punto que divide a la distribución de frecuencias en dos partes de tal forma que a su izquierda o por debajo de él se encuentre un determinado por ciento del total de observaciones.

El p -ésimo percentil de la muestra es un valor tal que al menos $100p\%$ de las observaciones están en o por debajo de ese valor, y cuando menos $100(1-p)\%$ están en o sobre ese valor. Esto no define exclusivamente a un percentil. Por simplicidad, si más de una observación satisface la definición, tomaremos su promedio.

Para calcular cualquier percentil los datos de la muestra tienen que estar ordenados, según su magnitud. Este ordenamiento puede ser ascendente o descendente.

Ejemplo:

Con los datos de la variable X : Calificaciones, calculemos los cuartiles.

Para ello el primer paso es ordenar los valores en forma ascendente los valores de la variable

Calificación: 66 67 67 68 69 70 70 71 72 72 73 73 74 74 75 75 76 76 77
77 77 77 78 78 79 79 80 80 81 82 82 84 84 84 85 86 86 87 88 88

El primer cuartil debe tener al menos $\frac{1}{4} * 40 = 10$ observaciones en o por debajo de su valor y al menos $\frac{3}{4} * 40 = 30$ en o mayores. Tanto el décimo como el decimoprimer valor más pequeño satisfacen el criterio, de modo que tomaremos su promedio.

$$C_1 = (72 + 73) / 2 = 72,5$$

El segundo cuartil o mediana, es el promedio de las observaciones ordenadas número 20 y número 21

$$C_2 = (77 + 77) / 2 = 77$$

El tercer cuartil es el promedio de las observaciones número 30 y número 31

$$C_3 = (82 + 82) / 2 = 82$$

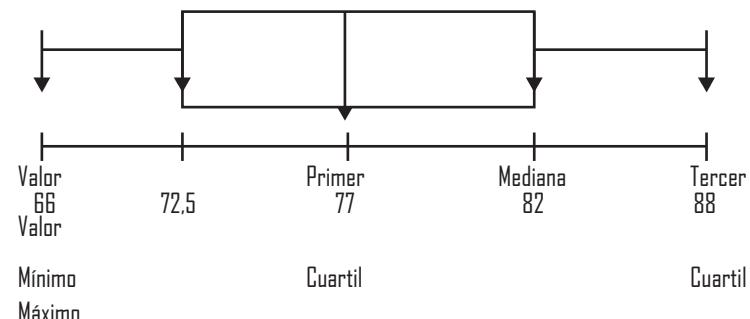
Por ejemplo, si calculamos el nonagésimo quinto percentil da una descripción útil de las calificaciones.

$$P_{0.95} = (87 + 88) / 2 = 87,5$$

Solo el 5% de los estudiantes tienen calificaciones superiores a 28,5 puntos.

1.5.2 Diagramas de caja.

El resumen de la información contenida en los cuartiles se visualiza en una presentación gráfica que se llama diagrama de caja. La mitad central de los datos, que va desde el primer hasta el tercer cuartil, se representa mediante un rectángulo. La mediana se identifica mediante una barra vertical dentro de esta caja. Una línea se extiende desde el tercer cuartil hasta el valor máximo y otra línea se extiende desde el primer cuartil hasta el mínimo (Hernández 1982).



Los diagramas de caja son de especial eficacia para retratar comparaciones entre conjuntos de observaciones. Son fáciles de comprender y tiene un gran impacto visual.

Con los datos del ejemplo de la variable X Calificaciones se obtuvo el siguiente diagrama de caja y bigote.

1.5.3 Medidas de dispersión.

El cálculo de las medidas de posición, por sí solas, no informan mucho si estas medidas no son acompañadas de otras que nos indiquen si existe mucha

variabilidad en la información, o si por el contrario, la masa de datos se encuentra concentrada alrededor de cierto valor.

Estas medidas permiten determinar el grado de acercamiento (alejamiento) que tienen los datos de la muestra respecto a una medida de tendencia central. Entre las medidas de dispersión están el rango, la varianza, la desviación estándar, el coeficiente de variación y el error estándar de la media.

El rango.

Es la medida de variación más simple que se utiliza y está dado por la diferencia entre el dato mayor y el dato menor de la muestra de tamaño n . Se denota por R y en símbolos es $R = X_{\max} - X_{\min}$

Cuanto más grande sea el rango, mayor será la dispersión de los datos de una distribución.

La varianza.

La varianza de una muestra de volumen n es la media aritmética del cuadrado de las desviaciones de cada dato respecto a la media de esa muestra.

$$\text{Se denota por } S^2 \text{ y su fórmula de cálculo es: } S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

Propiedades:

1. La varianza es siempre un número no negativo, es decir, será cero o un valor con signo positivo.
2. La varianza de una constante c es igual a cero.
3. La varianza de la suma de una variable y una constante es igual a la varianza de la variable.
4. La varianza del producto de una constante por una variable, es igual al producto del cuadrado de la constante por la varianza de la variable.

El valor de la varianza se expresa en unidades cuadráticas y su utilidad está dada en que da una medida del grado de desviación de los datos respecto a su promedio, estos datos son lineales, por lo que para eliminar esta dificultad se puede extraer la raíz cuadrada a la varianza, con lo que se obtiene otra medida de dispersión (Hernández 1982).

La desviación típica.

La desviación típica o estándar de una muestra de tamaño n es la raíz cuadrada positiva de la varianza. Se denota por S y en símbolos es: $S = +\sqrt{S^2}$

Esta medida es la que se interpreta. Mientras menor sea el valor de la desviación típica, menor será el grado de dispersión de los datos respecto a la media aritmética.

El coeficiente de variación.

El coeficiente de variación de una muestra de tamaño n es el cociente entre la desviación típica y la media aritmética de dicho muestra. Se denota por CV y en símbolos es: $CV = \frac{S}{\bar{X}}$. Con mucha frecuencia el valor de CV se multiplica por cien y se expresa en por ciento.

El coeficiente de variación es una medida muy propicia para comparar la variación entre dos conjuntos de datos que estén medidos en diferentes unidades, por ejemplo, una comparación entre la dispersión de las calificaciones y la dispersión del interés de los alumnos de la muestra.

El error estándar de la media.

El error estándar de la media de una muestra de tamaño n , es el cociente entre la desviación típica de la muestra y la raíz cuadrada del tamaño de esa muestra.

$$\text{Lo denotaremos por: } S_{\bar{X}} \text{ y su fórmula es } S_{\bar{X}} = \frac{S}{\sqrt{n}}$$

Autoevaluación:

- Analice las medidas descriptivas y de dispersión estudiadas en las diferentes variables del ejemplo.
- Obtenga, usando de cualquier paquete de programa Estadístico, las medidas descriptivas y de dispersión para los ejercicios de la auto evaluación anterior.

1.6 Distribuciones de frecuencias bivariadas.

Cuando se tabulan, de modo conjunto dos variables, la distribución de frecuencias se llama distribución bivariada. En ocasiones, se usan otras denominaciones para estas distribuciones; así, por la forma de su cuerpo, se les llaman tablas de "doble entradas" o "tablas de contingencia".

Para confeccionar estas tablas, se colocarán los "valores" de una de las variables en filas y los de la otra en columnas, ello se hace de modo indistinto. Se puede, utilizar las dos mediante agrupación o una de ellas de un modo simple y la otra en intervalos, para ello se seguirá las mismas reglas analizadas con anterioridad (Ritzer. 2003).

En el caso unidimensional se representaba las observaciones de la forma X_1, X_2, \dots, X_n , que es el que se había estudiado hasta ahora.

En el caso bivariado serán consideradas simultáneamente dos variables, o sea, serán estudiadas las distribuciones bidimensionales, las cuales serán denotadas de la forma (X, Y) , así por ejemplo si se observan simultáneamente (Field 2009).

- El número de hijos y el número de habitaciones de 50 núcleos familiares.
- La estatura y el peso de los estudiantes del grupo 4210.
- La edad y el ingreso de los profesores del Dpto. Estadística-Informática.

Esto es, de igual forma que en el caso unidimensional, las variables pueden ser discretas o continuas por lo que es factible analizar 2 variables discretas o dos variables continuas o una variable discreta y una continua a la vez.

Se estudiará de forma detallada como construir una tabla de frecuencia para variables bidimensionales discretas. Para ello es necesario elaborar una tabla denominada de DOBLE ENTRADA y que se forma escribiendo en el margen superior e izquierdo los distintos valores observados de cada una de las variables consideradas.

X_i	X_1	X_2	\dots	X_k	X_i	X_1	X_2	\dots	X_k
Y_j	n_{11}	n_{21}	\dots	n_{k1}	Y_j	f_{11}	f_{21}	\dots	f_{k1}
Y_1	n_{12}	n_{22}	\dots	n_{k2}	Y_2	f_{12}	f_{22}	\dots	f_{k2}
\vdots	\vdots	\vdots	\dots	\vdots	\vdots	\vdots	\vdots	\dots	\vdots
Y_m	n_{1m}	n_{2m}	\dots	n_{km}	Y_m	f_{1m}	f_{2m}	\dots	f_{km}

¿Qué representan i y j?

$i = 1, 2, \dots, k$ valores diferentes de la variable X

$j = 1, 2, \dots, m$ valores diferentes de la variable Y

n_{ij} = frecuencia absoluta conjunta, indica el número de repeticiones del valor X_i y del valor Y_j conjuntamente.

f_{ij} = frecuencia relativa conjunta, indica la fracción de repeticiones ó el por ciento de repeticiones del valor de las variables X_i y Y_j a la vez (Ritzer. 2003).

Ejemplo:

Considere las observaciones correspondientes a 25 laboratorios donde la primera variable (X_i) se refiere al número de virus detectados en un control y la segunda (Y_j) al número de técnicos que trabajan en él

No. de Técnicos Y_j	3 2 0 1 1 2 2 4 4 3 0 1 2 2 3 2 2 0 3 4 3 2 1 2 3	No.
de Virus X_i	1 3 1 2 2 2 3 4 3 2 1 2 2 1 3 2 1 2 4 3 3 2 2 2 3	
	Xi	

Construya la distribución de frecuencia bidimensional, para frecuencias absolutas y relativas. En este ejemplo se está trabajando con 2 variables discretas simultáneamente, recuerden que se podría trabajar con 2 variables continuas ó una continua y otra discreta.

¿Cómo se clasifican estas variables?

Primeramente se forma la tabla con los valores de la variable (el mismo tratamiento visto en variables discretas en el caso unidimensional). Es decir aquí la variable X toma los valores 1, 2, 3, y 4 y la variable Y toma los valores 0, 1, 2, 3, 4

Después se toman los pares, digamos el par (1,1) y se cuenta cuantas veces se repite y ese es el número que se pone en la tabla de doble entrada, en este caso es 2;

después se toma el par (2,1) y se hace lo mismo, se cuenta cuantas veces se repite, ninguna y se pone el cero, y así sucesivamente.

Y_j						n_{xj}
X_i	0	1	2	3	4	
1	2	0	2	1	0	5
2	1	4	5	1	0	11
3	0	0	2	3	2	7
4	0	0	0	1	1	2
n_{yji}	3	4	9	6	3	25

Como se aprecia

$$0 < n_{ij} < n$$

Se cumplen las mismas propiedades que para el caso unidimensional.

¿Cómo se interpretarían estas frecuencias absolutas?

n_{11} : Es elemento que está en la primera fila y primera columna significa que hay 2 laboratorios en los que trabaja un técnico y no se detectaron virus.

n_{23} : Es elemento que está en la segunda fila y tercera columna significa que hay 5 laboratorios en los que trabajan 2 técnicos y se detectaron 2 virus

De la misma forma que se presentó la tabla de doble entrada, con las frecuencias absolutas, se puede presentar con las frecuencias relativas, (recordar que la frecuencia relativa en el caso unidimensional era igual a $f_i = n_i/n$ en el caso bidimensional es igual), haciéndola a partir de la tabla de frecuencias absolutas.

El par (1,1) tiene frecuencia 2 por tanto sería $2/25 = 0.08$; el par (4,1) tiene frecuencia 1 entonces será $1/25 = 0.04$, y así sucesivamente. Se Divide entre 25 porque $n = 25$.

La tabla quedará de la siguiente forma:

Tabla de frecuencias relativas: Al igual que se planteó en la tabla anterior f_{yj} y f_{xi} son frecuencias marginales y que se estudiarán a continuación.

X_i	0	1	2	3	4	f_{xij}
Y_j	0.08	0	0.08	0.04	0	0.20
1	0.04	0.16	0.20	0.04	0	0.44
2	0	0	0.08	0.12	0.08	0.28
3	0	0	0	0.04	0.04	0.08
4	0	0	0	0.04	0.04	0.08
f_{yj}	0.12	0.16	0.36	0.24	0.12	1.00

Interpretación de las frecuencias relativas:

$f_{33} = 0.08$ indica que el 8% de los laboratorios tienen 2 Técnicos y se detectaron 3 virus.

$f_{45} = 0.04$ indica que el 4% de los laboratorios tienen 4 Técnicos y se detectaron 4 virus.

Las medidas más importantes y propias del caso bidimensional son: LA COVARIANZA Y el coeficiente de correlación. Las cuales serán estudiadas el tema de Correlación y Regresión.

CAPÍTULO II: MUESTREO Y ESTIMACIÓN

El muestreo es una herramienta de la investigación científica. Su función básica es determinar qué parte de una realidad en estudio (población o universo) debe examinarse con la finalidad de hacer inferencias sobre dicha población.

Al muestrear se reducen los costos, los gastos de recogida en recursos humanos, materiales y económicos y los de tratamiento de los datos serán menores, se logra mayor rapidez.

Cuando se muestrea los resultados pueden ser más exactos ya que al emplearse menor personal en la recogida de la información este puede ser más capacitado. En el caso que la acción de muestrear implique la destrucción de la unidad de muestreo resulta también más económica (Azorín 1972).

Población: Una población o universo es un conjunto finito o infinito de sujetos u objetos con una o más características cuyos integrantes tienen interés investigativo.

A cada uno de los integrantes de una población se le llama elemento de la población y al número total de ellos tamaño de la población. Se denota por N el tamaño de la población. La población puede ser finita o infinita en dependencia de la cantidad de elementos.

Cuando una medida descriptiva es obtenida para la población, recibe el nombre de parámetro; en tal caso, dicha medida caracteriza a esa población y para ella

cada parámetro es único. De forma general denotaremos a los parámetros por Θ . Particularmente se simbolizan con letras del alfabeto griego. Ejemplos de parámetros son la media poblacional (μ), la varianza poblacional (σ^2), la proporción poblacional (P), entre otros (Manly 1992).

Ejemplo 1:

Ejemplo de poblaciones:

- a. Los alumnos de la Educación Superior de Ecuador.
- b. Los alumnos de la Universidad Técnica de Cotopaxi.
- c. Los alumnos de la Maestría en Ciencias de la Educación Superior.

Censo: En ocasiones resulta posible estudiar cada uno de los elementos que componen la población, realizándose lo que se denomina un censo, es decir, el estudio de todos los elementos que componen la población.

Si la numeración de elementos, se realiza sobre la población estudiada, y no sobre la población teórica, entonces el proceso recibe el nombre de marco o espacio muestral.

Es importante cuando se va a realizar una investigación precisar cuál es el "marco" que abarca la población que se va a estudiar.

Muestra: Una parte o subconjunto de la población.

Característica: El signo o detalle que interesa estudiar.

Muestreo:

Se llama muestreo al procedimiento estadístico que se utiliza para seleccionar la muestra que será estudiada, es decir, es la recolección de información en la que se trabaja solo con una parte de la población.

En dependencia del tipo de muestreo empleado las muestras pueden ser probabilísticas y no probabilísticas. Elegir entre una muestra probabilística o una no probabilística, depende de los objetivos del estudio, del esquema de investigación y de la contribución que se piensa hacer con ella. Las muestras probabilísticas tienen muchas ventajas, quizás la principal es que puede medirse el tamaño del error en las predicciones. Para este tipo de muestra es necesario determinar el tamaño de la muestra para luego seleccionar los elementos muestrales.

2.1 Tipos de muestreo.

Muestreos no probabilísticos.

En los muestreos no probabilísticos, llamados también muestreos dirigidos, no es posible establecer a priori la probabilidad que tienen los miembros del universo, de ser seleccionados como parte de la muestra. El proceso de selección de los miembros de la muestra es subjetivo, a criterio y voluntad del investigador o del grupo de encuestadores. Su mayor inconveniente es la desconocida relación entre estimadores y parámetros, dificultando la estimación de estos últimos (Badii y Castillo. 2009).

¿Cuándo aplicar muestreo no probabilístico? Cuando se requiere una cuidadosa y controlada elección de sujetos con ciertas características especificadas previamente en el planteamiento del problema, cuando no hay un marco disponible para propósitos de muestreo y cuando se considera que no se requieren cifras exactas sobre la representatividad estadística de los resultados.

Debe tenerse bien claro que los resultados que se obtienen de muestras no probabilísticas son generalizables a la muestra en sí o a muestras similares. No son generalizables a la población.

Entre los diferentes tipos de muestreo no probabilístico se pueden mencionar:

- Muestreo por cuotas.
- Muestreo casual o fortuito.
- Muestreo de selección experta.
- Muestreo de conveniencia.

Muestreo por cuotas: También denominado en ocasiones "accidental". En este tipo de muestreo se fijan unas "cuotas" que consisten en un número de individuos que reúnen unas determinadas condiciones, por ejemplo: 20 estudiantes de 20 a 25 años, de sexo masculino y estudiantes universitarios residentes en Tegucigalpa. Se asienta generalmente sobre la base de un buen conocimiento de los estratos de la población y/o de los individuos más "representativos" o "adecuados" para los fines de la investigación. Mantiene, por tanto, semejanzas con el muestreo aleatorio estratificado, pero no tiene el carácter de aleatoriedad de aquél.

Una vez determinada la cuota se eligen los primeros que se encuentren que cumplan esas características. Este método se utiliza mucho en las encuestas de opinión.

Muestreo casual o fortuito: Aquí las muestras se integran por voluntarios o unidades maestrales que se obtienen en forma casual. Ejemplo: Un profesor investigador anuncia en su clase que va a hacer un estudio sobre motivación del universitario e invita a aquellos que acepten a someterse a determinadas pruebas.

Muestreo de selección experta: Denominado también como muestreo de juicio, es una técnica utilizada por expertos para seleccionar unidades representativas o típicas, según el criterio del experto; por ejemplo: la selección de un conjunto con determinadas características, para un experimento de laboratorio, o la selección de determinadas semanas del curso para llevar a cabo algunas evaluaciones.

Es importante hacer notar que en este caso los criterios de selección pueden variar de experto a experto, al determinar cuáles son las unidades de muestreo representativas de la población (Badii, y Castillo. 2009)..

Muestreo de conveniencia: Como su nombre lo indica son incluidos en la muestra los elementos de acuerdo con la conveniencia del investigador. Se justifica su empleo en la etapa exploratoria de la investigación como base para generar hipótesis.

Muestreos probabilísticos:

En un muestreo de tipo probabilístico, a partir de la muestra se pueden hacer inferencias sobre el total de la población. La selección de la muestra se puede hacer mediante un proceso mecánico similar al de una lotería, su equivalente práctico es la selección en las denominadas tablas de números aleatorios.

El tipo de muestreo probabilístico más importante es el muestreo aleatorio, en el que todos los elementos de la población tienen la misma probabilidad de ser extraídos; Aunque dependiendo del problema y con el objetivo de reducir los costos o aumentar la precisión, otros tipos de muestreo pueden ser considerados como veremos más adelante: muestreo sistemático, estratificado y por conglomerados.

Si el muestreo se realiza de tal manera que la unidad elemental se puede reemplazar (o devolver) a la población, de forma que pueda ser extraído de nuevo, tendremos

un muestreo con reemplazo. De una población de tamaño N se pueden seleccionar, con reposición, N elevado a la n muestras diferentes de tamaño (N^n).

Si la unidad elemental se retira de la población de manera que no puede volver a aparecer el muestreo es sin reemplazo, pudiendo determinarse el número de muestra diferentes de tamaño n de una población de tamaño N, mediante la expresión $\frac{n!}{N!(n - N)!}$.

Cuando el tamaño de la población que se investiga es grande, a las muestras que se han seleccionado sin reposición se les puede tratar, estadísticamente, con los métodos con los que se analizan las muestras con reposición.

Aunque en la selección de la muestra se haya empleado un muestreo probabilístico o aleatorio, como en la muestra no están incluidos todos los elementos de la población, es posible que se presente una diferencia entre el valor real del parámetro y el estimado lo que se conoce como "error de muestreo" o "error aleatorio", en símbolos: $e = \hat{\theta} - \Theta$. El valor de e podrá ser negativo -cuando el valor de la estimación sea menor que el del parámetro-, o positivo -en caso contrario-; pero también, e puede ser cero -si son iguales ambos valores-.

Sin embargo, en la práctica, esta particularidad no la llega a saber el investigador, ya que, "raras veces" conoce el valor del parámetro, pues no trabaja con todos los elementos de la población.

Por ejemplo si de una Universidad se obtiene el índice académico de los estudiantes y este resulta de 4,3 puntos; se selecciona aleatoriamente un grupo de estudiantes y se obtiene el índice y este resulta ser de 4 puntos, la diferencia entre ellos (0,3 puntos) es el error de muestreo, conocido también como "sesgo del muestreo".

El error de muestreo no es posible saberlo en cada caso específico, ya que por lo general, la población no es estudiada directamente, esto hace que se hayan desarrollado métodos estadísticos para "estimar" dicho sesgo, pero esto solo es posible hacerlo si se tiene una muestra aleatoria. El error aleatorio es la única desventaja que tiene el empleo de las muestras en la investigación. También pueden estar presentes otros errores, pero ellos dependen de las "habilidades" del investigador: no delimitar bien el marco de la población, no seleccionar adecuadamente el método de muestreo, extraviar o medir incorrectamente los datos, aplicar los instrumentos de investigación de modo indebido, etc. (Manly 1992).

2.2 La tabla de números aleatorios.

Una tabla de números aleatorios es una disposición, en filas y columnas, de dígitos, números del cero al nueve, de modo tal que estos números han sido ubicados al azar en dicha tabla.

Para emplearla en la selección de los elementos de una muestra se siguen los siguientes pasos:

1. Numeración de los elementos de la población desde 1 hasta N. Para hacer la numeración se tendrá en cuenta la cantidad de dígitos que tenga N; por ejemplo si N= 100, como tiene 3 dígitos, la numeración será 001, 002, 003,..., 045,..., 100.
2. Obtener el recorrido de los números aleatorios pudiendo seguirse cualquiera de los siguientes criterios:

- Recorrido desde 1 hasta N, (teniendo en cuenta lo que se planteó en 1.), es decir, ser menores o iguales que N: Se tomaran tan solo esos números, los que no cumplen con el requisito se desechan.

- Recorrido desde 1 hasta kN, donde kN es el mayor múltiplo de N que tiene la misma cantidad de dígitos que él, los que están por encima de kN se desechan. En el ejemplo con N = 100, kN es 900, se tomarían números aleatorios de 3 dígitos entre 001 y 900, transformando los que están por encima de 100.

Otra vía de transformación para el número que se encuentre en la tabla es restándole a dicho número el valor de N, pero siempre tomando como rango de transformación el anteriormente indicado. La transformación se realiza con el objetivo de no avanzar demasiado en la tabla.

