

ESTADÍSTICA DESCRIPTIVA

APLICADA A LAS CIENCIAS SOCIALES

NIVEL PRINCIPIANTE



compAs
Grupo de Capacitación e investigación pedagógica

JOHN HERRERA RIVERA

ESTADÍSTICA DESCRIPTIVA

APLICADA A LAS CIENCIAS SOCIALES

PRIMERA EDICIÓN



Estadística descriptiva aplicada a las ciencias sociales

Autor.

JOHN HERRERA RIVERA

Primera edición, julio 2017

Libro sometido a revisión de pares académicos.



Edición
Diagramación
Diseño
Publicación

Maquetación.

Grupo Compás

Cámara Ecuatoriana del Libro - ISBN-E: 978-9942-760-23-4

Guayaquil - Ecuador

Dedicado a:

Mi esposa e hijos por el apoyo y la comprensión demostrados durante el tiempo que duró la realización de este texto.

A mis padres por los valores morales que me inculcaron.

A 'mamá Mildred' por su ayuda invaluable y por acogerme como a un hijo.



Prólogo

Sin duda el autor presenta en este libro una forma práctica de aplicar la estadística descriptiva, aterrizando su investigación en un documento de fácil comprensión, pero con un alto impacto científico.

La aplicación de los métodos estadísticos hace que la ciencia se divida en varias alternativas para su desarrollo, en este libro el autor aplica investigación descriptiva específicamente en las ciencias sociales, lo que hace que este texto sea de actualidad y relevante ya que no muchos autores deciden trabajar sobre la investigación en el área de las ciencias sociales.

Su estructura es definitivamente una muestra clara del trabajo pensado en las necesidades de los investigadores del área, al dividir los capítulos hace que todo el documento tenga un orden lógico importante para su entendimiento.

Prefacio

En el año 2014 un video producido por el Banco Interamericano de Desarrollo (BID), en asociación con TEDEd, detalla 7 profesiones que tendrán mayor demanda en el futuro en América Latina y son:

1. Programador de Software
2. Asistente de salud
3. Ingeniero civil
4. Científico de alimentos
5. Ingeniero biomédico
- 6. *Analista estadístico***
7. Especialista en seguridad cibernética

De estas 7 profesiones podemos observar que en sexto lugar está el “Analista Estadístico”. El auge de esta profesión se debe a que, día a día se almacenan millones de datos y estos necesitan ser analizados de manera adecuada; lo cual conlleva a que las empresas necesiten de expertos que examinen los datos que tienen o deseen levantar.

En base a lo expuesto, el presente texto ha sido elaborado para que usted desarrolle la habilidad de manejar los datos de manera adecuada; desde el levantamiento, procesamiento, hasta el análisis descriptivo de la información.

Le doy la bienvenida al mundo estadístico; un mundo que espero que le agrade y le sirva para su desarrollo profesional.

Durante el estudio de este texto, usted contará con algunas definiciones y herramientas básicas para calcular y analizar datos, así como ejemplos y ejercicios para su mejor

comprensión y práctica; lo cual le permitirá diferenciarse y destacarse del resto de su competencia.

Recuerde que los datos por sí solos no representan nada; se requiere de un paso adicional para que esos datos se conviertan en información que sirva para tomar decisiones, pero hay que tener cuidado también que los datos durante el proceso de levantamiento sean realizados de manera adecuada y que verdaderamente recojan la información que se requiere. El paso adicional para convertir los datos en información es realizar su respectivo análisis usando como base la estadística, el cual nos permitirá realizar un análisis e interpretación adecuada de los mismos.

Para alcanzar el objetivo antes mencionado el presente documento ha sido dividido en los siguientes capítulos:

- I. Definiciones básicas
- II. Diseño de encuestas
- III. Muestreo
- IV. Procesamiento de datos
- V. Estadística Descriptiva
- VI. Probabilidad

Antes de iniciar este estudio, permítame sugerirle analizar e interiorizar la frase de Benjamín Franklin: “Dime y lo olvido, enséñame y lo recuerdo, involúcrame y lo aprendo”. De acuerdo a la opción que usted escoja de dicha frase, dependerá el beneficio que logre alcanzar con este texto.

Contenido General

Presentación	III
Prefacio	V
Contenido General	VII
Lista de Figuras	XI
Lista de Tablas	XIII
1. DEFINICIONES BÁSICAS	3
1.1 Introducción	3
1.2 Datos	3
1.3 Base de Datos	3
1.4 Información	4
1.5 Población Objetivo	4
1.6 Unidad de investigación	5
1.7 Censo	5
1.8 Muestra	5
1.9 Proceso Estocástico	5
1.10 Variables	6
1.11 Tipos de variables	6
1.12 Aleatoriedad	8
1.13 Variable aleatoria	8
2. DISEÑO DE ENCUESTAS	9
2.1 Introducción	11
2.2 Tipos de encuestas	12
2.3 Fases de la encuesta	13
2.4 Tipos de cuestionarios	14
2.5 Tipos de preguntas del cuestionario	15

2.6	Formato del cuestionario	17
2.7	Fiabilidad y validez del cuestionario	20
2.7.1	Fiabilidad del cuestionario	20
2.7.2	Validez del cuestionario	23
3.	MUESTREO	25
3.1	Introducción	27
3.2	Tipos de muestreo	28
3.2.1	Muestreo no probabilístico	28
3.2.2	Tipos de Muestreo no probabilístico	30
3.2.3	Muestreo probabilístico	31
3.2.4	Tipos de muestreo probabilístico	32
3.2.4.1	Muestreo aleatorio simple	32
3.2.4.2	Muestreo aleatorio Estratificado	36
3.2.4.3	Muestreo Aleatorio Sistemático	38
3.2.4.4	Muestreo por Conglomerados	38
3.2.5	Etapas del Muestreo	39
3.3	Para Ejercitar	40
4.	PROCESAMIENTO DE DATOS	45
4.1	Introducción	45
4.2	Codificación de las variables en la base de datos	47
4.3	Diseño de matriz de datos	47
5.	ESTADÍSTICA DESCRIPTIVA	51
5.1	Introducción	51
5.2	Estadística descriptiva	54
5.3	Tabla de frecuencias	56
5.3.1	Tabla de frecuencias para variables cuya escala de medición son de tipo intervalos o de razones.	57
5.3.2	Tabla de frecuencias para variables cuya	66

escala de medición es de tipo ordinal o nominal	
5.3.4 Para Ejercitar	69
5.4 Medidas De Tendencia Central	73
5.4.1 Media aritmética o simplemente media o promedio	73
5.4.2 Mediana	74
5.4.3 Moda	75
5.4.4 Para Ejercitar	76
5.5 Medidas de Variabilidad	78
5.7.1 Rango	78
5.7.2 Varianza	78
5.7.3 Desviación estándar	79
5.5.4 Para Ejercitar	79
5.6 Medidas de Posición No Centrales	80
5.7 Valores Atípicos o Valores Extremos	84
5.7.1 Para Ejercitar	86
5.8 Gráficos	88
5.8.1 Histogramas	88
5.8.2 Ojiva	89
5.8.3 Diagrama de cajas	90
6. PROBABILIDAD	93
6.1 Introducción	95
6.2 Probabilidad	95
6.3 Reglas de Probabilidad	97
6.3.1 Para Ejercitar	102
7. Usando Software para realizar Estadística Descriptiva	105
7.1 Introducción	107
7.2 Estadística descriptiva con SPSS	107
7.2.1 ¿Qué es SPSS?	107

7.2.2 Importar datos desde EXCEL	107
7.2.3 Realizando estadística descriptiva	113
7.3 Estadística descriptiva con RSTUDIO	119
7.3.1 ¿Qué es R?	119
7.3.2 ¿Qué es RStudio?	119
7.3.3 Instalando RSTUDIO	119
7.3.4 Importar datos desde EXCEL	124
7.3.5 Realizando estadística descriptiva	126
7.3.6 Histograma y Diagrama de cajas	131
7.3.7 Ingresando datos en RSTUDIO	133
7.4 Estadística descriptiva en EXCEL	134

Lista de figuras

1-1 ESCALAS DE MEDICIÓN	7
2-1 PASOS A SEGUIR DE UNA ENCUESTA	11
2-2 OBJETIVOS DE INVESTIGACIÓN	12
2-3 FASES DE LA ENCUESTA	13
2-4 MÓDULO DE UNA ENCUESTA	19
2-5 ENCUESTA USO DEL TIEMPO	19
3-1 TIPOS DE MUESTREO	28
4-1 HOJA ELECTRÓNICA	49
5-1 ÁREAS DE LAS ESTADÍSTICAS	53
5-2 COMPONENTES DE LA ESTADÍSTICA DESCRIPTIVA	55
5-3 TIPOS DE MODA	75
5-4 TIPOS DE MEDIDAS DE POSICIÓN NO CENTRAL	80
5-5 HISTOGRAMA DE NOTAS	88
5-6 OJIVA DE LAS NOTAS	89
5-7 EJEMPLO DE OJIVAS DE LAS NOTAS	90
5-8 EJEMPLO DE DIAGRAMA DE CAJAS	91
7-1 ABRIR ARCHIVO DESDE SPSS	109
7-2 ABRIR ARCHIVO TIPO EXCEL	109
7-3 SELECCIÓN DE HOJA DE TRABAJO	110
7-4 VENTANA DE INGRESO DE VARIABLES	110
7-5 TIPO DE VARIABLE	111
7-6 ETIQUETAS DE VALOR	112
7-7 VALORES PERDIDOS	112
7-8 MEDIDA	113
7-9 VISTA DE DATOS	113
7-10 ESTADÍSTICOS DESCRIPTIVOS	113
7-11 SELECCIÓN DE VARIABLES	114
7-12 OPCIONES DE ESTADÍSTICOS DESCRIPTIVOS	114
7-13 RESULTADOS	115
7-14 FRECUENCIAS	115
7-15 SELECCIÓN DE VARIABLES	116
7-16 SELECCIÓN DE ESTADÍSTICOS	116

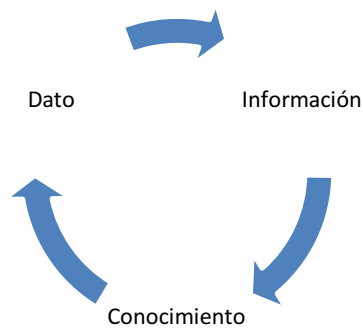
7-17 RESULTADOS ESTADÍSTICOS	117
7-18 TABLAS CRUZADAS	117
7-19 SELECCIÓN DE VARIABLES	118
7-20 RESULTADOS DE TABLAS CRUZADAS	118
7-21 PÁGINA WEB R	120
7-22 SELECCIÓN DE PAÍS	120
7-23 DESCARGAR R	120
7-24 INSTALAR R	121
7-25 DESCARGANDO R PARA WINDOWS	121
7-26 SELECCIONANDO IDIOMA	121
7-27 INSTALANDO R	122
7-28 INSTALACIÓN COMPLETA	122
7-29 ICONO DE R	122
7-30 PANELES RSTUDIO	123
7-31 GUARDANDO EXCEL COMO TIPO CSV	124
7-32 BUSCANDO ARCHIVO EXCEL	124
7-33 IMPORTANDO DESDE EXCEL	125
7-34 DATOS IMPORTADOS EN RSTUDIO	125
7-35 HISTOGRAMA	131
7-36 NOMBRE DE LOS ELEMENTOS DEL HISTOGRAMA	132
7-37 DIAGRAMA DE CAJAS	133
7-38 OPCIONES DE EXCEL	134
7-39 COMPLEMENTO DE EXCEL	135
7-40 HERRAMIENTA DE ANÁLISIS DE EXCEL	135
7-41 ANÁLISIS DE DATOS EN EXCEL	136
7-42 ESTADÍSTICA DESCRIPTIVA EN EXCEL	136
7-43 OPCIONES DE ESTADÍSTICA DESCRIPTIVA	137
7-44 RESULTADO ESTADÍSTICO EN EXCEL	138

Lista de Tablas

1-1 CLASIFICACIÓN DE LAS VARIABLES	7
2-1 TIPOS DE PREGUNTAS Y SUS CARACTERÍSTICAS	16
2-2 FORMATO DE CUESTIONARIO	18
2-3 ALFA DE CRONBACH	21
2-4 ALFA DE CRONBACH POR PREGUNTA	22
2-5 PRUEBA DE KMO Y BARTLETT	24
3-1 TIPOS DE MUESTREO NO PROBABILISTICO	31
3-2 ETAPAS DE MUESTREO	39
4-1 ESQUEMA DE UNA BASE DE DATOS	48
5-1 MÉTODO ESTADÍSTICO SEGÚN TIPO DE VARIABLE	56
5-2 MODELO DE LA TABLA DE FRECUENCIAS	57
5-3 NOTA DE ALUMNOS	61
5-4 NOTA DE ALUMNOS ORDENADOS	61
5-5 TABLA DE FRECUENCIAS	65
5-6 FORMATO DE TABLA DE FRECUENCIAS PARA DATOS ORDINALES O NOMINALES	67
5-7 ESTADO CIVIL DE LAS PERSONAS A NIVEL NACIONAL	68
5-10 CALIFICACIONES DE ALUMNOS	74

CAPÍTULO UNO

“La inteligencia consiste no sólo en el conocimiento, sino también en la destreza de aplicar los conocimientos en la práctica”
Aristóteles 384 AC-322AC





1. DEFINICIONES BÁSICAS

1.1 Introducción

Dentro de los procesos de investigación se resalta la observación de un fenómeno y la recolección de datos (Singh, 2006); pero no solamente es suficiente con capturar los datos sino que hay que procesarlos e interpretarlos (Durán, 2015), transformando los datos en información vital para la toma de decisiones.

Es importante resaltar que “no existe al momento, actividad constructiva que no sea la menos transversalmente afectada por la estadística” (Zurita, 2010).

El proceso investigativo en la mayoría de los casos requiere de la estadística; ya sea en la etapa de diagnóstico como en la etapa de comprobación de hipótesis; pero antes de iniciar nuestro aprendizaje es necesario conocer algunas definiciones básicas, las cuales serán utilizadas a lo largo de este texto.

1.2 Datos

“Es el conjunto de mediciones no procesadas, sean estas; numéricas (cuantitativas) o categóricas (cualitativas)” (Zurita, 2010); que han sido recolectadas a través de algún medio físico o electrónico.

1.3 Base de Datos

“Es un conjunto de datos relacionados entre sí, organizados y estructurados con información referente a algo” (Ramos, 2008). Por ejemplo se realizan encuestas a los habitantes de una

localidad y se levantan los siguientes datos: Edad, sexo, nacionalidad y nivel de educación. Estos datos levantados en la comunidad formarían una base de datos.

Cabe indicar que el conjunto de datos a investigar depende de la pregunta de investigación, puede que sea un conjunto de las personas con una enfermedad en particular, un conjunto de automóviles producidos por un fabricante en particular, una serie de visitas a un sitio web o un conjunto de transacciones con tarjetas de crédito, etc.

1.4 Información

“Es un conjunto de datos procesados que nos permite tomar decisiones racionales” (Zurita, 2010).

1.5 Población Objetivo

En estadística Población Objetivo hace referencia al conjunto de todos los elementos considerados en una investigación, de los cuales nos interesa obtener información o investigar algunas características. Cabe indicar que algunas personas confunden la población objetivo con el Universo. Un ejemplo de población objetivo lo encontramos en la “Encuesta Nacional sobre Relaciones Familiares y Violencia de Género contra las Mujeres”, realizada por el Instituto Nacional de Estadísticas y Censo cuyo objetivo es generar datos y producir información estadística sobre la magnitud de los distintos tipos de violencia de género (física, psicológica, sexual) que sufren o han sufrido las mujeres, define como población objetivo a todas las mujeres que tengan 15 años de edad o más, sin importar su estado civil o si han tenido o no una relación de

pareja, que residan habitualmente en las viviendas seleccionadas en la muestra.

1.6

Son los elementos de la población objetivo que se va a investigar y pueden ser de carácter social, económico, agropecuario, etc. Ej.: La Unidad de investigación de la “Encuesta Nacional sobre Relaciones Familiares y Violencia de Género”, son las viviendas seleccionadas y sus hogares.

1.7 Censo

Censo es la obtención de información específica de todas las unidades de información que componen una población o universo; o un inventario de las características que posee cada una de las unidades de información en un momento determinado.

1.8 Muestra

A la parte de la población, constituida por las unidades de información que realmente se investigan, se denomina muestra; la misma que debe ser representativa. Los elementos de una muestra deben ser seleccionados científicamente; para esto, es necesario construir muestras probabilísticas, utilizando técnicas basadas en la teoría de probabilidades.

1.9 Proceso Estocástico

Se denomina a todo proceso cuyo comportamiento no evidencia acciones predecibles del mismo, es decir, no se puede asegurar al 100% que ocurrirá un proceso.

1.10 Variables

Las variables son medidas o características de un elemento que pueden tomar varios valores, a diferencia de las constantes que son de un solo valor; en otras palabras, la razón por la que se denominan variables se debe a que las medidas o características investigadas varían de uno a otro elemento. Por ejemplo: Una variable sería la temperatura tomada en un determinado día y en donde se podrá observar que esta variable puede tener varios valores durante el transcurso de las horas en un determinado día.

1.11 Tipos de variables

Considere que dependiendo del problema a investigar, las variables pueden ser:

Según	Clasificación
Naturaleza	Cuantitativas: discretas; continuas
	Cualitativas: nominales; ordinales
Amplitud	Macrológicas, micrológicas
	Individuales, colectivas
Nivel de abstracción	Generales
	Intermedias
	Empíricas o indicadores
Escalas de medición	Nominales (Cualitativa)

Posición	Ordinales (Cualitativa)
	Intervalo (Cuantitativa)
	Razón (Cuantitativa)
	Independientes
	Dependientes
	Extraña o interviniente

TABLA 1-1 CLASIFICACIÓN DE LAS VARIABLES

Fuente: Elaboración propia

Tenga en consideración que los métodos estadísticos a emplearse dependerán de la escala de medición de los datos; los cuales se detallan en el siguiente gráfico:

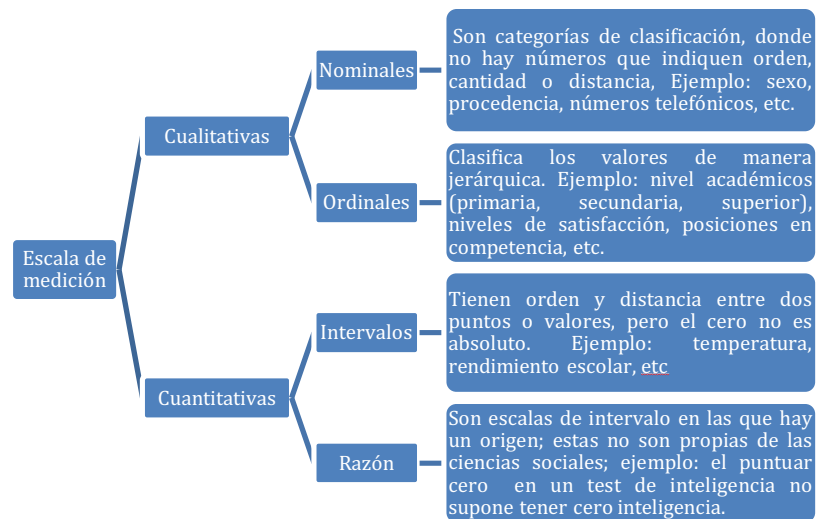


FIGURA 1-1 ESCALAS DE MEDICIÓN

Fuente: Elaboración propia

1.12 Aleatoriedad

La aleatoriedad es todo proceso cuyo resultado no es previsible, depende de un suceso fortuito, es decir no se puede determinar en ningún caso antes de que este se produzca. Por consiguiente, los procesos aleatorios son calculados a través de la probabilidad de ocurrencia. Por ejemplo el clima en un determinado día es aleatorio por cuanto un día puede ser soleado pero no se sabe si el siguiente día también lo será.

1.13 Variable aleatoria

Una variable aleatoria es una función que asigna eventos a números reales. Por ejemplo, los posibles resultados de lanzar un dado dos veces.

CAPÍTULO DOS DISEÑO DE ENCUESTAS

“Si los estadísticos teóricos hacen caso omiso al reto de enfrentar las encuestas multipropósito, entonces el vacío existente entre ellos y los estadísticos prácticos se hará cada vez más grande. El diseño y análisis de encuestas multivariantes debe ser una de las próximas áreas de mayor investigación.”

T. M. F. Smith (1976)





2. DISEÑO DE ENCUESTAS

2.1 Introducción

Para empezar este capítulo es necesario tener claro la diferencia entre encuesta y cuestionario.

Según (Grande, 2005), la Encuesta se puede definir como una técnica primaria de obtención de información sobre la base de un conjunto objetivo, coherente y articulado de preguntas, para ser analizadas. Para el Centro de Investigaciones Sociológicas de España, la encuesta “es una técnica de recogida de datos mediante la aplicación de un cuestionario a una muestra de individuos”. Según las Naciones Unidas, el Cuestionario es el instrumento de recopilación de datos para obtener información.

En resumen, la Encuesta es la técnica para obtener información, mientras que el cuestionario es el instrumento.

Antes de realizar una encuesta es necesario seguir ciertos pasos (ver Figura 2-1):

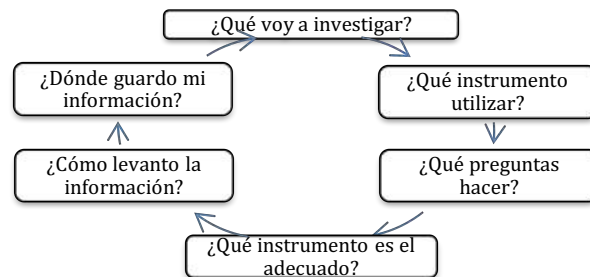


FIGURA 2-1 PASOS A SEGUIR DE UNA ENCUESTA

Fuente: Elaboración propia

La figura 2-1 empieza con la pregunta ¿qué voy a investigar?, la cual debe ser respondida con claridad antes de seleccionar el tipo de encuesta y el instrumento adecuado; en otras palabras se debe precisar el objetivo de la investigación y establecer si esta va a ser descriptiva, correlacional, etc. (ver Figura 2-2).

