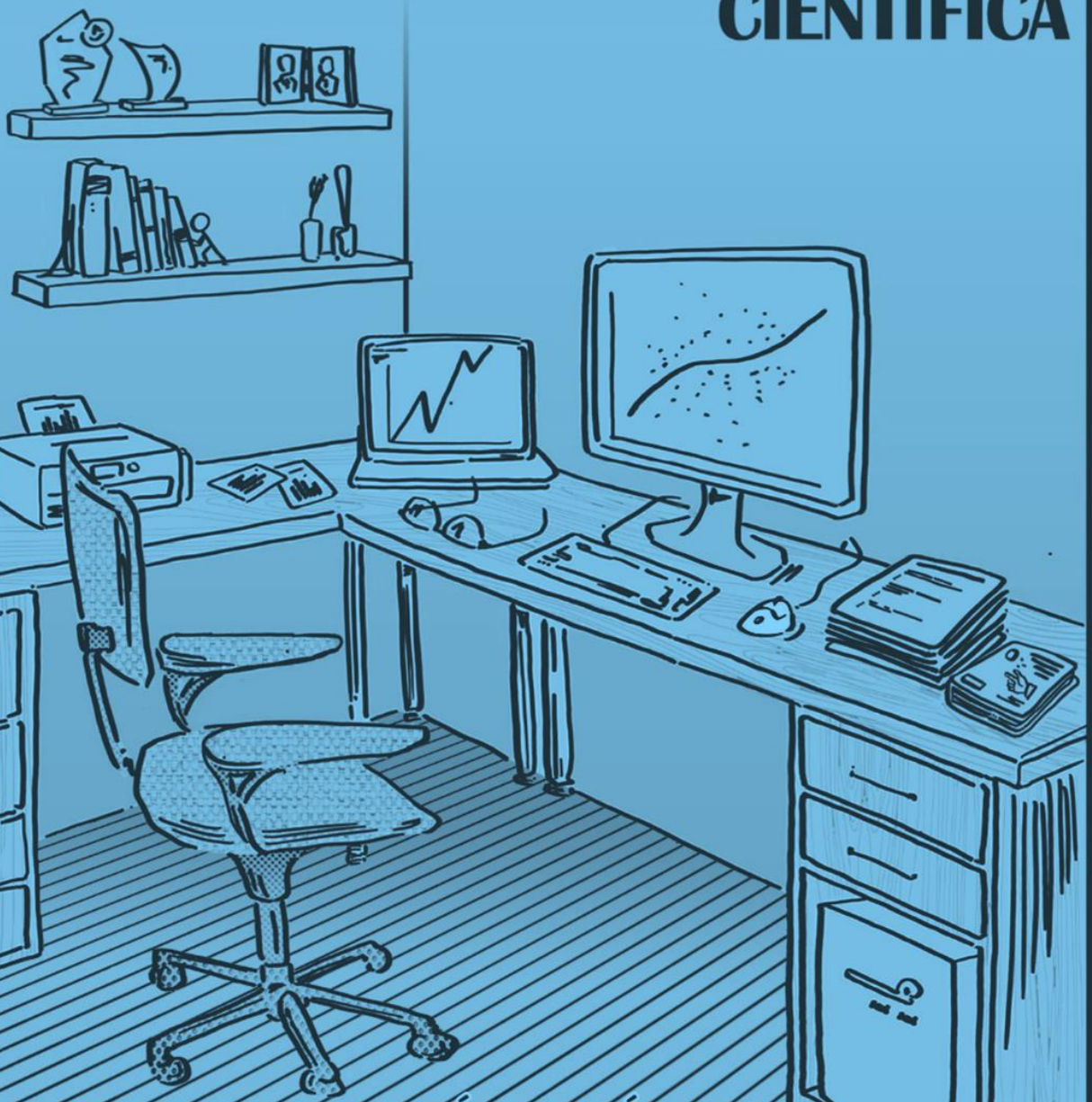
ESTADÍSTICA DESCRIPTIVA EN LA LÓGICA DE LA INVESTIGACIÓN CIENTÍFICA



EDUAR VÁSQUEZ SÁNCHEZ
GLORIA MARÍA ORTIZ BASAURI
EDUAR ANÍBAL VÁSQUEZ ORTIZ

EDITORIAL UNIVERSITARIA
UNIVERSIDAD NACIONAL PEDRO RUIZ GALLO

ESTADÍSTICA DESCRIPTIVA EN LA LÓGICA DE LA INVESTIGACIÓN CIENTÍFICA

EDUAR VÁSQUEZ SÁNCHEZ
GLORIA MARÍA ORTIZ BASAURI
EDUAR ANÍBAL VÁSQUEZ ORTIZ

EDITORIAL UNIVERSITARIA
UNIVERSIDAD NACIONAL PEDRO RUIZ GALLO

ESTADÍSTICA DESCRIPTIVA
EN LA LÓGICA DE LA
INVESTIGACIÓN CIENTIFÍCA

Eduar Vásquez Sánchez
Profesor principal de la Facultad de Ciencias Físicas y Matemáticas. Universidad Nacional Pedro Ruiz Gallo. Desarrolla las asignaturas de metodología de la investigación y estadística para la investigación científica.

Gloria María Ortiz Basauri

Profesora principal en la Facultad de Ciencias Físicas y Matemáticas. Universidad Nacional Pedro Ruiz Gallo. Desarrolla las asignaturas de lógica matemática, análisis matemático y variedades diferenciales.

Eduar Aníbal Vásquez Ortiz

Ingeniero Electrónico. Universidad Nacional Pedro Ruiz Gallo.
Mag. Procesamiento de señales e imágenes digitales. Pontificia Universidad Católica del Perú.
Profesor Asistente: Pontificia Universidad Católica del Perú.

Estadística descriptiva en la lógica de la investigación científica

Unidad de Editorial Universitaria
Universidad Nacional Pedro Ruiz Gallo