- Seleccionar de modo aleatorio, el arranque aleatorio en el bloque, es decir, la fila y la columna de la tabla de números aleatorios a partir de donde se comenzarán a tomar los números aleatorios.

- A partir del arranque aleatorio se comenzarán a tomar números aleatorios acorde con uno de los criterios anteriores. Si en la tabla, al llegar al final de la fila, no se ha completado la cantidad necesaria de números aleatorios, se continúa en la fila siguiente, y así sucesivamente, e incluso se puede seguir en el próximo bloque. De igual forma, si al llegar al final de la fila quedase algún número que no tenga la cantidad de dígitos que se requiere, se completa este con el (o con los dígitos) de la fila siguiente. Por otro

lado, si el número encontrado en la tabla es el 0 (00, 000,..., según el caso) por él se anotará el valor que tenga N.

En el caso en que en la tabla se encuentre un número que ya haya aparecido antes, si el muestreo es sin repetición, no se toma, de lo contrario, se tomará tantas veces como aparezca.

■ Despues de completar los n números aleatorios requeridos se busca, en la numeración del listado del paso I, cada uno de los elementos de la población a los que les corresponden estos números: esos elementos son los integrantes de la muestra aleatoria que será investigada en los que se podrán observar una o varias variables.

2.3 Tamaño de muestra.

Independientemente de lo planteado hasta aquí relacionado con el tamaño de la muestra, existen expresiones para calcularlo que desarrollaremos a continuación sin entrar en detalles, ni demostraciones.

La expresión para determinar el tamaño de la muestra depende de la precisión que se quiera. También hay que tener en cuenta si la población es finita o infinita. **Veamos cuestiones necesarias para determinar el tamaño de una muestra.**

Como ya se ha planteado de una población se pueden obtener una determinada cantidad de muestras posibles, (en dependencia del tipo de muestreo: con o sin reposición), en cada muestra se pueden obtener los estimadores media muestral, varianza muestral, desviación típica muestral, etc. Se tiene entonces un grupo de medias muestrales, (varianzas muestrales, etc), que han sido obtenidas a través de un muestreo aleatorio y por tanto esas medias muestrales pueden ser consideradas variables aleatorias y para toda variable aleatoria es posible conocer su distribución probabilística y sus parámetros (Badii et al 2014).

En el caso de la media muestral se ha demostrado que su distribución probabilística es la distribución Normal y que se encuentran bajo el área de la curva Normal, dentro de ± 2 desviaciones estándar con respecto a la media, el 95% de los casos, y, dentro de ± 3 desviaciones estándar con respecto a la media, el 99,7% de los casos.

Por otra parte, se entiende por nivel de confianza la probabilidad de que un

45

parámetro se encuentre entre dos límites y se denota $1 - \alpha$. Los niveles de confianza más utilizados son 90%, 95%, 98%, 99%. Para obtener esos límites de confianza se emplean expresiones que varían en dependencia del parámetro que se analiza y en esas expresiones están incluidos percentiles de probabilidades de las distribuciones normales, T Student y Chi-Cuadrada. Esos valores se buscan en tablas estadísticas.

También se ha planteado que existe diferencia entre el estimador y el parámetro y que a esto se le nombra error. Este error se puede dar en términos absolutos o en términos relativos, cuando se da en términos relativos el máximo valor admitido es 0,10. Este error máximo permisible se denota por d.

Es importante también el conocimiento que se tenga del fenómeno característica que se analiza. Este conocimiento permitirá plantear la probabilidad de éxito (p) asociada a esa característica, se denota por (q) la probabilidad de fracaso, teniendo presente que $p + q = 1$. Cuando no se conoce p se asume que su valor es 0,5. A partir de estas consideraciones se dan las siguientes expresiones para calcular tamaños de muestras.

Para poblaciones infinitas:

$$n = (9 * p * q) / d^2 \quad \text{Con una confiabilidad del 99\%}$$

$$n = (4 * p * q) / d^2 \quad \text{Con una confiabilidad del 95\%}$$

Ejemplo:

Se conoce que el 80% de los estudiantes expresan su satisfacción con los conocimientos elementales que tienen sobre las Nuevas Tecnologías de la Informática y las Comunicaciones. Se desarrolla una investigación y se necesita determinar qué cantidad de estudiantes hay que examinar para verificar esos conocimientos, si se está dispuesto a cometer un error de 0,05, con una confiabilidad del 95%.

Solución:

X: número de estudiantes con conocimientos de las NTIC.

$$p = 0,80 \quad q = 0,20 \quad (p + q = 1)$$

$$d = 0,05$$

$$1 - \alpha = 0,95$$

46 La expresión a emplear es $n = (4 * p * q) / d^2$

Sustituyendo:

$$n = (4 * 0,80 * 0,20) / (0,05)^2$$

$$n = 0,64 / 0,0025$$

n = 256 estudiantes.

Hay que examinar 256 estudiantes.

Expresiones para poblaciones finitas:

$$n = (9 * p * q * N) / d2 * (N - 1) + 9 * p * q$$

Para una confiabilidad del 99%

$$n = (4 * p * q * N) / d2 * (N - 1) + 4 * p * q$$

Para una confiabilidad del 95%

Ejemplo:

Supongamos que para el caso anterior se conoce que la población de estudiantes universitarios asciende a 10 000.

Entonces:

$$n = (4 * 0,80 * 0,20 * 10\,000) / (0,05)^2 (10\,000 - 1) + 4 * 0,8 * 0,2$$

$$n = 6400 / 24,9975 + 0,64$$

$$n = 6400 / 25,6375$$

$$n = 249,63 \approx 250 \text{ estudiantes}$$

Hay que evaluar 250 estudiantes.

2.4 El Muestreo Aleatorio Simple (M.A.S.)

Consideremos una población finita y homogénea en cuanto a la característica que se estudia de la que deseamos extraer una muestra. Cuando el proceso de extracción es tal que garantiza a cada uno de los elementos de la población la misma oportunidad de ser incluidos en dicha muestra, denominamos al proceso de selección muestreo aleatorio. Este tipo de muestreo es el que permite obtener muestras independientes. Para la selección de las muestras se emplea la tabla de números aleatorios (Badii et al 2014).

Existen expresiones para calcular el tamaño de la muestra teniendo en cuenta el parámetro que se va estimar.

2.4.1 El muestreo sistemático:

Este muestreo se utiliza cuando el volumen de la población que se estudia es finito y no muy grande, y además, se conoce que es homogénea en cuanto a la "variable que se investiga", tal y como ocurre en el M.A.S.

Exige, como el M.A.S. numerar todos los elementos de la población, pero en lugar de extraer n números aleatorios sólo se extrae uno. Se parte de ese número aleatorio i, que es un número elegido al azar (lo que se puede hacer empleando una tabla de números aleatorios), y los elementos que integran la muestra son los que ocupan los lugares i, i+k, i+2k, i+3k,...,i+(n-1)k, es decir se toman los individuos de k en k, siendo k el resultado de dividir el tamaño de la población entre el tamaño de la muestra: k = N/n. El número i que empleamos como punto de partida será un número al azar entre 1 y k. Este proceso se seguirá hasta completar el volumen de la muestra (García 1997).

Esta forma de seleccionar la muestra es más fácil que mediante la aplicación del M.A.S.; sin embargo, el tamaño de la muestra depende en gran medida del valor que se tome para k, por tanto, no es posible precisar antes de realizar el muestreo qué extensión tendrá la muestra.

El riesgo de este tipo de muestreo está en los casos en que se dan periodicidades en la población ya que al elegir a los miembros de la muestra con una periodicidad constante (k) podemos introducir una homogeneidad que no se da en la población. Imaginemos que estamos seleccionando una muestra sobre listas de 10 individuos en los que los 5 primeros son varones y los 5 últimos mujeres, si empleamos un muestreo aleatorio sistemático con k = 10 siempre seleccionaríamos o sólo hombres o sólo mujeres, no podría haber una representación de los dos sexos.

En este muestreo se tendrá en cuenta "no acomodar" el listado original de la población, es decir, se debe aceptar este tal y como resulta de su confección natural y espontánea.

Ejemplo:

Supongamos que la población tiene tamaño N igual a 1000 y se desea una muestra de tamaño n igual a 5. La fracción de muestro será 0,005 y el factor de elevación de 200 unidades en la población por cada elemento en la muestra. El muestreo sistemático consiste en:

1. Seleccionar un elemento al azar entre el primero y el que ocupa un lugar en la lista igual al factor de elevación. En el ejemplo seleccionaremos un

elemento al azar dentro de los 200 primeros en la lista. Para ello tomaremos un número aleatorio de tres cifras: si este número es menor de 200 seleccionamos el elemento que tenga ese orden; si es mayor de 200 lo desechamos y tomamos otro.

2. Completamos la muestra sumando el factor de elevación al primer valor obtenido y continuando de esta manera hasta completar el tamaño muestral.

Si existe algún tipo de ciclo en la lista podemos tener un sesgo de selección.

2.4.2 El muestreo estratificado

Con anterioridad hemos dicho que para aplicar el M.A.S. la población no puede ser muy grande, y además, tiene que ser homogénea: si no se cumpliera este último requisito, pero es factible dividirla en sub poblaciones o estratos que lo sean, entonces se optará por usar el muestreo aleatorio estratificado. Estos estratos deberán ser mutuamente excluyentes y exhaustivos, se debe tener en cuenta que todos los elementos de la población estén incluidos en uno, y solo en uno, de estos estratos, cuyos tamaños pueden ser diferentes (Badii et al 2014).

Se puede estratificar, por ejemplo, según la profesión, la especialidad que se estudia, el año de la carrera, el sexo, el estado civil, etc.

Cada estrato funciona independientemente, pudiendo aplicarse dentro de ellos el muestreo aleatorio simple o el estratificado para elegir los elementos concretos que formarán parte de la muestra. Empleando alguna de las diferentes técnicas se determina el tamaño de la muestra la que se distribuye por cada estrato. La distribución de la muestra en función de los diferentes estratos se denomina afijación, y puede ser de diferentes tipos:

Afijación Simple: A cada estrato le corresponde igual número de elementos maestrales.

Afijación Proporcional: La distribución se hace de acuerdo con el peso (tamaño) de la población en cada estrato. El tamaño de la muestra se distribuye proporcionalmente empleando la siguiente expresión:

$$ne = n \cdot (Ne/N)$$

Donde: ne: Tamaño de la muestra en el estrato.

n: Tamaño de la muestra

Ne: Tamaño del estrato

N: Tamaño de la Población

Afijación Óptima: Se tiene en cuenta la previsible dispersión de los resultados, de modo que se considera la proporción y la desviación típica. Tiene poca aplicación ya que no se suele conocer la desviación.

Ejemplo:

Supongamos que estamos interesados en estudiar el grado de aprendizaje de las Nuevas Tecnologías de la Informática y las Comunicaciones. A tal efecto seleccionamos una muestra de 250 estudiantes. (tamaño de muestra calculado anteriormente). Conocemos por los datos del Ministerio de Educación que de los 10 000 estudiantes de una ciudad, 6 000 están matriculados en la Enseñanza Primaria, 3 000 en la Enseñanza Media y 1 000 en la Media Superior. Como estamos interesados en que en nuestra muestra estén representados todos los tipos de enseñanzas, realizamos un muestreo estratificado empleando como variable de estratificación el tipo de enseñanza.

Si empleamos una afijación simple elegiríamos 200 niños de cada tipo de centro, pero en este caso parece más razonable utilizar una afijación proporcional pues hay bastante diferencia en el tamaño de los estratos. Por consiguiente, calculamos que proporción supone cada uno de los estratos respecto de la población para poder reflejarlo en la muestra.

$$\text{Enseñanza primaria: } nep = 250 * (6000/10\,000) = 150 \text{ estudiantes}$$

$$\text{Enseñanza media: } nm = 250 * (3000/10\,000) = 75 \text{ estudiantes}$$

$$\text{Enseñanza Media Superior: } nms = 250 * (1000/10\,000) = 25 \text{ estudiantes}$$

2.4.3 El muestreo por conglomerado.

Ya conocemos que para aplicar el M.A.S. la población no puede ser muy grande, y además, tiene que ser homogénea; por otra parte, si no se cumple este último requisito, pero es factible dividirla en sub poblaciones que lo sean, se utiliza el muestreo estratificado. En cambio, cuando tengamos una población que sea grande y homogénea, para "muestreala" se debe utilizar el muestreo aleatorio por conglomerados. El muestreo por conglomerados consiste en seleccionar aleatoriamente un cierto número de conglomerados (el necesario para alcanzar el tamaño muestral establecido) y en investigar después todos los elementos pertenecientes a los conglomerados elegidos. Cuando los conglomerados son áreas geográficas suele hablarse de "muestreo por áreas".

Los conglomerados deberán ser mutuamente excluyentes y exhaustivos: se debe tener en cuenta que todos los elementos de la población estén incluidos en uno, y solo en uno, de estos conglomerados, cuyos tamaños pueden ser diferentes.

Una observación queremos hacer finalmente sobre la importancia que tiene seleccionar la muestra de un modo correcto: en la literatura se recogen múltiples ejemplos de investigaciones invalidadas a causa de una incorrecta elección de la muestra; así como, también se dan fe de "pronósticos" no cumplidos porque fueron realizados sobre la base de la aplicación de un muestreo inadecuado (Cochran 1971).

2.5 Estimación.

Se llamará estimador, a cualquier función de "n" variables, donde después de sustituir en ella los valores muestrales, el resultado obtenido puede servir como sustituto del valor del parámetro poblacional. Se expresa por (sita circunflejo, este símbolo ^ circunflejo, denota estimación).

Como de una población de tamaño N, se pueden sacar muchas muestras, tantas como: $MN = n$ para muestras sin reposición y $Mn = N$ para muestras con reposición

Debe quedar claro que los estadísticos o medidas que se determinan en cada muestra, son variables aleatorias, que varían de una muestra a otra, aún de la misma población.

Ejemplo de estimadores: \bar{X} , S^2 , $\hat{\mu}$

Se denominará estimación al valor numérico concreto que resulta de un estimador, cuando se haga la sustitución de los datos muestrales, en el estimador.

Se llamará error de muestreo, a la diferencia entre el valor del estimador y del parámetro. (Es evidente que si se estima el parámetro poblacional, a partir de un estimador muestral, hay implícito un error, que es el error de muestreo).

$$e_m = \hat{\theta} - \theta \quad o \quad e_m = \bar{x} - \mu \quad \text{donde } e_m = \text{error de muestreo.}$$

Así:

e_m : Constituye una variable aleatoria, variará, de estimación a estimación. Pero además es un valor que no se puede conocer, pues habría que conocer el parámetro poblacional, y si se conociera éste, no habría necesidad de estimarlo (Barbancho 1982).

En la práctica existen dos tipos de estimación puntual y por intervalos, la primera es cuando se estima el parámetro, a través de un valor; y por intervalo a través de dos valores o un intervalo.

2.5.1 Estimación puntual

El objetivo que se persigue con esta estimación es obtener valores específicos del parámetro desconocido, el cual puede ser utilizado en su lugar.

Se trata pues de que para estimar los parámetros de la población:

- 1.- Elegir un buen estimador
- 2.- Calcular una estimación puntual que sustituya al parámetro desconocido.

Ahora ¿Cómo obtener un estimador si cualquier estadígrafo puede serlo? ¿Entre dos estimadores cual es el preferible? ¿Cuál debe ser el criterio de selección de estimadores?

Las ventajas y desventajas de los estimadores hay que juzgarlas, partiendo de las propiedades deseables para un estimador, que como es natural debe ser, que los valores posibles del estimador estén todo lo más cerca que se pueda del parámetro desconocido. Se debe destacar la necesidad de una buena evaluación pues se va a desarrollar u obtener con una muestra una estimación del parámetro, lo que evidentemente conlleva a un posible error, ya que la muestra no contiene exactamente la misma información que la población, siendo solamente un reflejo de ella y en ocasiones un reflejo bastante pálido (Beltrán and Peris 2013).

Para hablar de un buen estimador se definirá que las cualidades que este debe tener son:

- a.- Ser insesgado.
- b.- Ser consistente.
- c.- Ser eficiente.

Propiedades

- 1.- En el MAS la \bar{x} es un estimador consistente de μ :
- 2.- En el MAS la s^2 es un estimador consistente de σ^2 .
- 3.- Un estimador insesgado puede o no ser consistente.
- 4.- Todo estimador eficiente es consistente.

Se ha visto una de las formas de obtener resultados muestrales para generalizarlo a la población, que en estadística se conoce como inferencia estadística.

Hipotéticamente al usar el estadístico muestral para estimar el parámetro poblacional se debe examinar todas las muestras posibles que se pudieran obtener. Si en realidad se tuviera que hacer esta selección de todas las muestras posibles, a la distribución de los resultados se le conocería como una distribución muestral (Escudero 1994).

2.5.2 Distribuciones muestrales.

Ya se había dicho que si de una población cualquiera se tomaban todas las muestras posibles a través del MAS, de tamaño n , y si a todas ellas se les calculaba, la media muestral, se obtendrían valores diferentes de la media en cada muestra, y por tanto constituirían variables aleatorias, lo mismo pasaría con la varianza; por tanto se puede llegar a una conclusión muy importante:

Todo estimador es una variable aleatoria, y al ser variable aleatoria, tiene asociada: Característica numéricas o parámetros y distribución de probabilidad, por lo que a las distribuciones de probabilidad de estos estimadores se les denomina: distribución muestral (Biosca 1999).

Por tanto la distribución muestral del estimador se conforma a partir de las " n " muestras posibles tomadas de la población y en las cuales se determinó que por constituir variable aleatoria se le puede determinar su función, su esperanza y su varianza.

$$\text{Así } E(x_i) = \mu \quad V(x_i) = \sigma^2 / n$$

Estas características informan:

1.- El centro de la distribución poblacional y de la distribución muestral de media, coinciden

$$\mu(x) = \mu(x_i)$$

2.- Qué la varianza del estimador es n veces menor que la varianza de la población:

$$V(x_i) = \sigma^2 \quad V(x) = \sigma^2 / n$$

Lo que permite concluir que a medida que " n " aumenta los valores de la media muestral se concentran más alrededor de μ .

3.- Se sabe que la $V(x_i) = \sigma^2 / n$, y esto se podría escribir también como:

$V(x_i) = 1/n \sum (-\mu)^2$ y esta última expresión $\sum (-\mu)^2$, se conoce como error de estimación, por lo tanto: la desviación típica de la media va a indicar una medida del error promedio de estimación.

2.5.3 Distribución muestral de \bar{X} para σ^2 conocida

Hay un teorema que plantea:

Qué si "x" tiene una distribución normal, con media μ y varianza σ^2 y se selecciona una muestra aleatoria tamaño " n " por el procedimiento del MAS; entonces la media muestral tendrá una distribución normal con media μ y varianza σ^2/n .

Por tanto si $X \rightarrow N(\mu, \sigma)$ entonces $\bar{X} \rightarrow N(\mu, \frac{\sigma}{\sqrt{n}})$

y para calcular la probabilidad de cierto comportamiento de la media, se utilizará la variable estandarizada: $Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

¿Pero y si X no tiene una distribución normal?

Esto lo resuelve el Teorema Central del Límite en el que descansa, la gran importancia y el poder de aplicación de la distribución normal que plantea que:

Si X es una variable aleatoria con media μ y varianza σ^2 y \bar{X} es la media de una muestra aleatoria simple de tamaño " n ", entonces la variable

$(\bar{X} - \mu) / \sigma / \sqrt{n}$ tiene una distribución que se aproxima a la normal estandarizada a medida que $n \rightarrow \infty$

Esto es si $X \rightarrow N(\mu, \sigma)$ y $n \rightarrow \infty$ entonces $\bar{X} \approx N(\mu, \frac{\sigma}{\sqrt{n}})$

En la práctica se ha demostrado que siempre que $n \geq 30$ la aproximación a la normal es buena, por lo que se utilizará este criterio para considerar que $n \rightarrow \infty$

2.5.4 Distribución muestral de \bar{X} para σ^2 desconocida.

Recordar que cuando sea necesario estimar σ^2 se hace a través de s^2 (dividido por $n-1$ y no por n) que es un estimador insesgado, consistente y más eficiente.

Hay un teorema que plantea que si:

Si se tiene una población $N(\mu, \sigma)$ de la cual se ha extraído una muestra aleatoria de tamaño "n" y donde: $\frac{(\bar{X} - \mu)}{\sigma / \sqrt{n}} \rightarrow N(0,1)$

$\frac{(\bar{X} - \mu)}{\sigma / \sqrt{n}} \rightarrow N(0,1)$

$\chi^2(n-1) \rightarrow \chi^2(n-1)$ grados de libertad, donde la media y la varianza muestral son independientes se puede afirmar que:

$$\frac{(\bar{X} - \mu)}{S / \sqrt{n}} \rightarrow T_{(n-1)}$$

Así, si se quiere hallar probabilidad de cierto comportamiento de la media, cuando se desconozca la varianza de la población, se hace, si se cumple que la variable original $X \rightarrow N(\mu, ?)$ y $n < 30$ a través de t' student (formula anterior)

Ahora si $n > 30$ o cuando $n \rightarrow \infty$ la distribución t'student tiende a la normal estandarizada, esto es a $Z \rightarrow N(0, 1)$ y por tanto t se aproxima a través de Z.

Antes de hacer algún ejercicio se debe plantear que significan los grados de libertad, muy sencillamente.

La varianza de la muestra requiere del cálculo de:

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

Por lo tanto para calcular S^2 se necesita conocer primero \bar{X} . Por consiguiente se puede decir que solo $n - 1$ de los valores de la muestra está libre para variar. Es decir hay $n-1$ grados de libertad (Brunet 2000).

Se puede demostrar este concepto de la forma siguiente. Suponga que se tiene una muestra de cinco elementos con un media igual a 20. ¿Cuántos valores diferentes se necesitarían conocer antes de poder obtener el resto?

El hecho de que $n = 5$ y de que $= 20$ también indica que por lo tanto una vez que se conocen 4 valores el quinto no tendrá "libertad de variar" puesto que la suma tiene que ser 100. Digamos que 4 de los valores son: 18, 24, 19, y 16, el quinto solo puede ser 23 para que todos sumen 100.

2.5.5 Distribución muestral de la varianza (s^2)

Al estudiar S^2 se llega a la conclusión que S^2 no sigue una distribución normal, tiene una distribución asimétrica.

Hay un teorema que plantea qué:

Sea una población normal con media μ y desviación típica σ , entonces la expresión $(n-1)s^2/\sigma^2$ sigue una distribución χ^2 , con $n-1$ grados de libertad.

Recuerden que los grados de libertad de la distribución se representan por la

letra griega nu (ν) (ya se explicó anteriormente lo que expresaban).

Ejemplo

Calcule la probabilidad de que la varianza de una muestra de tamaño 21 obtenida de una población normal con media 5 y desviación típica 2

a.- Sea superior a 8

b.- Entre que dos valores se moverá S^2 con una probabilidad central de 0.95.