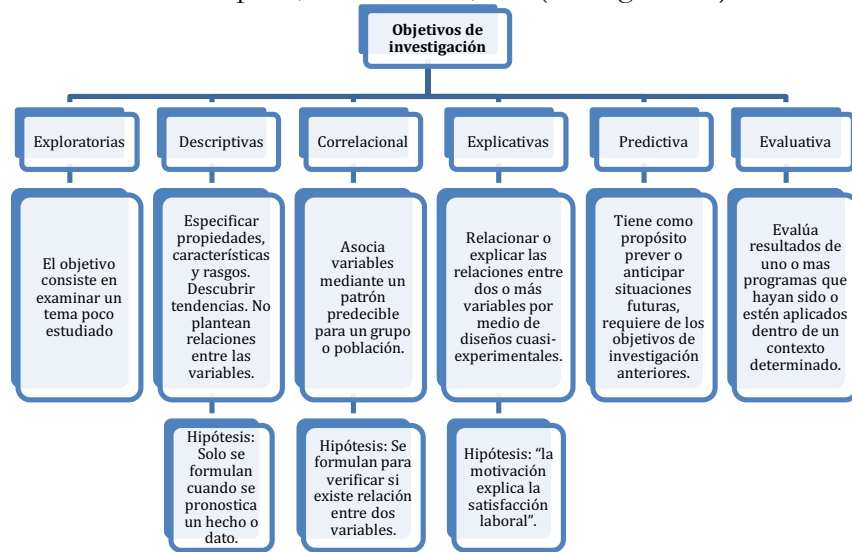


FIGURA 2-2 OBJETIVOS DE INVESTIGACIÓN

Fuente: Elaboración propia

2.2 Tipos de encuestas

Existen dos tipos de encuestas:

- Directa: Cara a cara.
- Indirecta: Correo, teléfono, redes sociales.



2.3 Fases de la encuesta

La encuesta por lo general tiene las siguientes fases:

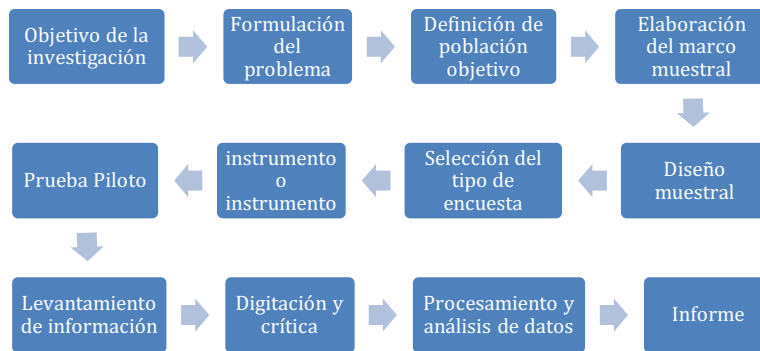


FIGURA 2-3 FASES DE LA ENCUESTA

Fuente: Elaboración propia

Teniendo claro el objetivo de la investigación se podrá formular el problema a investigar y posteriormente definir la población objetivo; esto permitirá delimitar el marco muestral y el diseño muestral de la investigación. Con lo antes planteado, se logrará determinar el tipo de encuesta a utilizar.

Tenga en consideración que previo a realizar el levantamiento de información se deberá efectuar una prueba piloto, lo cual permitirá mejorar el o los instrumentos; así como el tamaño de la muestra. Adicionalmente antes de desarrollar la prueba piloto se deberá hacer la planificación de la encuesta y considerar el tema de personal, movilización, recorrido de sector, etc. Además, se deberá obtener el cálculo del presupuesto de la encuesta, que dependerá del tamaño de la muestra, de la extensión del instrumento de la encuesta, encuestadores, implementos para la encuesta, etc.

Durante el proceso de levantamiento de información y digitación, se debe realizar la supervisión y crítica de la información levantada, así como, el control de calidad. Una vez terminada la digitación se procederá a la validación de los datos, previo al procesamiento y análisis de los mismos, para finalizar con el informe respectivo.

2.4 Tipos de cuestionarios

Existen tres tipos de cuestionarios:

- a) **Cuestionarios no estructurados:** Son utilizados cuando los investigadores tienen ligeros o casi nulos conocimientos de lo que van a investigar. La información que se levante puede servir para la elaboración de un cuestionario más estructurado con preguntas cerradas. Se utilizan por lo general en entrevistas a profundidad y grupos focales. Hay que considerar que las preguntas, por lo general, son abiertas y pueden aplicar ideas generadoras para investigar. Son utilizados en investigaciones exploratorias.
- b) **Cuestionarios semiestructurados:** Para el uso de este tipo de cuestionarios, el investigador deberá conocer algo del tema que va a investigar y tener una idea de la variedad de posibles repuestas de las variables; por ende, podrá contar con preguntas cerradas con la posibilidad de añadir otra respuesta diferente a lo indicado. Por lo general, son empleados en entrevistas personales como también en investigaciones exploratorias.

- c) **Cuestionarios estructurados:** Para el uso de este tipo de cuestionarios, el investigador deberá conocer el tema a investigar a través de investigaciones previas o similares; para lo cual, se utilizarán preguntas cerradas o escalas. Se utilizan en investigaciones descriptivas, correlacionales o explicativas.

De los 3 tipos de cuestionarios mencionados, el estructurado es fácil de responder, administrar y tabular; pero requiere de varias pruebas para reducir posibles errores.

2.5 Tipos de preguntas del cuestionario

En la tabla 2-1 se presentan los tipos de preguntas y sus características:

Tipos	Características
Abiertas	Recoge gran cantidad de información, que puede ser muy importante, si se registra textualmente como indica el encuestado y si el proceso de levantamiento ha sido adecuado; pero es muy difícil tabular y analizar las respuestas y requiere de mucho tiempo. Ejemplo: ¿Qué factor cree usted que incide en el aumento de la delincuencia en Guayaquil?
De Profundización	Son preguntas por lo general abiertas que buscan conseguir más información sobre una variable.

	Ejemplo: ¿Añadiría algo más a su respuesta?
De clarificación	Son preguntas que buscan explicar o aclarar las respuestas dadas con anterioridad. Ejemplo: ¿Puede explicarme su respuesta?
Cerradas dicotómicas	Son fáciles de preguntar y responder; pero dependiendo de la investigación, las respuestas pueden resultar incompletas. Ejemplo: si/no, falso/verdadero
Cerradas de respuestas múltiples	También son fáciles de preguntar y tabular; se debe asegurar de incluir todas las respuestas posibles. Ejemplo: Usted cree que el aumento de la delincuencia en la ciudad se deba a: a. Falta de empleo b. Falta de valores de la sociedad...

TABLA 2-1 TIPOS DE PREGUNTAS Y SUS CARACTERÍSTICAS

Fuente: Elaboración propia

Adicionalmente tenga en consideración los siguientes puntos:

- La estructuración y diseño del cuestionario depende en gran medida del tipo de encuesta y del público objetivo.
- El cuestionario debe ser codificado para su fácil ubicación en caso de error en la hoja electrónica.
- El cuestionario puede tener varios tipos de preguntas y varios tipos de variables, según su escala de medición, pero se sugiere que todas las variables sean del mismo tipo de escala.
- Las preguntas de calificación deben tener alguna escala.

- Las preguntas que requieren opinión, por lo general deben ser abiertas.
- El orden en el que se hacen las preguntas es importante; por lo general se debe comenzar con las preguntas más relevantes.
- Las preguntas deben ser sencillas y directas, con un lenguaje cotidiano y sin complejidad.
- Las preguntas no deben sugerir respuestas.
- De preferencia se debe agrupar las preguntas que tengan temas similares.
- Las preguntas deben contener una sola idea.
- Las preguntas y sus respuestas - de darse el caso - deben ser codificadas.

2.6 Formato del cuestionario

El cuestionario por lo general puede tener el siguiente formato:

Ítem	Descripción
Título	Nombre del cuestionario.
Autores	Nombre de investigadores o institución que realiza la investigación.
Presentación	Breve presentación sobre el objetivo del cuestionario, la cual debe ser insertada en el cuestionario, en caso de que lo vaya a llenar directamente el encuestado; pero si lo va realizar un encuestador, queda a decisión del investigador.
Ejemplo	Si el cuestionario lo llena el encuestado se debe poner un ejemplo de cómo llenarlo.
Agradecimiento	Si lo llena el encuestado, se debe poner el agradecimiento respectivo.

Preguntas y respuestas	Este apartado está estructurado por: a. Característica de la muestra: • Datos descriptivos de la encuesta: Ubicación geográfica de vivienda seleccionada, datos del hogar, etc. • Datos de identificación (principio o final): nombre del encuestado, género, edad, estado civil, etc. b. Objetivos: • Preguntas para despertar el interés del encuestado • Preguntas claves • Preguntas comprometedoras c. Ítems de control: personal responsable, fecha, hora, duración, observaciones, resultados de la encuesta, codificación de formulario, etc.
------------------------	--

TABLA 2-2 FORMATO DE CUESTIONARIO

Fuente: Elaboración propia

Se requiere tener en cuenta, los saltos de página o preguntas; y el manejo de bloques de preguntas. Recuerde que todo cuestionario debe tener su Instructivo de llenado.

Es necesario considerar, que las preguntas o alternativas planteadas deben ser codificadas, con la finalidad de que al digitar la información, esta se realice con facilidad y se reduzcan los errores de digitación.

A continuación se muestran unos ejemplos de cuestionarios:

FIGURA 2-4 MÓDULO DE UNA ENCUESTA
Fuente: Instituto Nacional de Estadística y Censo INEC.

FIGURA 2-5 ENCUESTA USO DEL TIEMPO
Fuente: Instituto Nacional de Estadística y Censo INEC.

2.7 Fiabilidad y validez del cuestionario

Una vez elaborado el cuestionario, es necesario comprobar la fiabilidad y validez del mismo.

2.7.1 Fiabilidad del cuestionario

La consistencia o fiabilidad del cuestionario representa la precisión de lo que se desea medir. Se considera fiable, cuando las medidas que se hacen con el cuestionario carecen de errores, son consistentes, confiables y precisas. Por ejemplo, un cuestionario será fiable si cada vez que se aplica a los mismos sujetos obtenemos los mismos resultados.

A continuación detallamos los métodos más conocidos para estimar la fiabilidad o consistencia de los cuestionarios:

- **Método de las dos mitades:** El cuestionario se divide en dos partes iguales, normalmente se toman los ítems pares por una parte y los impares por otra. Se correlacionan las puntuaciones obtenidas en ambas partes y el coeficiente resultante se corrige para una longitud doble (Confiabilidad de Pearson/Spearman-Brown, Confiabilidad Rulón, Confiabilidad Guttman).
- **Método test y re-test:** Se aplica el cuestionario a los mismos sujetos en dos ocasiones distintas y se correlacionan los resultados. Se debe tener en cuenta que el espacio de tiempo transcurrido, no sea ni muy corto ni muy largo (Correlación canónica).

- **Formas equivalentes:** Se aplican a los mismos sujetos, en dos ocasiones distintas, dos versiones paralelas del cuestionario; que contenga diferentes ítems pero tan equivalentes, como sea posible, en formato y dificultad, y que se correlacionen los resultados (Correlación canónica).
- **Consistencia interna:** Se estima la fiabilidad de un instrumento administrado, por única vez, a partir de algunos coeficientes. Entre los más empleados están: alfa de Cronbach, el ajuste por split-half usando las fórmulas de Spearman-Brown y la fiabilidad estimada a partir del patrón de respuesta propuesta por Kuder-Richardson (K-R20 y K-R21).

Un ejemplo de consistencia interna se presenta en la tabla 2-3, en donde el alfa de Cronbach es 0,821; esto indica que el cuestionario es fiable. Cabe resaltar que, algunos investigadores mencionan que si el valor del alfa de Cronbach es mayor a 0,7, el instrumento es fiable.

Alfa de Cronbach	Alfa de Cronbach basada en elementos estandarizados	N de elementos
,814	,821	13

TABLA 2-3 ALFA DE CRONBACH
Fuente: Elaboración propia usando Software SPSS

Una de las ventajas del alfa de Cronbach es que también se puede generar para cada pregunta, como observa en la tabla 2-4, en donde cada pregunta tiene calculado el valor del alfa en caso de que esta se eliminara, permitiendo medir cómo impactaría el eliminar una pregunta del cuestionario; como en

el caso de la pregunta 10 (P10); si se eliminara esta pregunta, el alfa de Cronbach aumentaría a 0,826, aunque este aumento no fuera tan significativo, el cuestionario seguiría siendo fiable.

	Media de escala si el elemento se ha suprimido	Varianza de escala si el elemento se ha suprimido	Correlación total de elementos corregida	Correlación múltiple al cuadrado	Alfa de Cronbach si el elemento se ha suprimido
P1	20,69	23,390	,624	,627	,787
P2	21,05	23,851	,497	,508	,798
P3	21,02	24,756	,448	,505	,802
P4	21,05	23,559	,597	,672	,790
P5	21,29	24,453	,664	,674	,790
P6	21,14	24,369	,509	,605	,797
P7	19,95	25,022	,388	,272	,807
P8	20,95	23,705	,620	,553	,789
P9	20,36	24,040	,422	,490	,805
P10	21,12	26,839	,132	,434	,826
P11	20,93	24,946	,403	,407	,805
P12	20,98	25,195	,340	,364	,811
P13	20,33	25,301	,330	,437	,812

TABLA 2-4 ALFA DE CRONBACH POR PREGUNTA

Fuente: Elaboración propia usando Software SPSS

2.7.2 Validez del cuestionario

Mientras la fiabilidad mide la precisión del cuestionario, la validez mide el grado en el que el cuestionario mide lo que se quiere medir. El poder predictivo de esta validez es el coeficiente de correlación obtenido entre las puntuaciones directas del cuestionario y las puntuaciones directas en el criterio.

A continuación detallamos los diversos tipos de validez:

- **Validez lógica:** Es el grado en que ‘parece’ que un cuestionario mide lo que quiere medir.
- **Validez de contenido:** Alude a la necesidad de garantizar que el cuestionario constituye una muestra adecuada y representativa de las variables que se pretenden evaluar con él.
- **Validez consensuada.** Existe un acuerdo entre expertos de que un determinado cuestionario es válido y mide lo que se pretende medir.
- **Validez de criterio.** Existe correlación entre las puntuaciones directas del cuestionario y las puntuaciones directas del criterio (índices de sensibilidad y especificidad).
- **Validez interna:** Los criterios utilizados habitualmente son la validez convergente y la validez divergente o discriminatoria (Peter, 1981). La validez convergente se refiere a que la escala de medición de cada concepto converja en la representación del constructo, que se

operacionaliza a partir de la relación entre la medida y otras medidas que evalúan el mismo constructo. La validez discriminatoria se refiere a que los conceptos que intervienen en la escala de medición no representen una única dimensión. Se emplean análisis factorial, relación entre los resultados obtenidos por la medida y otros constructos relacionados (KMO y prueba de Bartlett).

En la siguiente tabla se muestran los resultados de emplear análisis factorial a través del Test de Esfericidad de Bartlett para medir la relación entre las variables de un cuestionario. Si el valor de significancia de la tabla es menor al 5% afirmamos que los datos son válidos. En la investigación el nivel de significancia (0.000) permite afirmar los datos como válidos. Adicionalmente el valor de Índice Kaiser-Meyer-Olkin (0.718) afirma que es apropiado aplicar el modelo de análisis factorial a los datos.

Medida Kaiser-Meyer-Olkin de adecuación de muestreo		,718
Prueba de esfericidad de Bartlett	Aprox. Chi-cuadrado	197,653
	gl	78
	Sig.	,000

TABLA 2-5 PRUEBA DE KMO Y BARTLETT

Fuente: Elaboración propia usando Software SPSS

CAPÍTULO TRES

MUESTREO

"El muestreo no es la simple sustitución de una cobertura parcial para una cobertura total. Muestreo es la ciencia y el arte de controlar y medir la fiabilidad de la información estadística útil a través de la teoría de la probabilidad"

Deming (1950)





3. MUESTREO

3.1 Introducción

Durante el proceso de investigación se necesita contar con información de la población investigada y la mejor forma es realizar un Censo de la población; pero para realizarlo se requieren elementos tales como: costos altos, personal, tiempo, entre otros; y los mismos en la mayoría de los casos no se poseen. Si se realiza una investigación en la población de una provincia en la que se remite al Censo Poblacional realizado en el 2010, surge la duda en que si esos datos reflejan la realidad actual (2017), pues dependiendo de la dinámica de crecimiento poblacional y del sector a investigar ésta no puede ser tan fiable y el censo no tendrá muchas veces las variables que se necesitan para una investigación.

En base a lo antes expuesto, surge la interrogante de cómo realizar un levantamiento de información que no sea costoso, que sea representativo a la población sujeta de investigación; es aquí donde nace la necesidad de utilizar una herramienta muy utilizada en la investigación científica que se denomina muestreo.

El muestreo es una herramienta cuya función básica es seleccionar una parte de la población a ser investigada y que sea representativa para la población con la finalidad de realizar, una vez procesada la información, inferencias sobre dicha población.

3.2 Tipos de muestreo

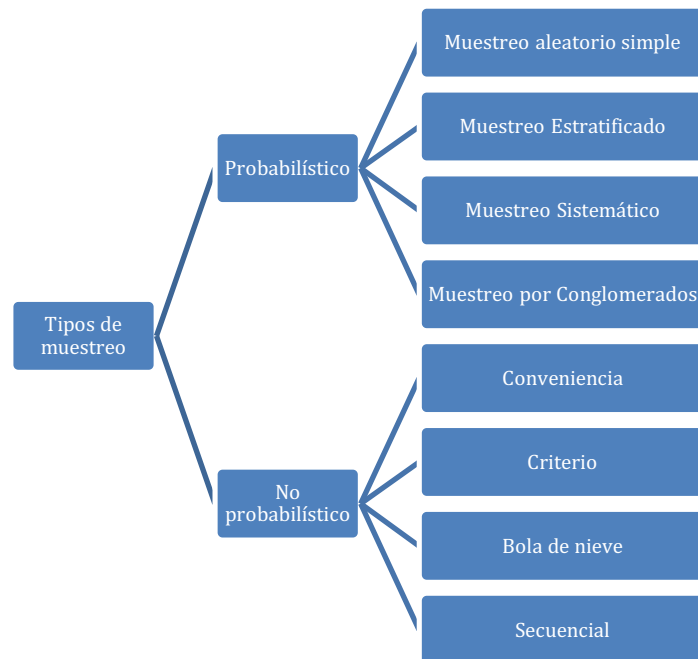


FIGURA 3-1 TIPOS DE MUESTREO

Fuente: Elaboración Propia

3.2.1 Muestreo no probabilístico

Las unidades muestrales de la investigación no son seleccionadas de manera aleatoria o al azar, sino que son elegidas por el investigador en base al criterio, juicio, prejuicio, experiencia, intereses, entre otros criterios. Este tipo de muestreo no es sujeto de calcular la confiabilidad del muestreo ni su margen de error por carecer de un marco metodológico que lo permita. Cabe resaltar, que al carecer de estos

parámetros de confiabilidad y error del diseño muestral, este tipo de muestreo no permitirá realizar la inferencia estadística a la población objeto del estudio, sino que el análisis deberá ser sujeto solamente a las unidades investigadas.

Ejemplo: Una emisora radial de Guayaquil, realiza una encuesta de opinión sobre la aprobación de la ley del Impuesto a la Herencia y la pregunta solo indica si está a favor o no. Se pide al oyente que llame y de su opinión; al final del proceso se dan los resultados y se indica que el 80% de los Guayaquileños no están a favor de la aprobación de dicha ley. Analizando la metodología utilizada por la emisora radial, se puede observar que no es un tipo de muestreo aleatorio y por ende, no se puede lograr la inferencia estadística hacia toda la ciudad; debido a que los que llamaron no son seleccionados al azar sino que son personas que escuchan la emisora (aquí se tiene un sesgo). De los que escuchan la emisora, no todos tienen un teléfono a la mano o el tiempo necesario para llamar (otro sesgo); entonces no se puede afirmar que esa es la opinión de los ciudadanos de Guayaquil.

Otro punto importante, es que el 80% correspondiente a 1 resultado de este ejemplo no indica cuántos oyentes votaron en su contra ya que ese 80% puede representar únicamente que 16 de 20 personas votaron en contra o 8 de 10 o 4 de 5 oyentes. Muchas veces el manejar porcentajes sin un adecuado argumento de los valores absolutos puede causar que se tome una mala decisión.

3.2.2 Tipos de Muestreo no probabilístico

Tipo	Característica	Ejemplos	Utilidad
Conveniencia: Las muestras se seleccionan según un criterio de accesibilidad o comodidad.	Comodidad en la selección de la muestra. Rapidez. Economía.	Encuesta a estudiantes en clase Encuestas en establecimientos comerciales.	- Diseño inicial de investigaciones. - Investigaciones exploratorias. - Formulación de hipótesis. - Validación de escalas. - Cálculos de varianza para afijación en el muestreo aleatorio.
Criterio: La muestra es elegida por un experto de acuerdo a su criterio, buscando las unidades más representativas.	Rapidez. Economía.	Seleccionar ciudades para hacer algún test de mercado o de producto Seleccionar una muestra de personas en una empresa.	- Se emplea cuando el tamaño de la muestra es pequeño.
Bola de nieve: Cada Unidad muestral es localizada por indicación de otra persona.	Comodidad en selección de la muestra No tiene por qué ser rápido No tiene por qué ser cómodo.	Identificar una muestra de dueños de perros pequineses. Identificar una muestra de aficionados al parapente.	- Se emplea cuando se trata de estudiar poblaciones pequeñas muy especializadas que son difíciles de localizar por no existir censos o ser inaccesibles.

Secuencial: La muestra inicial, reducida se incrementa hasta alcanzar las cuotas deseadas.	No tiene por qué ser rápido No tiene por qué ser cómodo.	Estudios de comportamiento del consumidor en general.	- Se aplica en investigaciones en su fase inicial. - Posteriormente se acude a muestreo aleatorio.
---	---	---	---

TABLA 3-1 TIPOS DE MUESTREO NO PROBABILISTICO

Fuente: Elaboración Propia

3.2.3 Muestreo probabilístico

El muestreo probabilístico tiene como base fundamental, que todos los individuos de la población objeto de nuestra investigación puedan formar parte de la muestra; es decir, cada uno tiene la misma probabilidad de formar parte de la muestra; y, la selección de los mismos debe ser de manera aleatoria. Para la mayoría de investigadores, este es el tipo de muestreo a utilizar en las investigaciones por ser riguroso y científico. Sin embargo, dependiendo de la investigación, del presupuesto y de la población a ser investigada, se deberá seleccionar el modelo de muestreo adecuado -ya sea probabilístico o no probabilístico- y se debe tener en consideración, que si se realiza adecuadamente la metodología de muestreo, levantamiento y procesamiento de información en el muestreo probabilístico, se podrá realizar la inferencia a toda la población; pero si realiza el muestreo no probabilístico solo podrá indicar que los resultados se sujetan únicamente a la individuos encuestados.