Calle Juan XXIII 391. Lambayeque.
Teléfono: 074 282081 www.unprg.edu.pe
fondoeditorial@unprg.edu.pe
Lambayeque. Perú.

Primera edición digital, diciembre 2023

Biblioteca Nacional del Perú.
ISBN: 978-9972-55-035-5
Depósito legal N°: 2024-05977

INDECOPI
Certificado de registro de obras literarias: 01331-2024

Autores: ©
Dr. Eduar Vásquez Sánchez
Dra. Gloria María Ortiz Basauri
Mag. Eduar Aníbal Vásquez Ortiz

© **Diseño y cubierta:**
Ing. Víctor Eduardo Vásquez Ortiz Ing.
Gloria del Milagro Salvador Vásquez

Publicación sometida a evaluación de pares académicos

DEDICATORIA

Dedicamos este trabajo a:

La memoria de nuestros padres:

Artidoro Vásquez Ramírez y Olinda Sánchez Bustamante,
Víctor Ortiz Tafur y Gosvinda Basauri Arce
por su amor infinito y sabias enseñanzas.

Nuestros hijos:
Por ser motor de nuestras vidas.

Nuestros Hermanos:
Por apoyar nuestros proyectos.

Nuestros sobrinos*:*
Por ser la esperanza de la familia.

El estadístico no puede
evadir la responsabilidad
de comprender el proceso
que aplica o recomienda.

Ronald Fisher

PRESENTACIÓN

En el presente libro se ofrece métodos y técnicas para el recojo de datos, la organización, el procesamiento y la presentación en tablas y gráficos. Los datos se producen al realizar la medición de las variables del objeto de investigación hecho que necesita saber la escala de medición de cualidades o cantidades.