Datos: $n=21$, $\mu = 5$, $\sigma = 2$

$$\begin{aligned} a.- P(S^2 > 8) &= 1 - P(S^2 < 8) = 1 - P[(n-1)s^2/\sigma^2 < 20(8)/4] = 1 - P(\chi^2_{(20)} < 160/4) \\ &= 1 - P(\chi^2_{(20)} < 40) = 1 - F_{\chi^2_{(20)}}(40) \\ &= 1 - 0.995 \\ &= 0.005 \end{aligned}$$

$$b.- P(S^2_a < S^2 < S^2_b) = 0.95$$

$0.025 \setminus 0.95 / 0.025$

$$\chi^2_a \quad \chi^2_b$$

Ahora bien estos se buscan: $\chi^2_a = X_{(0.025)}^2$ y $\chi^2_b = X_{(0.975)}^2$ que serían los valores que le corresponden a S^2_a y S^2_b , a partir de $\chi^2_{(n-1)} \sigma^2/(n-1) = S^2$

$$S^2_a = X_{(0.025)}^2 (4/20) = 9.59 (4/20) = 1.918$$

$$S^2_b = X_{(0.975)}^2 (4/20) = 34.2 (4/20) = 6.84$$

Por tanto se considera que los valores de S^2_a y S^2_b con una probabilidad central del 95% serán: $P(1.918 < S^2 < 6.84) = 0.95$

2.5.6 Error Máximo permisible

Como sabemos el error de muestreo (em) que está dado por la diferencia entre el estimador y el parámetro. Este error no es factible de determinar entre otras causas por no conocer el valor del parámetro, pero si se podría calcular una medida probabilística del error y que una vez obtenida una estimación puntual de un parámetro (Colera 2003). Es necesario determinar una medida probabilística de que el error no sea mayor que un determinado valor, que pudiera denotarse por "d" y que posteriormente se

definirá.

En el caso de μ aplicando propiedad de módulo:

$$P\left(\left|\bar{X} - \mu\right| \leq d\right) = P\left(-d \leq (\bar{X} - \mu) \leq d\right)$$

Representando los extremos del intervalo entre los cuales se mueve este error, con una probabilidad dada y que se representa por $1 - \alpha$

Ahora bien otra forma para obtener una medida probabilística del error, es la determinación del error máximo admisible, que se denota por "d" y que se define como:

Según teoremas:

$$\text{Si } X \rightarrow N(\mu, \sigma) \text{ entonces } \rightarrow N(\mu, \sigma / \sqrt{n}) \text{ y } d = Z_{(1-\alpha/2)} \sigma / \sqrt{n}$$

$$\text{Si } X \rightarrow N(\mu, \sigma^2) \text{ y } n > 30 \text{ entonces } \approx N(\mu, S / \sqrt{n}) \text{ y } d = Z_{(1-\alpha/2)} S / \sqrt{n}$$

$$\text{Si } X \rightarrow N(\mu, \sigma^2) \text{ y } n < 30 \text{ entonces } \rightarrow t\text{-student} \text{ y } d = t_{(1-\alpha/2)} S / \sqrt{n}$$

$$\text{Si } n \rightarrow \infty \Rightarrow n > 30 \text{ entonces } \approx N(P; \sqrt{pq / n}) \text{ y } d = Z_{(1-\alpha/2)} \sqrt{pq / n}$$

Y a partir de estos teoremas hay un corolario que plantea determinar el tamaño de la muestra "n", a partir del error máximo admisible, a través de un simple despeje.

$$1. - n = [Z_{(1-\alpha/2)} \sigma / d]^2 \quad 2. - n = [Z_{(1-\alpha/2)} S / d]^2$$

$$2. - n = [t_{(1-\alpha/2)} S / d]^2 \quad 4. - n = [Z_{(1-\alpha/2)}^2 (1-\alpha/2)pq] / d^2$$

Ejercicio.

La experiencia adquirida indica que las varillas de alambre producidas por cierta fábrica tienen una resistencia media a la ruptura de 400Kg y una desviación típica de 16 Kg. Se conoce que la resistencia de dichas varillas sigue una distribución normal, si se extrae una muestra de tamaño 16.

a.- Calcule la probabilidad de que el error en la estimación de μ no sea mayor de 8 Kg.

b.- Determine con una probabilidad de 0.99, ¿Cuál es el error máximo que se espera cometer al estimar μ , a través de la media muestral?

c.- Diga cuantas varillas deberán seleccionarse para que la media resultante tenga un error no mayor de 2 Kg. con una probabilidad de 0.95.

Solución.

a.- $X \rightarrow N(400, 16)$ entonces $\rightarrow N(400, 16 / \sqrt{16})$ por tanto:

$$\begin{aligned} P(|\bar{X} - \mu| \leq 8) &= P(-8/4 \leq Z \leq 8/4) = P(-2 \leq Z \leq 2) \\ &= F_Z(2) - F_Z(-2) \\ &= 0.9772 - 0.0228 \end{aligned}$$

= 0.9544 En el 95% de las muestras tamaños 16 el error que se puede cometer al estimar μ no va a ser mayor que 8.

b.- $d = Z_{(1-\alpha/2)} \sigma / \sqrt{n}$ entonces el valor de "d" será

$$d = 2.58 (4) = 10.32$$

Este valor de Z se encuentra en la tabla que está en la pag.17, que tiene sombreada las dos colas, a partir del valor que tenga α es decir $1 - \alpha = 0.99$ (nivel de confianza) por tanto $\alpha = 0.01$ buscando este valor en la tabla se obtendrá directamente el valor de Z, en la misma.

$$c. - n = [(Z_{(1-\alpha/2)} \sigma) / d]^2 = [1.96(16) / 2]^2 = 246 \text{ varillas.}$$

(Este valor de Z se obtiene buscando $\alpha=0.05$). Debe significarse que con una muestra de este tamaño se garantiza que el error en la estimación de μ , no sea mayor de 2 Kg con una probabilidad de certeza del 95

Se considera necesario puntualizar lo siguiente:

Se había planteado que siempre que se realiza una estimación puntual es necesario determinar una medida probabilística del error de muestreo:

$$P(|\bar{X} - \mu| \leq d) = P(-d \leq \bar{X} - \mu \leq d) = 1 - \alpha, \text{ es decir con una probabilidad } 1 - \alpha, \text{ el error de muestreo no será mayor que "d"}$$

¿Por qué? En la práctica como primer paso el investigador, al estimar μ , deberá prefijar el error máximo que está dispuesto a cometer con una probabilidad dada, es decir, al prefijarse "d" y "1 - α ", la investigación cumplirá con el requisito siguiente:

$$P(|\bar{X} - \mu| \leq d) = 1 - \alpha$$

El paso siguiente deberá ser, determinar el tamaño de muestra que satisfaga

la condición anterior.

De obviarse este paso se recomienda juzgar la precisión de la estimación obtenida, calculando el error máximo (d) con los datos muestrales y luego compararlo con el prefijado, o también se puede hacer a través de la probabilidad (Gracia 1997).

Si la " d " calculada es menor o igual que la prefijada la estimación cumple con los requisitos establecidos, por el investigador, de ahí que la estimación obtenida posea la precisión requerida. Por el contrario si la " d " calculada supera a la prefijada, tendrá que incrementarse el tamaño de muestra para aumentar la precisión hasta garantizar el requisito planteado.

Otra forma que pudiera hacerse, es utilizando la probabilidad, es decir se calcula la probabilidad teniendo en cuenta el error máximo que se está dispuesto a cometer y si la probabilidad resultante es menor que la prefijada, entonces la estimación de μ , no cumple con la precisión prefijada, si ésta probabilidad calculada es mayor o igual, entonces sí se puede decir que la estimación de μ cumple con la precisión prefijada (Ibáñez 2002).

2.6 Estimación por intervalos

Los intervalos de confianza se obtienen, partiendo de la distribución asociada al estimador del parámetro correspondiente. La estimación puntual, no permite medir cuán cercano está el valor determinado del parámetro, es decir no permite calcular la precisión de la estimación, ya que no se tiene ninguna indicación del posible error en la estimación puntual.

Sin embargo la estimación por intervalo o intervalo de confianza, en el que se da un intervalo cuyos extremos son variables aleatorias, y que de entre ellas se halla el parámetro a estimar con determinada probabilidad, nos permite medir el error que se comete al hacer la estimación (Martín 2004).

La probabilidad de que el intervalo contenga al parámetro a estimar es igual a $1 - \alpha$ y a esta probabilidad, se le llama nivel de confianza de la estimación por intervalo. Los valores de $1 - \alpha$, deben ser cercanos a 1 y sus valores más usuales son 0.95, 0.90, 0.99, en este orden, o lo que es lo mismo los valores más usuales de alfa son 0.05, 0.10, 0.01, no obstante se pueden usar otros niveles de confianza.

En general los intervalos de confianza de la media y la proporción se forman:
estimador \pm error máximo (d)

Debido a que por ser intervalos simétricos, el punto medio del intervalo coincide con el valor del estimador puntual

2.6.1 Intervalo de μ con σ^2 conocida.

Se sabe que si $X \rightarrow N(\mu, \sigma)$ entonces $\rightarrow N(\mu, \sigma/\sqrt{n})$ por lo tanto $d = Z(1-\alpha/2) \sigma / \sqrt{n}$ luego el intervalo será:
 $\pm d$ o lo que es lo mismo:

Y se plantea que con una probabilidad $(1 - \alpha)$ se encuentra en dicho intervalo el parámetro. Se debe aclarar que dado que para la normal estándar $Z(\alpha/2) = -Z(1-\alpha/2)$ se puede escribir indistintamente. Esta expresión (la del intervalo) representa un intervalo de extremos variables, ya que estos cambian en dependencia del valor que tome la media muestral.

En ellos se puede afirmar que $(1 - \alpha) 100\%$ de estos intervalos contendrá a μ , mientras que $\alpha 100\%$ restante serán intervalos que no contengan al verdadero valor de μ .

Concluyendo:

1. Al intervalo

Una vez sustituidos los valores en el intervalo de confianza de μ , será incorrecto decir con una probabilidad de $1 - \alpha$, se encuentra en dicho intervalo el parámetro.

2. A $Z(1-\alpha/2)$ se le denomina coeficiente de confianza.

3. A $1 - \alpha$ se le llama nivel de confianza.

4. A los extremos del intervalo se les da el nombre de límites de confianza.

Otro caso cuando $X \rightarrow N(\mu, \sigma)$ y $n > 30$ entonces $N(\mu, \sigma/\sqrt{n})$
y por tanto

Cuando σ^2 es desconocida:

Si $X \sim N(\mu, \sigma^2)$ entonces $(\bar{X} - \mu)/S/\sqrt{n} \rightarrow t(n-1)$ si $n < 30$

Entonces el intervalo será: $\pm t(1-\alpha/2)S/\sqrt{n}$

Si $X \rightarrow N(\mu, \sigma^2)$ $n > 30$ entonces $\rightarrow N(\mu, S/\sqrt{n})$

Entonces el intervalo será:

En el caso de la proporción, se sabe que para muestras grandes:

$\rightarrow N(P, \sqrt{pq/n})$ luego el intervalo será:

$$\pm Z(1-\alpha/2) \sqrt{pq/n}$$

Debe señalarse que cuando se va a determinar la muestra a través de "d", el error máximo, "n" es una función del valor deseado de P, y como este se desconoce, es decir es el que se está interesado en estimar, entonces el valor de "n" que se obtiene, es un valor conservador, es por ello que en estos casos se debe considerar $p = 1/2$ para obtener el tamaño de la muestra seleccionada (Martín 2004).

Se puede demostrar que para $0 \leq P \leq 1$, pq es un máximo cuando $p = 1/2$

2.6.2 Intervalo de confianza para la varianza poblacional

En este caso la formulación del intervalo se obtiene a través de una fórmula, es decir no se determina de la misma forma que los intervalos de μ y P , debido precisamente a que este no es un intervalo simétrico y por tanto el punto medio del intervalo de confianza, no coincide con el valor del estimador puntual.

Si $X \rightarrow N$ entonces el estadístico $(n-1)S^2/\sigma^2 \rightarrow \chi^2_{n-1}$ y se puede plantear que el intervalo de confianza será

$$P[(n-1)S^2/\chi^2(1-\alpha/2) \leq \sigma^2 \leq (n-1)S^2/\chi^2(\alpha/2)] = 1-\alpha$$

Y el intervalo de confianza de la Desviación Típica será, la raíz cuadrada positiva del intervalo de confianza de la varianza.

$$\sqrt{\frac{(n-1)S^2}{\chi^2(1-\alpha/2)}} \leq \sigma \leq \sqrt{\frac{(n-1)S^2}{\chi^2(\alpha/2)}}$$

CAPÍTULO III: PRUEBA DE HIPÓTESIS

Los problemas dójima de hipótesis: consisten en decidir entre solamente dos acciones, donde cada una de ellas está asociada a determinado estado de la naturaleza. Es decir, los posibles estados de la naturaleza se dividen en dos grupos que se recogen en dos hipótesis.

Una hipótesis estadística es, como cualquier otra hipótesis, la suposición de una cosa para sacar de ella una consecuencia. En los problemas de dójima de hipótesis solo una de las dos hipótesis es cierta, y nuestro problema consistirá en determinar cuál. Ello se hará a partir de la información que hayamos obtenido de los datos de una muestra (Valles 2000).

El problema a tratar estará relacionado con la resolución de este tipo de situación estadística donde cada parte en el proceso será crucial para la toma de decisiones en determinados puestos que definen la calidad de un producto en este tipo de empresa.

Las dos hipótesis en que se dividen los posibles estados de la naturaleza y de las cuales vamos a escoger una como la que realmente es cierta, reciben el nombre de hipótesis nula y alternativa.

➤ **Hipótesis nula:** Denotada por H_0 , es aquella hipótesis que siempre contiene la igualdad. En el caso de las pruebas paramétrica, puede tomar cualquiera de las siguientes formas: $>$, $<$, $=$.

➤ **Hipótesis alternativa:** Denotada por H_1 , es el complemento de la hipótesis nula, por lo que puede tomar cualquiera de las siguientes formas: $<$, $>$, \neq .

Para la solución de cualquier problema que requiera la utilización del método estadístico "Dójimas de Hipótesis", debe establecerse, en primer término, las hipótesis nulas y alternativa, entre las cuales se va a tomar una decisión.

Los dos conjuntos de valores (los de la hipótesis nula y alternativa) son exhaustivos. Además, un estado natural no puede formar parte a la vez de ambas hipótesis, nula o alternativa, es decir, los subconjuntos formados por los valores posibles del parámetro son excluyentes.

Un problema que preocupa inmediatamente es, en qué hipótesis se sitúa lo que uno quiere probar.

Regla de decisión.

La decisión acerca de cuál hipótesis es cierta y cuál falsa, no se puede hacer indiscriminadamente a través de la aplicación de cualquier procedimiento que nos lleve a aceptar como válida una de las dos posibles. Si pusieramos un ejemplo de decidir cuál de los tratamientos es mejor para curar cierta enfermedad, no podríamos tomar una decisión lanzando una moneda al aire y adjudicando cada uno de los posibles resultados a la aceptación de una de las dos hipótesis, pues en nada está vinculado este experimento con el problema que se estudia.

Una regla de decisión o dójima, es un procedimiento probabilístico que depende de observaciones realizadas (resultados) sobre experimentos estrechamente ligados al problema en estudio y que nos permite decidir si se rechaza o no una hipótesis previamente formulada (Borobia 2004).

El método de prueba de hipótesis debe brindar una regla de decisión con la que se determina cuál de las dos hipótesis debe ser aceptada, basándose en los valores de la muestra. Esta regla de decisión, en general, es de la forma siguiente:

- Si $(X_1, \dots, X_n) \in RC$, rechaza H_0 .
- Si $(X_1, \dots, X_n) \notin RC$, rechaza H_1 .

Donde RC es el conjunto de muestra total que de ser observadas, la regla de decisión sugiere que se rechace H_0 ; a esta región se denomina región crítica.

Siendo la regla de decisión un instrumento para decidir, en base a las observaciones, si se rechaza o no la hipótesis nula, en ella deben quedar perfectamente especificados para cuales valores de las observaciones rechazaremos la hipótesis nula y para cuáles no. De esta forma el espacio maestral o espacio de las posibles observaciones queda dividido en dos regiones: una región donde se rechaza H_0 y otra donde se acepta H_0 . Para definir esta región, denotamos por x el vector de las observaciones, esto es $X = (X_1, \dots, X_n)$.

En todos los problemas de hipótesis, el criterio de decisión se establece contrastando el valor del estadígrafo recomendado para el caso particular que se estudie con un valor que viene dado por el nivel de confianza con que se quiera tomar la decisión y la distribución probabilística del estadígrafo utilizado (Borobia 2004)..

Región crítica:

La región crítica de una dócima $\varphi(x)$ es el conjunto de valores de x que nos lleva a rechazar la hipótesis nula H_0 .

Otras definiciones dirían que la región critica (RC) o región de rechazo, es aquella región que incluye los valores del estadígrafo para los cuales se rechaza la hipótesis nula.

Una región critica ideal sería aquella que nos proporciona siempre la decisión correcta, o sea, que siempre que no se cumpliera H_0 incluyera al estadígrafo y, en cambio, siempre que H_0 se cumpliera no incluyera al estadígrafo; pero sabemos que esto es imposible, que siempre estamos tomando decisiones en presencia de la incertidumbre.

Región de aceptación:

La región de aceptación de una dócima $\varphi(x)$, es el conjunto de valores de x que nos llevan a la aceptación de la hipótesis nula.

Debemos tener presente que la decisión que toma se basa solo en la investigación de una muestra o subconjunto de la población y que esto le imprime a sus conclusiones un carácter probabilístico en el sentido de que nunca él sabe si la decisión aceptada como verdadera, lo es realmente o no. Esto es, el método de trabajo siempre nos deja abierta la posibilidad de aceptar como cierta una hipótesis falsa.

Resumiendo lo visto, en todo problema de prueba de hipótesis, se pueden cometer dos errores.

Error de tipo I:

Es el que se comete cuando aceptamos como cierta H_1 siendo H_0 la hipótesis verdadera. Es decir, el error que cometemos si rechazamos H_0 siendo cierto.

Error de tipo II:

Es el que se comete cuando aceptamos como cierta H_0 siendo H_1 la hipótesis verdadera. Para recordar fácilmente estas dos definiciones veamos la siguiente tabla:

	Se acepta H_0	Se acepta H_1
H_0 cierta	No se comete error	Error de tipo I
H_0 falsa	Error de tipo II	No se comete error

En este momento ya salta a la vista que una estrategia inmediata en la búsqueda de una buena regla de decisión debe ser tratar de minimizar la posibilidad de cometer ambos tipos de errores. Si se rechaza H_0 sólo es posible cometer un error de tipo I. Si se acepta H_0 sólo es posible cometer un error de tipo II.

Si se logra una disminución de la probabilidad de cometer un error de tipo I, en tanto que la probabilidad de cometer un error de tipo II aumenta. Se podría demostrar la imposibilidad de disminuir ambos errores simultáneamente al disminuir un error aumenta el otro (Wackerly et al 2000).

Interesa medir las magnitudes de esos errores y tratar de que esa magnitud sean las menores posibles, o sea, que la probabilidad de cometerlos sea suficientemente pequeña. Resulta imposible reducir ambas probabilidades de cometer errores tanto como se quiera, puesto que una disminución en una de ellas provoca, en general, un aumento de la otra.

La solución encontrada por los matemáticos consiste en fijar el valor de una de ellas, preferiblemente la de cometer error de connotación más graves a un nivel aceptablemente bajo, y tratar de hacer mínima la otra.

Con vista a verificar las notaciones y optimizar el método, se fija el contenido de las hipótesis H_0 y H_1 convenientemente, de modo que el error de tipo I sea el de consecuencia más grave y la probabilidad de cometerlo se fija en un valor suficientemente pequeño denotado por α aceptable para el investigador (Wackerly et al 2000).

En realidad se fija α de modo que:

$$P[\text{rechazar } H_0 \mid H_0 \text{ cierta}] \leq \alpha$$

Lo que se debe interpretar como que α , que recibe el nombre de nivel de significación, es la máxima probabilidad de cometer error de tipo I.

La probabilidad de cometer error tipo II se expresa:

$$P[\text{aceptar } H_0 \mid H_0 \text{ falsa}] \geq \beta$$

Riesgo de una dócima:

Hasta ahora hemos visto cómo, en cada decisión que tomemos está presente la posibilidad de equivocarnos. Pero veamos también que, de conocer la distribución de frecuencia del estadígrafo, es posible calcular la probabilidad de cada tipo de error. El poder conocer el riesgo que corremos de equivocarnos resulta indudablemente un factor muy importante a la hora de tomar una decisión (Cazau.2006)

En general tendremos que, $P(\text{error tipo I}) = P(\text{RC/ } H_0)$.

El nivel de significación de una dócima α es la probabilidad máxima de cometer un error de tipo I y se denota por α . Donde $P_{H_0}(H_1)$ denota la probabilidad de cometer un error de tipo I. Llamaremos riesgo β la probabilidad de error de tipo II, es decir, a la probabilidad de aceptar la hipótesis nula cuando en realidad se cumpla la hipótesis alternativa

α : es también conocido como nivel de significación de la dócima.

$1 - \beta$: se conoce como potencia de la dócima.

Un problema que enfrentará siempre el que necesite aplicar dócimas de hipótesis, será precisamente el de fijar los riesgos con que va a trabajar o el tamaño de muestra a utilizar.

Si las restricciones económicas fijan un tamaño de muestra, tendrá que decidir cómo balancear los riesgos α y β . Además, tendrá que determinar qué valor no detectado de la hipótesis alternativa puede causarle problemas graves, y a partir de una evaluación económica de los perjuicios que le ocasionaría cada uno de los dos posibles errores tomar una decisión.

La prueba de hipótesis que se refieren al valor que puede tomar un parámetro se divide en dos grandes grupos atendiendo a sus hipótesis, caso bilateral o caso unilateral.

El planteamiento de un problema de hipótesis consiste, como hemos visto, en establecer una hipótesis nula y una hipótesis alternativa. De acuerdo con los objetivos que se persigan con la hipótesis que se trate, será necesario formular un planteamiento bilateral o un planteamiento unilateral (Pérez 2004).

➤ **Caso bilateral:** Es el caso en que la hipótesis alternativa comprende tanto los valores mayores, como los menores.

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$

➤ **Caso unilateral:** Es el caso en que la hipótesis alternativa solo comprende los valores menores que , en algunos casos y a los valores mayores, en otros casos. Por lo que se forma:

$$H_0: \mu \leq \mu_0$$

$$H_1: \mu > \mu_0$$

3.1 Prueba de hipótesis para la media de una distribución normal con varianza conocida.

La distribución normal desde el punto de vista práctico tiene gran importancia. Desde el punto de vista teórico, ya sabemos que muchos problemas estadísticos encuentran una fácil solución cuando la distribución de la(s) variable(s) en estudio es una distribución normal y los métodos de prueba de hipótesis no constituyen una excepción a ello (Pick et al 1994).

Sea \bar{X} la media de una muestra simple aleatoria de tamaño n de la población. (Estimador de μ).

μ_0 es un número real.

α es el nivel de significación.

Z_p es el percentil p de la distribución normal estándar.

Resumen.

1.Hipótesis

$$H_0: \mu = \mu_0; H_1: \mu \neq \mu_0$$

$$H_0: \mu \leq \mu_0; H_1: \mu > \mu_0$$

$$H_0: \mu \geq \mu_0; H_1: \mu < \mu_0$$

2. Estadígrafo a emplear

$$U = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

3. Criterio de rechazo de H_0 , expresado en la región crítica.

$$|U| > Z_{\alpha/2}, U > Z_\alpha, U < -Z_\alpha$$

El procedimiento puede esquematizarse en principio, así:

1. Seleccionar un estadígrafo adecuado para tomar una decisión respecto al valor de μ .
2. Tomar un tamaño de muestra "n".
3. Evaluar el estadígrafo seleccionado, a partir de los datos obtenidos en la muestra tomada.
4. Comparar el estadígrafo con μ_0 . Si este difiere poco de μ_0 son iguales, se acepta H_0 ; si difieren mucho de μ_0 , se rechaza H_0 . Dicho de otra forma, se trata de establecer una RC para el estadígrafo de tal forma que si $|\bar{X} - \mu_0| > \alpha$ se rechace H_0 .