3.2.4 Tipos de muestreo probabilístico

3.2.4.1 Muestreo aleatorio simple

Para algunos investigadores este tipo de muestreo es el más fácil de realizar, ya que todos los individuos tienen la misma probabilidad de ser seleccionados. Sin embargo, el proceso de la selección del marco muestral, de la selección de los sujetos de investigación no necesariamente es fácil. La selección de los individuos que forman parte de la muestra puede realizarse a través de cualquier mecanismo que permita tener la misma probabilidad de que todos los individuos tengan las mismas opciones de ser seleccionados. Para el efecto, se utilizan tablas de números aleatorios. Microsoft Excel permite generar números aleatorios.

Para calcular el tamaño de la muestra hay que tomar en cuenta tres factores:

- La **confianza o el porcentaje de confianza** es el porcentaje de seguridad que existe para generalizar los resultados obtenidos. Esto quiere decir que un porcentaje del 100% equivale a decir que no existe ninguna duda para generalizar tales resultados, pero también implica estudiar la totalidad de los casos de la población. Comúnmente en las investigaciones sociales se busca entre el 90% y 95%.
- El **error o porcentaje de error** equivale a elegir una probabilidad de aceptar una hipótesis que sea falsa como si fuera verdadera, o la inversa: rechazar una hipótesis verdadera por considerarla falsa.

Comúnmente se aceptan entre el 4% y el 6% como error, tomando en cuenta de que no son complementarios la confianza y el error.

- La **variabilidad** es la probabilidad (o porcentaje) con la que se aceptó y se rechazó la hipótesis que se quiere investigar; tomada de alguna investigación anterior o de un ensayo previo a la investigación actual.

Cálculo de tamaño de la muestra si la variable a investigar es de carácter métrico

Para calcular el tamaño de la muestra debe considerar el tipo de variable principal de la investigación. Si el tipo de variable es métrica y conoce la varianza de la variable de la población a investigar, puede usar la siguiente fórmula:

$$n = \frac{N * Z_{\alpha/2}^2 * S^2}{(e^2 * N) + (Z_{\alpha/2}^2 * S^2)}$$

En donde:

- n: Es el tamaño de la muestra
- N: Es el tamaño de la población
- $Z_{\alpha/2}$: Coeficiente de confianza (este coeficiente es 1.96 si se utiliza el 95% de confianza)
- e: Margen de error seleccionado (puede ser 5%, 10%)
- S^2 : Varianza de la población (de la variable principal)

Si la población es grande (algunos sugieren, mayor a 100.000 unidades) la fórmula de muestreo se reduce a:

$$n = \frac{Z_{\alpha/2}^2 * S^2}{e^2}$$

Considere que durante la prueba piloto, se puede estimar la varianza de la variable y ser usada en la fórmula anterior.

Ejemplo: Se desea calcular el tamaño muestral para realizar un estudio a los habitantes de la Parroquia Rural El Morro, con un margen de error del 5% y un nivel de confianza del 95% y se conoce la varianza de la variable a investigar.

Entonces:

n: ¿?
 N: 5.019 habitantes según el Censo 2010
 $Z_{\alpha/2}$: 1,96 (95% de confianza)
 e: 0.05 (5% de error)
 S^2 : 1,44

Fórmula
$$n = \frac{N * Z_{\alpha/2}^2 * S^2}{(e^2 * N) + (Z_{\alpha/2}^2 * S^2)}$$

Substituyendo valores
$$n = \frac{5019 * 1,96^2 * 1,44}{(0,05^2 * 5019) + (1,96^2 * 1,44)}$$

Tamaño de la muestra
$$n = 1536$$

Cálculo del tamaño de la muestra, si la variable a investigar es de carácter cualitativo

Ahora, si el tipo de la variable principal es de carácter cualitativo, debe usar la siguiente fórmula:

$$n = \frac{N * Z_{\alpha/2}^2 * P * Q}{(e^2 * (N - 1)) + (Z_{\alpha/2}^2 * P * Q)}$$

En donde:

- n: Es el tamaño de la muestra
- N: Es el tamaño de la población
- $Z_{\alpha/2}$: Coeficiente de confianza (este coeficiente es 1.96 si se utiliza el 95% de confianza)
- e: Margen de error seleccionado
- P: Es la probabilidad de ocurrencia de la variable
- Q: Es la probabilidad de no ocurrencia de la variable (1-P)

Si la población es grande, la fórmula de muestreo se reduce a:

$$n = \frac{Z_{\alpha/2}^2 * P * Q}{e^2}$$

Tenga en consideración que algunas personas utilizan el valor P como 0.5 (que representa el 50% de ocurrencia) y por ende, Q sería 0.5 ($Q=1-P=1-0.5=0.5$); sin embargo esta opción se la debería utilizar, si no conozco la varianza de la variable de la población ni la probabilidad de ocurrencia del mismo.

Ejemplo: Se desea calcular el tamaño muestral para realizar un estudio a los habitantes de la Parroquia Rural El Morro, con un

margen de error del 5% y un nivel de confianza del 95% y se conoce que la probabilidad de ocurrencia de la variable a investigar es del 43%.

Entonces:

n: ¿?
 N: 5.019 habitantes según el Censo 2010
 $Z_{\alpha/2}$: 1,96 (95% de confianza)
 e: 0.05 (5% de error)
 P: 0.43 (43%)
 Q: 0.57 (=1-0.43)

$$\begin{aligned} \text{Fórmula} \quad n &= \frac{N * Z_{\alpha/2}^2 * P * Q}{(e^2 * (N - 1)) + (Z_{\alpha/2}^2 * P * Q)} \\ \text{Substituyendo} \quad n &= \frac{5019 * 1,96^2 * 0.43 * 0.57}{(0.05^2 * 5018) + (1,96^2 * 0.43 * 0.57)} \\ \text{valores} \end{aligned}$$

Tamaño de la muestra $n = 350$

3.2.4.2 Muestreo Aleatorio Estratificado

Cuando la población a estudiar está definida por varios grupos o estratos, de manera que, la variable de estudio es semejante en los componentes de un estrato y muy diferente entre los componentes de diferentes estratos, se utiliza el muestreo aleatorio estratificado. Por ejemplo, se desea estudiar la opinión de los estudiantes sobre la calidad de la educación en la

Universidad; por ende, se podría realizar un muestreo aleatorio simple que abarque toda la universidad y se tendría datos de la Universidad, pero no se podría analizar por facultad; para este caso se necesitará el muestreo estratificado, en donde los estratos serían las facultades.

Una vez definido los estratos de la población se debe calcular el tamaño de la muestra y se lo puede hacer como:

- **Muestreo estratificado por afijación proporcional:** el reparto de las unidades muestrales es distribuido entre cada estrato de acuerdo a la proporción que representan los estratos con respecto al universo. Por ejemplo, si se investigan las facultades de la Universidad y se sabe que la Facultad X contiene al 60% de los alumnos, la Facultad Y al 25% y la Facultad Z al 15% y el tamaño de la muestra es de 300 alumnos; la distribución debería ser 180 (300×0.60) alumnos en la Facultad X, 75 en la Facultad Y y 45 para la Facultad Z.
- **Muestreo estratificado por afijación uniforme:** para cada estrato, este tipo de reparto consiste en asignar a cada estrato el mismo número de unidades muestrales. Este tipo de afijación favorece a estratos pequeños pero a estratos grandes los perjudicará; por lo que se sugiere se realice en poblaciones cuyos estratos sean de tamaño similar. Por ejemplo, si se investigan las facultades de la Universidad y se sabe que la Facultad X contiene al 60% de los alumnos, la Facultad Y al 25% y la Facultad Z al 15% y el tamaño de muestra es de 300

alumnos, la distribución debería ser 100 alumnos para cada Facultad.

- **Muestreo estratificado óptimo:** este tipo de muestreo define el tamaño de la muestra en cada estrato, en base a la proporción de la desviación estándar de las variables objeto de estudio; es decir, que si un estrato tiene mayor variabilidad, el tamaño de la muestra será mayor con respecto a los otros estratos.

3.2.4.3 Muestreo Aleatorio Sistemático

Este tipo de muestreo indica que las unidades muestrales serán seleccionadas de manera sistemática o por intervalos o ciclos, a excepción del primer elemento de la muestra que se escoge aleatoriamente.

Para determinar el intervalo o ciclo de selección se divide el tamaño de la población para el tamaño de la muestra y el resultado indicará, cada cuantos elementos se deberán seleccionar en la unidad muestral.

Ejemplo, si en una población de 1500 unidades se desea realizar muestreo sistemático y el tamaño de muestra es de 300 unidades, se divide $1500/300$ para obtener el valor del intervalo que en este caso es 5. Se selecciona aleatoriamente el primer elemento y se escoge el elemento 185, siguiendo el intervalo de cada 5 elementos, el siguiente número en orden de selección será el 190, luego 195 y así sucesivamente se seleccionaran las unidades muestrales hasta completar las 300.

3.2.4.4 Muestreo por Conglomerados

Este tipo de muestreo es lo opuesto al muestreo estratificado, debido a que, primero se seleccionan al azar grupos de elementos de una población, para luego estudiar todos los elementos de cada grupo. Por ejemplo; una Universidad tiene 18 Facultades que abarcan 65 carreras, los conglomerados estarían conformados por las facultades y las carreras; de las cuales, se seleccionarían algunas facultades y dentro de estas algunas carreras.

3.2.5 Etapas del Muestreo

En el siguiente cuadro, se resumen los procesos del diseño muestral en 3 etapas: antes, durante y después.

Etapas	
Antes	– Objetivo de estudio – Variables – Población – Marco Muestral (mapas, planos, ubicación de viviendas, etc) – Instrumento de medición – Plan de levantamiento de información – Capacitación encuestadores – Prueba piloto – Plan de análisis
Durante	– Procedimiento de selección – Tamaño de la muestra (muestra y muestra de reposición) – Nivel de confianza y probabilidad de error – Determinar probabilidad de selección

Después	– Crítica y validación de cuestionario – Codificación – Digitación – Crítica y Validación de base de datos
---------	---

TABLA 3-2 ETAPAS DE MUESTREO

Fuente: Elaboración Propia

3.3 PARA EJERCITAR

- Se desea realizar una encuesta sobre el proceso de ingreso a la Universidad de Guayaquil en el 2017, para lo cual, la población a investigar es de 5.766 estudiantes que ingresaron a primer nivel. Determinar el tamaño de la muestras usando muestreo aleatorio simple.
- Dado el siguiente cuadro de datos, elija una muestra estratificada de tamaño $n=4$ de esta población.

Alumno	¿Trabaja?	Alumno	¿Trabaja?
Luis	No	Aurelio	Si
Diana	Si	Ricardo	Si
Joselyn	Si	Fátima	No
Mayra	Si	Elton	No
Ana	No	Eder	Si
Yesenia	Si	Lourdes	No
Marco	No	Luz	No
Lorena	Si	Gabriela	No
Dionisio	No	Richard	Si
Alejandro	No	Bryan	Si

- En la Parroquia El Morro de Guayaquil habitan 700 hombres adultos, 800 mujeres adultas y 500 menores de edad. Se quiere seleccionar una muestra de 180 personas, utilizando, para ello,

muestreo estratificado con afijación proporcional. ¿Cuál será la composición que debe tener dicha muestra?

- En la Parroquia El Morro de Guayaquil habitan 2000 personas. Se quiere seleccionar una muestra de 10 personas, utilizando para ello, muestreo sistemático. Indique el número de las personas seleccionadas.
- De un listado de 453 estudiantes de la Escuela de Sociología se desea obtener una muestra de 5 personas mediante muestreo aleatorio sistemático. Describa el proceso y simule dicha muestra sistemática indicando el número de las personas seleccionadas.



CAPÍTULO CUATRO PROCESAMIENTO DE DATOS

“El auténtico genio consiste en la capacidad
para evaluar información incierta, aleatoria y
contradictoria.”

Winston Churchill





4. PROCESAMIENTO DE DATOS

4.1 Introducción

Hay que tener en consideración que cuando se trata de datos, hay que ver si están procesados o no. Los datos no procesados que provienen de la fuente original, sin las modificaciones realizadas por el analista de datos, son generalmente difíciles de usar para el análisis, ya sea porque es grande o tiene problemas que deben ser corregidos; por ende, previo a realizar el análisis de datos, se deberá realizar una validación o un procesamiento de depuración de los datos primarios que hayan sido levantados, para que puedan ser analizados posteriormente y realizar los modelos estadísticos respectivos. Tenga en cuenta que los datos primarios pueden necesitar ser procesados una vez; y todos los pasos deben ser registrados, mientras que los datos procesados, por el contrario, son datos que están listos para su análisis.

Durante el procesamiento de datos se pueden presentar casos como la unión de algunas bases de datos, subconjuntos de un conjunto determinado de variables, la transformación de algunos de los datos o la eliminación de valores atípicos. Tenga en consideración que es posible que existan normas para el procesamiento, en función del tipo de datos que se está utilizando, y por eso, siempre que sea posible, se deberán aprovechar esas normas; independientemente de lo que el procesamiento haya utilizado. Todas estas normas deberán ser registrados para que futuras investigaciones puedan utilizar estos datos procesados y realizar su réplica a partir de los mismos datos en bruto. Imagine que se investiga la estatura en

cms de 6 alumnos de un determinado curso de la Escuela de Sociología y los datos son: 163, 175, 165, 168, 191, 119. Al observar los datos y realizar los análisis respectivos, parece que no existen inconvenientes en los mismos; sin embargo, dentro de la observación que se realizó al momento del levantamiento de información no se detectaron alumnos que midieran menos de 150 cms y si se revisan los datos, existe un dato que indica 119 cms, lo cual lleva a revisar el cuestionario donde se registró la información y verificar la información y este indica 191 cms, esto implica corregirlo en la base. Si hubiera realizado el análisis sin depurar la información al realizar el análisis del promedio de estatura este fuera 163,5 cms, pero corrigiendo el dato este sería 174,5 cms de estatura. ¿Cuál sería el valor real? ¿Si tuviera que decidir en base al resultado obtenido y tomara el valor de los datos sin procesar, tomaría una buena decisión?

John Tukey, un famoso analista de datos dijo: “Los datos no pueden contener la respuesta. La combinación de algunos datos y un deseo para una respuesta no asegura que una respuesta razonable se puede extraer de un cuerpo dado de datos. Así que es importante tener en cuenta que si usted está tratando de responder a una pregunta específica, que es la base para la mayoría de los análisis de datos fiables, puede que no tenga los datos para responder a esa pregunta”.

Cabe resaltar que no importa cuán extenso sean los datos, ya que en algunas ocasiones es necesario dar un paso atrás y revisar si los datos responden a la pregunta planteada. Este es el primer paso importante para hacer el correspondiente análisis de los datos.

4.2 Codificación de las variables en la base de datos

Por lo general, las investigaciones manejan una gran cantidad de datos, por lo que se requiere hacer un análisis compacto, y que sea comprensible para las personas. Para el efecto, hay que tener en cuenta la manera en que se interpreten los datos en el archivo previo a su tabulación.

Cuando la variable es pequeña se puede escribir en forma completa, pero en el caso de que sea extensa, es necesario representarla por una letra mayúscula. Por ejemplo: se utiliza la letra A mayúscula para representar la altura o una P para el peso. En caso de haber más datos del mismo tipo, se emplearán subíndices; por ejemplo, tres alturas de las distintas personas, se las podrán etiquetar con un subíndice (P1, P2 y P3).

4.3 Diseño de matriz de datos

Previo al procesamiento de los datos, se debe usar algún software para almacenarlos; estos pueden ser Excel, SPSS, STATA, ACCESS u otros, para ingresar la información que luego se procesará. Para realizar el registro de estos en la base, tenga en consideración los siguientes puntos:

- Cada columna de la base será utilizada para cada variable.
- En cada columna se registrará solamente una variable específica. Por ejemplo; hay personas que en una variable ponen edad y peso cuando lo óptimo sería una columna para edad y otra columna para el peso.
- Cada fila corresponderá a cada formulario levantado.
- Tenga en consideración que existirán algunas variables que deben ser codificadas previo a ser registradas en la base de

datos. Por ejemplo; la variable sexo la puede codificar como 1 para mujeres o 2 para hombres, o M y H.

La siguiente tabla emula el diseño de la base de datos de un estudio clínico realizado en el año 2012 a pacientes que estaban internados en hospitales públicos, y su nivel de satisfacción con respecto a la atención que recibieron. Se puede observar que cada columna representa a las variables (tipo de enfermedad, fecha de ingreso, fecha de egreso y nivel de satisfacción) y cada fila representa a las observaciones (paciente 1, paciente 2, paciente 3, etc).

# Paciente	Tipo de enfermedad	Fecha de ingreso	Fecha de egreso	Nivel de satisfacción
1	A	01/05/2012	03/05/2012	1
2	B	03/07/2012	08/07/2012	3
3	A	19/02/2012	01/05/2012	4
4	A	10/08/2012	03/09/2012	1
5	C	01/07/2012	03/07/2012	2
.....				
.....				

TABLA 4-1 ESQUEMA DE UNA BASE DE DATOS

Fuente: Elaboración Propia

Se debe considerar que en cada tabla o archivo se almacenan los datos sobre un tipo de observación, por ejemplo, en un estudio clínico, no se debe incluir en la misma tabla la información sobre los pacientes y la de los hospitales. El

objetivo es separar los datos de tal de manera que resulte fácil responder las preguntas durante el análisis de los datos.

	A	B	C	D	AA	AC	AE	AF	AG	AH	AI	AJ	AK
2	cod encuesta	FECHA DE DIGITACION	3. Cabece	4. Comunidad	5.1 tipo de vivienda	6.1 es: material	7.1 material	7.2 otro material	8.1 material	8.2 otro material	9.1 material	9.2 otro material	10.1 estado
3	24-9-20-11-14	17/12/2014	chongan	chongan	viv	1	3	2		1		2	2
4	8-2-18-11-14	17/12/2014	chongan	chongan		1	8	2		1		1	8
5	5-5-19-11-14	17/12/2014	chongan	chongan		1	3	2		1		2	2
6	5-8-19-11-14	17/12/2014	chongan	chongan		1	3	2		1		1	2
7	23-9-17-11-14	17/12/2014	chongan	chongan		1	3	1		1		1	1
8	20-2-18-11-14	19/12/2014	chongan	chongan		4	1	2		2		5	2
9	23-2-19-11-14	19/12/2014	chongan	chongan		1	3	2		1		2	2

FIGURA 4-1 HOJA ELECTRÓNICA

Fuente: Elaboración Propia

Tenga en cuenta que es necesario realizar la validación de la hoja electrónica y verificar que no existan errores. La validación consiste en revisar cada columna a través de los filtros para cotejar que exista un valor no adecuado a la respuesta de la pregunta. Un ejemplo de error, es que en la columna sexo exista la opción 3, cuando solo se tienen dos opciones. De darse el caso de errores, hay que revisar el cuestionario digitado para la corrección respectiva.



CAPÍTULO CINCO

ESTADÍSTICA DESCRIPTIVA

No existe la suerte. Sólo hay preparación adecuada o inadecuada para hacer frente a una estadística.

Robert Heinlein





5. ESTADÍSTICA DESCRIPTIVA

5.1 Introducción

La estadística es la ciencia formal que estudia la recolección, análisis e interpretación de datos de un universo o una muestra representativa y es utilizada ya sea para la toma de decisiones o para explicar condiciones regulares o irregulares de algún fenómeno o estudio aplicado, de ocurrencia en forma aleatoria o condicional. Es la herramienta fundamental que permite llevar a cabo el proceso relacionado con la investigación científica. Es transversal a una amplia variedad de disciplinas, desde la física hasta las ciencias sociales, desde las ciencias de la salud hasta el control de calidad.

La estadística se divide en dos grandes áreas:

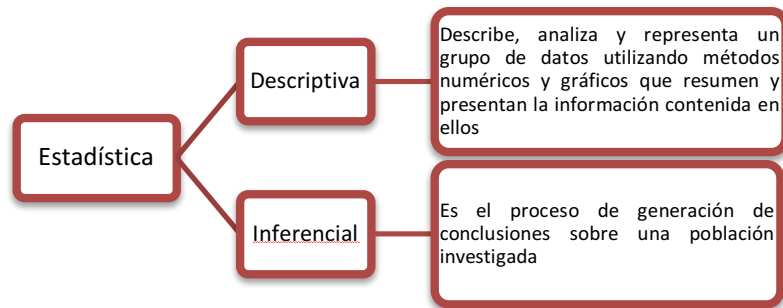


FIGURA 5-1 ÁREAS DE LAS ESTADÍSTICAS

Fuente: Elaboración Propia

5.2 Estadística descriptiva

Describe, analiza y representa un grupo de datos utilizando métodos numéricos y gráficos que resumen y presentan la información contenida en ellos. Puede definirse como aquel método que contiene la recolección, organización, presentación y resumen de una serie de datos. El mencionado resumen puede hacerse con tabulación, gráfico o numérico. El análisis que se realiza se limita en sí mismo a los datos recolectados y no se puede realizar inferencia alguna o generalizaciones algunas, acerca de la población de donde provienen esos datos estadísticos.

El proceso que sigue la estadística descriptiva para el estudio de una cierta población consta de los siguientes pasos:

1. Selección de las variables a ser estudiadas.
2. Obtención del valor de cada individuo investigado en las variables seleccionadas.
3. Elaboración de tablas de frecuencias, mediante la adecuada clasificación de los individuos.
4. Obtención de parámetros estadísticos, números que sintetizan los aspectos más relevantes de una distribución estadística, entre las cuales están las medidas de tendencia central y las de variabilidad.
5. Representación gráfica de los resultados (elaboración de gráficos estadísticos).