Los instrumentos para la medición de variables deben ser válidos y confiables, establecer exactitud y precisión en las mediciones según sea la naturaleza de la variable, la técnica del muestreo que se use para la toma de los datos o el modelo estadístico para el análisis de datos. Las distribuciones de frecuencia de la variable que se estudia para poder realizar un correcto análisis e interpretación se organizarán de acuerdo a su definición conceptual y propiedades estadísticas de los datos.

Las estadísticas para describir el comportamiento de una variable son fundamentalmente las de centralización y dispersión muy usadas en las investigaciones de nivel diagnóstico, las regresiones y correlaciones, así como las series de tiempo adecuadamente tratadas son muy útiles para realizar predicciones y retrodicciones. Finalmente se presenta una introducción a los números índices, orientando lo que sería el contenido mínimo de una estadística en la investigación básica.

El contenido mínimo de la estadística descriptiva que se presenta en este libro corresponde a la organización básica de las tablas estadísticas que se prepararían en la descripción y análisis de las variables a nivel de la muestra que se necesita para el planteamiento de las hipótesis en la estadística inferencial.

Presentamos este libro a fin de que los investigadores se apropien con facilidad de los conceptos básicos de la Estadística Descriptiva.

Los autores

ÍNDICE

CAPÍTULO I

LAS VARIABLES EN LA INVESTIGACIÓN CIENTÍFICA

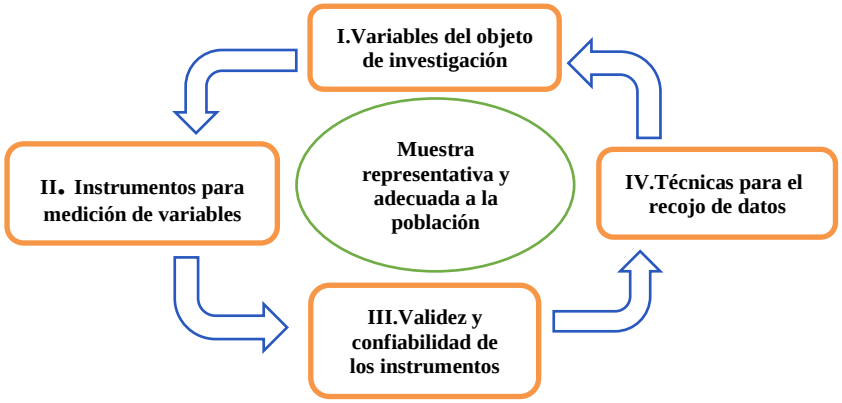
El **objeto de la investigación científica** se define en términos de **variables**; por lo que resulta de vital importancia, en el tratamiento estadístico de los datos, identificarlas y conocer los **instrumentos** para medirlas, así como, la **validez** y **confiabilidad** de los mismos, las **técnicas** para obtener información y los tipos de **muestreo**.

Los datos en la estadística constituyen la materia prima, se generan a partir de la medición de las variables que son susceptibles de medición, pues para producir datos confiables para investigar las propiedades y

atributos de las variables es necesario un adecuado tratamiento estadístico. La figura 1.1 muestra el trabajo con las variables.

Figura 1.1

Tratamiento estadístico de las variables en la investigación.



1.1 VARIABLES

Son características o propiedades cualitativas o cuantitativas, de la unidad en estudio, que se pueden medir. Si la variable cambia cualitativamente es categórica, subjetiva o lógica, su medición es indirecta a través de indicadores y requiere instrumentos documentales; si cambia cuantitativamente es numérica, objetiva o física y requiere instrumentos físicos para medirlas.

1.2 TIPOS DE VARIABLES

a) **Variables por la naturaleza de su medición**

Variable cualitativa.

La variable cualitativa es una propiedad de la unidad de estudio, puede ser ordinal o nominal

- **Variable nominal.**

Es una propiedad de la unidad de estudio que no se puede ordenar numérica o jerárquicamente, solo clasificar, ejemplo tipos de empresas, números de camisetas de jugadores, vías de administración de vacunas, tipos de investigación, marcas de autos, procedencia de las personas etc.