3.2 Prueba de hipótesis para la media de una distribución normal con varianza desconocida.

El caso más frecuente en las aplicaciones prácticas es aquel en el que no se conoce la varianza poblacional, la prueba de hipótesis sobre la media de una distribución normal, con varianza desconocida, es similar a cuando la varianza es conocida, puesto que el estadígrafo es muy parecido, con la única diferencia de que al no conocer la varianza, se emplean (Aliaga and Gunderson 1998).

Sea \bar{X} a media de una muestra siempre aleatoria de tamaño n de la poblacional. (Estimador de μ).

μ_0 es un número real.

α es el nivel de significación.

$t_p(k)$ es el percentil p de la distribución.

t - de student con k grados de libertad. ($k=1,2,3,\dots$)

s^2 la varianza muestral

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

$$T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$$

T es un estadígrafo cuya distribución es t- de estudent con $n-1$ grado de libertad bajo la suposición $\mu = \mu_0$ y se obtiene sustituyendo por su estimador s^2 en la formulación μ de estadígrafo Z de la dómica anterior.

Es importante decir que la distribución t se aproxima a una distribución normal a medida que crece el tamaño de muestra y por ejemplo para valores de $n=500,200$ y hasta 100; no resulta significativo la diferencia entre ambas. Es por ello que solo se utiliza la distribución t cuando el tamaño de la muestra no es grande. Por esto en muchos libros se define esta hipótesis como un método para muestras pequeñas (Aliaga and Gunderson 1998).

Nuevamente aquí, como en todo problema de hipótesis se sigue el esquema enumerado en la hipótesis con varianza conocida analizado anteriormente.

Resumen de hipótesis para la media con varianza desconocida.

1. Hipótesis.

$$H_0 : \mu = \mu_0 ; H_1 : \mu \neq \mu_0$$

$$H_0 : \mu \leq \mu_0 ; H_1 : \mu > \mu_0$$

$$H_0 : \mu \geq \mu_0 ; H_1 : \mu < \mu_0$$

2. Estadígrafo a emplear:

$$T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$$

3. Criterio de rechazo de H_0 expresado en la región crítica.

$$|t| > t_{(\alpha/2 : n-1)} ; |t| > t_{(\alpha : n-1)} ; t < -t_{(\alpha : n-1)}$$

Ejemplo

La fábrica "Bucanero" productora de cerveza, se encuentra inmersa en la implantación de un nuevo sistema de gestión de la calidad basado en las normas ISO 9000:2000, para lo cual el equipo consultor especializado, ha tomado información de

algunas partes claves del proceso que influyen directamente con la calidad del producto final y que un buen control de los mismos, evitaría gastos excesivos. Entre las partes claves analizadas, se encuentran: el proceso de enfriamiento de la cerveza, el contenido neto de cada lata y la cantidad de lasca (cobertura para recubrir la lata y evitar el contacto con el aluminio). En dicha investigación se observó lo siguiente:

➤ En el proceso de fabricación de la cerveza, debe someterse a una temperatura de enfriamiento de 120C. Se conoce que esta variable sigue una distribución normal y que cada 3 minutos se extrae una muestra aleatoria para medir el parámetro, el cual, si es menor, es dirigido a un área en la que se le elevará la temperatura hasta el nivel establecido para luego envasarla en latas; si es mayor, se separa y posteriormente es añadido a la primera parte del proceso de enfriamiento para que alcance su temperatura normada, y sea luego dirigido al área de envasado y si es igual pasa directamente al área de envasado; o sea que la temperatura debe ser estrictamente del nivel establecido. Se debe decir además que el equipo empleado para esta labor fue traído recientemente por una inversión realizada y presenta un panel de control en el cual el obrero perteneciente al puesto de trabajo, está aprendiendo a manejarlo solo, por lo que en ocasiones quizás puedan existir errores en las mediciones. Para ver si la cantidad de latas que inicialmente no alcanzan la temperatura normada es significativa, se realizó un muestreo en esta parte del proceso en el que se obtuvo la información de 5 horas de trabajo de un día laborable, obteniéndose los resultados que se muestran a continuación:

12	12	12	12	12	12	12	12	12	12
12	12	13	12	12	12	12	12	15	12
14	11	12	12	13	12	14	12	12	12
12	12	12	12	12	12	11	12	12	10
12	12	12	12	14	12	12	12	14	12
12	12	15	12	12	12	12	12	12	12
12	12	12	11	12	12	12	14	12	13
12	12	14	12	15	12	12	12	12	12
10	12	12	12	12	12	12	13	11	12
12	12	12	13	13	15	12	12	12	15

➤ Para verificar si el contenido promedio envasado cumple lo estipulado (355 ml.), el departamento de calidad también realizó un muestreo en esta sentido, con vista a prestar un servicio de excelencia y así evitar las quejas por insatisfacciones por parte de los clientes (comercializadores). De acuerdo con el proceso de llenado, la cantidad en mililitros sigue una distribución normal. El control realizado tomó una muestra aleatoria cada seis minutos debido a que el proceso lo requiere de esta forma, todo esto en el último pedido que realizó el CIMEX, lo cual demoró un día laborable (8 horas), obteniéndose los siguientes resultados:

354	355	352	353	355	355	355	352
355	355	355	355	355	353	355	355
352	354	355	355	355	355	355	355
355	355	352	354	352	354	355	354
352	354	352	355	355	355	354	355
355	355	353	355	352	355	355	352
353	355	355	352	354	352	355	355
355	355	353	355	355	353	352	354
355	355	355	355	352	352	355	355
352	355	352	352	353	355	352	352

Resolución

Al problema que nos enfrentamos, al aplicar décima de hipótesis en el control de la calidad en cuestión, será precisamente el de fijar los riesgos con que se va a trabajar. Tendremos que determinar qué valor no detectado de la hipótesis alternativa puede causar problemas graves, y a partir de una evaluación económica de los perjuicios que le ocasionarían cada uno de los dos posibles errores al tomar una decisión. Haciendo una valoración general se puede decir que es posible cuantificar los gastos en que incurrirían si las latas llevan más de 4 onzas de recubrimiento en su interior, y seguramente es posible cuantificar las pérdidas o reclamaciones en que se incurrirían si las latas no llevan 355 mililitros de cerveza y llevan menos, además de la pérdida de prestigio por incumplir las normas establecidas. Pero también pudiera ser valorada la calidad en el proceso de enfriamiento al que son sometidas las latas, si como resultado se obtuvieran que la mayoría no tengan 120C como temperatura inicial.

Después de haber valorado económicoamente lo que implicaría cada uno de los resultados que se pudieran obtener, pasemos a identificar los datos y definir las hipótesis, que es, por el planteamiento del problema; a continuación debe establecerse el criterio de decisión, es decir, los valores del estadígrafo a utilizar para lo cual se acepta una u otra hipótesis.

Análisis del proceso de enfriamiento de la cerveza.

Datos

$X \rightarrow$ Temperatura máxima de enfriamiento inicial de cada cerveza.

$\bar{X} \rightarrow$ Temperatura promedio de enfriamiento inicial de las cervezas.

$\mu \rightarrow$ Temperatura media de enfriamiento inicial que debe tener cada cerveza.

$$X \sim N(\mu = 12, \sigma^2)$$

Cada 3 minutos 1 muestra * 5 horas = 100 muestras

n = 100

Formulación del planteamiento

Como se desea saber si la temperatura de enfriamiento inicial de la cerveza es igual a 12 °C o no, ya que por diversos motivos se torna necesario que esto sea así, pues de lo contrario sería perjudicial tanto que la media sea superior como inferior a lo prefijado, se plantea entonces:

$$H_0: \mu = 12 \text{ } ^\circ\text{C}$$

$$H_1: \mu \neq 12 \text{ } ^\circ\text{C}$$

Este planteamiento será el que nos permitirá llegar a una respuesta. Si aceptamos H_0 , afirmamos que la temperatura de enfriamiento inicial de la cerveza es igual a 12 °C; si rechazamos H_0 afirmamos que no es 12 °C. En este caso, se trata de un caso bilateral donde la alternativa incluye valores mayores y/o menores que 12 °C.

Cálculo de la media aritmética (promedio).

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

$$\bar{X} = \frac{12 * 77 + 11 * 4 + 10 * 2 + 14 * 6 + 13 * 6 + 15 * 5}{100}$$

$$\bar{X} = \frac{924 + 44 + 20 + 84 + 78 + 75}{100}$$

$$\bar{X} = \frac{1225}{100}$$

$$\bar{X} = 12.25$$

Cálculo de la varianza

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

$$\sigma^2 = \frac{77(12 - 12.25)^2 + 4(11 - 12.25)^2 + 2(10 - 12.25)^2 + 6(14 - 12.25)^2 + 6(13 - 12.25)^2 + 5(15 - 12.25)^2}{100}$$

$$\sigma^2 = \frac{4.8125 + 6.25 + 10.125 + 18.375 + 3.375 + 37.8125}{100}$$

$$\sigma^2 = \frac{80.75}{100}$$

$$\sigma^2 = 0.8075$$

Cálculo de la desviación típica

$$\sigma = \sqrt{\sigma^2}$$

$$\sigma = \sqrt{0.8075}$$

$$\sigma = 0.8986 \quad \text{Evaluación del estadígrafo}$$

$$U = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{12.25 - 12}{0.8986 / \sqrt{100}}$$

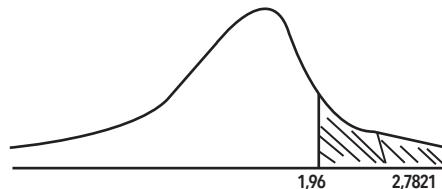
$$U = 2.7821$$

Cálculo de la región crítica

$$\alpha = 0.05$$

$$\alpha/2 = 0.025$$

$$Z_{\alpha/2} = Z_{(0.025)} = 1.96$$



Comparación y decisión

$$\text{Rechazo } H_0 \text{ si } |U| > Z_{\alpha/2}$$

$$2.7821 > 1.96$$

Por lo tanto, como la muestra pertenece a la región crítica, el esquema de decisión de la prueba de hipótesis señala que debe rechazarse H_0 , por lo que se pudo comprobar que la cantidad de cervezas con temperatura inicial distinta de 12 °C es significativa.

Auto evaluación

Realice el análisis del contenido de cada lata de cerveza.

3.3 Dócimas de hipótesis para la varianza de una distribución normal.

Las dócimas para la varianza no presentan diferencias con las dócimas para la media en cuanto a la metodica general a seguir. También aquí será necesario formular las hipótesis, establecer los riesgos, la región crítica, tomar la muestra, evaluar un estadígrafo, comparar el estadígrafo con la región crítica y tomar una decisión.

Las diferencias vienen dadas por las características del estadígrafo a utilizar.

Planteamiento:

Aquí tendremos 3 planteamientos posibles:

$$H_0: \sigma^2 = \sigma_0^2 \quad H_1: \sigma^2 \neq \sigma_0^2$$

$$H_0: \sigma^2 \geq \sigma_0^2 \quad H_1: \sigma^2 < \sigma_0^2$$

$$H_0: \sigma^2 \leq \sigma_0^2 \quad H_1: \sigma^2 > \sigma_0^2$$

es decir, un caso bilateral y dos casos unilaterales, donde pueden igualmente

utilizarse σ ó σ^2 en la formulación de la hipótesis.

Caso bilateral:

El planteamiento bilateral es:

$$H_0: \sigma = \sigma_0$$

$$H_1: \sigma \neq \sigma_0$$

El procedimiento para establecer la región crítica y realizar la dócima se basará en la propiedad que tiene el estadígrafo $(n-1)S^2/\sigma_0^2$ de seguir una distribución χ^2 .

La decisión se tomara a partir de la información brindada por una muestra. Se calculará S^2 en la muestra; si S^2 resulta mucho más grande que σ_0^2 o mucho más pequeño llegaremos a la conclusión de que $\sigma \neq \sigma_0$.

El estadígrafo decisional será denotado por χ^2 y vendrá dado por la expresión:

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

Este estadígrafo seguirá una distribución χ^2 con $n-1$ grados de libertad cuando $\sigma = \sigma_0$ donde σ^2 es la varianza de la variable "x" estudiada, o sea, de $x \sim N(\mu, \sigma^2)$.

La región crítica más adecuada para docimar la varianza de una distribución normal es, cuando se conocen α y "n", y se está en el caso bilateral, la siguiente:

$$\text{RC: } \chi^2 > \chi_{\alpha/2; n-1}^2 \quad \text{y} \quad \chi^2 < \chi_{1-\alpha/2; n-1}^2$$

donde χ^2 es el estadígrafo y su cálculo viene dado por:

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

y $\chi_{\alpha/2; n-1}^2$ y $\chi_{1-\alpha/2; n-1}^2$ son los límites críticos, y sus valores respectivos se obtienen a partir de las expresiones:

$$P(\chi^2 > \chi_{\alpha/2; n-1}^2) = \alpha/2 \quad \text{y} \quad P(\chi^2 < \chi_{1-\alpha/2; n-1}^2) = \alpha/2$$

Casos unilaterales.

Los planteamientos unilaterales son:

$$\begin{array}{ll} 1) H_0: \sigma \geq \sigma_0 & 2) H_0: \sigma \leq \sigma_0 \\ H_1: \sigma < \sigma_0 & H_1: \sigma > \sigma_0 \end{array}$$

El estadígrafo es único para los 2 casos unilaterales y es el mismo utilizado para el caso bilateral, es decir:

$$X^2 = \frac{(n - l) S^2}{\sigma_0^2}$$

Las regiones críticas respectivas son:

- 1) $X^2 < X_{1-\alpha}^2; n-l$
- 2) $X^2 > X_\alpha^2; n-l$

Pruebas no paramétricas

La prueba χ^2 (chi-cuadrado) es una prueba no paramétrica de utilidad cuando las muestras consideradas proceden de poblaciones que no están normalmente distribuidas. Puede utilizarse en numerosos tipos de problemas.

Aquí se exponen sólo tres de los más comunes.

1. En el primero de ellos se aplica la prueba χ^2 para comparar dos muestras y determinar si en realidad existe una diferencia significativa entre ellas.
2. En el segundo, la prueba χ^2 se aplica para comprobar si la distribución de proporciones de una muestra se ajusta a la esperada según una distribución teórica.
3. El tercer y último tipo de problemas que se analiza es el de las tablas de contingencia. Consiste en aplicar la prueba χ^2 para comprobar si existe asociación entre dos variables cualitativas medidas en una muestra (Pett 1997).

La manera clásica de estudiar las diferencias entre frecuencias esperadas y observadas

Acontecimiento	A1	A2	A3	An
Frecuencia esperada	E1	E2	E3	En
Frecuencia observada	O1	O2	O3	On

Las frecuencias esperadas $E_i = n p_i$

Donde n es el tamaño de la muestra y p_i la probabilidad de ocurrencia de cada uno de los eventos o acontecimientos A_i .

El estadígrafo de prueba para cuantificar la discrepancia entre ambas.

$$X^2 = [(O_1 - E_1)^2 / E_1] + [(O_2 - E_2)^2 / E_2] + [(O_3 - E_3)^2 / E_3] + \dots + [(O_n - E_n)^2 / E_n]$$

O sea,

$$X^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

De la fórmula surge que si $X^2 = 0$, la frecuencia esperada coincide exactamente con la observada y entonces la teoría predice perfectamente los acontecimientos. Si existe una diferencia cualquiera entre ambas frecuencias, será $X^2 > 0$, valor que irá aumentando si no es producto del azar, hasta alcanzar valores significativos que permitan rechazar la hipótesis nula de la igualdad entre la teoría y la realidad.

Ejemplo

A usted se le encomienda la responsabilidad de decidir si un método de generación de dígitos aleatorios es bueno o no.

Para ello deben aplicarse procedimientos estadísticos que comprueben:

- Que los dígitos no tengan correlación serial.
- Que los dígitos sean equiprobables.

Particularmente desea desarrollarse el experimento estadístico que compruebe la segunda restricción a través de una dócima de bondad de ajuste.

Los dígitos posibles son por supuesto: 0, 1, 2, 3, ..., 9 y para que la serie generada cumpla lo estipulado antes y pueda considerarse aleatoria, la ocurrencia de cada uno debe tener asociada la misma probabilidad.

Se parte de una muestra de 100 dígitos generales que se pueden resumir en la siguiente tabla:

Dígito	0	1	2	3	4	5	6	7	8	9
Observ	12	9	9	10	10	8	9	11	10	12

Ejecute la dócima con un nivel de significación del 10%.

Solución

$$H_0 : p_i = \frac{1}{10}$$

$$H_1 : \text{alguno} \left(p_i \neq \frac{1}{10} \right)$$

$$E_i = 100 \cdot \frac{1}{10} = 10$$

$$\begin{aligned} \chi^2 &= \frac{(12 - 10)^2}{10} + \frac{(9 - 10)^2}{10} + \frac{(9 - 10)^2}{10} + \frac{(10 - 10)^2}{10} + \frac{(10 - 10)^2}{10} + \frac{(8 - 10)^2}{10} + \frac{(9 - 10)^2}{10} + \\ &+ \frac{(11 - 10)^2}{10} + \frac{(10 - 10)^2}{10} + \frac{(12 - 10)^2}{10} = 1.6 \end{aligned}$$

$$RC = \{ \chi^2 > \chi^2_{(\alpha, k-1)} \} = \{ 1.6 > 14.684 \}$$

Por tanto $\chi^2 \notin RC \Rightarrow \chi^2 \in RA \Rightarrow \text{Acepto } (H_0)$

R: / No hay elementos suficientes para rechazar H_0 , por tanto el método de generación de dígitos aleatorios es bueno.

Prueba para verificar si una variable sigue una determinada distribución con parámetros desconocidos.

$$RC = \{ \chi^2 > \chi^2_{(\alpha, k-e-1)} \}$$

$$\chi^2 = \sum_{i=1}^k \left(\frac{(n_i - E_i)^2}{E_i} \right)$$

Donde e es el número de parámetros que se estiman.

Estas pruebas se pueden verificar fácilmente apartir de las salidas de máquina de cualquier programa estadístico.

Ventajas de los Métodos No Paramétricos

1. Los métodos no paramétricos pueden ser aplicados a una amplia variedad de situaciones porque ellos no tienen los requisitos rígidos de los métodos paramétricos correspondientes. En particular, los métodos no paramétricos no requieren poblaciones normalmente distribuidas.
2. Diferente a los métodos paramétricos, los métodos no paramétricos pueden frecuentemente ser aplicados a datos no numéricos, tal como el género de los que contestan una encuesta.
3. Los métodos no paramétricos usualmente involucran simples computaciones que los correspondientes en los métodos paramétricos y son por lo tanto, más fáciles para entender y aplicar (Moses 1952).

Desventajas de los Métodos No Paramétricos

1. Los métodos no paramétricos tienden a perder información porque datos numéricos exactos son frecuentemente reducidos a una forma cualitativa.
2. Las pruebas no paramétricas no son tan eficientes como las pruebas paramétricas, de manera que con una prueba no paramétrica generalmente se necesita evidencia más fuerte (así como una muestra más grande o mayores diferencias) antes de rechazar una hipótesis nula.
3. Cuando los requisitos de la distribución de una población son satisfechos, las pruebas no paramétricas son generalmente menos eficientes que sus contrapartes paramétricas, pero la reducción de eficiencia puede ser compensada por un aumento en el tamaño de la muestra.

La prueba χ^2 (chi-cuadrado) es una prueba no paramétrica de utilidad cuando las muestras consideradas proceden de poblaciones que no están normalmente distribuidas. Puede utilizarse en numerosos tipos de problemas (Moses 1952).

Aquí se exponen sólo tres de los más comunes.

1. En el primero de ellos se aplica la prueba χ^2 para comparar dos muestras y determinar si en realidad existe una diferencia significativa entre ellas.
2. En el segundo, la prueba χ^2 se aplica para comprobar si la distribución de proporciones de una muestra se ajusta a la esperada según una distribución teórica.

3. El tercer y último tipo de problemas que se analiza es el de las tablas de contingencia. Consiste en aplicar la prueba χ^2 para comprobar si existe asociación entre dos variables cualitativas medidas en una muestra.

La manera clásica de estudiar las diferencias entre frecuencias esperadas y observadas es usar el estadígrafo para cuantificar la discrepancia entre ambas.

Acontecimiento	A1	A2	A3	An
Frecuencia esperada	E1	E2	E3	En
Frecuencia observada	O1	O2	O3	On

$$\chi^2 = [(O_1 - E_1)^2 / E_1] + [(O_2 - E_2)^2 / E_2] + [(O_3 - E_3)^2 / E_3] + \dots + [(O_n - E_n)^2 / E_n]$$

o sea,

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

De la fórmula surge que si $\chi^2 = 0$, la frecuencia esperada coincide exactamente con la observada y entonces la teoría predice perfectamente los acontecimientos. Si existe una diferencia cualquiera entre ambas frecuencias, será $\chi^2 > 0$, valor que irá aumentando si no es producto del azar, hasta alcanzar valores significativos que permitan rechazar la hipótesis nula de la igualdad entre la teoría y la realidad.

La distribución muestral del valor χ^2 se aproxima mucho a una del tipo Chi-cuadrado, modelo que se usa para validar hipótesis. Los grados de libertad se calculan para dos casos posibles:

1. Hipótesis Extrínseca: es una hipótesis externa a los datos. No se necesita de estos para obtener los parámetros poblacionales en el cálculo de las frecuencias observadas. Por ejemplo, el lanzamiento de una moneda o un dado en teoría de juegos, las leyes de Mendel, etc.

$$v = n - 1$$

2. Hipótesis Intrínseca: es una hipótesis interna a los datos. Se necesitan los datos para sacar los parámetros poblacionales. Por ejemplo, si se trabaja bajo el supuesto de normalidad: la media y la varianza de la población se estiman con los datos muestrales, habrá $r = 2$ grados de libertad perdidos.

$$v = n - r - 1$$

En el caso de las extrínsecas es $v = n - 1$ porque con las primeras $n - 1$ muestras, se puede determinar la restante, entonces como una de ellas se obtiene a partir de las demás, se pierde solo un grado de libertad (Badii et al 2014).

Caso 3:

Prueba para verificar si una variable sigue una determinada distribución con parámetros desconocidos.

$$H_0: X \sim \chi^2(\alpha; k - e - 1)$$

$$\chi^2 = \sum_{i=1}^k \left(\frac{(n_i - E_i)^2}{E_i} \right)$$

Donde e es el número de parámetros que se estima.