La estadística descriptiva se divide en cinco partes:

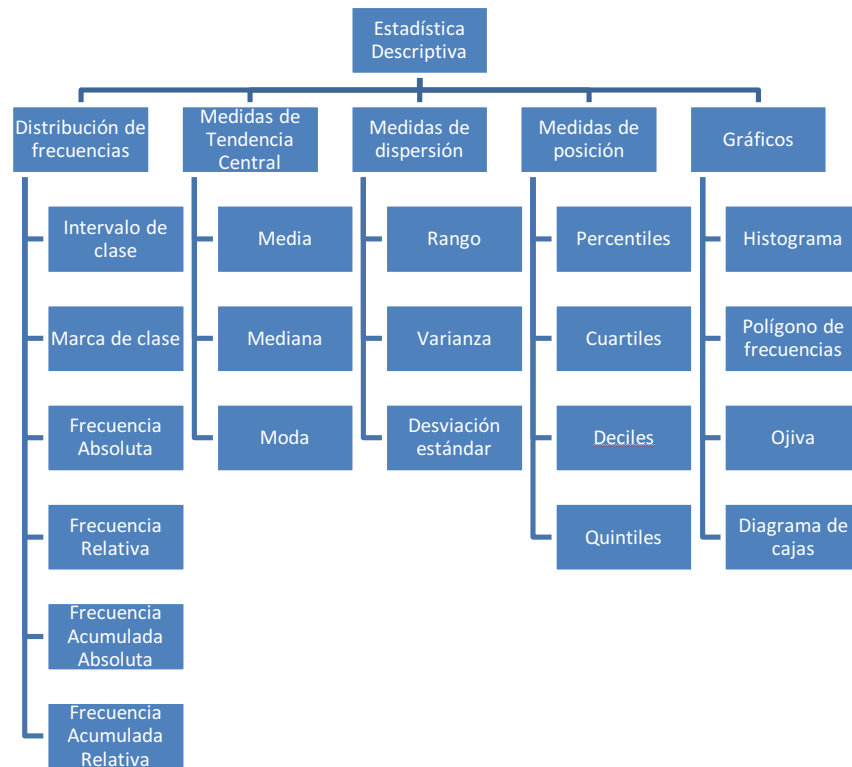


FIGURA 5-2 COMPONENTES DE LA ESTADÍSTICA DESCRIPTIVA

Fuente: Elaboración Propia

Recuerde que los métodos estadísticos a emplearse dependerán de la escala de medición de los datos. En la siguiente tabla se detalla el tipo de método estadístico a emplear, según el tipo de variable.

Escalas de medición	Descripción	Método estadístico
Nominales	Variables categóricas, no tiene orden ni distancia	✓ Frecuencia ✓ Moda
Ordinales	Variables categóricas con orden pero no tienen distancias	✓ Frecuencias ✓ Moda ✓ Mediana
Intervalos	Variable que tiene orden y distancia. El cero no es absoluto	✓ Frecuencia ✓ Moda ✓ Mediana ✓ Media
Razones o proporciones	Variables que tienen distancia, orden y el cero es absoluto	✓ Frecuencia ✓ Moda ✓ Mediana ✓ Media ✓ Varianza

TABLA 5-1 MÉTODO ESTADÍSTICO SEGÚN TIPO DE VARIABLE

Fuente: Elaboración Propia

A continuación se detalla cada una de las cinco partes en que se divide la estadística descriptiva.

5.3 Tabla de frecuencias

Al realizar una investigación en la cual se levantan datos, muchas veces se topa con la pregunta ¿y ahora qué hago con todos estos datos? ó ¿Por dónde empiezo para procesar los datos levantados?. Para contestar estas preguntas se debe recordar que la meta de la estadística descriptiva es representar los datos de manera sencilla y práctica que permita descubrir patrones en los datos levantados. Una de las maneras más

fáciles de empezar a descubrir estos patrones es a través de las frecuencias de los datos que no es más que contar los valores particulares de una variable de los datos levantados. Estos valores se pueden agrupar en una tabla que se denomina “Tabla de frecuencias”. Tenga en consideración que la tabla de frecuencias puede ser usada para variables de tipo de escala ordinal, nominal, de intervalos o de razones.

5.3.1 Tabla de frecuencias para variables cuya escala de medición son de tipo intervalos o de razones.

La tabla de frecuencias está compuesta por:

- Intervalos de frecuencia o clase
- Marca de clase
- Frecuencia absoluta (f_a).
- Frecuencia relativa (f_r).
- Frecuencias acumuladas absolutas (FA).
- Frecuencia acumulada relativa (FR).

#	Intervalos de frecuencia o clase	Marco de clase	Frecuencia absoluta (f_a).	Frecuencia relativa (f_r)	Frecuencias acumuladas absolutas (FA)	Frecuencia acumulada relativa (FR)
1	$[L.I.1 \quad LS1)$	$(L.I.1 + LS1)/2$	f_{a1}	$f_{r1} = f_{a1}/n$	$FA1 = f_{a1}$	$FAR1 = FA1/n$
2	$[L.I.2 \quad LS2)$		f_{a2}	$f_{r2} = f_{a2}/n$	$FA2 = f_{a1} + f_{a2}$	$FAR2 = FA2/n$
...	$[\quad)$					

TABLA 5-2 MODELO DE LA TABLA DE FRECUENCIAS

Fuente: Elaboración Propia

Antes de realizar el detalle de cada uno de los elementos que forman parte de la tabla de frecuencias y cómo se deben realizar los procesos de cálculo, se deben seguir los siguientes pasos:

1. Ordenar los datos de menor a mayor valor
2. Calcular la cantidad de intervalos necesario para elaborar la tabla de frecuencias. Para determinar el número de intervalos se recomienda utilizar las siguientes fórmulas:

$$\#intervalos = \sqrt{n} \text{ (Raíz cuadrada de } n\text{);}$$

ó

$$\#intervalos = 1 + 3,3 * \log (n);$$

donde n es la cantidad de datos que se analizarán; estas fórmulas no necesariamente darán valores enteros, por lo cual habrá que redondearlos. Tener en consideración que estas fórmulas no son una camisa de fuerza, son solo una guía y cada vez que se utilicen dependerá de los datos que se analicen.

Ejemplo: Si se tienen 25 datos, la tabla deberá tener 5 intervalos ($\sqrt{25}=5$). Si fueran 100, se necesitaría una tabla de frecuencias con 10 intervalos.

Cabe resaltar, que las cantidades de intervalos no siempre serán los adecuados en la tabla; en ese caso se deberá hacer un análisis de los datos y estimar qué cantidad de intervalos son los óptimos. Ejemplo si se tienen 100 datos entre 1 y 3, si se sigue al pie de la letra los pasos anteriores se tendría una tabla de

frecuencias de 10 intervalos, lo cual implicaría que 7 de los 10 tendrían 0 como frecuencias absolutas.

3. Calcule la distancia de cada intervalo. Se puede ayudar con el siguiente procedimiento:

$$di = \frac{\text{Valor máximo} - \text{Valor mínimo}}{\text{Cantidad de intervalos}}$$

4. Elabore los intervalos, usando los siguientes símbolos para cada intervalo: al inicio “[” y al final “)”, por ejemplo [a,b) lo cual indica que este intervalo está entre los valores de a y valores antes de b. Este proceso se repite para cada intervalo. Recuerde que la distancia de cada intervalo debe ser la misma.

A continuación se enuncian todos los elementos que conforman la tabla de frecuencias.

Intervalos de frecuencia o clase: En esta columna se registran los intervalos que agrupan los datos. Algunos investigadores sugieren que la cantidad de intervalos sea más de 3 y menos de 10, esto permitirá observar los patrones de los datos de una manera adecuada, y lo cual se podrá apreciar cuando se grafique el Histograma.

En los intervalos se deben considerar tres puntos importantes:

1. Cada intervalo debe tener la misma amplitud o distancia entre su punto inicial y final;
2. El primer intervalo debe incluir el valor mínimo de los datos;

3. El último intervalo debe incluir el valor máximo de los datos, esto implica que el valor final del último intervalo no necesariamente debe ser el valor máximo de los datos, este puede ser mayor, pero nunca menor.

Marca de clase: Es el valor medio de los intervalos de clase. Se calcula sumando el límite superior con el límite inferior de cada intervalo dividido para dos. Los polígonos de frecuencias se grafican con los datos de los intervalos de clase y marca de clase.

Frecuencia absoluta: Contabiliza el número de datos de cada intervalo de frecuencia o clase. Su símbolo es **fa**.

Frecuencia relativa: Representa la proporción de la frecuencia absoluta con respecto a toda la población. Esta se calcula al dividir la frecuencia absoluta para el total de datos y por lo general, se la representa en porcentajes. Se utiliza para graficar el Histograma. Su símbolo es **fr**.

Frecuencias acumuladas absolutas: Acumulan las frecuencias absolutas de cada intervalo. Su símbolo es **FA**.

Frecuencia acumulada relativa: Divide los valores obtenidos en la frecuencia acumulada absoluta para el total de datos de la muestra. Se utiliza para graficar la ojiva. Su símbolo es **FR**.

En base a los siguientes datos correspondientes a la Calificación sobre 100 obtenida por 25 alumnos de la materia de estadística 1 en su primera evaluación; se construye la tabla de frecuencias.

Alumno	Nota	Alumno	Nota	Alumno	Nota	Alumno	Nota	Alumno	Nota
1	80	6	80	11	70	16	95	21	65
2	35	7	82	12	95	17	30	22	65
3	45	8	90	13	85	18	26	23	73
4	32	9	65	14	59	19	50	24	35
5	78	10	75	15	50	20	57	25	34

TABLA 5-3 NOTA DE ALUMNOS

Fuente: Elaboración Propia

Paso 1: Ordenar los datos.

Alumno	Nota	Alumno	Nota	Alumno	Nota	Alumno	Nota	Alumno	Nota
1	26	6	35	11	59	16	73	21	82
2	30	7	45	12	65	17	75	22	85
3	32	8	50	13	65	18	78	23	90
4	34	9	50	14	65	19	80	24	95
5	35	10	57	15	70	20	80	25	95

TABLA 5-4 NOTA DE ALUMNOS ORDENADOS

Fuente: Elaboración Propia

Paso 2: Calcular la cantidad de intervalos mediante la raíz cuadrada de n (n es la cantidad de datos, que en este ejemplo son 25 datos).

$$\#intervalos = \sqrt{25} = 5$$

Paso 3: Calcular la distancia de cada intervalo, restando el valor máximo de los datos (95) y el valor mínimo (26), dividido para el total de intervalos (5).

$$Distancia = \frac{datomáximo - datomínimo}{\# de intervalos} = \frac{95 - 26}{5} = 13,8$$

Intervalos de clase: La distancia de cada intervalo es de 13,8 unidades; y se puede redondear a 14. El primer intervalo tiene como límite inferior el valor mínimo de mis datos y como superior será la distancia del intervalo más el límite inferior. Este proceso se lo realiza para calcular el resto de intervalos. El primer intervalo tendría el dato 26 como límite inferior y 40 como límite superior (26+14); el segundo intervalo empieza con 40 y termina con 54 y así sucesivamente.

Intervalos de clase
[26 – 40)
[40 – 54)
[54 – 68)
[68 – 82)
[82 – 96]

Marca de clase: Una vez determinados los intervalos de clase, se calcula la marca de clase de cada intervalo.

$$Marca de clase = \frac{límite inferior + límite superior}{2} = \frac{26 + 40}{2} = 33$$

Frecuencia absoluta (fa): Dentro de cada intervalo se cuenta la cantidad de datos que estén entre el límite inferior y antes del límite superior del mismo; en el ejemplo el primer intervalo $[26 - 40)$ se cuentan los datos que estén entre 26 y antes de 40; es decir se cuentan los datos 26, 30, 32, 34, 35, 35 que en total son 6 en este intervalo. Cabe resaltar que si existe un dato igual al límite superior del intervalo que se está desarrollando, no se lo cuenta, puesto que en el siguiente intervalo va a estar repetido y se alteraría el total de datos al contabilizarlo dos veces.

Frecuencia absoluta (fa).
6
3
5
6
5

Tener en cuenta que la suma total de la frecuencia absoluta debe sumar el total de datos que se están analizando. En el ejemplo de las notas de los alumnos, la sumatoria de valores deberá ser 25 ($6+3+5+6+5=25$), que correspondería al número de datos del ejercicio.

Frecuencia relativa (fr)
24% ($6/25=0,24$)
12% ($3/25=0,12$)
20% ($5/25=0,2$)
24% ($6/25=0,24$)
20% ($5/25=0,2$)

Frecuencia relativa (fr): Recordar que la frecuencia relativa es la proporción de cada intervalo con respecto al total de datos, y se lo calcula dividiendo la frecuencia absoluta para el total de datos. En el primer intervalo, la frecuencia relativa es 24% ($6/25=0,24$); para el resto de intervalos el proceso es igual. Tener presente que la suma de porcentajes de todos los intervalos debe dar como resultado el 100%. Para evitar omitir información relevante, que generalmente se

refleja cuando se trabaja con una gran cantidad de datos, se sugiere utilizar dos decimales en los porcentajes. Ejemplo: En una muestra de 5000 unidades donde la proporción de un

intervalo es 28,4% pero solo se coloca 28%; al momento de analizar se podrá observar que el 28% de 5000 es 1400 unidades, pero si se calcula el 28,4% el resultado será 1420 unidades, existiendo una diferencia de 20 unidades que no estarán siendo sujetas a estudio por redondear el porcentaje u omitir los decimales.

Frecuencia Acumulada (FA): Para el cálculo de esta frecuencia, el primer intervalo tendrá el mismo valor que la frecuencia absoluta, pero para en el segundo intervalo, el valor se obtendría sumando la frecuencia absoluta del primer intervalo más el valor de la frecuencia absoluta del segundo intervalo ($6+3=9$); para el resto de intervalos el proceso es igual. Considerar que el último valor del FA debe coincidir con el total de datos con los que se está trabajando.

FA
6
9
14
20
25

(FR)
24% ($6/25=0,24$)
36% ($9/25=0,36$)
56% ($14/25=0,56$)
80% ($20/25=0,80$)
100% ($25/25=1,00$)

Frecuencia acumulada relativa (FR): Este proceso se realiza igual al de la frecuencia relativa, tomando esta vez como numerador el valor de la Frecuencia Acumulada Absoluta y como denominador el total de datos. Considerar que el último intervalo debe dar como resultado el 100%.

La tabla de frecuencia quedaría de la siguiente manera:

#	Intervalo de clase	Marca de clase	fa	fr	FA	FR
1	[26 – 40)	33	6	24%	6	24%
2	[40 – 54)	47	3	12%	9	36%
3	[54 – 68)	61	5	20%	14	56%
4	[68 – 82)	75	6	24%	20	80%
5	[82 – 96)	89	5	20%	25	100%

TABLA 5-5 TABLA DE FRECUENCIAS

Fuente: Elaboración Propia

Este trabajo no termina con el cálculo de la tabla de frecuencias sino que se debe realizar el análisis de la misma utilizando los valores de las frecuencias relativas (fr) y de las frecuencias acumuladas relativas (FR), junto con los intervalos.

Análisis fr

El 24% de los estudiantes del curso de Estadística 1 de la Escuela de Sociología tienen nota entre 26 y hasta antes de 40/100.

El 12% de los estudiantes del curso de Estadística 1 de la Escuela de Sociología tienen nota entre 40 y hasta antes de 54/100.

El 20% de los estudiantes del curso de Estadística 1 de la Escuela de Sociología tienen nota entre 54 y hasta antes de 68/100.

Cabe resaltar que en un informe no es necesario analizar todas las frecuencias relativas sino más bien las que reflejen valores llamativos.

Análisis FR

Para este análisis no se considera el primer intervalo por cuanto el valor es igual al primer intervalo analizado en la frecuencia relativa (24%) y tampoco se analiza el último intervalo por cuanto no es correcto decir que el 100% de la población tiene calificaciones entre 26 y 96 / 100.

El 36% de los estudiantes del curso de Estadística 1 de la Escuela de Sociología tienen nota entre 26 y menos de 54/100.
El 56% de los estudiantes del curso de Estadística 1 de la Escuela de Sociología tienen nota entre 26 y menos de 68/100.
El 80% de los estudiantes del curso de Estadística 1 de la Escuela de Sociología tienen nota entre 26 y menos de 82/100.

5.3.2 Tabla de frecuencias para variables cuya escala de medición es de tipo ordinal o nominal

La tabla de frecuencias para variables de tipo de escala ordinal o nominal está compuesta por:

- Valores de la variable o modalidades del atributo
- Frecuencia absoluta (fa).
- Porcentaje.
- Porcentaje acumulado.

#	Valores de la variable	Frecuencia absoluta (fa).	Porcentaje	Porcentaje acumulado
1	Valor 1	fa1	$P1=(fa1/n)*100$	$PA1=p1$
2	Valor 2	fa2	$p2=(fa2/n)*100$	$PA2=p2+p1$
...	[)			

TABLA 5-6 FORMATO DE TABLA DE FRECUENCIAS PARA DATOS ORDINALES O NOMINALES

Fuente: Elaboración Propia

Valores de las variables: En esta columna colocar los diversos tipos de atributos que tiene la variable; por ejemplo, el sexo de las personas investigadas (hombre o mujer).

Frecuencia absoluta: Contabilizar el número de datos de cada intervalo de frecuencia o clase. Se simboliza con fa.

Porcentaje: Se calcula igual que en la frecuencia relativa de las variables de tipo de intervalo o razón. Se dividen los valores obtenidos en la frecuencia absoluta para el total de datos de la muestra y luego se lo multiplica por 100%. Se utilizan para graficar el Histograma.

Porcentaje acumulado: Se calcula igual que en la frecuencia acumulada relativa de las variables de tipo de intervalo o razón. Considerar que el último intervalo deberá sumar 100%.

En la siguiente tabla se detalla el estado civil de las personas de la Provincia del Pichincha. En la primera columna se anotan los atributos de la variable estado civil (casado, soltero, etc.), en la segunda columna está la frecuencia absoluta de cada atributo, luego el porcentaje de cada atributo con respecto al total de la

Provincia y por último se encuentra la columna de Porcentaje Acumulado.

Estado Civil	Frecuencia absoluta	Porcentaje	Porcentaje Acumulado
Casado/a	801.113	40,14%	40,14%
Unido/a	261.701	13,11%	53,26%
Separado/a	73.963	3,71%	56,96%
Divorciado/a	58.577	2,94%	59,90%
Viudo/a	64.285	3,22%	63,12%
Soltero/a	735.944	36,88%	100,00%
Total	1.995.583	100,00%	

TABLA 5-7 ESTADO CIVIL DE LAS PERSONAS A NIVEL NACIONAL

Fuente: INEC CENSO 2010, Elaboración Propia

El análisis en este caso sería: Según el Censo Nacional del 2010, el 40,14% de las personas de Pichincha mayores de 12 años está casada, porcentaje comparado a nivel nacional (32,49%) es alto y supera casi tres veces a la Provincia de Esmeraldas (15,34%).

Cabe resaltar que, por lo general, se sugiere trabajar con dos decimales, debido a que si se analiza una población grande y no se consideran los decimales se omitirán valores con respecto a la cantidad de personas.

Para finalizar este capítulo es necesario recordar que dentro del análisis se debe considerar la tabla de distribución de frecuencias, así como los gráficos que se generan a partir de la tabla.

5.3.4 Para Ejercitar

La siguiente tabla examina el porcentaje de estado civil “Unido” de las Parroquias Urbanas de Guayaquil, según el Censo 2010.

Parroquia Urbana	Estado Civil Unido/a
Pedro Carbo	12,00 %
Rocafuerte	12,80 %
Nueve de Octubre	15,06 %
Sucre	15,14 %
Roca	15,22 %
Olmedo	15,58 %
Ayacucho	15,99 %
Bolívar	16,65 %
Chongón	17,76 %
García Moreno	19,19 %
Urdaneta	19,53 %
Letamendi	20,33 %
Tarqui	23,08 %
Ximena	23,31 %
Febres Cordero	23,60 %
Pascuales	34,24 %

1. En base a los datos sobre el porcentaje de estado civil “Unido” de las Parroquias Urbanas de Guayaquil, seleccione la tabla de distribución de frecuencias correcta.

A.

TABLA DE DISTRIBUCIÓN DE FRECUENCIAS				
CLASE	fa	fr	FA	FR
[12 - 17)	8	50,00%	8	50,00%
[17 - 22)	4	25,00%	12	75,00%
[22 - 27)	2	12,50%	14	87,50%
[27 - 32)	1	6,25%	15	93,75%
[32 - 37)	1	6,25%	16	100,00%

B.

TABLA DE DISTRIBUCIÓN DE FRECUENCIAS				
CLASE	fa	fr	FA	FR
[12 - 17,56)	8	50,00%	8	50,00%
[17,56 - 23,12)	5	31,25%	13	81,25%
[23,12 - 28,68)	2	12,50%	15	93,75%
[28,68 - 34,24)	1	6,25%	16	100,00%

C.

TABLA DE DISTRIBUCIÓN DE FRECUENCIAS				
CLASE	fa	fr	FA	FR
[12 - 17)	8	50,00%	8	50,00%
[17 - 22)	4	25,00%	12	75,00%
[22 - 27)	3	12,50%	14	87,50%
[27 - 32)	0	6,25%	15	93,75%
[32 - 37)	1	6,25%	16	100,00%

D.

TABLA DE DISTRIBUCIÓN DE FRECUENCIAS				
CLASE	fa	fr	FA	FR
[12 - 18)	9	56,25%	9	56,25%
[18 - 24)	6	37,50%	14	93,75%
[24 - 30)	0	0,00%	15	93,75%
[30 - 36)	1	6,25%	16	100,00%

2. En el siguiente ejercicio de edades, seleccione el análisis correcto de la tabla de distribución de frecuencias.

TABLA DE DISTRIBUCIÓN DE FRECUENCIAS				
CLASE	fa	fr	FA	FR
[12 - 17)	8	50,00%	8	50,00%
[17 - 22)	4	25,00%	12	75,00%
[22 - 27)	2	12,50%	14	87,50%
[27 - 32)	1	6,25%	15	93,75%
[32 - 37)	1	6,25%	16	100,00%

A	El 25% de los datos están entre 12 y 22 años de edad	
B	El 75% de los datos están entre 12 y 22 años de edad	
C	El 87,5% de los datos están entre 22 y 27 años de edad	
D	El 6,25% de los datos tienen menos de 32 años de edad	

3. La frecuencia relativa de la tabla de frecuencias se la calcula:

A	Sumando todos los datos y dividirlo para el total de datos	
B	Dividiendo la frecuencia absoluta para el total de datos	
C	Dividiendo el 50% de los datos para el total de datos	
D	Dividiendo el 75% de los datos para el total de datos	

4. Realice la tabla de frecuencias y el análisis de los datos del índice de jefatura femenina (IJF) por Provincias, tomadas del Censo 2010.