- **Variable ordinal.**

Es una propiedad de la unidad de estudio, que se ordena jerárquicamente, ejemplos: grados de instrucción, grado de cáncer, nivel de logro de aprendizaje, grado militar, estrato social, grados académicos, estadios clínicos de tumores, etapas de la vida del ser humano, etc.

Variable Cuantitativa.

Es una propiedad de la unidad de estudio que se expresa en números naturales o reales. Puede ser:

- **Variable discreta.**

Es aquella cuya medición se hace con los números naturales. Ejemplo: número de alumnos en una clase, número de programas académicos de un instituto, número de hijos, número de asignaturas de un plan de estudios, etc.

- **Variable continua.**

Es aquella que se expresa por un número real. Ejemplos, volumen de biomasa de especies marinas, temperatura corporal, peso del recién nacido, longitud de una pieza metálica, tiempo de espera en una consulta, gasto de combustible por una unidad de transporte, resistencia de un material de construcción etc.

b) **Variables en el objeto de investigación.**

Pueden ser cualitativas o cuantitativas. El objeto de investigación puede estar dado por una sola variable o por dos o más variables estableciendo una relación.

Esta relación es de asociación, si son cualitativas; de correlación, si son cuantitativas. Si una variable produce cambios en la otra, la relación es de tipo causal, reconociéndose las variables independiente, dependiente e interviniente.

Variable independiente.

La que se puede manipular de acuerdo al interés del investigador en una relación causal, se denota con **X**

Variable dependiente.

La que se modifica según los cambios de la variable independiente, se denota con **Y.**

Variable interviniente.

Aquella que modifica el grado de relación entre las variables. Se denota con **Z**

Ejemplo 1.

En la investigación: ¿Cómo influye el Método de Pólya en el aprendizaje de Geometría diferencial, en los estudiantes de matemática pura? En este trabajo hay una influencia del Cociente Intelectual de los estudiantes, constituyéndose en variable interviniente, siendo la

independiente: el Método de Pólya y la dependiente: el aprendizaje de Geometría Diferencial.

Ejemplo 2.

La temperatura superficial de mar influye directamente en el volumen de biomasa de especies marinas. Aquí la variable dependiente, el volumen de biomasa de especies marinas, variable independiente: temperatura superficial de mar y la variable interviniente sería el fenómeno de El Niño.

Variables por el número de sus dimensiones.

Pueden ser unidimensionales o simples y multidimensionales o complejas; siendo, la dimensión un aspecto o componente de la variable, que en su operacionalización permite acercarse al plano empírico.

Variables simples o unidimensional.

Cuando la dimensión es la misma variable, cuando tiene una sola característica o un solo atributo, la tabla 1.1 que sigue muestra ejemplos de variables simples cuantitativas y cualitativas.

Tabla 1.1
Variables simples cuantitativas y cualitativas

Variables simples cuantitativas		
Variable	**Dimensión**	**Instrumento**
Temperatura	Temperatura	Termómetro
Peso	Peso	Balanza
Talla	Talla	Cinta métrica
Tiempo	Tiempo	Cronómetro
Variables simples cualitativas		
Variable	**Dimensión**	**Instrumento**
Nivel de instrucción de las personas	Nivel de instrucción de las personas	Cuestionario
Ocupación	Ocupación	Cuestionario
Asistencia a clase	Asistencia a clase	Registro

Variable Compleja o multidimensional.

Son variables que tienen varias características, propiedades o atributos, toma en cuenta el contexto de la investigación, Abreu (2012), da como ejemplos lo siguiente: "[si la variable es] clase social [de una persona, pueden tomarse] nivel económico, [grado] de instrucción […] como dimensiones, […] Para la inteligencia, según Gardner se tiene, inteligencia verbal, matemática, artística, intrapersonal, interpersonal, kinestésica, etc." (p. 3). El aprendizaje de una materia para el docente puede ser simple como aprobado o desaprobado, pero en la psicología puede ser complejo, con sus dimensiones: biológica, cognitiva y social.