Ejercicio:

El instituto de meteorología Provincial de Holguín desea registrar las precipitaciones ocurridas durante el periodo de 465 días para tener el control exacto del comportamiento de las lluvias en este territorio además esta institución se interesa por verificar si la variable "control de precipitaciones" sigue una distribución de Poisson con un nivel de significación de un 5%

Solución:

$$H_0: X \sim P$$

$$H_1: X \neq P$$

xi	ni	pi	Ei	ni-Ei	(ni-Ei)2	(ni-Ei)2/Ei
2	1	0.185	86.025	85.025	7229.25063	84.0366245
3	2	0.2158	100.347	98.347	9672.13241	96.3868617
3	3	0.2158	100.347	97.347	9476.43841	94.4366888
5	4	0.1822	61.463	57.463	3301.99637	53.7233192
2	5	0.185	86.025	81.025	6565.05063	76.3156132
3	6	0.2158	100.347	94.347	8901.35641	88.7057551
4	7	0.1888	87.792	80.792	6527.34726	74.3501374

4	8	0.1888	87.792	79.792	6366.76326	72.5209958
2	9	0.185	86.025	77.025	5932.85063	68.9665867
5	10	0.1322	61.463	51.463	2648.44037	43.0899951
2	11	0.185	86.025	75.025	5628.75063	65.4315679
2	12	0.185	86.025	74.025	5479.70063	63.698932
4	13	0.1888	87.792	74.792	5593.84326	63.7170046
4	14	0.1888	87.792	73.792	5445.25926	62.0245497
2	15	0.185	86.025	71.025	5044.55063	58.6405187
2	16	0.185	86.025	70.025	4903.50063	57.0008791
3	17	0.2158	100.347	83.347	6946.72241	69.2270064
3	18	0.2158	100.347	82.347	6781.02841	67.5757961
4	19	0.1888	87.792	68.792	4732.33926	53.903992
5	20	0.1322	61.463	41.463	1719.18037	27.9709804
2	21	0.185	86.025	65.025	4228.25063	49.1514167
3	22	0.2158	100.347	78.347	6138.25241	61.1702633
5	23	0.1322	61.463	38.463	1479.40237	24.0698041
5	24	0.1322	61.463	37.463	1403.47637	22.8344918
5	25	0.1322	61.463	36.463	1329.55037	21.6317194
5	26	0.1322	61.463	35.463	1257.62437	20.4614869
3	27	0.2158	100.347	73.347	5379.78241	53.6117912
3	28	0.2158	100.347	72.347	5234.08841	52.1598893
3	29	0.2158	100.347	71.347	5090.39441	50.7279182
3	30	0.2158	100.347	70.347	4948.70041	49.315878
				Total	1746.85846	

Como el parámetro " λ " de la distribución de Poisson es desconocido tengo que estimarlo con \bar{X} .

$$\bar{X} = (2(1) + 3(2) + 3(3) + 5(4) + 2(5) + 3(6) + 4(7) + 4(8) + 2(9) + 5(10) + 2(11) + 2(12) + 4(13) + 4(14) + 2(15) + 2(16) + 3(17) + 3(18) + 4(19) + 5(20) + 2(21) + 3(22) + 5(23) + 5(24) + 5(25) + 5(26) + 3(27) + 3(28) + 3(29) + 3(30)) / 465$$

$$\bar{X} = 1630 / 465 = 3.5\lambda$$

Buscamos este último valor en la tabla obteniendo como resultado $\alpha = 0.05$.

$$RC = \{X_0 > X^2(\alpha; k-e-l)\}$$

$$RC = \{1746.85846 > X^2(0.05; 30-1-1)\}$$

$$RC = \{1746.85846 > X^2(0.05; 28)\}$$

$$RC = \{1746.85846 > 16.9\}$$

El estadígrafo de prueba pertenece a la región de crítica por lo que se rechaza la Hipótesis nula (H_0); la variable de control precipitaciones no sigue una distribución de Poisson.

3.4 Pruebas X^2 de bondad de ajuste. Pruebas de kolmogorov-smirnov de bondad de ajuste.

En este informe se realizará una recopilación de la información suficiente sobre los temas Pruebas X^2 de Bondad de Ajuste y Pruebas de Kolmogorov-Smirnov, los cuales se tratarán con sus casos respectivos y se analizarán ejemplos de cada uno.

Pruebas X^2 de Bondad de Ajuste

En esta prueba se trata, el caso en que se conoce como se distribuye la variable y las hipótesis se refieren al parámetro de la distribución. Las hipótesis nulas (H_0), contempla siempre la igualdad de la distribución de la variable en estudio, con una distribución determinada y para comprobar si se cumple o no, H_0 , se calcula la frecuencia empírica a partir de los datos, comparándose esta frecuencia con la distribución específica en H_0 (Elorza 2007).

La decisión se basa en determinar si las diferencias se deben a la aleatoriedad o a que coinciden la distribución de la variable, con la especificada en H_0 . A la distribución especificada en H_0 , con la que se compara la de la variable que se estudia, se le llama Teórica.

Esta prueba tiene tres casos los cuales se analizarán a continuación:

3.4.1.1 Criterio de Pearson:

Sea el espacio muestral de una variable dividido en K eventos exhaustivos y excluyentes A_1, \dots, A_k y P_{01}, \dots, P_{0k} sus probabilidades respectivas según la distribución teórica de que se trate.

frecuencia teórica o esperada la cual se calcula a través de la siguiente fórmula: $E_i = n \cdot P_{0i}$

Las hipótesis en general son:

$$H_0 : P_i = P_{0i} \quad (i=1, \dots, k)$$

H_1 : Al menos una de las igualdades anteriores no se cumple.

Si H_0 es cierta, se concluye que cuando n es demasiado grande $E_i \approx O_i$.

Para mejor aproximación debe cumplirse que $E_i \geq 5$ y $K \geq 5$

Como medida del grado de aproximación entre E_i y O_i es:

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^k \frac{O_i^2}{E_i} - n$$

Donde:

$$\sum_{i=1}^k O_i = \sum_{i=1}^k E_i = n$$

Luego la región crítica es: $\chi^2 \geq \chi^2_{\alpha; k-1}$

Ejemplo resuelto:

A Ud. se le encomienda la responsabilidad de decidir si un método de generación de dígitos aleatorios es bueno o no para asignar a pacientes psiquiátricos.

Para ello deben aplicarse procedimientos estadísticos que comprueben:

Que los dígitos no tengan correlación lineal.

Que los dígitos sean equiprobables.

Particularmente desea desarrollarse el experimento estadístico que compruebe la segunda restricción a través de una décima de bondad de ajuste.

Los dígitos posibles son, por supuesto: 0, 1, 2, ..., 9 y para que la serie generada cumpla lo estipulado antes y pueda considerarse aleatoria, la ocurrencia de cada uno debe tener asociada la misma probabilidad.

Se parte de una muestra de 100 dígitos generales que se pueden resumir en la siguiente tabla:

Dígito	0	1	2	3	4	5	6	7	8	9
Observaciones	12	9	9	10	10	8	9	11	10	12

Ejecute la décima con un nivel de significación del 10%.

Solución

El problema trata de probar si un conjunto de datos se ajusta (o se puede considerar proveniente) de una ley de probabilidad que en este caso particular será de característica constante: todos los dígitos, que son 10, deben tener igual probabilidad y por tanto ésta debe ser igual a $1/10 = 0,10$.

Así, la hipótesis general.

$$H_0 : P_i = P_{0i}$$

$$H_1 : P_i \neq P_{0i}$$

Se particulariza en:

$$H_0 : P_i = 0,10$$

$$H_1 : P_i \neq 0,10 \text{ (Alguno)}$$

Donde

P_i : Probabilidad del i -ésimo dígito. Con la nomenclatura usual.

O_i : Frecuencia observada del i -ésimo dígito.

$E_i = n$: Frecuencia esperada del i -ésimo dígito.

Luego en todos los casos: $E_i = (100)(0,10) = 10$;

Clase	O_i	E_i	$(O_i - E_i)^2$	$(O_i - E_i)^2 / E_i$
0	12	10	4	0,4
1	9	10	1	0,1
2	9	10	1	0,1
3	10	10	0	0
4	10	10	0	0
5	8	10	4	0,4
6	9	10	1	0,1
7	11	10	1	0,1
8	10	10	0	0
9	12	10	4	0,4
				1,6

$$\text{Estadígrafo: } \chi^2 = \sum_{i=1}^{10} \frac{(O_i - E_i)^2}{E_i} = 1.6$$

Nótese que se cumple que $E_i \geq 5$; $k \geq 5$ y que no hay parámetros estimados con la muestra ($\theta = 0$)

$$RC : \chi^2 > \chi^2 \alpha; k-1$$

$$\chi^2 > \chi^2 \alpha; 10:9$$

$$1.6 > 14.684$$

Luego $\chi^2 \leq RC$, por tanto se acepta $H_0 : P_1 = 0.10$ y se puede decidir estadísticamente que la serie generada es equiprobable.

3.4.1.2. Pruebas χ^2 de bondad del ajuste de una distribución teórica tipo, con parámetros conocidos:

Esta aplicación se basa simplemente en considerar una partición conveniente en el espacio muestral de la variable y a través de la distribución teórica tipo, calcular las posibilidades teóricas de estos eventos.

El caso más usual es en el que se quiere verificar la bondad del ajuste de una distribución normal, con parámetros conocidos. También se pudiera estudiar el caso en que se presente la distribución binomial o de Poisson y el tratamiento es similar. Solo basta especificar que es posible, en estos casos directos, considerar cada evento contenido solamente uno o varios valores y no como un intervalo, que es como se trata en el caso de la normal u otra distribución de variable aleatoria continua (O'Reilly y Rueda 1999).

En el caso en que se tiene una variable aleatoria continua la partición efectuada en el espacio muestral de la variable, se hace utilizando intervalos disjuntos y exhaustivos.

Ejemplo 1:

$$H_0 : F(x) = 1/5 \quad \text{para } x = 1, 2, \dots, 5$$

Observemos que pudieramos expresar esta hipótesis de modo similar al caso anterior.

Ejemplo 2:

$$H_0 : X \sim N(4, 4)$$

El análisis del ejemplo 1 nos indica que es posible resolver la duda aplicando el

mismo procedimiento que en el caso 1. En situaciones como el del ejemplo 2 debe hacerse una adaptación al procedimiento mediante la formación de clases o intervalos que consideramos de forma similar a como lo hacímos con los eventos del caso 1.

En este caso 2 también debe lograrse que las frecuencias esperadas sean todas mayores o iguales que 5 y deben establecerse más de 5 categorías ($E_i \geq 5$ y $k \geq 5$).

La segunda condición la tendremos en cuenta al definir las clases y la primera puede lograrse aumentando " n ". En muchos casos en los cuales en una primera tentativa no se logra que todas las E_i sean mayores o iguales que 5, puede resolverse la situación sin incrementar el tamaño de la muestra si se redefinen las clases convenientes; si la frecuencia esperada de una clase no es superior a 5, deberá combinarse con otra u otras casillas hasta que la condición quede satisfecha.

A modo de conclusión podemos decir que:

Si H_0 especifica totalmente la ley de probabilidad de una variable aleatoria y además $E_i \geq 5$, $k \geq 5$ entonces podrán usarse como estadígrafo y región crítica:

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \quad \text{y } \chi^2 \geq \chi^2 \alpha; k-1$$

Ejemplo resuelto:

En relación con la ilustración siguiente, pruébese con un nivel de significación de 0.01 si los datos pueden considerarse como una variable aleatoria que tiene la distribución de Poisson con

$$\lambda = 4.6.$$

Desempeño en el programa de entrenamiento				
Exito en el trabajo	Menor que el promedio	Promedio	Superior al promedio	Total
Deficiente	23 16.8	60 52.6	29 42.6	112
Mediano	28 25.0	79 78.5	60 63.5	167
Muy bueno	9 18.1	49 56.9	63 46.0	121
Total	60	188	152	400

Solución:

1. H_0 : La variable aleatoria tiene una distribución de Poisson con $\lambda = 4.6$.

H_1 : La variable aleatoria no tiene distribución de Poisson con $\lambda = 4.6$.

2. Nivel de confianza: $\alpha = 0.01$

3. Criterio:

Se rechaza la hipótesis nula si $X^2 > 16.919$, el valor de X^2 para $k - 1 = 10 - 1 = 9$ grado de libertad, donde X^2 está dada por la fórmula anterior (el número de grados de libertades $10 - 1 = 9$ dado que solo una cantidad, la frecuencia total de los 400, es necesario en los datos observados para calcular las frecuencias esperadas).

4. Cálculos: sustituyendo en la fórmula para X^2 se obtiene

$$X^2 = \frac{(18 - 22.4)^2}{22.4} + \frac{(47 - 42.8)^2}{42.8} + \dots + \frac{(9 - 10.0)^2}{10.0} + \frac{(8 - 8.0)^2}{8.0} = 6.749$$

5. Decisión: dado que $X^2 = 6.749$ no sobrepasa 16.919, la hipótesis nula no puede rechazarse; se concluye que la distribución de Poisson con $\lambda = 4.6$ proporciona un buen ajuste.

3.4.1.3 Pruebas X^2 de bondad del ajuste de una distribución teórica tipo, con parámetros desconocidos:

Este es el caso más frecuente en nuestro campo de aplicaciones. La primera característica de este caso es que la hipótesis plantea un tipo de distribución sin especificar completamente sus parámetros.

Ejemplo 3:

H_0 : $X \sim \text{Poisson}$ (no especifica el valor de λ)

H_0 : $X \sim N$ (no especifica los valores de μ y σ^2)

H_0 : $X \sim N(4; \sigma^2)$ (no especifica el valor de σ^2)

Para establecer la región crítica tendremos en cuenta la siguiente propiedad:

La dómica X^2 es aplicable aun cuando las probabilidades de las categorías dependen de los parámetros desconocidos, siempre que estos parámetros se sustituyan por las estimaciones apropiadas y se rebaje un grado de libertad por cada parámetro estimado.

Además se mantiene las restricciones en relación con E_i y k .

A modo de conclusión podemos decir que:

Si H_0 de una dómica específica la ley de probabilidad de una variable aleatoria, excepto "e" parámetros de la misma y, además, $E_i \geq 5$ para toda i y $k \geq 5$ entonces podrá usarse como estadígrafo y región crítica:

$$X^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \quad \text{y} \quad X^2 \geq X^2_{\alpha; k - \alpha - 1}$$

Ejemplo propuesto:

En un establecimiento de servicio se está haciendo una investigación sobre la cola de usuarios y es necesario saber si la cantidad de usuarios que arriban en una hora puede considerarse una variable con distribución Poisson.

Con este fin, durante 50 horas, se cuantifican los arribos y posteriormente se agrupan por frecuencias.

Arribos h	≤ 3	4	5	6	7	8	9	10	11	12	13	14	≥ 15
Frecuencia	0	2	2	6	9	7	11	1	5	4	2	1	0

Responda con $\alpha = 0.05$ si puede o no considerarse Poisson esta variable muestreada.

3.5 Pruebas de Kolmogorov - Smirnov para la bondad de ajuste.

La prueba de Kolmogorov - Smirnov (prueba K-S), se clasifica dentro de las llamadas pruebas de la bondad de ajuste, porque mediante su empleo, es posible determinar si una muestra proviene o no de una población que sigue una distribución previamente especificada. La prueba K-S, tiene como características esenciales la aplicación a datos continuos y su posible utilización para cualquier tamaño de muestra. Se basa en la tendiente aproximación de las frecuencias acumulativas observadas a las frecuencias acumulativas bajo el supuesto de cierta distribución, si esta distribución es la verdadera (Spinelli y Stephens 1997).

3.5.1 Pruebas de Kolmogorov - Smirnov para una muestra.

La prueba unimuestral se refiere a la concordancia entre una distribución acumulada observada de valores maestrales y una función de distribución continua determinada; es una prueba de bondad de ajuste.

Es en general más eficiente que la χ^2 para la bondad de ajuste en muestras pequeñas, y puede emplearse en muestras muy pequeñas donde la prueba χ^2 no se aplica. Los valores de la distribución acumulada de una muestra aleatoria de tamaño n y una distribución teórica determinada (Pérez et al 2009).

Para decidir si esta diferencia es mayor de lo que razonablemente puede esperarse con un nivel de significancia determinado, se buscan los valores críticos de D en la tabla (Valores críticos de D^*)

Supóngase que se quiere probar la hipótesis de que la función desconocida $F(x)$ es una función conocida $F_c(x)$.

$H_0 : F(x) = F_c(x)$ para toda x .

$H_1 : F(x) \neq F_c(x)$ para alguna x .

En la práctica, la función $F_c(x)$ podría ser, por ejemplo, una función normal con cierta media y varianza conocidas.

Sea D_n la máxima diferencia entre los valores $S_n(x)$ y $F_c(x)$, o sea:

$$D_n = \text{Máx} | S_n(x) - F_c(x) |$$

Como $S_n(x)$ se aproxima a la verdadera función de distribución de X ; entonces la probabilidad de que D_n tome valores altos es muy pequeña si $F_c(x)$ es la verdadera función de distribución.

En resumen, la prueba K-S es una prueba para variable aleatoria continua que usa todos los datos contenidos en la muestra y que se puede aplicar para cualquier tamaño de muestra, por cuanto se basa en la distribución exacta de D_n . Para valores de n (35 se puede usar una aproximación sencilla a la distribución de D_n que viene indicada en la propia tabla donde se busca D_n).

Ejemplo:

En una fábrica fueron observadas 7 máquinas destinadas a la producción de un mismo artículo A, observándose el tiempo que trabajaba cada una ininterrumpidamente. Los datos de encuentran en la siguiente tabla. Puede suponerse que la variable aleatoria X sigue una distribución $N(2,1)$.

Máquina	1	2	3	4	5	6	7
No de h de Trabajo	0	1	2	3	4	5	6

Solución:

La función $S_7(x)$ tendrá la siguiente forma:

$$S_7(x) = \begin{cases} 1/7 & \text{si } 0(X(1)) \\ 2/7 & \text{si } 1(X(2)) \\ 3/7 & \text{si } 2(X(3)) \\ 4/7 & \text{si } 3(X(4)) \\ 5/7 & \text{si } 4(X(5)) \\ 6/7 & \text{si } 5(X(6)) \end{cases}$$

1 más de 6

Valores de X	0	1	2	3	4	5	6	7
$S_7(x)$	0,1428	0,2856	0,4284	0,5712	0,7140	0,8568	1	1
$P(X=F_c(x))$	0,0228	0,1587	0,5	0,8413	0,9772	0,9986	0,9999	0,9999
Según $N(2,1)$								
Intervalos	[0,1)	[1,2)	[2,3)	[3,4)	[4,5)	[5,6)	mas de 6	mas de 6
Dif.extr.inf	0,12	0,1269	0,0716	0,2701	0,2632	0,1418	0,001	
Dif.extr.sup	-0,0159	-0,2144	-0,4129	-0,4060	-0,2846	-0,1431	-0,0001	

Como toda función de distribución es no decreciente y la función de distribución normal es una función continua, para buscar la máxima diferencia entre $S_7(x)$ y el valor de la función de distribución normal en X , solo hace falta encontrar la diferencia entre los valores de $S_7(x)$ y los valores de la función de distribución normal indicada en los extremos de los intervalos de X señalados en $S_7(x)$, porque cualquier otra diferencia para los X que están en un mismo intervalo siempre será menor.

Cuando en un intervalo como el [3,4] la curva normal queda enteramente por encima de $S_7(x)$ no cabe duda que la máxima diferencia se obtiene en el extremo superior del intervalo. Si $S_7(x)$ queda por encima de la curva normal entonces la máxima diferencia se hallara en el extremo inferior. Cuando ambas curvas se cruzan en un intervalo entonces la máxima diferencia entre las dos puede estar en cualquiera de los dos extremos.

Como regla general para este caso, podemos hallar la diferencia en los dos extremos para todos los intervalos como se ilustra:

La máxima diferencia observada es

$$\text{Máx } |S_7(x) - F_c(x)| = 0,4129$$

Fijando $\alpha = 0,05$ encontramos en la tabla correspondiente un valor D7 (0,95) = 0,486 > 0,4129. Por tanto no se puede rechazar H_0 , esto es, que no se puede rechazar la distribución de X es N (2,1). En este caso antes de tomar una decisión definitiva sería conveniente aumentar el tamaño de la muestra, pues es extraordinariamente pequeña.

3.5.2 Dócima de Kolmogorov – Smirnov para la comparación de dos poblaciones

La Dócima de Kolmogorov – Smirnov para la comparación de dos poblaciones sirve para determinar si dos muestras independientes provienen de poblaciones con una misma distribución de probabilidad. La hipótesis alternativa puede ser una hipótesis bilateral donde solo se plantee una diferencia entre las distribuciones o una hipótesis unilateral donde se prediga que la diferencia se da en una determinada dirección. La condición impuesta de ambas muestras sean independientes, es de necesario cumplimiento para el uso de esta dócima (Marini et al. 1999).

La dócima K-S para la comparación de dos poblaciones, sigue la zona aplicada en el caso de la bondad de ajuste y ahora se compararan las dos distribuciones muéstrales que resultan de la agrupación de los datos.

La dócima t para la comparación de medias tiene un sustituto en la dócima K-S para la comparación de dos poblaciones, pero esta última no exige la normalidad de los datos.

La prueba K-S para esta comparación se basa en la comparación de la distribuciones empíricas de frecuencias acumulativas formadas con las observaciones procedentes de las dos poblaciones.

La comparación de dos muestras independientes pueden ser efectuadas al comprobar si la distribución de las variables coinciden. Es decir si:

$$H_0 : F(t) = G(t) \text{ para todo } t$$

La idea de esta prueba descansa en el mismo principio de la prueba para una

muestra.

$$|F(t) - G(t)| = 0$$

Para toda t entonces tomando

$$N(t) = \text{número de observaciones de } X \text{ menores e iguales que } t$$

$$M(t) = \text{número de observaciones de } Y \text{ menores e iguales que } t$$

$$F_n(x) = N(t) / n$$

$$G_m(x) = M(t) / m$$

Deben ser similares por tanto:

$$D = \text{Max } |F_n(t) - G_m(t)|$$

Debe ser pequeña. A partir de la distribución correspondiente de este estadígrafo, no aceptamos H_0 si m y n son menores que 25 y $K_{mn} = mnD > K(n, m, \alpha)$ donde $K(n, m, \alpha)$ aparece en la tabla correspondiente a la prueba de Kolmogorov – Smirnov para las dos muestras.

Ejemplo:

El estudio de los lectores de una biblioteca es efectuado para establecer si hombres y mujeres tienen la misma distribución de tiempo de permanencia en ella. Los resultados de 5 lectores seleccionados fueron:

Hombres 2,3 4,8 1,2 0,3 4,2

Mujeres 1,3 5,4 3,3 1,9 1,4

¿Aceptaría que son iguales con $\alpha = 0,05$?

Solución:

Frecuencias y diferencias en la distribución empírica de lectores por sexo

t	0,3	1,2	1,3	1,4	1,9	2,3	3,3	4,4	4,8	5,4
F _s (t)	1/5	2/5	2/5	2/5	2/5	3/5	3/5	4/5	1	1
G _s (t)	0	0	1/5	2/5	3/5	3/5	4/5	4/5	4/5	1

$D(5,5) = 2/5$ por lo que $K(5,5) = 5.5 \cdot 0,4 = 10$. En la tabla obtenemos que:

$K(5, 5, 0, 0, 5) = 25$, por lo que aceptamos que son iguales

3.6 Distribuciones empíricas de frecuencia

La estadística es de gran importancia en la vida de todo profesional, principalmente para los investigadores de las ciencias sociales los cuales la utilizan durante casi toda su vida profesional. Con el objetivo de reforzar e incrementar los conocimientos y habilidades referentes al primer tema, Distribuciones Empíricas de Frecuencias, se ha orientado este trabajo en el cual se realizarán diversos cálculos estadísticos como por ejemplo: cálculos de medias, modas, medianas, varianzas, desviaciones típicas, etc.