#	Provincia	IJF % 2010	#	Provincia	IJF % 2010
1	AZUAY	59	13	MANABI	34
2	BOLIVAR	45	14	MORONA SANTIAGO	41
3	CANAR	56	15	NAPO	31
4	CARCHI	33	16	PASTAZA	41
5	COTOPAXI	37	17	PICHINCHA	38
6	CHIMBORAZO	46	18	TUNGURAHUA	38
7	EL ORO	38	19	ZAMORA CHINCHIPE	38
8	ESMERALDAS	42	20	GALAPAGOS	34
9	GUAYAS	43	21	SUCUMBIOS	33
10	IMBABURA	41	22	ORELLANA	29
11	LOJA	44	23	SANTO DOMINGO	40
12	LOS RIOS	35	24	SANTA ELENA	31

5.4 Medidas De Tendencia Central

A partir de este capítulo se analizará la importancia de resumir toda la información de una variable solamente en algunos puntos de análisis, que en muchos de los casos, se los conoce como "las estadísticas de resumen"; la cual está dividida en tres grandes segmentos:

- i. Medidas de tendencia central
- ii. Medidas de Variabilidad
- iii. Medidas de posición no centrales.

5.4.1 Media aritmética o simplemente media o promedio

Es la medida más utilizada y comprende el proceso de sumar todos los valores que forman parte de una variable y dividirlo para el número total de datos de esa variable. Permite identificar y ubicar el punto alrededor del cual suelen reunirse los valores. Esto es conocido como la estimación puntual de una distribución. Cabe resaltar, que la media es susceptible a valores que son demasiado altos con respecto al conjunto de valores que forman parte de la variable investigada, lo cual podría llevar a tomar decisiones erradas sobre el promedio de los datos.

Ejemplo: *La media de los datos de la tabla 5.7 es 62,04 y resulta de sumar todos los valores (1.551) dividido para el número total de notas (25).*

Análisis: *El promedio de nota de los alumnos de esta clase es de 62,04 sobre 100.*

Alumno	Nota	Alumno	Nota	Alumno	Nota	Alumno	Nota	Alumno	Nota
1	80	6	80	11	70	16	95	21	65
2	35	7	82	12	95	17	30	22	65
3	45	8	90	13	85	18	26	23	73
4	32	9	65	14	59	19	50	24	35
5	78	10	75	15	50	20	57	25	34

TABLA 5-8 CALIFICACIONES DE ALUMNOS

Fuente: Elaboración Propia

5.4.2 Mediana

La mediana es la puntuación media en una distribución; en otras palabras es la ubicación central dentro de un conjunto de datos ordenados. Esto quiere decir, que el conjunto de datos iguales o menores que la mediana supondrá, será el 50% de los datos.

Ejemplo: La mediana de los datos de la tabla 5.7 es 65. Si usted ordena los datos podrá observar que el dato que se encuentra en el centro de todos es el que está ubicado en el puesto 13 de los 25 puestos.

Análisis: El 50% de las notas de los alumnos tienen menos de 65/100.

Tenga en cuenta que si la cantidad de datos es impar, la mediana será el dato que está ubicado en el centro de todos los datos; mientras que si la cantidad de datos es par se debe tomar los dos datos que están ubicados en el centro del conjunto de datos, sumar ambos y dividirlo para dos (o sacar el promedio de ambos).

Ejemplo: Si se eliminara el último dato de la tabla 5.7, se obtendría 24 datos y los puntos medios estarían en la ubicación 12 (calificación 65) y 13 (calificación 65) el promedio entre ambas sería 65, es decir la mediana.

Cabe resaltar que la mediana no es susceptible a datos que sean demasiado altos con respecto al conjunto de datos que forman parte; sin embargo no toma en consideración a todos los datos como lo hace la media, sino que solo toma los datos ubicados en el centro del conjunto de datos.

5.4.3 Moda

Se denomina moda al valor de una variable que se presenta con mayor frecuencia en un grupo de datos. Dependiendo de la cantidad de valores que se repiten con mayor frecuencia y en igual cantidad la variable investigada puede ser: Unimodal, Bimodal o Multimodal.

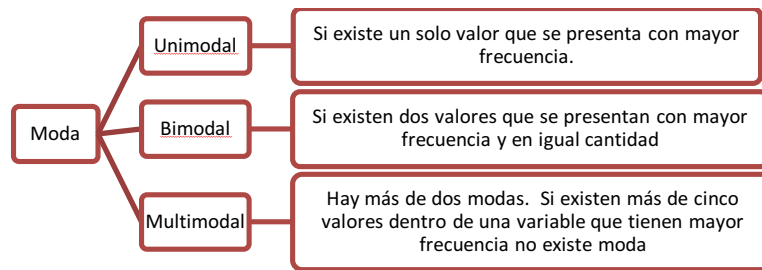


FIGURA 5-3 TIPOS DE MODA

Fuente: Elaboración Propia

Ejemplo: La moda de los datos de la tabla 5.7 es 65, porque es el dato que se repite más veces (3 veces).

Análisis: Las notas de los alumnos es unimodal pues indica que tres alumnos tienen como nota 65/100.

5.4.4 Para Ejercitar

La siguiente la tabla indica el % de estado civil “Unido” de las Parroquias Urbanas de Guayaquil en el 2010.

Parroquia Urbana	Estado Civil Unido/a
Pedro Carbo	12,00 %
Rocafuerte	12,80 %
Nueve de Octubre	15,06 %
Sucre	15,14 %
Roca	15,22 %
Olmedo	15,58 %
Ayacucho	15,99 %
Bolívar	16,65 %
Chongón	17,76 %
García Moreno	19,19 %
Urdaneta	19,53 %
Letamendi	20,33 %
Tarqui	23,08 %
Ximena	23,31 %
Febres Cordero	23,60 %
Pascuales	34,24 %

1. Escoger el análisis correcto de los datos obtenidos:

A	El promedio de% de Estado civil “Unido” de las parroquias urbanas de Guayaquil es 17,21% y el promedio de la mediana del % de Estado civil “Unido” de las parroquias urbanas de Guayaquil es 18,72%	
B	El promedio de % de Estado civil “Unido” de las parroquias urbanas de Guayaquil es 16,65% y el promedio de la mediana del 50% de Estado civil “Unido” de las parroquias urbanas de Guayaquil es la mediana es 18,72%	

C	El promedio de % de Estado civil “Unido” de las parroquias urbanas de Guayaquil es 18,72% y el 50% de las parroquias urbanas de Guayaquil tienen menos del 17,21% de Estado civil “Unido”	
D	El promedio de % de Estado civil “Unido” de las parroquias urbanas de Guayaquil es 18,72% y el 50% de las parroquias urbanas tienen menos del 16,65% de Estado civil “Unido”	

2. La media representa:

A	El 50% de los datos	
B	El promedio de los datos	
C	El 25% de los datos	
D	El 75% de los datos	

3. La mediana representa:

A	El 50% de los datos	
B	El promedio de los datos	
C	El 25% de los datos	
D	El 75% de los datos	

4. Si existe un valor aberrante esta altera la:

A	Media	
B	Cuartiles	
C	Mediana	
D	No altera nada	

5. Calcule las medidas de tendencia central de los datos presentados en el índice de jefatura femenina (IJF) por Provincias, tomadas del Censo 2010. También realice el análisis de los resultados obtenidos. ¿Qué opina?

5.5 Medidas de Variabilidad o dispersión

Las medidas de variabilidad muestran la dispersión de una distribución de datos, indicando por medio de un valor, cómo los diferentes datos de una variable están dispersos o concentrados con respecto a la media y mientras mayor sea ese valor, indica que es mayor la variabilidad o dispersión de los datos de la distribución. Las medidas principales que veremos son el Rango, la varianza y la desviación estándar.

5.5.1 Rango

El rango o recorrido estadístico es la diferencia entre el valor máximo y el valor mínimo en un grupo de números aleatorios. Su simbología es $Rang$ e indica el número de unidades de escala de medición que se necesita para incluir los valores máximos y mínimos; en otras palabras, la extensión total de los datos.

Ejemplo: *El rango de los datos de la tabla 5.7 es 69 (valor máximo es 95 menos; el valor mínimo es 26). El análisis será: La amplitud de calificación de los estudiantes es de 69 unidades.*

5.5.2 Varianza

La varianza se define como la media de las diferencias cuadráticas de los datos con respecto a su media aritmética y se simboliza con σ^2 . Por lo general es utilizada en los análisis inferenciales y con fines descriptivos, se utiliza la desviación estándar.

Ejemplo: *La varianza de los datos de la tabla 5.7 es 449,95 y es el resultado de la suma del total de los datos menos su media elevados al cuadrado (11248,96) y dividido para el total de datos (25).*

5.5.3 Desviación estándar

La varianza no permite analizar los datos, ya que se mide en unidades elevadas al cuadrado. Para dar una mejor lectura a este valor, se define otra medida de dispersión, que es la desviación típica o desviación estándar, que se calcula sacando la raíz cuadrada de la varianza, este valor es el promedio de desviación de las puntuaciones con respecto a la media, es decir cuánto se desvía en promedio de la media un conjunto de datos.

Ejemplo: La varianza de los datos de la tabla 5.7 es 449,95 y la raíz cuadrada de esta es 21.21, que equivale a la Desviación Estándar.

Análisis: La dispersión de los datos es alta (21.21).

Existen autores que afirman que si la desviación estándar pasa más de 10 unidades, la dispersión es alta.

5.5.4 Para Ejercitar

1. Las medidas de dispersión están conformadas por:

A	Media, moda	
B	Desviación estándar	
C	Mediana, media	
D	Cuartiles	

2. La desviación estándar representa:

A	Dispersión de los datos	
B	La variación entre la media y la mediana	
C	La frecuencia absoluta	
D	Los cuartiles	

3. Calcule las medidas de variabilidad de los datos presentados en el índice de jefatura femenina (IJF) por Provincias, tomadas del Censo 2010, con su respectivo análisis.

ii.

5.6 Medidas De Posición No Centrales

Las medidas de posición no centrales también se conocen como Cuantiles. Es una técnica que agrupa en intervalos iguales la distribución de datos. Tenga en cuenta que para utilizar esta técnica es necesario que los datos estén ordenados de menor a mayor.

Las medidas de posición más conocidas son:

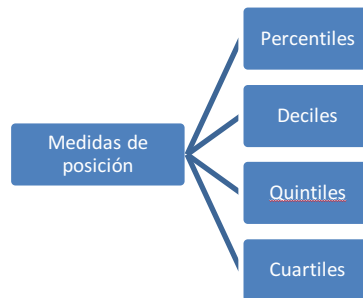


FIGURA 5-4 TIPOS DE MEDIDAS DE POSICIÓN NO CENTRAL

Fuente: Elaboración Propia

Percentiles: Agrupa la distribución de datos en 100 intervalos iguales y cada intervalo contendrá el 1% de los datos.

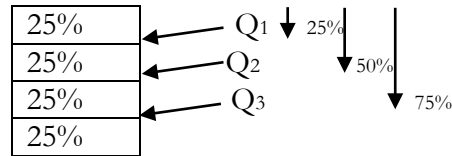
Deciles: Agrupa la distribución de datos en 10 intervalos; cada intervalo contiene el 10% de los datos.

Quintiles: Agrupa la distribución de datos en 5 intervalos; cada uno contendrá el 20% de los datos.

Cuartiles: Agrupa la distribución de datos en 4 intervalos iguales y cada uno contendrá el 25% de los datos. Siendo esta medida de posición la más utilizada.

Cuartiles

Al dividir en 4 partes iguales la distribución de datos generamos tres límites superiores en cada intervalo, conocidos como: Primer Cuartil (Q_1), Segundo Cuartil (Q_2) y Tercer Cuartil (Q_3). Cabe resaltar que el Q_2 es también la mediana.



Para calcular los cuartiles (y demás cuantiles) se puede utilizar las siguientes fórmulas (Zurita, 2010):

$x_m = x_{0,a(n+1)}$ → Ubicación o posición del dato

$x_{(i,a)} = x_i + 0, a * (x_{(i+1)} - x_i)$ → Valor del dato

Dónde:

n = total de datos

m = es el orden del dato del cuartil

x_i = es la posición del dato a ser utilizado

Ejemplo

Tomando los datos Ordenados del Censo 2010 sobre el índice de jefatura femenina por Provincia se realiza el cálculo de los cuartiles:

#	Provincia	IJF 2010
X(1)	Orellana	29
X(2)	Napo	31
3	Santa Elena	31
4	Carchi	33
5	Sucumbíos	33
X(6)	Manabí	34
X(7)	Galápagos	34
8	Los Ríos	35
9	Cotopaxi	37
10	El Oro	38
11	Pichincha	38
X(12)	Tungurahua	38
X(13)	Zamora Chinchipe	38
14	Santo Domingo	40
15	Imbabura	41
16	Morona Santiago	41
17	Pastaza	41
X(18)	Esmeraldas	42
X(19)	Guayas	43
20	Loja	44
X(21)	Bolívar	45
22	Chimborazo	46
23	Cañar	56
24	Azuay	59

Primero se determina el cuartil a hallar, que en este caso será el primer cuartil Q_1 (25% de los datos); luego usamos la formula $x_m = x_{0,a(n+1)}$; en donde $n=24$ datos, entonces:

$$x_m = x_{0,25(24+1)} = x_{6,25}$$

El $x_{6,25}$ significa que el dato en la posición 6,25 sería el primer cuartil, pero en la tabla el puesto 6,25 estaría entre las Provincias de Manabí (Puesto 6) y Galápagos (Puesto 7). Para calcular el valor de este puesto se utiliza la siguiente fórmula:

$$x_{(i,a)} = x_i + 0, a * (x_{(i+1)} - x_i)$$

donde $i=6$ y $a=25$ (25%)

$$x_{(6,25)} = x_6 + 0,25 * (x_{(6+1)} - x_6),$$

en la tabla el puesto x_6 es para Manabí que tiene como I.J.F=34 y el puesto $x_{(6+1)} = x_{(7)}$ es de Galápagos con I.J.F=34, reemplazando en la fórmula se obtiene:

$$x_{(6,25)} = 34 + (0,25 * (34 - 34)) = 34$$

El primer cuartil $Q_1 = 34\%$.

Análisis: El 25% de las provincias del Ecuador tienen menos del 34% de Índice de Jefatura Femenina.

El segundo y tercer cuartil (Q_2 y Q_3) se las calculan de manera similar al primer cuartil y sus análisis son similares.

Segundo Cuartil Q_2 o mediana (50% de datos)	Tercer Cuartil Q_3 (75% de datos)
$x_m = x_{0,a(n+1)}$ $n=24$ y $a=50$	$x_m = x_{0,a(n+1)}$; $n=24$ y $a=50$
$x_m = x_{0,50(24+1)} = x_{12,5}$	$x_m = x_{0,75(24+1)} = x_{18,75}$
Posición 12,5 está entre Provincias de Tungurahua (38) y Zamora Chinchipe (38)	Posición 18,75 está entre Provincias de Esmeraldas (42) y Guayas (43)
$x_{(i,a)} = x_i + 0, a * (x_{(i+1)} - x_i)$	$x_{(i,a)} = x_i + 0, a * (x_{(i+1)} - x_i)$
$x_{(12,5)} = x_{12} + 0,5 * (x_{(12+1)} - x_{12})$	$x_{(18,75)} = x_{18} + 0,75 * (x_{(18+1)} - x_{18})$
$x_{(12,5)} = x_{12} + 0,5 * (x_{(12+1)} - x_{12})$	$x_{(18,75)} = x_{18} + 0,75 * (x_{(18+1)} - x_{18})$
$x_{(12,5)} = 38 + 0,5 * (38 - 38)$	$x_{(18,75)} = 42 + 0,75 * (43 - 42)$
$x_{(12,5)} = 38$	$x_{(18,75)} = 42,75$
<i>El resultado indica que el 50% de las provincias del Ecuador tienen menos del 38% de Índice de Jefatura Femenina.</i>	<i>El resultado indica que el 75% de las provincias del Ecuador tienen menos del 42,75% de Índice de Jefatura Femenina.</i>

La diferencia entre el tercer y primer cuartil se denomina rango intercuartil $RIQ = Q_3 - Q_1$ y proporciona información sobre la variabilidad del 50% de los datos alrededor de la mediana. Los cuartiles son la base para realizar el diagrama de cajas.

5.7 Valores Atípicos o Valores Extremos

En un grupo de datos puede que existan valores atípicos o extremos. Un **valor atípico** es una observación que es numéricamente distante al resto de los datos. Las estadísticas derivadas de los conjuntos de datos que incluyen valores atípicos serán frecuentemente engañosas, dependiendo del tipo de investigación.

Por ejemplo, si tienen las siguientes edades (18, 18, 25, 26, 25, 55, 60), se puede observar que existen dos casos que son más altos que el resto. El promedio de edad de todos los datos es 32,42 años; y si se calcula el promedio sin contar las edades altas (55, 60), la media es 22,4 años, como se puede apreciar la media cambió y se puede observar que es susceptible a valores atípicos.

Algunos investigadores sugieren eliminar los valores atípicos para tener una medición más real de los datos; sin embargo esta decisión debe estar sujeta al tema que se está investigando, pues los datos atípicos pueden dar mayor información y requiere mayor atención.

Para determinar qué valores son atípicos se utilizan los cuartiles en las siguientes fórmulas, los mismos que indicarán las cotas inferiores o superiores que permitan decir que valores son atípicos.

$$\text{Cota Inferior: } Q_1 - 1,5 \cdot (Q_3 - Q_1)$$

$$\text{Cota Superior: } Q_3 + 1,5 \cdot (Q_3 - Q_1)$$

#	Provincia	IJF 2010
1	Orellana	29
2	Napo	31
3	Santa Elena	31
4	Carchi	33
5	Sucumbíos	33
6	Manabí	34
7	Galápagos	34
8	Los Ríos	35
9	Cotopaxi	37
10	El Oro	38
11	Pichincha	38
12	Tungurahua	38
13	Zamora Chinchipe	38
14	Santo Domingo	40
15	Imbabura	41
16	Morona Santiago	41
17	Pastaza	41
18	Esmeraldas	42
19	Guayas	43
20	Loja	44
21	Bolívar	45
22	Chimborazo	46
23	Cañar	56
24	Azuay	59

En el ejemplo que se desarrolla para los cuartiles, se tiene $Q_1 = 34\%$ y $Q_3 = 42,75\%$, entonces las cotas serán:

Cota Inferior:

$$Q_1 - 1,5*(Q_3 - Q_1) = [34 - 1,5*(42,75-34)] = 20,88\%$$

Cota Superior:

$$Q_3 + 1,5*(Q_3 - Q_1) = [42,75 + 1,5*(42,75-34)] = 55,88\%$$

Todo valor que esté por debajo de la cota inferior (20,88%) o que esté por encima de la cota superior (55,88%) serán considerados valores atípicos o extremos, y para los datos de Índice de Jefatura Femenina podemos observar que en la cota inferior no hay valores atípicos por cuánto el valor mínimo de estos datos es 29%. Para el caso de la cota superior se puede observar que existen dos valores atípicos, ya que sobrepasan la cota superior.

Para el análisis se indica que existen valores atípicos inferiores o superiores y se

Menciona cuántos y cuáles son.

5.7.1 Para Ejercitar

La siguiente tabla refleja el % de estado civil “Unido” de las Parroquias Urbanas de Guayaquil, Censo 2010.

Parroquia Urbana	Estado Civil Unido/a
Pedro Carbo	12,00 %
Rocafuerte	12,80 %
Nueve de Octubre	15,06 %
Sucre	15,14 %
Roca	15,22 %
Olmedo	15,58 %
Ayacucho	15,99 %
Bolívar	16,65 %
Chongón	17,76 %
García Moreno	19,19 %
Urdaneta	19,53 %
Letamendi	20,33 %
Tarqui	23,08 %
Ximena	23,31 %
Febres Cordero	23,60 %
Pascuales	34,24 %

Estadística para las Ciencias Sociales

1. El análisis correcto de los cuartiles primero y tercero obtenidos en base a los datos del Estado civil “Unido” es:

A	El 25% de Estado civil “Unido” de las parroquias urbanas de Guayaquil es menos del 15,18% y el 75% de Estado civil “Unido” de las parroquias urbanas de Guayaquil es menos del 21,71%	
B	El 25% de la parroquias urbanas de Guayaquil, tienen menos del 15,18% de Estado civil “Unido” y el 75% de la parroquias urbanas de Guayaquil, tienen menos del 21,71% de Estado civil “Unido”	
C	El 25% de Estado civil “Unido” de las parroquias urbanas de Guayaquil es menos del 15,14% y el 75% de Estado civil “Unido” de las parroquias urbanas de Guayaquil es menos del 20,33%	
D	l El 25% de la parroquias urbanas de Guayaquil, tienen menos del 15,14% de Estado civil “Unido” y el 75% de la parroquias urbanas de Guayaquil, tienen menos del 20,33% de Estado civil “Unido”	

5.8 Gráficos

5.8.1 Histogramas

El histograma es un gráfico que se desarrolla a partir de la tabla de frecuencias utilizando el Intervalo y la frecuencia relativa o porcentajes. Permite observar cuál es la tendencia de los datos, si estos tienen algún tipo de sesgo o se agrupan en algún intervalo o presentan una tendencia.

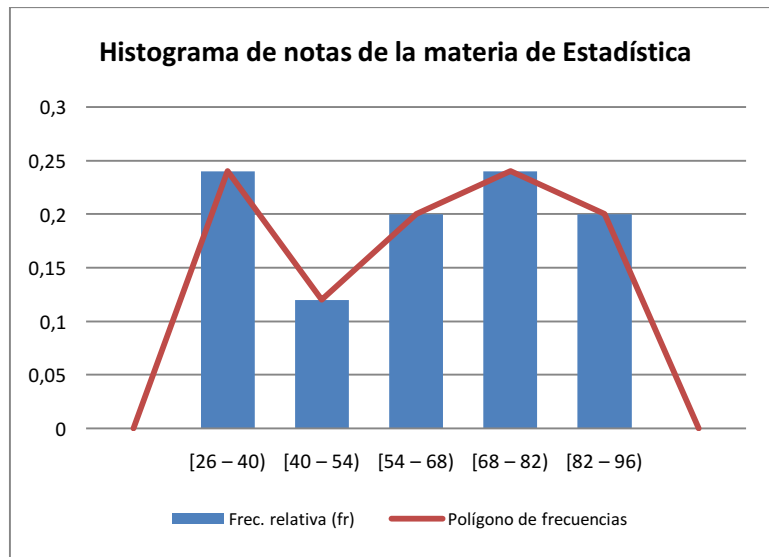


FIGURA 5-5 HISTOGRAMA DE NOTAS

Fuente: Elaboración Propia

En la figura anterior se puede observar cómo los datos tienen mayor agrupación a partir de las notas mayores a 54, el mismo que está resaltado por el polígono de frecuencias que resulta de utilizar la marca de clase y la frecuencia relativa de la tabla de frecuencias.