Las variables: estilos de aprendizaje según Kolb, agresividad de género, calidad de agua son otros ejemplos de variable compleja. En la tabla 1.2 se muestran variables complejas de acuerdo a su dimensión e instrumentos de medición.

Tabla 1.2
Variables complejas cuantitativas y cualitativas

Variables complejas cuantitativas		
Variable	**Dimensiones**	**Instrumento de medición**
Masa corporal	Peso	Balanza
	Talla	Metro
Presión arterial	Presión sistólica	Estetoscopio
	Presión diastólica	Estetoscopio
Velocidad	Espacio	Cinta métrica
	Tiempo	Cronómetro
Variables complejas cualitativas		
Variable	**Dimensiones**	**Instrumento de medición**
Calidad de servicio	Elementos Tangibles	Escala Servqual
	Fiabilidad	
	Capacidad de respuesta	
	Seguridad	
	Empatía	
Desarrollo Humano	Salud	Esperanza de vida
	Educación	Tasa de alfabetización de adultos
	Riqueza	PBI per cápita
Rendimiento Académico	Cognitivo	Prueba
	Procedimental	Lista de Cotejo
	Actitudinal	Test de actitudes
Nivel socio-económico	Salario	Cuestionario
	Educación	
	Vivienda	
	Acceso a servicios	
	Ocupación	

1.3 INDICADORES DE LAS VARIABLES

De acuerdo a Rojas (2007):

> Variable es un concepto tomado de las matemáticas y significa la propiedad que tienen las personas, los hechos, fenómenos y procesos de tomar ciertos valores cualitativos o cuantitativos. […] [Ejemplo]: sexo, edad, estatura, peso, estado civil, clase social, afiliación política y nacionalidad. […]
>
> Los indicadores representan [una característica específica] de una variable o de una dimensión de ésta. El indicador nos "indica" la situación de una variable. Por ejemplo, la fiebre es un indicador cualitativo y el dato 40º es un indicador cuantitativo de la variable "enfermedad". […]
>
> [Existen] variables [que] contienen indicadores más complejos que otros, por ejemplo, los indicadores "nivel de escolaridad" y "estilo de vida".

Los indicadores sirven para elaborar los instrumentos para la toma de datos y describen el comportamiento de las variables, por ejemplo, en el contexto educativo los indicadores pueden ser de proceso, percepción o rendimiento.

Indicadores de proceso.

Permiten medir el proceso de un evento, suceso o actividad educativa mientras se desarrolla.

Ejemplo.

Los tutores programan sus asignaturas según un modelo establecido por el centro educativo y para obtener un indicador de su avance, realizan una autoevaluación del desarrollo de la enseñanza y la expresan como porcentaje.

Indicadores de percepción.

Son las diferentes opiniones que tienen los observadores de los procesos educativos.

Ejemplos.

1. Grado de satisfacción del profesorado con la organización horaria de la docencia.
2. Grado de satisfacción del alumnado con el clima de convivencia del centro.

Indicadores de rendimiento.

Son los descriptores estadísticos o caracterizaciones de logros en los hechos o procesos educativos.

Ejemplos.

1. Porcentaje de inserción del alumnado con contrato laboral.
2. Porcentaje de aprobados en selectividad sobre los matriculados en 2º de bachillerato; para ambos indicadores su criterio de aceptación sería del 20% y 75% respectivamente.

Las propiedades que los indicadores deben cumplir son:

1. **Simplicidad.** Abarcan un número reducido de dimensiones, mayormente solamente una.
2. **Comparabilidad.** Hacen posible la aplicación de las escalas de medición.
3. **Comunicación.** Permiten conocer la situación real de las variables que se investigan.