Problema:

En la empresa se quiere realizar un estudio de la producción, específicamente; en el puesto de trabajo de envasado pues se han presentado problemas con el cumplimiento del plan de producción, por lo que se necesita analizar el tiempo que tarda un obrero en hacer una pieza, para ello se tomó una muestra de 50 tiempos de una población de 200 (Ver anexos).

Solución:

Definición de la variable continua

X: Tiempo que dura la realización de una pieza en minutos (variable continua).

Para resolver el problema se utilizó un Muestreo Aleatorio Simple (MAS), pues la población es homogénea, pues se está analizando la producción de una jornada laboral, específicamente, de tazas sanitarias, y todos los datos (de tiempo) son accesibles. Luego teniendo definido el tipo de muestreo a aplicar, se enumeraron los elementos de la población, los cuales fueron medidos en la empresa junto al obrero, luego se buscó en el libro de tabla, página 124, columna I, fila I. Primero se seleccionaron números de tres cifras (menores que 200) ya que la población cuenta con tres cifras valga la redundancia.

Resolución por datos agrupados:

Para la resolución del problema planteado por datos agrupados se realizaron las operaciones siguientes:

1. Valor mínimo ($X_{\text{máx}}$): 30,23

Valor máximo ($X_{\text{máx}}$): 33,99

2. Rango (R):

$$R = X_{\text{máx}} - X_{\text{min}} = 33,99 - 30,23 = 3,76$$

3. $K=10$

$$C = \frac{R}{K} = \frac{3,76}{10} = 0,376 \approx 0,38$$

Donde K: es la cantidad de intervalos

C: es la amplitud del intervalo

4. Rango de la tabla (R_T):

$$R_T = C \cdot K = 0,38 \cdot 10 = 3,8$$

$$R_T - R = 3,8 - 3,76 = 0,04$$

$$X_{\text{min}} - 0,02 = 30,23 - 0,02 = 30,21$$

$$X_{\text{máx}} + 0,02 = 33,99 + 0,02 = 34,01$$

Donde estos valores van a ser el L.I del primer intervalo y el L.S del último intervalo respectivamente.

Distribuciones Empíricas de Frecuencias:

	Clases	n_i	N_i	f_i	F_i	X_i	X_{n_i}	$X_{n_i}^2$
1	[30.21-30.59]	3	3	0.06	0.06	30.40	91.20	2772.48
2	[30.59-30.97]	8	11	0.16	0.22	30.78	246.24	7579.27
3	[30.97-31.35]	2	13	0.04	0.26	31.16	62.32	1941.89
4	[31.35-31.73]	6	19	0.12	0.38	31.54	189.24	5968.63
5	[31.73-32.11]	10	29	0.20	0.58	31.92	319.20	10188.86
6	[32.11-32.49]	0	29	0	0.58	32.30	0	0
7	[32.49-32.87]	4	33	0.08	0.66	32.68	130.72	4271.93
8	[32.87-33.25]	3	36	0.06	0.72	33.06	99.18	3278.89
9	[33.25-33.63]	12	48	0.24	0.96	33.44	397.88	13418.80
10	[33.63-64.01]	2	50	0.04	1.00	33.82	67.64	2287.85
	Total	50		1.00			294.10	1603.42
								51708.6

Donde:

n_i : Frecuencia absoluta simple $n = \sum n_i$

N_i : Frecuencia absoluta acumulada $N_i = \sum_{i=1}^n n_i$

f_i : Frecuencia relativa simple $f_i = \frac{n_i}{n}$

F_i : Frecuencia relativa absoluta $F_i = \sum_{i=1}^p f_i$

X_i : Marca de clase o punto medio del intervalo

Procesamiento de los datos:

Media o promedio:

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{249.10}{50} = 4.982 \cong 4.98$$

El resultado obtenido quiere decir que el valor medio o promedio de los tiempos de realización de las tazas sanitarias es de 4.98.

Varianza:

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^n (X_i - 4.98)^2}{49} = \frac{289.49}{49} \cong 5.91$$

Este resultado indica que la varianza de los tiempos de realización de las tazas sanitarias es 5.91 aproximadamente.

Desviación típica o estándar:

$$S = \sqrt{S^2} = \sqrt{5.91} = 2.43$$

Este resultado indica que la desviación típica o estándar de los tiempos de realización de las tazas sanitarias es 2.43 aproximadamente.

Coeficiente de variación:

$$C_v = \frac{S}{\bar{X}} = \frac{2.43}{4.98} \cong 0.49$$

Este resultado indica que el coeficiente de variación de los tiempos de realización de las tazas sanitarias es 0.49 aproximadamente.

Moda:

Es la clase que tiene una mayor frecuencia que en algún caso pueden ser hasta dos las clases modales o sea la muestra puede ser unimodal o bimodal.

$$M_o = L_i + \frac{d_1}{d_1 + d_2} \cdot C = 33.25 + \frac{0.06}{0.06 + 0.1} \cdot 0.38 = 33.39$$

Donde:

d_1 : Diferencia sin consideración de signos entre la frecuencia (f_i) de la clase modal y la de la clase precedente.

d_2 : Diferencia sin consideración de signos entre la frecuencia (f_i) de la clase modal y la de la clase siguiente.

C : amplitud del intervalo.

L_i : Límite inferior del primer intervalo.

Mediana:

Es única y siempre existe y constituye el punto central.

$$M_e = L_i + \left[\frac{\frac{n}{2} - \sum f_i}{f_{\text{mediana}}} \right] \cdot C = 31.35 + \left[\frac{\frac{50}{2} - 1}{6} \right] \cdot 0.38 = 32.996 \cong 33.00$$

Donde:

f_{mediana} : Frecuencia absoluta de la clase modal.

Interpretación de un valor de cada tipo de frecuencia de la tabla de Distribuciones empíricas de frecuencia:

$n_2 = 8$: quiere decir que 8 de las tazas tienen un tiempo de fabricación entre 30.59 y 30.97 minutos.

$f_2 = 0.16$: quiere decir que el 16% de los tiempo de fabricación de las tazas están entre 30.59 y 30.97 minutos.

$N_2 = 11$: quiere decir que 11 de la tazas tienen un tiempo de fabricación entre 30.59 y 30.97 minutos.

$F_2 = 0.22$: quiere decir que el 22% de las tazas tienen un tiempo de fabricación hasta 30.97 minutos

Los deciles, cuartiles y algunos percentiles (relaciones):

Algunos Percentiles:

$$P_{(0.10)} - ? \dots 0.10 \cdot 50 = 5$$

Como 5 representa el 10% de 50, entonces:

$$P_{[0.10]} = \frac{32.69 + 33.11}{2} = \frac{65.80}{2} = 32.90$$

$$P_{(0.20)} - ? \dots 0.20 \cdot 50 = 10$$

Como 10 representa el 20% de 50, entonces:

$$P_{[0.20]} = \frac{31.70 + 33.23}{2} = \frac{64.93}{2} = 32.47$$

$$P_{(0.30)} - ? \dots 0.30 \cdot 50 = 15$$

Como 15 representa el 30% de 50, entonces:

$$P_{[0.30]} = \frac{30.52 + 30.44}{2} = \frac{60.96}{2} = 30.48$$

$$P_{(0.40)} - ? \dots 0.40 \cdot 50 = 20$$

Como 20 representa el 40% de 50, entonces:

$$P_{[0.40]} = \frac{30.52 + 30.44}{2} = \frac{60.96}{2} = 30.48$$

$$P_{(0.50)} - ? \dots 0.50 \cdot 50 = 25$$

Como 25 representa el 50% de 50, entonces:

$$P_{[0.50]} = \frac{31.38 + 31.37}{2} = \frac{62.75}{2} = 31.375 \approx 31.38$$

$$P_{(0.60)} - ? \dots 0.60 \cdot 50 = 30$$

Como 30 representa el 60% de 50, entonces:

$$P_{[0.60]} = \frac{32.43 + 32.10}{2} = \frac{64.53}{2} = 32.265 \approx 32.27$$

$$P_{(0.70)} - ? \dots 0.70 \cdot 50 = 35$$

Como 35 representa el 70% de 50, entonces:

$$P_{[0.70]} = \frac{32.23 + 33.60}{2} = \frac{65.89}{2} = 32.945 \approx 32.95$$

$$P_{(0.80)} - ? \dots 0.80 \cdot 50 = 40$$

Como 40 representa el 80% de 50, entonces:

$$P_{[0.80]} = \frac{32.29 + 30.23}{2} = \frac{62.52}{2} = 31.26$$

$$P_{(0.90)} - ? \dots 0.90 \cdot 50 = 45$$

Como 45 representa el 90% de 50, entonces:

$$P_{[0.90]} = \frac{30.61 + 31.06}{2} = \frac{61.67}{2} = 30.835 \approx 30.84$$

Deciles:

1er decil: $P_{(0.10)} = 32.90$

Es decir, que el 10% de las tazas tienen un tiempo de fabricación por debajo de los 32.90 minutos.

2do decil: $P_{(0.20)} = 32.47$

Es decir, que el 20% de las tazas tienen un tiempo de fabricación por debajo de los 32.47 minutos.

3er decil: $P_{(0.30)} = 30.48$

Es decir, que el 30% de las tazas tienen un tiempo de fabricación por debajo de los 30.48 minutos.

4to decil: $P_{(0.40)} = 32.51$

Es decir, que el 40% de las tazas tienen un tiempo de fabricación por debajo de los 32.51 minutos.

5to decil: $P_{(0.50)} = 31.38$

Es decir, que el 50% de las tazas tienen un tiempo de fabricación por debajo de los 31.38 minutos.

6to decil: $P_{(0.60)} = 32.27$

Es decir, que el 60% de las tazas tienen un tiempo de fabricación por debajo de los 32.27 minutos.

7mo decil: $P_{(0.70)} = 32.95$

Es decir, que el 70% de las tazas tienen un tiempo de fabricación por debajo de los 32.95 minutos.

8vo decil: $P_{(0.80)} = 31.26$

Es decir, que el 80% de las tazas tienen un tiempo de fabricación por debajo de los 31.26 minutos.

9no decil: $P_{(0.90)} = 30.84$

Es decir, que el 90% de las tazas tienen un tiempo de fabricación por debajo de los 30.84 minutos.

$$\text{3er cuartil: } Q_{[0.75]} = \frac{33.38 + 31.17}{2} = 32.275 \approx 32.28$$

Este resultado indica que las tres cuartas partes de las tazas tienen un tiempo de fabricación por debajo de los 32.28 minutos.

Resolución por datos no agrupados.

Intervalos de Confianza para Tiempo

95.0% intervalo de confianza para la media: 31.9248 +/- 0.309957 [31.6148; 32.2348]

95.0% intervalo de confianza para la desviación típica:

[0.911048; 1.35908]

Resumen Estadístico para Tiempo

Frecuencia = 50	Rango = 3,76
Media = 31.9248	Primer cuartil = 31,06
Mediana = 32,025	Segundo cuartil = 32,69
Media geométrica = 31,9065	Rango intercuartílico = 1,63
Varianza = 1,18949	Coeficiente de variación = 3,41628%
Desviación típica = 1,09064	Suma = 1596,24
Error estándar = 0,15424	Mínimo = 30,23

Cuartiles:

$$\text{1er cuartil: } Q_{[0.25]} = \frac{33.23 + 32.02}{2} = 32.625 \approx 32.63$$

Este resultado indica que la cuarta parte de las tazas tienen un tiempo de fabricación por debajo de los 32.63 minutos.

$$\text{2do cuartil: } Q_{[0.50]} = \frac{31.38 + 31.37}{2} = 31.375 \approx 31.38$$

Este resultado indica que el 50% de las tazas tienen un tiempo de fabricación por debajo de los 31.38 minutos.

Percentiles:	Deciles:
1,0% = 30,23	10% = 30,23
5,0% = 30,24	20% = 30,24
10,0% = 30,44	30% = 30,44
25,0% = 31,06	40% = 31,06
50,0% = 32,025	50% = 32,025
75,0% = 32,69	60% = 32,69
90,0% = 33,305	70% = 33,305
95,0% = 33,66	80% = 33,66
99,0% = 33,99	90% = 33,99

Tabla de Frecuencias para Tiempo

Clase	Limite Inferior	Limite Superior	Marca	Frecuencia	Frecuencia Relativa	Frecuencia Acumulada	Frecuencia Acum. Rel.
Menor o igual		30,0		0	0,0000	0	0,0000
1	30,0	30,5	30,25	6	0,1200	6	0,1200
2	30,5	31,0	30,75	6	0,1200	12	0,2400
3	31,0	31,5	31,25	6	0,1200	18	0,3600
4	31,5	32,0	31,75	6	0,1200	24	0,4800
5	32,0	32,5	32,25	12	0,2400	36	0,7200
6	32,5	33,0	32,75	2	0,0400	38	0,7600
7	33,0	33,5	33,25	9	0,1800	47	0,9400
8	33,5	34,0	33,75	3	0,0600	50	1,0000
9	34,0	34,5	34,25	0	0,0000	50	1,0000
10	34,5	35,0	34,75	0	0,0000	50	1,0000
mayor		35,0		0	0,0000	50	1,0000

Referencias Bibliográficas.

- Kim, J. D., & Mueller, C. W. (1978). Factor Analysis: Statistical Methods and Practical Issues from the Series: Quantitative Applications in the Social Sciences: London: Sage University Paper.
- Kerlinger, F.N. (1975) Investigación del comportamiento: técnicas y metodología. México: Nueva Editorial Interamericana.
- Krathwohl, D. R. (1998). Methods of Educational and Social Science Research: An Integrated Approach: M Waveland Press, Inc. (Second Edition).
- Marascuilo, L.A. & Serlin, R.C. Statistical Methods for the Social and Behavioral Sciences. W.H. Freeman and Company, Nueva York, 1988.
- T. Rivas Moya, M.A. Mateo, F. Ríus Díaz, M. Ruiz, (1991). Estadística Aplicada a las Ciencias Sociales: Teoría y Ejercicios (EAC). Secretariado de Publicaciones de la Universidad de Málaga, Málaga.
- Quivy, R. y Van Campenhoudt, L. (2000). Manual de investigación en ciencias sociales. México: Noriega.
- Peña, D. y Romo, J (1997). Introducción a la estadística para las ciencias sociales. Madrid: McGraw-Hill.
- Ibáñez, J. (1993). El análisis de la realidad social. Métodos y técnicas de investigación. Varios autores. Alianza Universidad Textos. Madrid (5^a ed.)
- Solanas, A. et al (2002). La Enseñanza de la Estadística en las Ciencias del Comportamiento a Inicios del Siglo XXI. Metodología de las Ciencias del Comportamiento 4, no. 2, 157-183.
- Glass, G. y Stanley J. (1980). Métodos Estadísticos aplicados a las Ciencias Sociales. Prentice Hall, Madrid.
- J. Amón. (1980). Estadística para Psicólogos: I Estadística Descriptiva. Pirámide, Madrid.
- Hernández L. O. (1982). Elementos de Probabilidad y Estadística, Fondo de Cultura Económica, México, 1979; 2nd.
- Ritzer, Ferris J. (2003). Estadística para las Ciencias Sociales (McGraw-Hill, México).

- Field, Andy (2009). *Discovering Statistics Using SPSS for Windows*. Third Edition (Sage, London).
- Azorín Poch, F. (1972) *Curso de Muestreo y Aplicaciones*. Aguilar, Madrid.
- Manly, B.F.J. (1992) *The Design and Analysis of Research Studies*. Cambridge University Press, Cambridge.
- Badii, M.H. y J. Castillo. (2009). *Muestreo Estadística: Conceptos y Aplicaciones*. UANL, Monterrey.
- Badii, M.H., Guillen, A. y Abreu, J.L. *Tamaño Óptimo de Muestra en Ciencias Sociales y Naturales Optimal Simple Size (OSS) in Social and Natural Sciences*. International Journal of Good Conscience. 9(2)41-51. Agosto 2014. ISSN 1870-557X
- García Ferrando, Manuel (1997). *Socioestadística*. Alianza Editorial, Madrid.
- Cochran, William. (1971). *Técnicas de Muestreo*. Editorial CECSA. México
- Barbancho, A. G. (1982). *Estadística Elemental Moderna*. Ed. Ariel Economía.
- Beltrán, J. y Peris, M. J. (2013). *Introducció a l'estadística aplicada a les ciències socials*. Servei de Publicacions de la UJI . Collecció Sapientia.
- Escudero Vallés, R. (1994). *Métodos estadísticos aplicados a la economía*. Ed. Ariel Economía.
- Biosca, A., Espinet, M. J., Fandos, M. J., Jimeno, M. y Villagrà, J. (1999). *Matemáticas aplicadas a las Ciencias Sociales II*. Barcelona: Edebé.
- Brunet, I., Belzunegui, A. y Pastor, I. (2000.) *Les tècniques d'investigació social i la seva aplicació*. Universitat Rovira i Virgili.
- Colera, J., García, R. y Oliveira, M. J. (2003.) *Matemàtiques aplicades a les Ciències Socials*. Madrid: Anaya.
- Gracia, F., Mateu, J. y Vindel, P. (1997). *Problemas de Probabilidad y Estadística*. Valencia. Tilde.
- Ibáñez, M. V. y Simó, A. (2002). *Apuntes de Estadística para Ciencias Empresariales*. Castellón. UJI.
- Martín Pliego, J. (2004) *Introducción a la Estadística Económica y Empresarial*. Ed. AC. Colección Plan Nuevo.
- Valles, Miguel S. (2000). *Técnicas cualitativas de investigación social*. Madrid. Síntesis.
- Borobia Raquel. (2004). La hipótesis en estudios cualitativos. El caso de la inducción analítica en una investigación sobre adolescencia. *Revista Pilquen. Sección Ciencias Sociales*. Año VI. N° 6.
- Wackerly, Dennis D., William Mendenhall III y Ricard L. Sheaffer. (2000). *Estadística matemáticas con aplicaciones*, 6a ed., Biblioteca de Matemáticas, Thomson, México.
- Cazau, Pablo. (2006). *Introducción a la investigación en ciencias sociales*. Tercera Edición. Buenos Aires. Marzo.
- Pérez López, C. (2004). *Técnicas de análisis multivariante de datos con SPSS*. Madrid. Pearson
- Pick, Susan y López, Ana Luisa. (1994). *Cómo investigar en ciencias sociales*. 5^a ed. México. Ed. Trillas.
- Aliaga, M. y Gunderson B. (1998). *Interactive Statistics*. Edition Preliminary. Prentice Hall. Inc.
- Pett, M.A. (1997). *Nonparametric statistics for health care research*. Thousand Oaks, Cal: Sage Publications Inc.
- Moses, L.E. *Non-parametric statistics for psychological research*. Psychol Bull 1952; 49: 122-43.
- Badii, M.H., Guillen, A. Lugo Serrato, O.P. y Aguilar Garnica, J.J. *Correlación no-paramétrica y su aplicación en la investigaciones científica* non-parametric correlation and its application in scientific research. International Journal of Good Conscience. 9(2)31-40. Agosto 2014. ISSN 1870-557X
- Elorza, H. (2007). *Estadística para las ciencias sociales, del comportamiento y de la salud*. México: CENGAGE Learning.
- F. O'Reilly and R. Rueda. (1999). *Tests of fit for discrete distributions based on the probability generating function*. Comm. Statist. Sim. Comp. 28(1), 259-274.
- J. Spinelli and M.A Stephens. (1997) *Cramér-von mises tests of fit for the Poisson distribution*. Can. Jour. Statisti. 25(2), 257-268.
- Pérez Juste, R., García Llamas, J.L., Gil Pascual, J.A. y Galán González, A. (2009). *Estadística aplicada a la Educación*. Madrid. UNED - Pearson.
- Marini, Elisabetha, Racugno, Walter y Borgognini Tarli, Silvana M. (1999). *Univariate estimates of sexual dimorphism: the effects of intrasexual variability*. American Journal of Physical Anthropology. No. 109, pp. 501-508

ANEXOS

Anexo 1.

Población correspondiente a los tiempos de realización de 200 piezas en la sección de llenado.

1	2	3	4	5	6	7	8	9	10
33,82	31,54	31,38	32,65	30,74	31,27	33,13	33,21	31,28	31,75
11	12	13	14	15	16	17	18	19	20
32,18	31,53	31,02	30,15	33,59	30,69	33,09	32,69	30,76	30,23
21	22	23	24	25	26	27	28	29	30
30,16	30,26	32,61	31,26	30,76	32,5	31,44	31,2	31,17	31,92
31	32	33	34	35	36	37	38	39	40
30,69	33,93	30,36	31,03	32	31,68	33,74	30,57	30,6	32,75
41	42	43	44	45	46	47	48	49	50
31,7	32,29	31,95	33,13	31,82	30,5	32,02	31,06	32,12	31,4
51	52	53	54	55	56	57	58	59	60
32,79	32,71	33,43	31,32	32,04	31,35	32,76	33,09	33,58	31,06
61	62	63	64	65	66	67	68	69	70
32,31	31,18	30,29	32,5	32,82	33,38	31,26	32,21	33,5	31,38
71	72	73	74	75	76	77	78	79	80
32,1	30,05	30,97	31,67	30,91	30,93	33,98	32,03	31,47	31,38
81	82	83	84	85	86	87	88	89	90
31,71	33,57	31,54	31,63	30,51	31,92	31,68	33,11	30,61	30,62
91	92	93	94	95	96	97	98	99	100
31,06	33,11	31,96	32,9	33,39	31,74	31,16	31,43	32,86	33,99
101	102	103	104	105	106	107	108	109	110
33,2	33,09	30,44	33,23	30,65	30,48	30,44	31,87	33,9	31,04
111	112	113	114	115	116	117	118	119	120
30,13	32,38	30,58	30,98	30,4	31,89	32,76	33,21	32,27	33,23
121	122	123	124	125	126	127	128	129	130
30,65	32,84	31,29	30,29	32,95	33,28	32,69	32,1	33,6	32,44
131	132	133	134	135	136	137	138	139	140

30,52	31,37	33,42	30,78	31,51	33,69	33,87	31,34	31,65	33,7
141	142	143	144	145	146	147	148	149	150
33,66	31,73	32,93	32,29	31,46	33,19	31,37	33,23	30,65	33,33
151	152	153	154	155	156	157	158	159	160
31,74	31,8	31,38	30,17	32,23	32,91	32,15	31,64	33,92	33,93
161	162	163	164	165	166	167	168	169	170
33,04	30,44	33,66	31,6	31,98	30,4	33,89	32,78	31,98	30,88
171	172	173	174	175	176	177	178	179	180
31,48	31,39	32,5	30,21	32,1	33,79	30,76	31,31	33,45	32,8
181	182	183	184	185	186	187	188	189	190
33,16	30,24	31,38	32,76	32,1	32,23	32,43	31,24	32,78	32,69
191	192	193	194	195	196	197	198	199	200
33,3	32,2	30,4	32,24	30,23	30,44	30,14	30,38	32,35	30,52

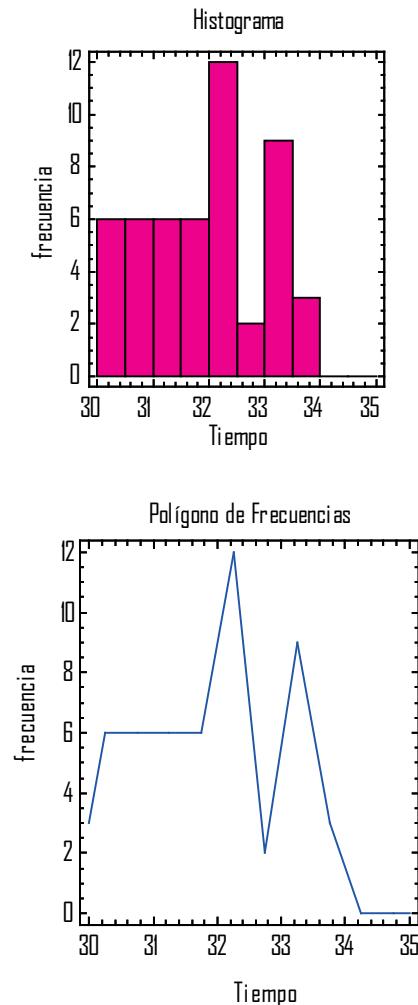
Anexo 2

Muestra correspondiente a los tiempos de realización de 50 piezas seleccionadas de una población de 200 (ver anexo 1).