5.8.2 Ojiva

La ojiva se elabora a partir de la frecuencia acumulada y se utiliza para determinar si las frecuencias en cada clase crecen de una manera acelerada o no. En el gráfico de la ojiva se puede observar que las notas crecen de una manera moderada.

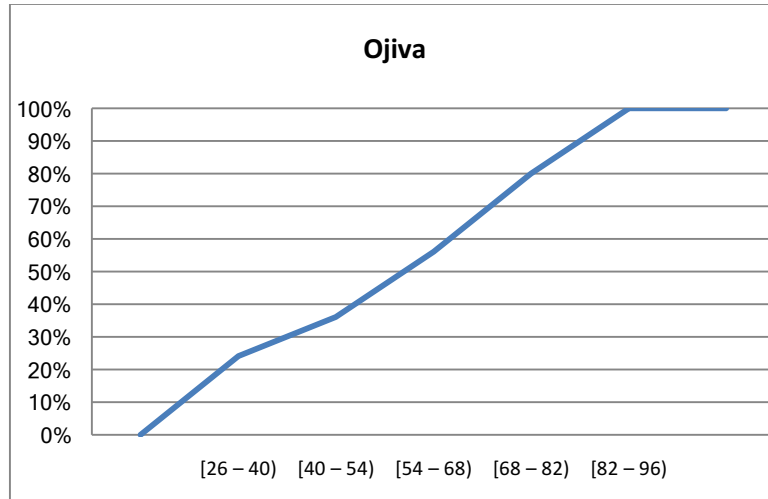


FIGURA 5-6 OJIVA DE LAS NOTAS

Fuente: Elaboración Propia

En el siguiente gráfico se han alterado los datos para poder observar ejemplos de ojivas cuyo crecimiento de los datos pueden ser acelerados (ejemplo 1) o de manera lenta (ejemplo 2).

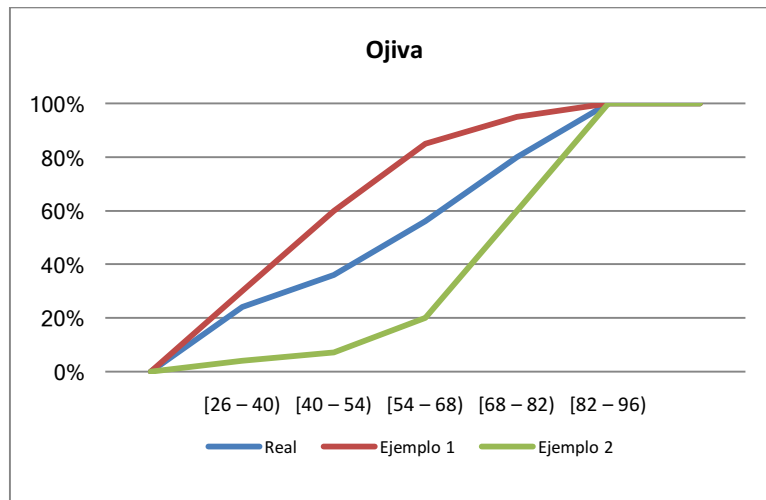


FIGURA 5-7 EJEMPLO DE OJIVAS DE LAS NOTAS
Fuente: Elaboración Propia

5.8.3 Diagrama de cajas

El diagrama de cajas se elabora a partir de los cuartiles y permite observar cómo se agrupan los datos; y la caja que se representa abarca el 50% de los datos desde el primer cuartil hasta el tercer cuartil. Los tramos que une el valor mínimo con el primer cuartil y el tercer cuartil con el valor máximo se denominan bigotes, pero si existen valores extremos estos serán graficados como puntos y el bigote que une Q3 con el valor máximo se acortará hasta donde está el límite para indicar que a partir de él, los valores que sobrepasen son valores atípicos.

En el gráfico del diagrama de cajas elaborado a partir de los datos de estado civil unido de las parroquias urbanas de Guayaquil se observa que el valor mínimo es 12% el máximo es 34,24%, el primer cuartil (Q1) es 15,16%, el segundo cuartil

(Q2 o mediana) es 17,21% y el tercer cuartil es 23,39%. Adicionalmente se puede observar que dentro de este grupo de datos existe un valor atípico (34,24%) y que es graficado como un punto.

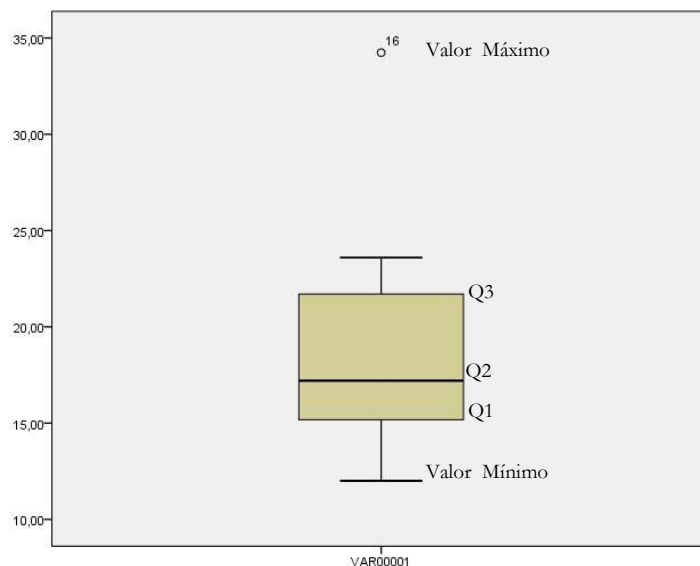


FIGURA 5-8 EJEMPLO DE DIAGRAMA DE CAJAS

Fuente: Elaboración Propia

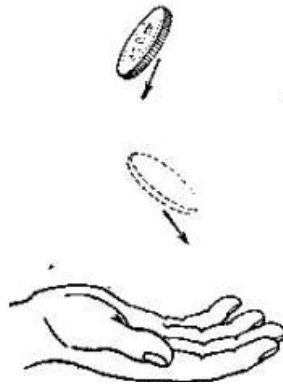


CAPÍTULO SEIS

PROBABILIDAD

“Para el hombre solo la muerte ocurre con probabilidad uno, todo lo demás es estocástico y sujeto de variabilidad”

Gaudencio Zurita





6. PROBABILIDAD

6.1 Introducción

El pensamiento “para el hombre solo la muerte ocurre con probabilidad uno, todo lo demás es estocástico y sujeto de variabilidad”, ratifica ver que la muerte es segura para todo ser humano, aunque no se conozca la fecha de este evento (variable aleatoria) ni cómo va a ocurrir; para el resto de eventos de nuestra vida la probabilidad de que ocurra es variable y depende de varios factores tanto internos como externos del ser humano; ejemplo: si se quiere graduar de una carrera universitaria en un lapso de 5 años, deberá calcular la probabilidad de ocurrencia de este evento, el cual podría ser uno (100%), si es aplicado en los estudios, si las normativas de la institución no cambian, si los docentes son los adecuados para la enseñanza, si su salud física y mental no decae, si las posibilidades económicas lo permiten y si aprueba todas las materias; pero si algunos de estos factores no se cumplen, la probabilidad de graduarse comienza a decrecer.

El presente capítulo trata sobre probabilidades de ocurrencia de eventos, cómo generarlas y cómo analizarlas de una manera más sencilla; también se contará con ejercicios de prácticas tomados de datos reales.

6.2 Probabilidad

La probabilidad se define como “el número real $P(E)$ de que un evento E ocurra, bajo las condiciones impuestas en el

experimento”, (Zurita, 2010), en otras palabras es un valor o número que se asigna al suceso o aparición de un evento.

Un ejemplo típico y sencillo es lanzar una moneda al aire y anotar el evento E; pero antes de que caiga la moneda se tendría dos eventos posibles E (cara, sello); basados en estos dos eventos la probabilidad de que sea cara será de 0.5 (50%)

$$P(\text{cara}) = \frac{\# \text{ de caras de la moneda}}{\# \text{ total de eventos}} = \frac{1}{2} = 0.5$$

La ecuación anterior proviene de la Regla de Laplace que permite calcular la probabilidad de un suceso, siempre que los eventos tengan la misma probabilidad de ocurrir; y para calcular la probabilidad de un evento A este se obtiene de dividir el número de resultados que forman el evento A para el número de eventos totales con que se cuenta.

Regla de Laplace: $P(A) = \frac{\# \text{ de eventos favorables a A}}{\# \text{ total de eventos}}$

Otro ejemplo es calcular la probabilidad de que al lanzar un dado este salga 5, para ello se debe hacer las siguientes preguntas:

- ¿Cuántos eventos pueden darse en un dado?
- ¿Cuántos 5 pueden existir en el dado?

El total de eventos (6) que se pueden dar en un dado son E(1, 2, 3, 4, 5, 6); y si se observa el dado apenas tiene un 5 en sus eventos; entonces la P(5) es:

$$P(5) = \frac{\# \text{ de 5 en el dado}}{\# \text{ total de eventos}} = \frac{1}{6} = 0.1667 \text{ o } 16,67\%$$

La probabilidad es un valor que siempre va a estar entre 0 (no ocurre el evento) y 1 (siempre ocurrirá). Si el valor de la probabilidad es mayor que 1 o es negativo, esto indica que algo se hizo mal. Por lo general, la presentación del resultado de la probabilidad es en decimales (0.5 o 0.1667) pero para el análisis es más sencillo trabajar en porcentajes, para lo cual se multiplica el resultado obtenido de la probabilidad por 100 (similar al cálculo de la frecuencia relativa de la tabla de frecuencias). En el caso de los dados la $P(5)$ es del 16,67%, lo que significa que si se lanza el dado 100 veces 16,67 (o 17) veces saldrá el número 5 o si se lanza 1000 veces saldrá 166,7 (o 167) veces el 5.

6.3 Reglas de Probabilidad

Regla rango de la probabilidad: Dado un evento E , la probabilidad de que ocurra ese evento está entre 0 y 1.

$$0 \leq P(E) \leq 1$$

En otras palabras, esta regla nos indica que la probabilidad nunca va a ser negativo ni mayor a 1.

Regla de la probabilidad 1: La suma de todas las probabilidades de un evento E , siempre será 1.

$$\sum P(E) = 1$$

Ejemplo: El dado tiene 6 eventos $E(1, 2, 3, 4, 5, 6)$, la probabilidad de cada uno de los eventos es:

$$P(1)=P(2)=\dots=P(6)=\frac{1}{6}$$

Al sumar cada probabilidad el resultado será siempre 1

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

Regla del complemento: La probabilidad del complemento de un evento E, es igual a la diferencia entre 1 menos la probabilidad del evento.

$$P(E^c) = 1 - P(E)$$

Ejemplo: Se lanza un dado ¿cuál sería la probabilidad de 5 complemento $P(5^c)$?

$$P(5^c) = 1 - P(5) = 1 - \frac{1}{6} = \frac{5}{6}$$

Regla de suma de eventos mutuamente no excluyentes: Dado dos eventos E1 y E2, la probabilidad de que ocurra E1 o E2, es igual a la suma de la probabilidad de cada evento menos la probabilidad de la intersección de los dos eventos.

$$P(E1 \cup E2) = P(E1) + P(E2) - P(E1 \cap E2)$$

Cabe resaltar que al referirse de no excluyentes es que los dos eventos pueden ocurrir a la vez.

Regla de suma de eventos mutuamente excluyentes: Dado dos eventos E1 y E2, la probabilidad de que ocurra E1 o E2 es igual a la suma de la probabilidad de cada evento.

$$P(E1 \cup E2) = P(E1) + P(E2)$$

Cabe resaltar que al referirse a excluyentes, los dos eventos no pueden ocurrir a la vez; es decir que la intersección de ambos eventos es 0

$$P(E1 \cap E2) = 0$$

Ejemplo: Se lanza un dado ¿cuál sería la probabilidad que salga 5 o 6?

$$P(5 \cup 6) = P(5) + P(6) - P(5 \cap 6)$$

$$P(5 \cup 6) = \frac{1}{6} + \frac{1}{6} - 0 = \frac{2}{6} = \frac{1}{3} = 0,333 = 33,33\%$$

$P(5 \cap 6)$ es 0 por cuanto nunca el dado va a salir 5 y 6 a la vez.

Regla de probabilidad condicional de dos eventos dependientes: Dado dos eventos $E1$ y $E2$, se denomina probabilidad condicional a la probabilidad de que ocurra el evento $E1$, dado que ya ocurrió el evento $E2$ y se calcula:

$$P(E1/E2) = \frac{P(E1 \cap E2)}{P(E2)}$$

Tenga en consideración que $P(E2)$ es mayor que cero.

Ejemplo: A una reunión asisten 20 personas, de las que 11 son hombres y 9 mujeres. Si 5 hombres y 3 mujeres llevan gafas y escogemos una persona al azar, calcula la probabilidad de que:

- sea una mujer y lleve gafas,
- sea una mujer sabiendo que lleva gafas.

	Gafas	Sin Gafas	
Hombre	5	6	11
Mujer	3	6	9
TOTAL	8	12	20

a) De las 20 personas 3 son mujeres con gafas luego la probabilidad de que la persona elegida sea una mujer con gafas es:

$$P(\text{MUJERES} \cap \text{GAFAS}) = 3/20 = 0,15$$

b) En este caso el suceso "ser mujer" está condicionado por el suceso "llevar gafas", elegimos una persona de entre las que llevan gafas. De las 8 que hay, 3 son mujeres, por tanto la probabilidad es:

$$P(\text{MUJERES} / \text{GAFAS}) = \frac{P(\text{M} \cap \text{G})}{P(\text{G})} = \frac{3/20}{8/20} = \frac{3}{8}$$

Regla de probabilidad condicional de dos eventos independientes: Dado dos eventos E1 y E2, se denomina probabilidad condicional a la probabilidad de que ocurra el evento E1, dado que ya ocurrió el evento E2 y se calcula:

$$P(E1/E2) = P(E1)$$

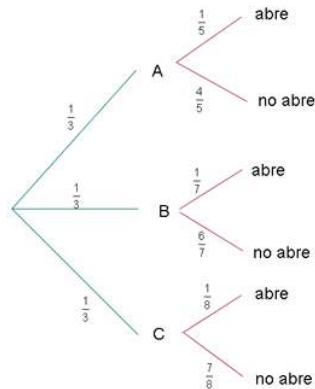
Regla de multiplicación para dos eventos dependientes E1 y E2: Dado dos eventos E1 y E2, la probabilidad de que ocurra ambos eventos a la vez $P(E1 \cap E2)$ es igual a la multiplicación de la probabilidad del evento E1 por la probabilidad del evento E2, dado que ocurrió el evento E1.

$$P(E1 \cap E2) = P(E1) P(E2/E1)$$

Ejemplo: En una casa hay tres llaveros A, B y C; el primero con cinco llaves, el segundo con siete y el tercero con ocho, de las que sólo una de cada llavero abre la puerta trasera. Se escoge al azar un llavero y, de él una llave para abrir la puerta trasera.

¿Cuál será la probabilidad de que el llavero escogido sea el tercero y la llave no abra?

1. Realizar los diferentes eventos posibles



2. Calcular la probabilidad:

$$P(\text{Llavero C} \cap \text{no abre}) = P(\text{Llavero C}) * P(\text{no abre}) = \frac{1}{3} * \frac{7}{8} = 0.2917$$

Regla de multiplicación para dos eventos independientes

E1 y E2: Dado dos eventos E1 y E2, la probabilidad de que ocurra ambos eventos a la vez $P(E1 \cap E2)$ es igual a la multiplicación de las probabilidades de ambos eventos.

$$P(E1 \cap E2) = P(E1) P(E2)$$

6.3.1 Para Ejercitar

1. Una urna tiene ocho bolas rojas, 5 amarilla y siete verdes. Si se extrae una bola al azar calcular la probabilidad de:

- Sea roja.
- Sea verde.
- Sea amarilla.
- No sea roja.
- No sea amarilla.

2. Se extrae una bola de una urna que contiene 4 bolas rojas, 5 blancas y 6 negras, ¿cuál es la probabilidad de que la bola sea roja o blanca? ¿Cuál es la probabilidad de que no sea blanca?

3. En una clase hay 10 alumnas rubias, 20 morenas, cinco alumnos rubios y 10 morenos. Un día asisten 45 alumnos, encontrar la probabilidad de que un alumno:

- Sea hombre.
- Sea mujer morena.
- Sea hombre o mujer.

4. Se lanzan dos dados al aire y se anota la suma de los puntos obtenidos. Se pide:

- La probabilidad de que salga el 7.
- La probabilidad de que el número obtenido sea par.
- La probabilidad de que el número obtenido sea múltiplo de tres.

5. Una clase consta de seis niñas y 10 niños. Si se escoge un comité de tres al azar, hallar la probabilidad de:

- Seleccionar tres niños
- Seleccionar exactamente dos niños y una niña
- Seleccionar por lo menos un niño
- Seleccionar exactamente dos niñas y un niño

6. En una casa hay tres llaveros A, B y C; el primero con cinco llaves, el segundo con siete y el tercero con ocho, de las que sólo una de cada llavero abre la puerta trasera. Se escoge al azar un llavero y, de él una llave para abrir la puerta trasera. Y si la llave escogida es la correcta, ¿cuál será la probabilidad de que pertenezca al llavero A?

7. Un estudiante cuenta, para un examen con la ayuda de un despertador, el cual consigue despertarlo en un 80% de los casos. Si oye el despertador, la probabilidad de que realiza el examen es 0.9 y, en caso contrario, de 0.5.

- Si va a realizar el examen, ¿cuál es la probabilidad de que haya oído el despertador?
- Si no realiza el examen, ¿cuál es la probabilidad de que no haya oído el despertador?



CAPÍTULO SIETE

USANDO SOFTWARE PARA ESTADÍSTICA DESCRIPTIVA

“El modo más efectivo de fortalecer nuestra comunidad es diseminar la comprensión del valor de la libertad -enseñar a más gente la inaceptabilidad moral del software no-libre. La gente que valore la libertad es, a largo plazo, su mejor y esencial defensa.”

Richard Stallman





7. USANDO SOFTWARE PARA ESTADÍSTICA DESCRIPTIVA

7.1 Introducción

Existen en el mercado varios software que permiten realizar procesos estadísticos y en la mayoría de los casos son similares como: SPSS, STATA, MINITAB, PSPP, R, EXCEL, etc.

Esta unidad le permitirá aprender el uso básico de tres software importando una base de datos desde Excel: SPSS, R y Excel.

7.2 Estadística descriptiva con SPSS

7.2.1 ¿Qué es SPSS?

SPSS son las siglas de Statistical Package for the Social Sciences, que en su traducción sería “Paquete Estadístico para las Ciencias Sociales”. Se trata de un software estadístico que se emplea muy a menudo en las ciencias sociales y por las empresas y profesionales de investigación de mercados. En la versión 12, este programa era capaz de operar hasta con 2 millones de registros y 250.000 variables. Este software no es gratuito.

7.2.2 Importar datos desde EXCEL

Por lo general, las bases de datos están en EXCEL y se importara a SPSS. Para realizar este proceso se siguen los siguientes pasos:

1. Se crea una base en EXCEL y se guardan; los datos son:

#	Sector	sexo	Edad	Calificación
1	FLOR DE BASTION	MUJER	27	7
2	FLOR DE BASTION	HOMBRE		8
3	FLOR DE BASTION	HOMBRE	22	5
4	FLOR DE BASTION	MUJER	21	9
5	FLOR DE BASTION	MUJER	1	10
6	FLOR DE BASTION	HOMBRE	43	2
7	FLOR DE BASTION	MUJER	69	3
8	FLOR DE BASTION	MUJER	51	4
9	FLOR DE BASTION	HOMBRE	55	9
10	FLOR DE BASTION	MUJER	31	8
11	FLOR DE BASTION	MUJER	11	7
12	FLOR DE BASTION	HOMBRE	9	6

2. Se abre SPSS, se escoge “Abrir otro archivo” (Figura 7-1); si es la primera vez que se va a abrir el archivo de EXCEL. Si se trabajó con anterioridad queda grabado un acceso directo al archivo y solo tendrá que hacer click en el nombre del archivo.

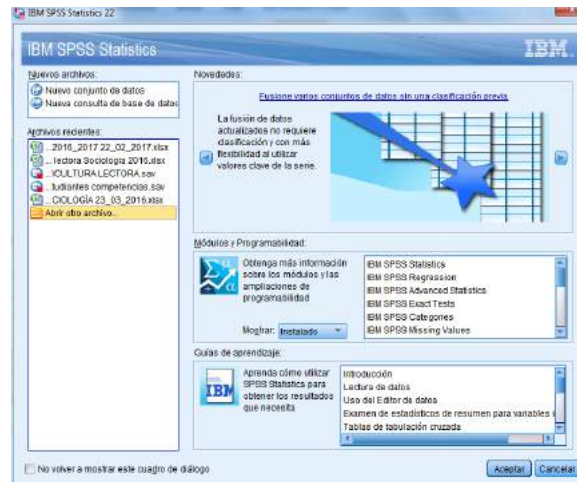


FIGURA 7-1 ABRIR ARCHIVO DESDE SPSS

3. Se busca el lugar donde está guardado el archivo en EXCEL y en la opción “Archivos de tipo” se seleccionan los archivos de tipo “Datos de Excel” y se presentarán los archivos de Excel y se selecciona el archivo creado.

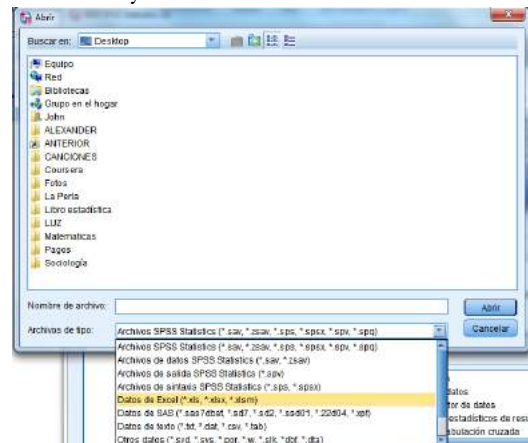


FIGURA 7-2 ABRIR ARCHIVO TIPO EXCEL

4. Una vez abierto el archivo, se presenta la ventana “Apertura de origen de datos de Excel”, en donde se señalará la hoja con la cual se va a trabajar (Figura 7-3). Se selecciona “Aceptar”.