Ejemplos.

- Presenta una temperatura axilar mayor de 39°5’.
- Investiga un tema diferente una vez al año.
- El perfil profesional responde a las necesidades del entorno.
- Obtiene una nota de 17 sobre 20 puntos.
- Resuelve 4 de 5 problemas del contexto utilizando operaciones algebraicas.
- Tasa bruta de natalidad por 1 000 habitantes.
- Oferta de bienes y servicios de la Empresa X en el 2021.
- Tenencia de activos fijos para satisfacer obligaciones a corto plazo.
- Usar información del aprendizaje de los estudiantes para guiar la planificación.

Para elaborar indicadores en el campo educativo se articulan tres elementos, un verbo que indica la capacidad específica; lo que se desea medir que viene a ser el contenido y las condiciones; ejemplos:

- Identifica (**capacidad específica**) las ideas principales y secundarias (**contenido**) en un texto expositivo **(condiciones**).
- Fórmula (**capacidad específica**) problemas de compra venta (**contenido**), utilizando las cuatro operaciones con números decimales (**condiciones**).

1.4 ESCALAS DE MEDICIÓN DE VARIABLES

Una escala es un patrón de medición con la que se hace corresponder el valor de una variable. Si la variable es cualitativa o categórica, la escala a elegir será nominal u ordinal; si la variable es cuantitativa o numérica, la escala será de intervalos o de razón, estas escalas de medición de variables fueron propuestas por el psicólogo Stevens (1946).

Escala nominal.

Con esta escala se identifica y clasifica variables asignándoles un nombre o código, solo se cuenta y compara, estableciendo la relación de igualdad contrastándose hipótesis de asociación entre variables.

Ejemplos.

Nacionalidad: peruana, ecuatoriana, brasileña, española, francesa, rusa.
Estado civil: soltero, casado, divorciado, conviviente, viudo.
Uso de anteojos: Sí, No.
Profesión: arquitectura, contabilidad, derecho, medicina, psicología.
Número de cédula de identidad: 18876454, 13221978, etc.

Escala ordinal.

Establece un orden en las categorías de la variable que se mide, donde las relaciones de tipo "mayor que" o "menor que" están implícitas. Para las variables que se miden con esta escala se puede calcular la mediana, la correlación de Spearman y r de Kendall; para contrastar hipótesis se aplican pruebas no paramétricas, de Wilcoxon, test de signo, de Kolmogórov-Smirnov.

Ejemplos.
Nivel de Instrucción: Primaria, secundaria, superior.
Categorías de profesores universitarios: auxiliar, asociado, principal.
Grado de desnutrición: Leve, moderada, grave.
Orden de llegada a la meta de los atletas: Primer puesto, segundo puesto, etc.
Preferencia a productos de consumo: Débil, indiferente, fuerte.

Estado de una fruta: Verde, madura, podrida.
Niveles de ansiedad: Leve, moderado, severo.
Grado de acuerdo: muy en desacuerdo, en desacuerdo, indiferente, de acuerdo, muy de acuerdo.

Escala de intervalos.

En esta escala se establecen intervalos de igual amplitud, en la que el valor cero no significa que no existe el atributo, simplemente es una localización arbitraria en la escala, por lo que se admiten valores negativos, nulos (ceros) o positivos. Las mediciones de las variables con esta escala se apoyan en la distribución normal, los modelos para contrastar hipótesis son paramétricos tales como: El promedio, la desviación estándar, r de Pearson, t de Student, ANOVA etc.

Ejemplos.

Temperatura ambiental: -10°C, 0°C, 10°C, 20°C
Longitud o latitud cartográfica: -50º, 0º, 20º

Escala de razón.

La escala de razón o de proporciones admite valores numéricos a partir del cero, que indica la ausencia de la característica a medir. Las mediciones de las variables que se realizan con escala de razón, se analizan con modelos paramétricos, pueden usarse los diferentes tipos

de promedios, como el aritmético, geométrico, armónico o la variación relativa o coeficiente de variación relativa etc.