1	2	3	4	5	6	7	8	9	10
33,99	30,52	33,93	33,09	32,69	33,11	31,32	30,24	31,75	31,7
11	12	13	14	15	16	17	18	19	20
33,23	30,14	33,23	32,02	30,52	30,44	32,23	32,29	30,44	33,2
21	22	23	24	25	26	27	28	29	30
31,82	31,63	30,74	33,23	31,38	31,37	32,21	33,43	30,4	32,43
31	32	33	34	35	36	37	38	39	40
32,1	30,58	33,11	31,06	32,23	33,66	32,52	33,38	31,17	32,29
41	42	43	44	45	46	47	48	49	50
30,23	31,98	32,61	32,1	30,61	31,06	32,5	32	33,2	32,03

Anexo3:

Gráficos correspondientes al Histograma y al Polígono de Frecuencias.



Ejemplo Resuelto.

En un Hospital Pediátrico de Ecuador, específicamente en la consulta de neurofisiología, se desea analizar los índices de concurrencia a la consulta (es decir, cantidad de pacientes que asisten diariamente) y para ello se decidió tomar una muestra de 50 días.

111

Este trabajo se hace con la finalidad de conocer el número exacto de los pacientes que visitan a diario esta consulta pues se desea construir una sala de espera acorde con esta cantidad para brindarle con la máxima calidad la asistencia médica necesaria.

Solución:

Definiendo la variable:

X: cantidad de pacientes que asisten diariamente a la consulta de neurofisiología en el Hospital Pediátrico

Para la resolución del problema anteriormente expuesto se ha decidido utilizar un muestreo aleatorio sistemático (M.S.A) debido a las siguientes razones:

Este muestreo se utiliza cuando el volumen de la población que se estudia es finito y no muy grande, y además, se conoce que es homogénea en cuanto a la "variable que se investiga", tal y como ocurre en el M.A.S.

Para realizar este muestreo se siguen los siguientes pasos:

1.- Realizar un listado de los elementos de la población y numerar consecutivamente, desde uno hasta n, a cada elemento de ella.

2.- De entre los k primeros números del listado de la población, tomar uno al azar lo que se puede hacer empleando una tabla de números aleatorios.

El valor de k se decide de la siguiente forma: así por ejemplo, si n=50 y N=200 entonces como $k = \frac{N}{n}$ se obtiene que k=4, ahora seleccionamos al azar un número entre 1 y 4, fue elegido el 3.

3.- A partir del número seleccionado al azar en el paso anterior, se comienza a conformar la muestra. El primer elemento de dicha muestra será aquel que en el listado original le corresponde el número aleatorio seleccionado (en este caso el 3); ahora a dicho número le adicionamos el valor de k y al elemento de la población que le corresponda esa suma, será el siguiente integrante. Este proceso se seguirá hasta completar el volumen de la muestra.

Utilidad: Control de la Calidad

► Ventaja: Es rápido, práctico y no requiere de personal altamente calificado.

► Desventaja: Los resultados obtenidos pueden estar viciados por factores subjetivos.

112

Resolución por datos agrupados:

Para la resolución del problema planteado por datos agrupados se realizaron las operaciones siguientes:

Valor mínimo de x : 5

Valor máximo de x : 25

Rango (R) $R = V_{\text{máx}} - V_{\text{min}} = 25 - 5 = 20$

$$R = 20$$

Estableciendo a $K = 5$ (Número de intervalos o clases)

$$C = \frac{R}{K} = \frac{20}{5} = 4 \quad \text{Donde } C: \text{tamaño o amplitud del intervalo.}$$

$$R_t = C * K = 4 * 5 = 20$$

$$R_t - R = 20 - 20 = 0$$

X mínimo	5
X máxima	25

Distribución Empírica de Frecuencias

	Intervalos o clases	n _j	N _j	f _j	F _j	x _j	X _j *n _j	X _j ² *n _j
1	5-Sept	1	1	0.02	0.02	7	7	49
2	Sept-13	1	2	0.02	0.04	11	11	121
3	13-17	14	16	0.28	0.32	15	210	3150
4	17-21	22	38	0.44	0.76	19	418	7942
5	21-25	12	50	0.24	1.00	23	276	6348
	Total	50	107	1.00	2.14	75	922	17610

Donde:

n_j: frecuencia observada simple,

N_j: frecuencia absoluta acumulada,

F_j: frecuencia relativa acumulada,

f_j: frecuencia relativa simple.

x_j: marca de clase o punto medio del intervalo.

Fórmulas:

$$\sum n_j = n \quad N_j = \sum_i^n n_j \quad f_j = \frac{n_j}{n} \quad \sum f_j = 1 \quad F_j = \sum_i^j f_j = \frac{N_j}{n}$$

Procesamiento de los datos

Media o promedio

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_j * n_j = \frac{922}{50} = 18.44 \approx 19$$

El resultado obtenido nos expresa que el valor medio o promedio de los pacientes atendidos en un día en la consulta de neurofisiología del Pediátrico es el siguiente: 19

Varianza

$$S^2 = \frac{\sum x_j^2 * n_j - \frac{(\sum x_j * n_j)^2}{n}}{n - 1} = \frac{17610 - \frac{850084}{50}}{49} = \frac{608.32}{49} = 12.4147 \approx 12.41$$

Este resultado indica que la varianza de los pacientes atendidos en un día en la consulta de neurofisiología del Pediátrico es el siguiente: 12.41.

Desviación típica o estándar

S = $\sqrt{S^2} = \sqrt{12.41} = 3.5228$ 3.52. Este resultado indica que la desviación típica o estándar de los pacientes atendidos en un día en la consulta de neurofisiología del Pediátrico es el siguiente es de 3.52.

Cv: coeficiente de variación

$$Cv = \frac{S}{\bar{x}} * 100 = \frac{3.52}{18.44} * 100 = 0.1908 * 100 = 19.08 \%$$

Este resultado muestra el coeficiente con el que varían los datos de los pacientes atendidos en un día en la consulta de neurofisiología del Pediátrico es de 19.08 %.

Moda: es la clase que tiene una mayor frecuencia que en algún caso pueden ser hasta dos las clases modales o sea la muestra puede ser unimodal o bimodal.

$$114 \quad Mo \approx Lmo + \left(\frac{dl}{dl + d2} \right) * C \quad Lmo: \text{límite inferior de la clase modal.}$$

d2: diferencia sin consideración de signos entre la frecuencia de la clase modal y la de la clase siguiente.

d1: diferencia sin consideración de signos entre la frecuencia de la clase modal y la de la clase precedente.

C: amplitud del intervalo.

$$Mo \approx l_7 + \left(\frac{8}{8+10} \right) * 4 \approx l_7 + 1.77$$

Este resultado de moda obtenido evidencia que el índice que más se repite en la muestra analizada es 18.77.

Mediana: es única y siempre existe, constituye el punto central.

$$Me \approx L_m + \left[\frac{\frac{n+1}{2} - S}{nm} \right] * C \approx l_7 + \left[\frac{\frac{51}{2} - 16}{22} \right] * 4 \approx l_7 + 1.72$$

nm: frecuencia de la clase modal.

Lm: límite inferior de la clase modal.

S: suma de las nj de las clases anteriores.

El resultado obtenido indica que el punto medio de los pacientes atendidos en un día en la consulta de neurofisiología del Pediátrico es el siguiente es de 18.72, por debajo del cual están el 50% de los datos y por encima, de igual forma, el otro 50%.

Interpretación de algunos de los estadígrafos de posición calculados, presentes en la tabla de distribución empírica de frecuencias.

nj (03) = 14 quiere decir que en 14 de los 50 días se atendieron entre 13-17 pacientes.

Nj (02) = 2 expresa que en 2 de los días se atendieron de 5 a 13 pacientes o hasta 13 pacientes.

fj (04) = 0.44 indica que en el 44% de los días se atendieron de 17-21 pacientes.

Fj (03) = 0.32 el resultado expresa que en el 32% de los días se atendieron de 5-17 o hasta 17 pacientes.

Los deciles, cuartiles y algunos percentiles (relaciones)

Percentiles: el p.ésimo percentil es un valor tal que al menos 100p% de los datos están por debajo de ese valor, y cuando menos 100(l-p)% están en o sobre ese valor.

Deciles: Dividen los datos en diez partes iguales o sea son puntos de división resultantes.

1er decil: contiene el 10% de los datos menores que él y a la vez es el décimo percentil.

2do decil: contiene el 20% de los datos menores que él y a la vez es el décimo segundo percentil.

3er decil: contiene el 30% de los datos menores que él y a la vez es el décimo tercer percentil.

4to decil: contiene el 40% de los datos menores que él y a la vez es el décimo cuarto percentil.

5to decil: contiene el 50% de los datos menores que él, además es a la vez es el quincuagésimo percentil y el segundo cuartil que coincide con la mediana.

(Esto se repite de forma similar para los demás deciles)

Cuartiles: Dividen los datos en cuatro partes iguales y al igual que los deciles son puntos de división resultantes.

1er cuartil: contiene el 25% de los datos menores que él y a la vez es el 25avo percentil de la muestra.

2do cuartil: contiene el 50% de los datos menores que él o sea es el punto medio de la muestra, mediana, y a la vez es el quincuagésimo percentil y el quinto decil.

3er cuartil: contiene el 75% de los datos menores que él y a la vez es el 75avo percentil de la muestra.

Cálculos:

Para apoyar el cálculo se confecciona una tabla con los datos de la muestra en orden ascendente y debidamente enumerados (ver anexo 3)

Algunos percentiles

$$1\% = 5$$

$$25\% = 17$$

$$50\% = 20$$

$$75\% = 21$$

$$99\% = 25$$

Deciles:

$$1er decil = [P_{0.10}] = \frac{3.22 + 3.24}{2} = 3.23$$

Este resultado indica que el 10% de los índices de los estudiantes de Ing. Industrial de Iro a 3er año de esa misma facultad en la Uho Oscar Lucero M. durante el curso 2005-2006 están por debajo de 3.23.

$$3er decil = [P_{0.30}] = \frac{3.25 + 3.33}{2} = 3.29$$

Este resultado indica que el 30% de los índices de los estudiantes de Ing. Industrial de Iro a 3er año de esa misma facultad en la Uho Oscar Lucero M. durante el curso 2005-2006 están por debajo de 3.29.

$$4to decil = [P_{0.40}] = \frac{3.49 + 3.51}{2} = 3.50$$

$$6to decil = [P_{0.60}] = \frac{3.92 + 3.99}{2} = 3.955 \approx 3.96$$

Este resultado indica que el 60% de los índices de los estudiantes de Ing. Industrial de Iro a 3er año de esa misma facultad en la Uho Oscar Lucero M. durante el curso 2005-2006 están por debajo de 3.96.

$$7mo decil = [P_{0.70}] = \frac{4.75 + 4.80}{2} = 4.78$$

$$8vo decil = [P_{0.80}] = \frac{4.75 + 4.75}{2} = 4.75$$

Este resultado indica que el 80% de los índices de los estudiantes de Ing. Industrial de Iro a 3er año de esa misma facultad en la Uho Oscar Lucero M. durante el curso 2005-2006 están por debajo de 4.75 y solo el 20% se encuentra por encima.

$$9no decil = [P_{0.90}] = \frac{4.90 + 4.95}{2} = 4.93$$

Cuartiles:

$$1er cuartil = Q_1 = \frac{3.22 + 3.24}{2} = 3.23 = [P_{0.25}]$$

El resultado indica que la cuarta parte de los índices de los estudiantes de Ing. Industrial de Iro a 3er año de esa misma facultad en la Uho Oscar Lucero M. durante el curso 2005-2006 están por debajo de 3.23.

$$2do cuartil = Q_2 = \frac{3.77 + 3.80}{2} = 3.78 = [P_{0.50}] \text{ (mediana)}$$

Este resultado indica que el 50% de los índices de los estudiantes de Ing. Industrial de Iro a 3er año de esa misma facultad en la Uho Oscar Lucero M. durante el curso 2005-2006 están por debajo de 3.78 y por encima de este # está el otro 50% constituyendo el mismo la mediana o punto medio de los mismos.

$$3er cuartil = Q_3 = \frac{4.35 + 4.35}{2} = 4.35 = [P_{0.75}]$$

El resultado indica que las tres cuartas partes de los índices de los estudiantes de Ing. Industrial de Iro a 3er año de esa misma facultad en la Uho Oscar Lucero M. durante el curso 2005-2006 están por debajo de 4.35.

Resolución por datos no agrupados:

	X_i	n_i	N_i	f_i	R	x^2
1	5	1	1	0.02	0.02	25
2	10	1	2	0.02	0.04	100
3	14	3	5	0.06	0.1	196x3=588
4	15	3	8	0.06	0.16	225x3=675
5	16	4	12	0.08	0.24	256x4=1024
6	17	4	16	0.08	0.32	289x4=1156
7	18	5	21	0.1	0.42	324x5=1620
8	19	3	24	0.06	0.48	367x3=1101
9	20	3	27	0.06	0.54	400x3=1200
10	21	11	38	0.22	0.76	441x11=4851
11	22	7	45	0.14	0.90	484x7=3388
12	23	3	48	0.06	0.96	529x3=1587
13	24	1	49	0.02	0.98	576
14	25	1	50	0.02	1	625
Sumatoria	249	50		1		18516

$$\text{Promedio o Media } x = \frac{\sum xi}{n} = 18.88 \approx 19$$

$$\text{Mediana (como la muestra es par) } me = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} = \frac{20 + 20}{2} = 20$$

Moda: 21

$$\text{Varianza } S^2 = \frac{\sum x^2 - \frac{(\sum xi)^2}{n}}{n-1} = \frac{18516 - \frac{891136}{50}}{49} = 14.15$$

$$\text{Desviación Estándar } S = \sqrt{S^2} = \sqrt{14.15} = 3.76$$

$$\text{Coeficiente de Variación } Cv = \frac{S^2}{x} = \frac{14.15}{18.88} = 0.75$$

Percentiles

10,0% = 10,0

25,0% = 15,0

50,0% = 18,5

75,0% = 22,0

99,0% = 25,0

Resumen Estadístico para Col_1

Frecuencia = 14	Mínimo = 5,0
Media = 17.7857	Máximo = 25,0
Mediana = 18,5	Rango = 20,0
Moda = 21	Primer cuartil = 15,0
Media geométrica = 16,6378	Segundo cuartil = 22,0
Varianza = 30,9505	Rango intercuar. = 7,0
Desviación típica = 5,56332	Coef. de variación = 31,2797%
Error estándar = 1,48686	Suma = 249,0

Conclusiones

Luego de la investigación estadística llevada a cabo se determinó que el número promedio de pacientes que asiste diariamente a esta consulta es 19, con una confianza del 95%, por lo que podemos agregar que el valor ofrecido en la resolución del problema es válido y fiable. Esto se pudo garantizar trabajando con dos formas diferentes de agrupar los datos recogidos y de procesarlos.

Se ofrecen, además, valores de las frecuencias observadas de asistencia así como de los porcentajes de mayor importancia para generar las conclusiones del trabajo y se expresan sus significados a continuación permitiendo así una completa explicación que ayude al entendimiento del informe p

Anexo 3.

Población seleccionada correspondiente a la cantidad de personas atendidas diariamente en la consulta de neurofisiología.

001	002	003	004	005	006	007	008	009	010
25	18	22	17	23	6	15	23	10	23
011	012	013	014	015	016	017	018	019	020
23	15	14	6	18	22	19	20	23	16
021	022	023	024	025	026	027	028	029	030
20	23	17	20	18	21	20	17	18	14
031	032	033	034	035	036	037	038	039	040
18	6	19	21	22	23	23	22	5	19
041	042	043	044	045	046	047	048	049	050
20	22	19	22	21	20	18	17	18	22
051	052	053	054	055	056	057	058	059	060
21	18	22	5	14	20	15	16	15	20
061	062	063	064	065	066	067	068	069	070
18	23	18	22	22	22	16	20	23	20
071	072	073	074	075	076	077	078	079	080
25	21	20	22	21	23	22	18	22	23
081	082	083	084	085	086	087	088	089	090

22	22	22	25	18	21	22	23	22	16
091	092	093	094	095	096	097	098	099	100
15	17	21	20	17	16	23	5	16	10
101	102	103	104	105	106	107	108	109	110
16	11	14	14	14	16	17	15	20	21
111	112	113	114	115	116	117	118	119	120
16	17	14	14	10	9	10	15	17	18
121	122	123	124	125	126	127	128	129	130
21	21	21	23	22	23	21	19	21	18
131	132	133	134	135	136	137	138	139	140
22	19	23	18	18	21	14	13	16	11
141	142	143	144	145	146	147	148	149	150
10	12	21	12	18	19	21	20	20	19
151	152	153	154	155	156	157	158	159	160
21	21	21	22	21	21	20	25	24	21
161	162	163	164	165	166	167	168	169	170
21	20	19	18	19	21	23	24	21	20
171	172	173	174	175	176	177	178	179	180
20	21	20	22	21	23	22	18	14	23
181	182	183	184	185	186	187	188	189	190
22	22	22	25	18	21	21	23	22	16
191	192	193	194	195	196	197	198	199	200
19	20	23	24	21	15	17	18	20	9

Anexo 4.

Muestra seleccionada de la cantidad de pacientes que asisten a la consulta de neurofisiología del Pediátrico Ecuatoriano

01	02	03	04	05	06	07	08	09	10
22	15	23	18	23	17	20	18	22	5
11	12	13	14	15	16	17	18	19	20
19	18	21	14	15	18	16	25	21	22
21	22	23	24	25	26	27	28	29	30
22	22	15	17	16	14	17	16	10	17
31	32	33	34	35	36	37	38	39	40
21	21	22	18	16	21	21	21	21	24
41	42	43	44	45	46	47	48	49	50
19	23	20	21	14	22	21	19	21	20

Anexo 5

Tabla de la muestra obtenida ordenada de forma ascendente, utilizada en el cálculo de algunos percentiles, algunos deciles y los cuartiles.

01	02	03	04	05	06	07	08	09	10
5	10	14	14	14	15	15	15	16	16
11	12	13	14	15	16	17	18	19	20
16	16	17	17	17	17	18	18	18	18
21	22	23	24	25	26	27	28	29	30
18	19	19	19	20	20	20	21	21	21
31	32	33	34	35	36	37	38	39	40
21	21	21	21	21	21	21	21	22	22
41	42	43	44	45	46	47	48	49	50
22	22	22	22	22	23	23	23	24	25

Diagramas de Frecuencias Absolutas o Frecuencias Observadas

Diagrama de Barras de Cant Pac

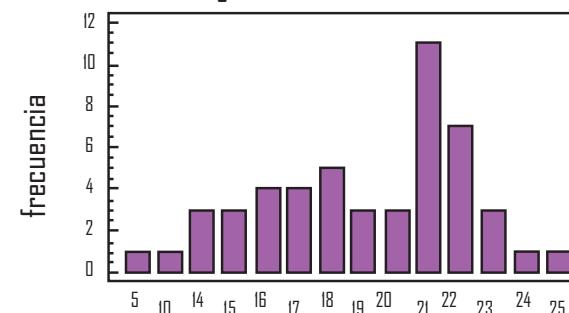


Diagrama de Sectores de Cant Pac

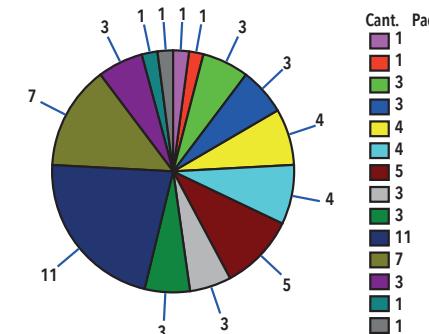
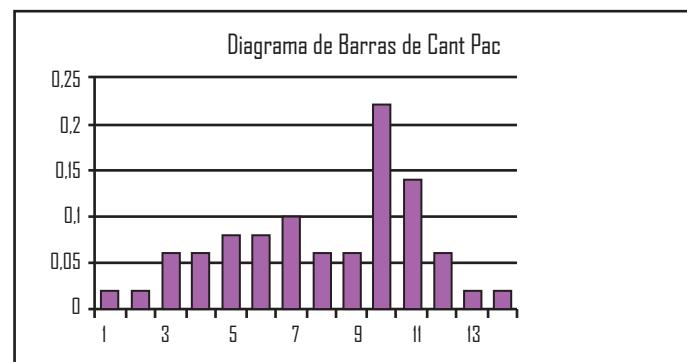
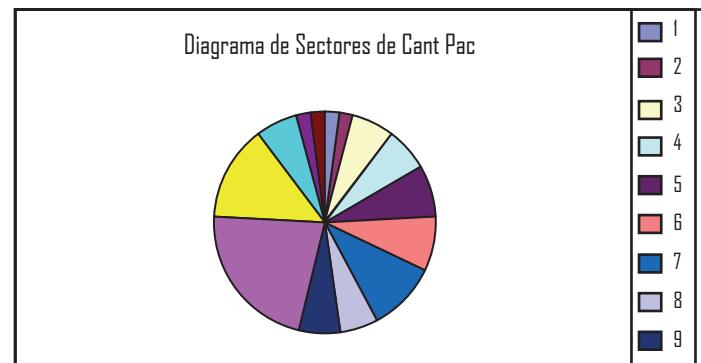


Diagrama de Frecuencias Relativas Simples



Anexo 6.

ALGUNOS FUNDAMENTOS MATEMÁTICOS EN LA TEORÍA DE LA CONFIABILIDAD.

La tarea de diseñar y fabricar un producto, cada día se hace más complejo, por la propia complejidad de los productos, la agresividad de las condiciones ambientales a que se ven sometido los productos, los costos de producción y mantenimiento, la fiabilidad (confiabilidad) del producto.

Nos dedicaremos a mostrar algunos elementos de las teorías matemáticas útiles en el estudio de la confiabilidad y el tiempo de vida útil.

Definición: definiremos como fiabilidad de un producto, la PROBABILIDAD de que el producto funcione dentro de límites dados al menos durante un período de tiempo en condiciones de trabajo específicas.

Lo anterior, dice que un producto puede funcionar de manera satisfactoria bajo determinadas condiciones, pero no funcionar satisfactoriamente cuando las condiciones cambian, que el rendimiento del producto para un fin, no garantiza un rendimiento adecuado en otro. "Confiabilidad es calidad en el tiempo".

La definición nos pone en contacto con los primeros elementos matemáticos, que debemos conocer al estudiar confiabilidad, Teoría de la Probabilidad.

Espacio muestral: Conjunto de todos los resultados de un experimento, S .