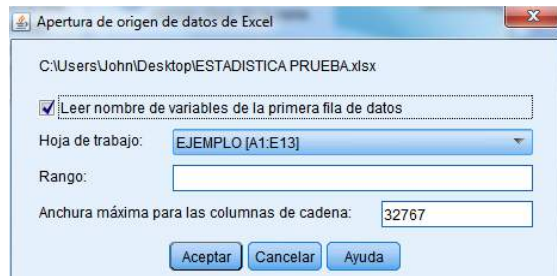


FIGURA 7-3 SELECCIÓN DE HOJA DE TRABAJO

5. Se activan dos ventanas “Resultado1” y “Sin título1”. Esta última contiene dos viñetas que dan acceso a las opciones “Vista de datos” y “Vista de variables”.

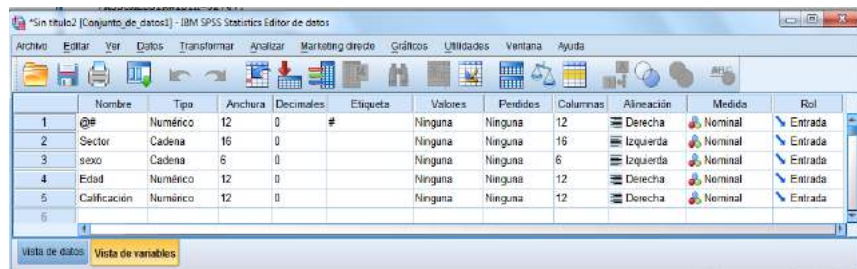


FIGURA 7-4 VENTANA DE INGRESO DE VARIABLES

6. En “Vista de variables” se observan las opciones:
 - **Nombre:** Se presentan el nombre de las variables que se importaron desde Excel. Se pueden cambiar si se requiere. Esta opción no permite que los caracteres estén separados, es decir si una variable lleva por nombre

“nivel de instrucción”, esta opción no la aceptará y los unirá automáticamente (niveldeinstrucción) o le pondrá valores como V1, V2, etc. Si se requiere poner o describir la variable se lo puede hacer en “Etiqueta”.

- **Tipo:** Indica el tipo de nuestra variable (Numérico, Cadena, etc.). Si se hace click se abrirá una ventana en donde se podrá seleccionar el tipo de variable.

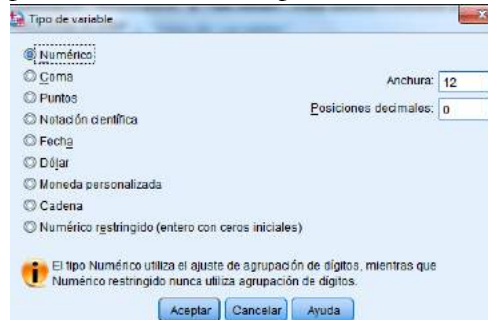


FIGURA 7-5 TIPO DE VARIABLE

- **Anchura:** Indica el ancho de la variable. Por ejemplo, si es de tipo de cadena se podría la opción del número de caracteres que se requiere que estén en esta variable.
- **Decimales:** Se pone el número de decimales que desea que estén presentes en nuestras variables.
- **Etiqueta:** Se describe la variable o se coloca el nombre completo de la variable o pregunta.
- **Valores:** Esta opción permite poner valores a los datos en caso de que estos estén codificados. Por ejemplo, si la variable es sexo y el código 1 representa hombre y el 2 mujeres, aquí se codifica y SPSS cuando realice la tabulación de los datos en la presentación de los resultados, lo hará con los nombres asignados.



FIGURA 7-6 ETIQUETAS DE VALOR

- **Perdidos:** SPSS maneja dos tipos de valores perdidos; el primero es perdido por el sistema cuando no existen datos; es decir, casillas vacías; el programa detecta automáticamente los valores perdidos por el sistema y los omite; y el segundo corresponde a los datos perdidos definidos por el usuario (No sabe, No responde o No aplica), estos valores perdidos deben ser definidos al programa o de lo contrario los cálculos que se realizarán contarán con estos valores, lo cual puede afectar severamente los resultados.

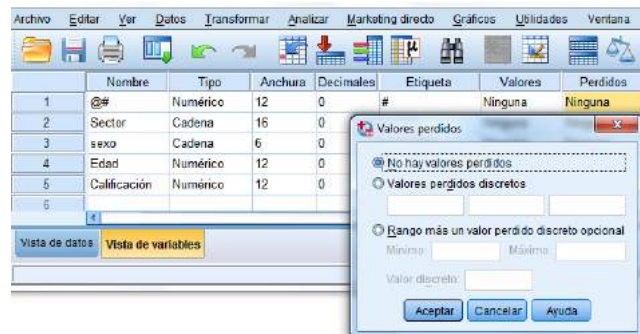


FIGURA 7-7 VALORES PERDIDOS

- **Medida:** Permite definir si la variable es de tipo Escala, Ordinal o Nominal

Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida
@#	Numérico	12	0	#	Ninguna	Ninguna	12	Derecha	Nominal
Sector	Cadena	16	0		Ninguna	Ninguna	16	Izquierda	Escala
sexo	Cadena	6	0		Ninguna	Ninguna	6	Izquierda	Ordinal
Edad	Numérico	12	0		Ninguna	Ninguna	12	Derecha	Nominal
Calificación	Numérico	12	0		Ninguna	Ninguna	12	Derecha	Nominal

FIGURA 7-8 MEDIDA

7. En “Vista de datos” se observan los datos de la base importada o creada.

	@#	Sector	sexo	Edad	Calificación
1		1 FLOR DE BASTION	MUJER	27	7
2		2 FLOR DE BASTION	HOMBRE		8
3		3 FLOR DE BASTION	HOMBRE	22	5
4		4 FLOR DE BASTION	MUJER	21	9
5		5 FLOR DE BASTION	MUJER		6

FIGURA 7-9 VISTA DE DATOS

7.2.3 Realizando estadísticas descriptivas

1. Una vez importada la base desde Excel, se va a la ventana de “Analizar”, se selecciona “Estadística Descriptiva” y luego “Descriptivos”.

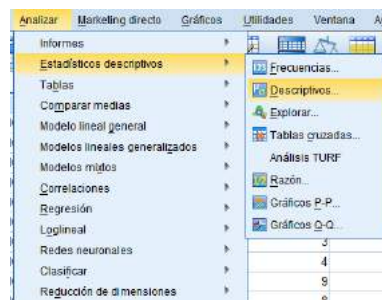


FIGURA 7-10 ESTADÍSTICOS DESCRIPTIVOS

2. En “Descriptivos” se presenta una ventana que permitirá seleccionar las variables que se desean analizar, para lo cual,

se selecciona la variable o variables y con la flecha se ubica en el cuadro “Variable” o arrastra las variables hasta este cuadro.

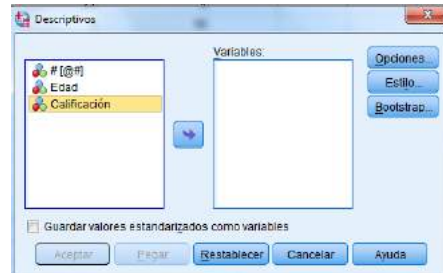


FIGURA 7-11 SELECCIÓN DE VARIABLES

3. Luego, se va a opciones y se activará otra ventana en donde se escogerá qué tipo de estadística se desea para la variable (Calificación), se seleccionan las opciones que se quiere y se regresará a la otra ventana poniendo “Continuar” y después aceptar.

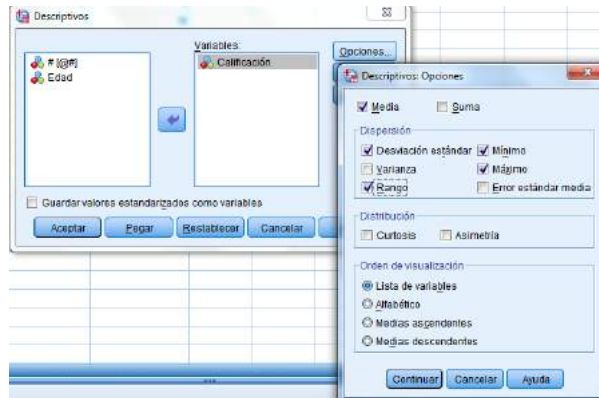


FIGURA 7-12 OPCIONES DE ESTADÍSTICOS DESCRIPTIVOS

4. Los resultados se presentan en la ventana “Resultado1” como un cuadro de datos.

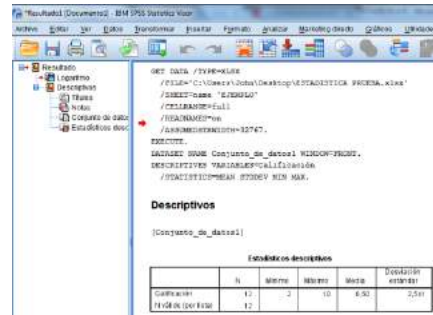


FIGURA 7-13 RESULTADOS

- Otro de los comandos para obtener resultados se denomina “Frecuencias”. Se va a la ventana de “Analizar”, y se selecciona “Estadística Descriptiva” y luego “Frecuencias”.

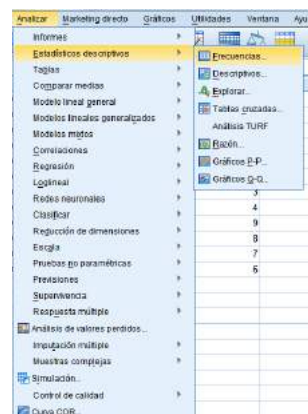


FIGURA 7-14 FRECUENCIAS

- En “Frecuencia” se presenta una ventana que permitirá seleccionar las variables que se desea analizar, para lo cual se selecciona la variable o variables y con la flecha se las ubica en el cuadro “Variable” o se arrastran las variables hasta este cuadro.

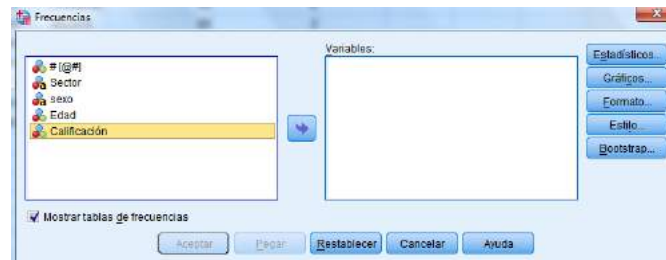


FIGURA 7-15 SELECCIÓN DE VARIABLES

7. Luego se va a “Estadísticos” y se activará otra ventana en donde seleccionaremos que tipo de estadística deseamos para nuestra variable (Calificación), seleccionamos las opciones que queremos y regresamos a la otra ventana poniendo “Continuar” y después ponemos aceptar.

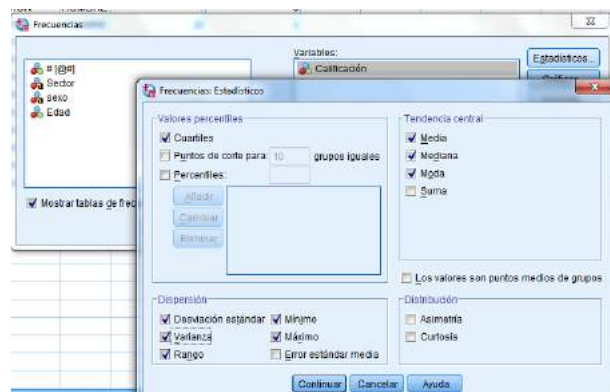


FIGURA 7-16 SELECCIÓN DE ESTADÍSTICOS

8. Los resultados se presentan en la ventana “Resultado1” como un cuadro de datos.

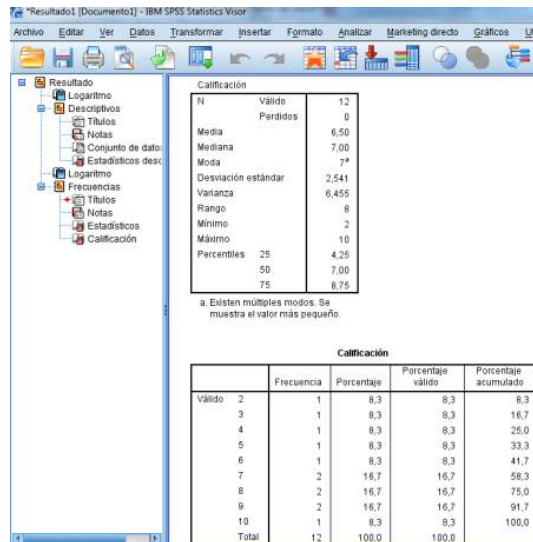


FIGURA 7-17 RESULTADOS ESTADÍSTICOS

9. Se puede cruzar variables como por ejemplo sexo con notas. Para esto vamos a la ventana de “Análisis”, seleccionamos “Estadística Descriptiva” y luego “Tablas cruzadas”.

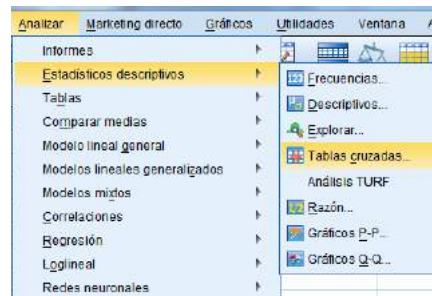


FIGURA 7-18 TABLAS CRUZADAS

10. En la ventana “Tablas cruzadas” se selecciona las variables con las que se va a trabajar y se ubican en las opciones que se deseen presentar por filas y columnas.

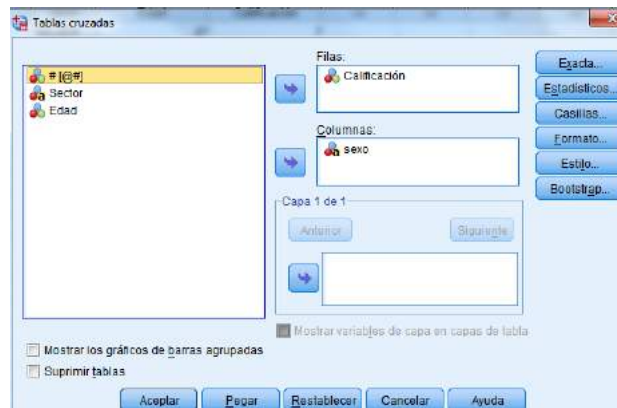


FIGURA 7-19 SELECCIÓN DE VARIABLES

11. La tabla cruzada se presenta en la ventana “Resultado1”.

Tablas cruzadas

Resumen de procesamiento de casos					
	Casos				
	Válido		Perdidos		Total
	N	Porcentaje	N	Porcentaje	N
Calificación * sexo	12	100,0%	0	0,0%	12

Calificación * sexo tabulación cruzada				
Recuento		sexo		Total
		HOMBRE	MUJER	
Calificación	2	1	0	1
	3	0	1	1
	4	0	1	1
	5	1	0	1
	6	1	0	1
	7	0	2	2
	8	1	1	2
	9	1	1	2
	10	0	1	1
Total		5	7	12

FIGURA 7-20 RESULTADOS DE TABLAS CRUZADAS

7.3 Estadística descriptiva con RSTUDIO

7.3.1 ¿Qué es R?

R es un conjunto integrado de servicios de software para la manipulación de datos, cálculo y representación gráfica y fue desarrollado en los Laboratorios Bell (antes de AT & T, ahora Lucent Technologies) por John Chambers y colegas.

R ofrece una amplia variedad de técnicas gráficas estadística (lineal y no lineal de modelado, pruebas estadísticas clásicas, análisis de series temporales, clasificación, agrupamiento, etc); está disponible como software libre bajo los términos de la Free Software Foundation 's Licencia Pública General de GNU en forma de código fuente. Se compila y se ejecuta en una amplia variedad de plataformas UNIX y sistemas similares (incluyendo FreeBSD y Linux), Windows y MacOS.

R, está diseñado en torno a un cierto lenguaje informático, y permite a los usuarios añadir funcionalidad adicional mediante la definición de nuevas funciones.

7.3.2 ¿Qué es RStudio?

RStudio es un entorno de desarrollo integrado (IDE) para R (lenguaje de programación) que hace más fácil el manejo del software R. Es decir permite manejar R de manera más dinámica.

7.3.3 Instalando R STUDIO

1. Primero debemos instalar R y lo puede descargar gratuitamente en la página web: <https://www.r-project.org/about.html>



FIGURA 7-21 PÁGINA WEB R

- Una vez que está cargada la página haga click en la opción “CRAN” y se abrirá otra página, en donde buscará el país y la entidad que tiene el acceso al programa. Para nuestro caso vamos a “Ecuador” y se tendrá la opción de CRAN de la ESPOL, hacemos click en este link.



FIGURA 7-22 SELECCIÓN DE PAÍS

- Se carga la página con el link: <http://cran.espol.edu.ec/> en donde se podrá descargar el programa según el modelo del computador. En este caso se descarga “Download R for Windows”

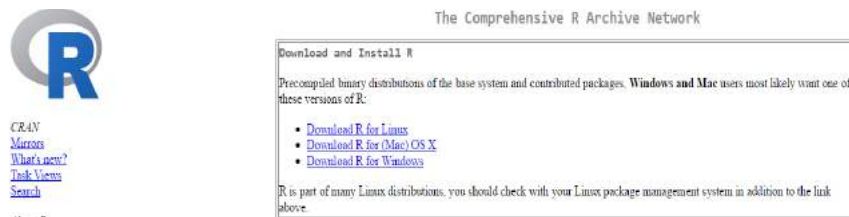


FIGURA 7-23 DESCARGAR R

4. Se cargará la página que se verá a continuación y en donde se seleccionará “install R for the first time”

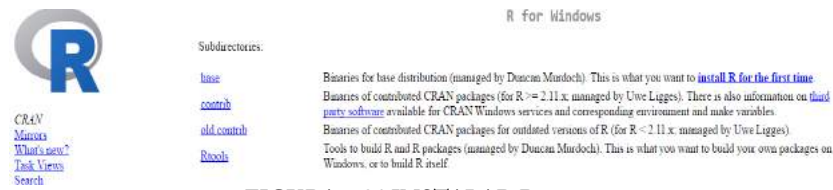


FIGURA 7-24 INSTALAR R

5. Luego se cargará la página que se verá a continuación y en donde se seleccionará “Download R 3.4.0 for Windows R” y se descargará un archivo que se debe activarlo.



FIGURA 7-25 DESCARGANDO R PARA WINDOWS

6. Al activarse el programa se selecciona la opción “Español”, y se presiona aceptar.

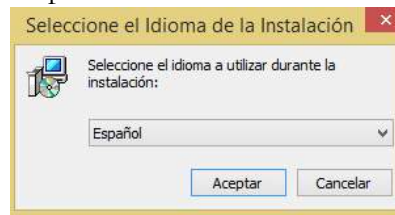


FIGURA 7-26 SELECCIONANDO IDIOMA

7. En las ventanas que salgan, se selecciona “Siguiente” hasta que empiece a ejecutarse el instalador y aparezca la ventana que indique que se ha completado la instalación.



FIGURA 7-27 INSTALANDO R



FIGURA 7-28 INSTALACIÓN COMPLETA

8. En el escritorio de su computadora estará el acceso directo del programa. Hacer click y el programa se ejecutará.



FIGURA 7-29 ICONO DE R

- Una vez instalado R se instala RStudio, para lo cual se debe ir a la página Web de RSTUDIO en www.rstudio.com, y hacer click en “Download RStudio”, luego click en “Desktop” y se habrá instalado RStudio.
- Una vez instalado R STUDIO, se abrirá una ventana que estará dividida en cuatro paneles:
 - El panel del lado izquierdo superior es el script donde se presentarán los datos.
 - El panel del lado izquierdo inferior “Console” es la consola donde se pondrán los comandos para calcular la estadística de las variables y en donde se presentarán los resultados.
 - En el panel del lado derecho superior está el historial del RStudio y su entorno.
 - En el panel del lado derecho inferior se visualizan los archivos con los cuales se ha trabajado.

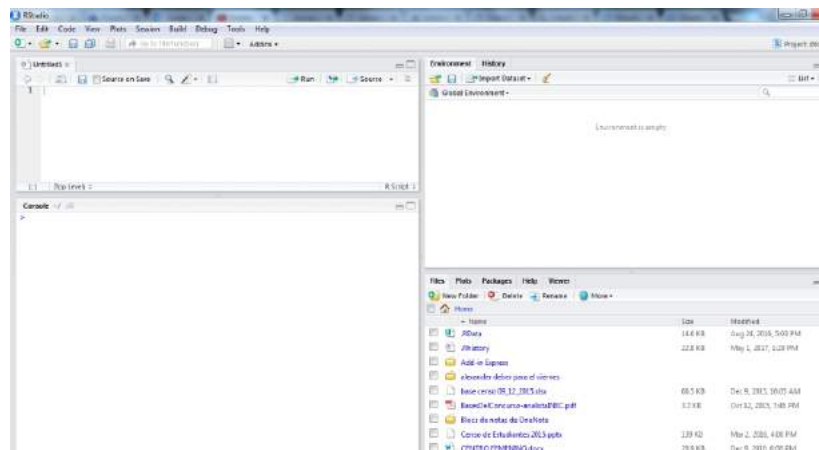


FIGURA 7-30 PANELES RSTUDIO

7.3.4 Importar datos desde EXCEL

Para importar de EXCEL a RStudio se realiza lo siguiente:

1. Trabajar con el mismo archivo de Excel que se creó para SPSS; y se guarda como tipo CVS (MS-DOS) o como CVS(Delimitado por comas).

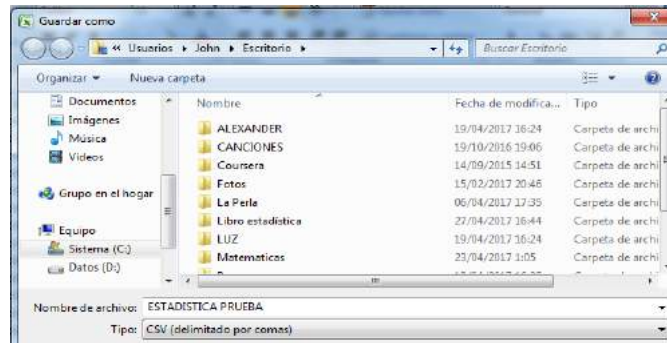


FIGURA 7-31 GUARDANDO EXCEL COMO TIPO CSV

2. En R Studio ir a Import Dataset y seleccionar “From Local File”, para que abra una ventana en donde se buscará el archivo de EXCEL y dar click en abrir.