Ejemplo: El M.E.E. quiere construir 2 nuevas hidroeléctricas (H) y quiere indicar cuantas hidroeléctricas (H) están en la Provincia de Cotopaxi (C) y cuántas en la Provincia de Guayas (G). Escribir S .

Solución: C y G toman valores 0, 1, 2. Sea (C, G) par ordenado
 $S = \{(1,0), (0,1), (1,1), (0,2), (2,0), (0,0)\}$

Evento: cualquier parte de S . Cualquier $E \subseteq S$, incluye S y \emptyset .

Ejemplo:

- a) Cotopaxi y Guayas tienen la misma cantidad de Hidroeléctricas: $E_1 = \{(0,0), (1,1)\}$
- b) Cotopaxi y Guayas no fueron tomadas en cuenta: $E_2 = \emptyset$
- c) Cotopaxi no recibió H : $E_3 = \{(0,1), (0,2), (0,0)\}$
- d) Cotopaxi recibe al menos una H : $E_4 = \{(1,0), (1,1), (2,0)\}$

Si E_3 y E_4 no tienen elementos en común, se llaman eventos mutuamente excluyentes,

$$Si E_3 \cup E_4 = S$$

Consideraremos n el número de elementos de (S) y por (e) el número de elementos de cualquier $E \subseteq S$

Definición: Si los n elementos de S son igualmente posible y ocurren, e son considerados éxitos, entonces llamaremos probabilidad que ocurra "un" éxito,

$$P(E) = \frac{e}{n}$$

Ejemplo:

a) $P(E_1) = \frac{2}{6} = \frac{1}{3} \approx 33.3\%$

b) $P(E_2) = P(\emptyset) = 0$

c) $P(E_3) = \frac{3}{6} = \frac{1}{2} = 50\%$

d) $P(E_4) = \frac{3}{6} = \frac{1}{2} = 50\%$

Algunas propiedades de Probabilidad:

Dado S y $E \subseteq S$, la $P(E)$ cumple:

1. $0 \leq P(E) \leq 1$

2. $P(S) = 1; P(\emptyset) = 0$

3. Sean E_1 y E_2 eventos de S , MUTUAMENTE EXCLUYENTES, entonces:

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

4. Sean E_1 y E_2 eventos independientes de S , entonces:

$$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2)$$

Las propiedades 3 y 4 son generalizadas para n eventos de S.

5. Sea el complemento de E (E' contiene todos los elementos de S que no están en E), entonces $P(E') = 1 - P(E)$

Demostración de las propiedades:

1. Sea $E \subseteq S$, tal que $0 \leq e \leq n$, entonces:

Si $0 < e < n$, $P(E) = \frac{e}{n} < 1.0 < P(E) < 1$
Si por ser

2. Si $e = n$, $P(E) = \frac{n}{n} = 1 = P(S)$, por ser $E = S$

Si $e = 0$, $P(E) = \frac{0}{n} = 0 = P(\emptyset)$, por se $E = \emptyset$

3. $P(E_1) = \frac{e_1}{n}, P(E_2) = \frac{e_2}{n}; P(E_1 \cup E_2) = P(E_1) + P(E_2); \frac{e_1 + e_2}{n} = \frac{e_1}{n} + \frac{e_2}{n} = \frac{e_1 + e_2}{n}$

4. $P(E_1 \cap E_2) = P(E_1) \cdot P(E_2); \frac{e_1 \cdot e_2}{n^2} = \frac{e_1}{n} \cdot \frac{e_2}{n} = \frac{e_1 \cdot e_2}{n^2}$

5. Consideraremos E y E'

$$P(E \cup E') = P(S)$$

$$P(E) + P(E') = 1$$

$$P(E') = 1 - P(E)$$

Dado que el complemento de S es, de 5 se tiene:

$$P(\emptyset) = 1 - P(S) = 1 - 1 = 0$$

Todo producto en general puede ser considerado un sistema S de n componentes independientes conectados, en serie, paralelo o ambas combinadas.

Sistema en serie: El sistema deja de funcionar si al menos uno de sus n componentes falla.

Sistema en paralelo: El sistema deja de funcionar si sus n componentes falla.

Determinemos la confiabilidad para cada conexión:

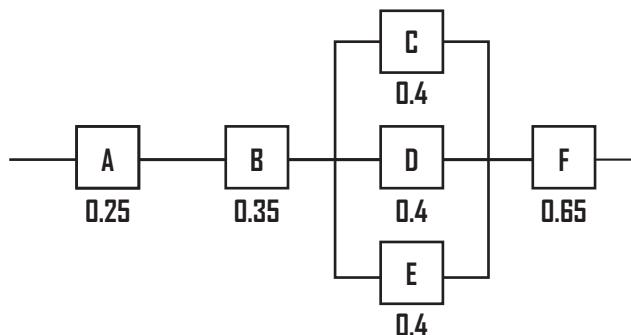
En la conexión en serie la confiabilidad de un componente no afecta la confiabilidad de los otros, entonces por la definición de confiabilidad, la probabilidad de que S funcione es igual al producto de la probabilidad de funcionamiento de cada uno de los n componentes. Aplicando 4 generalizada.

$$C_s = P_s = \prod_{i=1}^n P_i = P_1 P_2 \dots P_{n-1} P_n = \prod_{i=1}^n C_i \quad \text{ec.6}$$

En la conexión en paralelo el sistema falla si sus n componentes fallan. Entonces la confiabilidad del sistema es la probabilidad de que el sistema falle, $P_s' = 1 - P_s$, aplicando ec.5 y ec.6 obtenemos.

$$C_s' = 1 - P_s = 1 - \prod_{i=1}^n (1 - P_i) = 1 - \{(1 - P_1)(1 - P_2) \dots (1 - P_n)\} = 1 - \prod_{i=1}^n (1 - C_i) \quad \text{ec.7}$$

Ejemplo: dado el sistema, los valores representan los valores de confiabilidad de cada componente. Determinar la confiabilidad del sistema.



El sistema está compuesto por conexiones en serie y en paralelo.

$$C_{CDE} = 1 - (1 - 0.4)^3 = 0.784$$

$$C_s = C_A C_B C_{CDE} C_F = (0.25)(0.35)(0.784)(0.65) = 0.04459$$

DISTRIBUCIÓN DEL TIEMPO DE FALLA

FALLA: Cuando el producto deja de realizar satisfactoriamente la función para la que fue creada.

TIEMPO DE FALLA: Tiempo hasta que el producto falla.

Para estudiar el tiempo de falla, debemos estudiar la Razón de Falla que caracteriza la distribución del tiempo de falla.

Vamos a recordar algunos conceptos que necesitaremos, durante todo el estudio.

VARIABLE ALEATORIA C: Es una función definida sobre el espacio muestral S.

Para cada valor de VAC sobre el espacio muestral, se le hace corresponder su valor único de probabilidad $f = S \rightarrow [0,1]$, que denominaremos función densidad de probabilidad de la VAC.

Esta función $f(x)$ cumple las siguientes condiciones:

$$1) f(x) \geq 0; \forall x \in D_f$$

$$2) \int_{-\infty}^{\infty} f(x) dx = 1, \text{ con estas condiciones cumple con las propiedades}$$

1), 2) y 3) de probabilidad.

Definimos una función $F(x)$ como la probabilidad de que la VAC con $f(x)$ tome un valor menor o igual a x ($P(f(x) \leq x)$) en general:

$$F(x) = \int_{-\infty}^x f(x) dx, \text{ FUNCIÓN DE DISTRIBUCIÓN DE LA VAC.}$$

CÁLCULO DE LA RAZÓN DE FALLA

Entonces, la probabilidad de que el componente falle $[0, t]$, viene dado por la

$$F(t) = \int_0^t f(x) dx \quad (*)$$

La confiabilidad que el componente dure más del tiempo t ,

$$C(t) = 1 - F(t) \quad \text{ec. 8}$$

DETERMINEMOS LA RAZÓN DE FALLA:

La probabilidad que el componente falle $[t, t + \Delta t]$ dado que el componente duró más de t , viene dada:

$$\frac{F(t+\Delta t) - F(t)}{C(t)} : \text{ multiplicando por } \frac{1}{\Delta t} \text{ y calculando } \lim_{\Delta t \rightarrow 0}$$

$$\lim_{\Delta t \rightarrow 0} \frac{F(t+\Delta t) - F(t)}{\Delta t} \bullet \frac{1}{C(t)} = \frac{F'(t)}{C(t)} = Z(t)$$

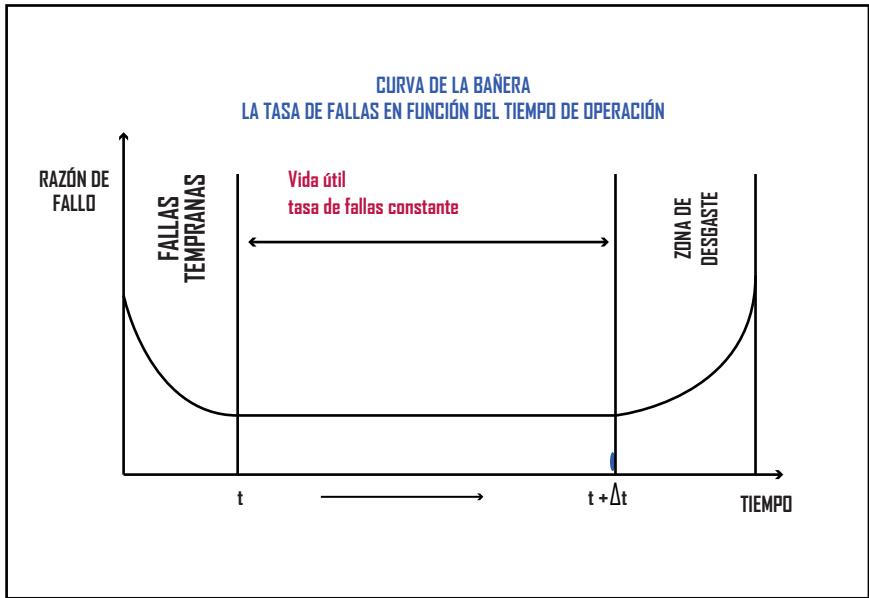
$$Z(t) = \frac{F'(t)}{C(t)} \quad \text{ec. 9} \quad \text{RAZÓN DE FALLA}$$

$$Z(t) = \frac{f(t)}{1-F(t)} \quad \text{ec. 10} \quad \text{RAZÓN DE FALLA}$$

Derivando (*) $F'(t) = f(t)$ (**)
y sustituyendo (8) en (9) obtenemos:

Razón de falla en términos de las funciones densidad y función distribución del tiempo de falla.

Una curva que caracteriza a la razón de falla, es la curva de la bañera:



Esta curva también expresa el comportamiento de la mortalidad humana, la primera parte representa la mortalidad infantil y la tercera curva representa la tercera edad.

Derivando (8) y sustituyendo (9) obtenemos:

$$C'(t) = -F'(t); C'(t) = -Z(t)C(t) \quad \text{ec. II}$$

Resolviendo (II) que representa una ecuación diferencial ordinaria de primer orden en Variables Separables.

$$\frac{dC(t)}{dt} = -Z(t)C(t)$$

$$\frac{dC(t)}{C(t)} = -Z(t)dt, \text{ integrando en } [0, t], \text{ ambos miembros}$$

$$\ln C(t) = - \int_0^t Z(t)dt, \text{ aplicando Euler a ambos miembros}$$

$$C(t) = e^{- \int_0^t Z(t)dt} \quad \text{ec. I2}$$

De (8), (I0) y (I2) obtenemos:

$$Z(t)e^{- \int_0^t Z(t)dt} = f(t) \quad \text{ec. I3} \quad \text{ECUACIÓN GENERAL PARA EL TIEMPO DE FALLA}$$

Consideraremos $Z(t) = \alpha = \text{constante positiva}$

Sustituyendo en (I3):

$$\alpha e^{-\alpha t} = f(t) \quad \text{ec. I4} \quad \text{ECUACIÓN EXPONENCIAL para el tiempo de falla, con razón de falla constante}$$

I4).- Expresa una idealización, solamente muestra en la gráfica de la bañera las fallas por accidente, el período de vida útil del componente, pues se considera la Razón de Falla constante. No expresa nada relativo a las partes creciente y decreciente de la curva.

Para obtener un método que exprese de forma más próximo a la realidad, tenemos:

$$Z(t) = \alpha \beta t^{\beta-1} \quad \text{ec. I5; } t > 0; \alpha, \beta \in \mathbb{R}^+$$

$$Z(t) = \begin{cases} \beta = 1, & \text{parte constante} \\ \beta < 1, & \text{parte decreciente} \\ \beta > 1, & \text{parte creciente} \end{cases}$$

Sustituyendo (I5) en (I3)

$$\alpha \beta t^{\beta-1} e^{- \int_0^t \alpha \beta dt} = f(t); \text{ resolviendo la integral}$$

$$-\alpha \beta \int_0^t t^{\beta-1} dt = -\frac{\alpha \beta t^\beta}{\beta} \Big|_0^t = -\alpha t^\beta \text{ sustituyendo}$$

FUNCIÓN DE WEIBULL, α, β, t positivos para el tiempo de falla

MODELO DE WEIBULL EN PRUEBAS DE VIDA.

Un método eficaz y ampliamente utilizado para resolver problemas en Teoría de Confiabilidad es la prueba de vida.

Para realizar ésta prueba, se selecciona de forma aleatoria n componentes y se someten a prueba bajo condiciones específicas y se observan los tiempos de fallos de cada componente.

Las pruebas de vida se pueden clasificar en prueba con reemplazo y prueba sin reemplazo, prueba acelerada. La prueba acelerada permite reducir el tiempo y el número de componentes a ser utilizadas en la prueba. Cuando se emplea esta prueba se aconseja emplear MÉTODOS ESTADÍSTICA de predicción y optimización, dentro de los cuales se encuentran, EL AJUSTE DE CURVA utilizar mínimo 20 componentes.

El modelo de Weibull en pruebas de vida, describe la manera adecuada los tiempos de falla de los componentes.

Sean:

$$1. \quad f(t) = \alpha\beta t^{\beta-1} e^{-\alpha t^\beta}; \quad t, \alpha, \beta > 0 \quad \text{la función del tiempo de falla de Weibull}$$

$$2. \quad Z(t) = \alpha\beta t^{\beta-1}, \quad \text{Función Razón de falla de Weibull}$$

De (1) y (2) obtenemos la función de confiabilidad de Weibull como:

$$3. \quad C(t) = \frac{f(t)}{Z(t)} = e^{-\alpha t^\beta}$$

El tiempo medio de falla del modelo de Weibull se calcula resolviendo la integral.

$$M = \alpha\beta \int_0^\infty t t^{\beta-1} e^{-\alpha t^\beta} dt, \text{ hagámos el cambio de variable}$$

$$\mu = \alpha t^\beta; \quad \alpha^{\frac{1}{\beta}} \mu^{\frac{1}{\beta}} = t; \quad \alpha^{\frac{1}{\beta}} \frac{1}{\beta} \mu^{\frac{1}{\beta}-1} d\mu = dt$$

$$M = \int_0^\infty \mu \alpha^{\frac{1}{\beta}} \frac{1}{\beta} \mu^{\frac{1}{\beta}-1} e^{-\mu} d\mu = \alpha^{\frac{1}{\beta}} \int_0^\infty \mu^{\frac{1}{\beta}} e^{-\mu} d\mu$$

$$M = \alpha^{\frac{1}{\beta}} \Gamma\left(\frac{1}{\beta}\right) \quad (4) \quad \text{TIEMPO MEDIO DE FALLA DE WEIBULL}$$

ALGUNAS PROPIEDADES DE LA FUNCIÓN

$$1. \quad \Gamma\left(\frac{1}{\beta}\right) > 0$$

$$1) \quad \Gamma\left(1 + \frac{1}{\beta}\right) = \frac{1}{\beta} \Gamma\left(\frac{1}{\beta}\right); \quad \frac{1}{\beta} > 0$$

$$2) \quad \Gamma(1) = \Gamma(2) = 1$$

$$3) \quad \text{Para } 0 < \frac{1}{\beta} < 1 \quad \Gamma\left(\frac{1}{\beta}\right) \cdot \Gamma\left(1 - \frac{1}{\beta}\right) = \frac{\pi}{\operatorname{sen} \pi \frac{1}{\beta}}$$

Como se puede observar para todas las ecuaciones que hemos obtenido es necesario determinar las constantes positivas α y β .

Consideraremos la ecuación de confiabilidad de Weibull en la forma siguiente:

$$C(t) = e^{-(\alpha t)^\beta}$$

$$\ln C(t) = -(\alpha t)^\beta$$

$$-\ln C(t) = (\alpha t)^\beta$$

$$\ln C(t)^{-1} = \beta \ln \alpha t$$

$$\ln \left(\ln \frac{1}{C(t)} \right) = \beta \ln \alpha + \beta \ln t; \text{ esto es una recta en función del ln t con pendiente } \beta$$

$$a + bt = \beta \ln \alpha + \beta \ln t;$$

$$\text{para que la igualdad se verifique tiene que cumplirse} \quad b = \beta; \quad a = \beta \ln \alpha; \quad e^{\frac{a}{b}} = \alpha$$

Tenemos que determinar los coeficientes de la recta. Puede emplearse el método de máxima verosimilitud. Emplearemos otro camino:

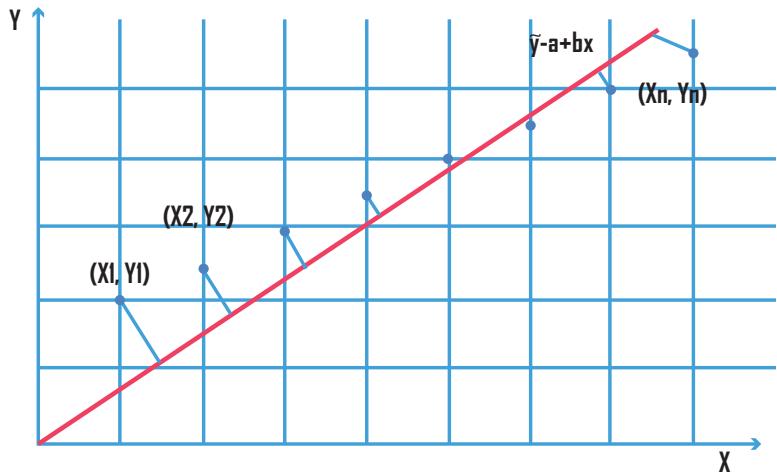
Determinar la recta de mejor ajuste. Método de Mínimos Cuadrados:

Sea dada una tabla que relaciona el tiempo de fallo de componentes respecto a los valores de humedad.

i	1	2	...	n-1	n
t _i	t ₁	t ₂	...	t _{n-1}	t _n
H _i	H ₁	H ₂	...	H _{n-1}	H _n

Determinar la distancia mínima de los puntos (x_i, y_i) a una recta $H = a + bx$. Se necesita determinar las constantes a y b .

Representamos los datos en un sistema de coordenadas X-Y.



Determinar el error mínimo (distancia mínima) de los puntos (x_i, y_i) a la recta $\tilde{y} = a + bx$

$$\text{MIN } \sum_{i=1}^n e_i = \text{MIN } \sum_{i=1}^n (y_i - (a + bx_i))^2; \quad \sum_{i=1}^n (y_i - (a + bx_i))^2 \text{ una función } f(a, b)$$

Problema: Determinar $\text{MIN } f(a, b)$

$$\text{C.N.E de MIN, las } \frac{\partial f}{\partial a} = 0, \quad \frac{\partial f}{\partial b} = 0$$

$$\begin{cases} \frac{\partial f}{\partial a} = 2 \sum_{i=1}^n (y_i - (a + bx_i)) (-1), = 0 \\ \frac{\partial f}{\partial b} = 2 \sum_{i=1}^n (y_i - (a + bx_i)) (-x_i), = 0 \end{cases}$$

133

$$\begin{cases} - \sum_{i=1}^n y_i + \sum_{i=1}^n a + \sum_{i=1}^n bx_i = 0 \\ - \sum_{i=1}^n y_i x_i + \sum_{i=1}^n ax_i + \sum_{i=1}^n bx_i^2 = 0 \end{cases}$$

$$\begin{cases} a \sum_{i=1}^n 1 + b \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \\ a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i x_i \end{cases}$$

$$\begin{cases} an + b \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \\ a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i x_i \end{cases}$$

(13) S.E.L.N.H.

Escribimos la ec. 13 en forma matricial

$$\begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n y_i x_i \end{pmatrix}$$

ec. 14

Resolvemos la ec. 14 por el método de Gauss.

$$\begin{pmatrix} n & \sum_{i=1}^n x_i & \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 & \sum_{i=1}^n y_i x_i \end{pmatrix}$$

Multiplicando la fila 1 por $\frac{\sum x_i}{n}$ y restando la fila 2, obtenemos:

$$\begin{pmatrix} n & \sum_{i=1}^n x_i & \sum_{i=1}^n y_i \\ 0 & \frac{1}{n} \sum_{i=1}^n x_i & \sum_{i=1}^n x_i - \frac{1}{n} \sum_{i=1}^n y_i x_i \end{pmatrix}$$

ec. 15

134

Escribiendo la ec. 15 en forma de ec. 13, obtenemos:

$$\begin{cases} a n + b \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \\ b \frac{\sum_{i=1}^n x_i}{n} \sum_{i=1}^n x_i - \sum_{i=1}^n x_i^2 = \frac{\sum_{i=1}^n x_i}{n} \sum_{i=1}^n y_i - \sum_{i=1}^n y_i x_i \end{cases} \quad \text{ec. 16}$$

Despejando b de la ec. 16, obtenemos:

$$b = \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i - \sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i \sum_{i=1}^n x_i - \sum_{i=1}^n x_i^2}$$

Sustituyendo b en la primera ecuación, y despejando a de la ecuación. 16, obtenemos:

$$a = \frac{1}{n} \left\{ \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \left(\frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i - \sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i \sum_{i=1}^n x_i - \sum_{i=1}^n x_i^2} \right) \right\}$$

Sustituyendo a y b en la ec. 11 y ec. 12 respectivamente, obtenemos los coeficientes .

i	x_i	y_i	$x_i y_i$	x_i^2
1	x_1	y_1	$x_1 y_1$	x_1^2
2	x_2	y_2	$x_2 y_2$	x_2^2
3	x_3	y_3	$x_3 y_3$	x_3^2
n	x_n	y_n	$x_n y_n$	x_n^2

$\sum_{i=1}^n$ Resultados

Algunos problemas que pueden ser presentados:

- Determinar el tiempo de garantía de un producto.
- Estimación de la confiabilidad de un producto.
- Pronóstico de vida de un producto después del tiempo de garantía.
- Comparar dos o más prototipos de un producto.

Anexo 7.

Hoja electrónica con la resolución de los ejercicios de Autoevaluación, disponible en:

https://docs.google.com/a/utc.edu.ec/forms/d/e/1FAIpQLSfudbk40hAPPQ5605vaLBa4gtpD_waLZifXXMvL00kvw5UMkA/viewform

o solicítelo al correo electrónico veronica.tapia@utc.edu.ec

Anexo 8.

Versión interactiva del libro disponible en:

https://docs.google.com/a/utc.edu.ec/forms/d/e/1FAIpQLSfudbk40hAPPQ5605vaLBa4gtpD_waLZifXXMvL00kvw5UMkA/viewform

o solicítelo al correo electrónico veronica.tapia@utc.edu.ec