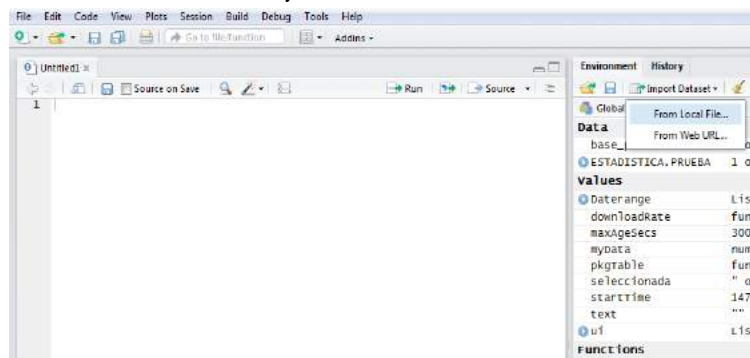


FIGURA 7-32 BUSCANDO ARCHIVO EXCEL

3. Se abrirá la ventana “Import Dataset”; luego seleccionar “Import” para abrir la base.

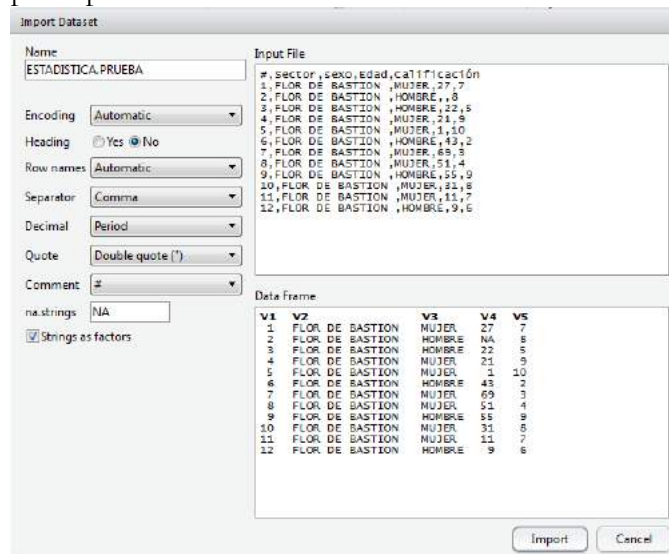


FIGURA 7-33 IMPORTANDO DESDE EXCEL

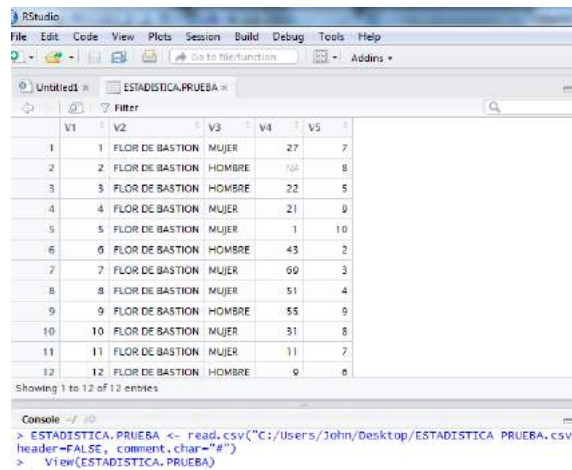


FIGURA 7-34 DATOS IMPORTADOS EN RSTUDIO

7.3.5 Realizando estadística descriptiva

A continuación se enunciarán algunos comandos para realizar la estadística descriptiva de los datos:

1. **attach()**: Para trabajar con la base de datos hay que separar las variables; para esto, en la Consola se deberá utilizar el comando `attach()`, que nos permitirá leer por separado las variables y dentro del paréntesis se colocará el nombre de la base de datos que se tenga.

```
>attach(ESTADISTICA.PRUEBA)
```

2. **max()**: Este comando permite ver el valor máximo de la variable. Para esta base la variable V5 es la calificación obtenida por los alumnos. Al activar el comando y dar enter saldrá el resultado que es 10.

```
> attach(ESTADISTICA.PRUEBA)
> max(V5)
[1] 10
```

3. **summary()**: Este comando da un resumen estadístico de la base de datos o de las variables a analizar. Este resumen consta de:

- **Min:** Valor mínimo
- **1st Qu.:** Primer cuartil
- **Median:** Mediana o segundo cuartil
- **Mean:** Promedio o media
- **3rd Qu.:** Tercer cuartil
- **Max.:** Valor máximo

Para obtener el resumen que se tiene en la base se pone en la consola: `summary(ESTADISTICA.PRUEBA)` y se muestra una tabla con el resumen de cada una de las variables

```
> summary(ESTADISTICA.PRUEBA)
      V1      V2      V3      V4
Min.   : 1.00  FLOR DE BASTION :12  HOMBRE:5  Min.   : 1.00
1st Qu.: 3.75                                1st Qu.:16.00
Median : 6.50                                Median :27.00
Mean   : 6.50                                Mean   :30.91
3rd Qu.: 9.25                                3rd Qu.:47.00
Max.   :12.00                                Max.   :69.00
      NA's :1
      V5
Min.   : 2.00
1st Qu.: 4.75
Median : 7.00
Mean   : 6.50
3rd Qu.: 8.25
Max.   :10.00
```

En cambio para sacar el resumen estadístico de la variable sería `summary(V5)`

```
> summary(V5)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 2.00   4.75   7.00   6.50   8.25  10.00
```

- **Min:** La nota mínima de calificación es 2.00
- **1st Qu.:** El 25% de los datos tienen menos de 4.75 como nota
- **Median:** El 50% de los datos tienen menos de 7 como nota
- **Mean:** El promedio de nota es 6.50
- **3rd Qu.:** El 75% de los datos tienen menos de 8.25 como nota
- **Max.:** La nota máxima de calificación es 10.00

4. **mean()**: Este comando calcula el promedio o la media de la variable.

```
> mean(v5)
[1] 6.5
```

5. **median()**: Calcula la mediana de la variable.

```
> median(v5)
[1] 7
```

6. **var()**: Calcula la varianza de la variable.

```
> var(v5)
[1] 6.454545
```

7. **sd()**: Calcula la desviación estándar de la variable.

```
> sd(v5)
[1] 2.54058
```

8. **range()**: Calcula el rango de la variable, presentando el valor mínimo y máximo.

```
> range(v5)
[1] 2 10
```

9. **Frecuencias**: Para calcular la frecuencia de un grupo de datos es necesario realizar los siguientes pasos:

- Se debe transformar la variable en una de tipo tabla con el comando `Table()` y asignarla a una nueva variable: `Frecuencia<- table(V5)`
- Se ejecuta la nueva variable “frecuencia” y aparecerá en la parte superior los diversos datos de la tabla y en la

parte inferior la cantidad de cada dato (frecuencia absoluta).

```
> frecuencia<-table(v5)
> frecuencia
v5
 2  3  4  5  6  7  8  9 10
1  1  1  1  1  2  2  2  1
```

- Para presentar los valores como frecuencia relativa se debe transformar la variable “frecuencia” en otra tabla con el comando `prop.table()` y se obtendrá la frecuencia relativa de cada dato.

```
> frecuenciarelativa<-prop.table(frecuencia)
> frecuenciarelativa
v5
      2      3      4      5      6      7
0.08333333 0.08333333 0.08333333 0.08333333 0.08333333 0.16666667
      8      9     10
0.16666667 0.16666667 0.08333333
```

- Si se quieren los valores de la frecuencia relativa en porcentajes, se multiplica por 100 la variable creada (frecuenciarelativa) para hallar la frecuencia. Se puede asignar a otra variable o a la misma variable creada y luego ejecutar la variable.

```
> frecuenciarelativa<-frecuenciarelativa*100
> frecuenciarelativa
v5
      2      3      4      5      6      7      8
8.333333 8.333333 8.333333 8.333333 8.333333 16.666667 16.666667
      9     10
16.666667 8.333333
```

- Si quiere presentar los valores en forma de vectores o columnas se utilizará el comando `data.frame()` y se lo asignará a una nueva variable (marco1). Este paso se puede hacer para la frecuencia absoluta y la relativa

(marco2). En la gráfica siguiente se pueden observar las notas y sus frecuencias en columnas.

```
> marco1<-data.frame(frecuencia)
> marco2<-data.frame(frecuenciarelativa)
> marco1
  V5 Freq
1  2     1
2  3     1
3  4     1
4  5     1
5  6     1
6  7     2
7  8     2
8  9     2
9 10     1
> marco2
  V5 Freq
1  2 8.333333
2  3 8.333333
3  4 8.333333
4  5 8.333333
5  6 8.333333
6  7 16.666667
7  8 16.666667
8  9 16.666667
9 10 8.333333
```

- Si se quiere poner nombre a las columnas de las tablas se utilizará el comando `name()` y se mencionará la variable a la cual se le van a poner nombre (marco1) y se indica qué nombre van a tener las columnas. Luego se ejecuta nuevamente la variable.

```
> names(marco1)<-c("Nota", "Frecuencia Absoluta")
> names(marco2)<-c("Nota", "Frecuencia Relativa")
> marco1
  Nota Frecuencia Absoluta
1    2                  1
2    3                  1
3    4                  1
4    5                  1
5    6                  1
6    7                  2
7    8                  2
8    9                  2
9   10                  1
> marco2
  Nota Frecuencia Relativa
1    2          8.333333
2    3          8.333333
3    4          8.333333
4    5          8.333333
5    6          8.333333
6    7         16.666667
7    8         16.666667
8    9         16.666667
9   10          8.333333
```

- Si se quiere unir ambas tablas, se utiliza el comando `merge()`, el cual se asignará a una nueva variable (`tablafrecuencia`). En el comando `merge()` se mencionan los dos marcos, se indica la columna que los va a unir (`Nota`).

```
> tablafrecuencia<-merge(marco1, marco2, by="Nota")
> tablafrecuencia
```

	Nota	Frecuencia Absoluta	Frecuencia Relativa
1	10	1	8.333333
2	2	1	8.333333
3	3	1	8.333333
4	4	1	8.333333
5	5	1	8.333333
6	6	1	8.333333
7	7	2	16.666667
8	8	2	16.666667
9	9	2	16.666667

7.3.6 Histograma y diagrama de cajas

1. **hist()**: Nos da el histograma de la variable seleccionada.

```
> hist(v5)
```

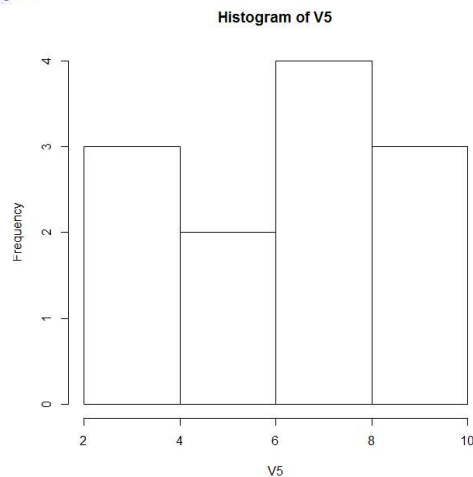


FIGURA 7-35 HISTOGRAMA

Si se quiere poner nombre al histograma y a los ejes, realizamos los siguientes pasos:

1. Poner comando `hist()` y la variable a graficar.
2. Dentro del comando, separado por una coma agregar comando `main()` en donde pondrá entre comilla el nombre del histograma.
3. A continuación, también separado por una coma pondrá el comando `xlab` en donde colocará el nombre del eje de las X
4. Con el comando `ylab`, colocará el nombre del eje de las Y

```
> hist(v5, main='Nota de Estadística', xlab='Nota', ylab='Frecuencia')
```

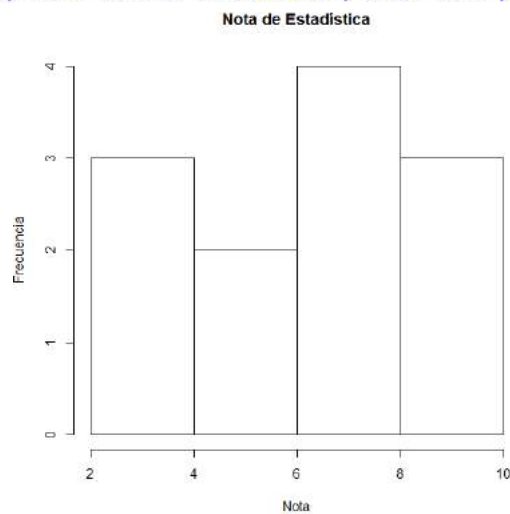


FIGURA 7-36 NOMBRE DE LOS ELEMENTOS DEL HISTOGRAMA

2. **boxplot()**: Presenta el diagrama de cajas de la variable seleccionada. También se puede presentar dos o más diagramas de cajas a la vez, con solo agregar las variables al mismo comando, ejemplo `boxplot(V4,V5)`

```
> boxplot(v5)
```

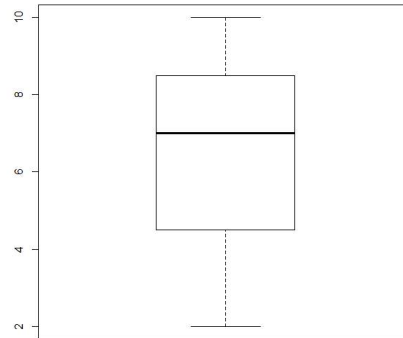


FIGURA 7-37 DIAGRAMA DE CAJAS

7.3.7 Ingresando datos en RSTUDIO

Otra manera de trabajar con R si se cuenta con pocos datos y no se quiere trabajar con una base, es ingresando directamente los datos en RSTUDIO, para lo cual se deben seguir los siguientes pasos:

1. En la “Console” ingresamos los datos asignando a una variable “X” los datos dentro de un conjunto `c()` y separados por comas:

```
Console ~/
> x=c(8,9,10,5,6,7,8)
> mean(x)
[1] 7.571429
```

2. Se puede trabajar con todos los comandos vistos anteriormente, de la misma manera que cuando se importó el archivo de EXCEL, con la diferencia de que la variable siempre va a ser el nombre que se puso, como en el ejemplo anterior la variable es “X”, entonces la media se calcularía con el comando `mean(variable)`.
3. En caso de trabajar con decimales en RStudio el símbolo es el punto ‘.’

Ejemplo: notas=c(5.5,6.5,8,9.75,10,7.85,2.45)

```
> notas=c(5.5,6.5,8,9.75,10,7.85,2.45)
> mean(notas)
[1] 7.15
```

7.4 Estadística descriptiva con EXCEL

Para realizar estadística descriptiva en Excel, se utiliza un comando denominado “Análisis de datos” y se encuentra en “Datos”; pero si no está activado, se realizan los siguientes pasos para activarlo:

1. Ir a Opciones de Excel.

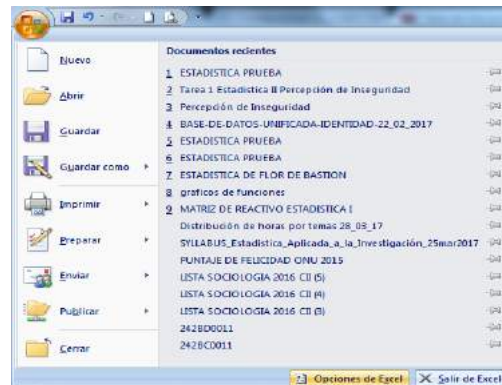


FIGURA 7-38 OPCIONES DE EXCEL

2. Ir a complementos, luego en la ventana “Administrar” verificar si está direccionado a “Complemento de Excel” y hacer click en “Ir”.

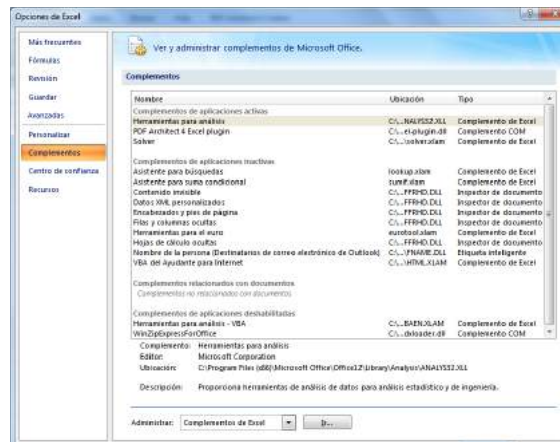


FIGURA 7-39 COMPLEMENTO DE EXCEL

3. Aparecerá la ventana complemento y se deberá seleccionar “Herramienta para análisis” y poner aceptar.

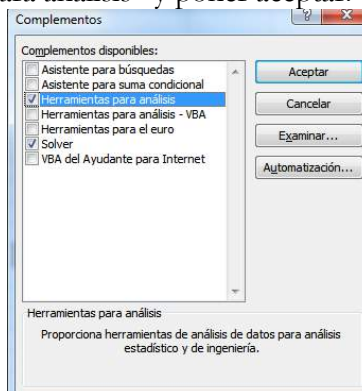


FIGURA 7-40 HERRAMIENTA DE ANÁLISIS DE EXCEL

4. Se escoge la opción Datos, para verificar que está activado el comando “Análisis de datos”

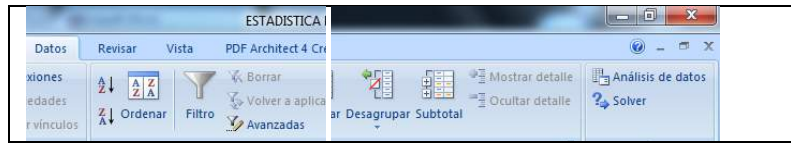


FIGURA 7-41 ANÁLISIS DE DATOS EN EXCEL

Una vez efectuada dicha verificación se podrá realizar la estadística descriptiva de los datos; con los datos creados con anterioridad en Excel.

1. Hacer click en “Análisis de datos” y se abrirá la ventana para seleccionar “Estadística Descriptiva” y presionar aceptar.

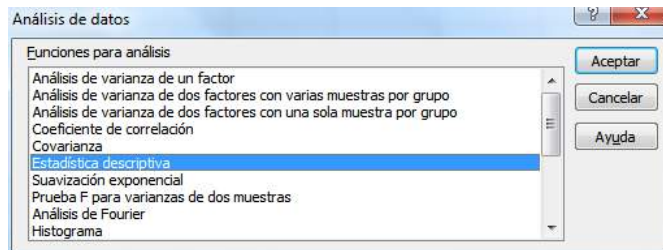


FIGURA 7-42 ESTADÍSTICA DESCRIPTIVA EN EXCEL

2. En la ventana “Estadística Descriptiva” aparece “Rango de entrada”, hacer click y la variable que se va a analizar; si esta variable tiene encabezado, se activa “Rótulos en la primera fila”. Luego, en “Rango de salida” se selecciona donde se presentan los resultados, en la misma hoja o los mismos que pueden ser en una nueva hoja. Y por último se activa la opción “Resumen de estadística” y luego se activa aceptar.

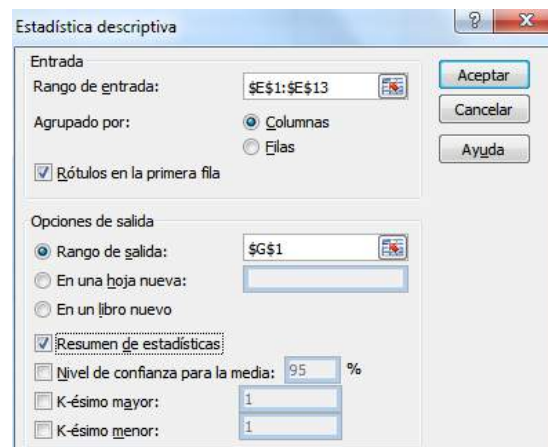
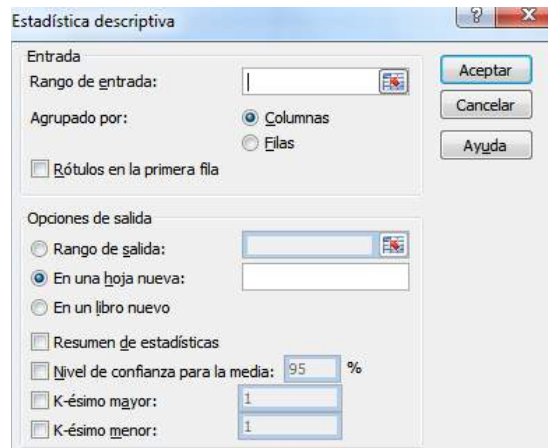


FIGURA 7-43 OPCIONES DE ESTADÍSTICA DESCRIPTIVA

3. En la figura siguiente se observa el resultado estadístico del dato, en donde se encuentra la media, error típico, mediana, moda, desviación estándar, etc.

	A	B	C	D	E	F	G	H
1	#	Sector	sexo	Edad	Calificación		Calificación	
2	1	FLOR DE BASTION	MUJER	27	7			
3	2	FLOR DE BASTION	HOMBRE		8		Media	6.5
4	3	FLOR DE BASTION	HOMBRE	22	5		Error típico	0.733402201
5	4	FLOR DE BASTION	MUJER	21	9		Mediana	7
6	5	FLOR DE BASTION	MUJER	1	10		Moda	7
7	6	FLOR DE BASTION	HOMBRE	43	2		Desviación estándar	2.540579748
8	7	FLOR DE BASTION	MUJER	69	3		Varianza de la muestra	6.454545455
9	8	FLOR DE BASTION	MUJER	51	4		Curtosis	-0.845784567
10	9	FLOR DE BASTION	HOMBRE	55	9		Coefficiente de asimetría	-0.478985716
11	10	FLOR DE BASTION	MUJER	31	8		Rango	8
12	11	FLOR DE BASTION	MUJER	11	7		Mínimo	2
13	12	FLOR DE BASTION	HOMBRE	9	6		Máximo	10
14							Suma	78
15							Cuenta	12

FIGURA 7-44 RESULTADO ESTADÍSTICO EN EXCEL

Una vez obtenido los resultados se pueden realizar los análisis descriptivos.



BIBLIOGRAFÍA

- Bruce Levin, M. &. (2015). *Statistics for Lawyers*. Pennsylvania: Springer.
- David S. Matteson, D. &. (2015). *Statistics and Data Analysis for Financial Engineering*. Ithaca, NY: Springer.
- Holland, R. M. (2015). *Statistical Analysis and Data Display*. Philadelphia: Springer.
- III, J. E. (2015). *Political Analysis Using R*. Suiza: Springer.
- Zurita, G. (2010). *Probabilidad y Estadística*. Guayaquil: Escuela Superior Politécnica del Litoral.

John Aníbal Herrera Rivera es profesor de la Carrera de Sociología y Ciencias Políticas de la Facultad de Jurisprudencia y Ciencias Sociales y Políticas de la Universidad de Guayaquil. Es profesional en Estadística Informática, especializado en Administración y Dirección de Empresas. Se ha desempeñado como consultor de proyectos e investigaciones estadísticas; así como también ha sido instructor de cursos estadísticos.

ISBN: 978-9942-760-23-4



9 789942 760234