Ejemplos de variables que se miden con escalas de razón.

Longitud: 0 m, 34 m, 17 m, etc.
Peso: 0 kg, 12,60 kg, 58 kg, etc.
Tiempo: 0 s, 45 s, 72 s etc.
Ingresos: S/ 0,00; S/ 33,00; S/ 7 599,00 etc.
Volumen de producción: 562 t; 621 t; 6 591 t etc.
Altura de personas: 1,78m; 1,96m; 1,99m etc.
Velocidad de un auto en la carretera: 233 km/h; 287 km/h etc.
Número de goles marcados en un partido: 0, 1, 2, 3, etc.

1.5 OPERACIONALIZACIÓN DE VARIABLES

Las variables son abstracciones de las características reales de los objetos y procesos, operacionalizarlas es hacer posible su medición concretizándolas en datos.

La operacionalización de las variables permite delimitar el contenido teórico de las variables y establecer la base conceptual para la elaboración de los instrumentos de medición de las variables y dar el soporte de la validez de contenido. En la tabla siguiente se aprecia

Tabla 1.3
Operacionalización de una variable

Variable	Dimensiones	Indicadores	Instrumento
Inteligencia	Inteligencia Verbal	Comprensión rápida de textos. Vocabulario amplio. Sistematización.	**Cuestionarios.** **Escalas** **Tests**
	Inteligencia Abstracta	Abstracción. Conceptualización. Creatividad.	
	Inteligencia Numérica	Cálculo rápido de operaciones aritméticas. Resolución de problemas.	

Tabla 1.4
Operacionalización de dos variables

Variables	Dimensiones	Indicadores	Instrumentos
Comprensión lectora	Literal	Entender el significado de los términos. Hacer textos descriptivos	**Tests** **Cuestionarios**
	Parafrástica	Establecer relaciones. Esquematizar un texto.	
	Inferencial	Predecir los resultados. Entrever relaciones causales	
	Criterial	Juzgar un texto bajo determinado enfoque Emitir juicio frente a comportamientos	
Aprendizaje académico	Problematización	Identificar situaciones problemáticas cuestionamiento, curiosidad, motivación.	**Tests** **Cuestionarios**
	Organización del Conocimiento	Conectar lo aprendido, significar algo. Integración informativa.	
	Procesamiento de la Información	Operacionalizar variables Realizar inferencias, comparaciones, clasificaciones. Operaciones mentales para desarrollar conocimientos y habilidades.	
	Aplicación de la Información	Tratamiento de problemas reales o probables. Operar conceptos, investigar casos.	
	Meta Cognición	Resolver controversias. Aprender a aprender.	

CAPÍTULO II

INSTRUMENTOS PARA LA RECOLECCIÓN DE DATOS

El **dato,** resultado de la medición de las variables cualitativas o cuantitativas, son representaciones simbólicas, numéricas, alfabética, algorítmica etc. con datos se describen hechos empíricos, sucesos y entidades. Los datos procesados constituyen la **información** que permite el conocimiento completo del objeto que se investiga.

Los instrumentos para medir variables y producir datos son recursos de los que se vale el investigador para acercarse a los hechos, sucesos o fenómenos, conecta la teoría con la práctica.

Los cuestionarios cumplen este doble papel de medir y producir el dato, claro está en base a este instrumento se realiza las encuestas entrevistas, incluso para el registro de las observaciones dependiendo de la naturaleza de la variable, la observación en muchos casos es audiovisuales, fotografías, radiografías etc.

Los instrumentos pueden ser **físicos o documentales.**

2.1 INSTRUMENTOS FÍSICOS

Los instrumentos físicos utilizados para medir variables objetivas deben ser exactos y precisos.

Exactitud.

Un instrumento es exacto cuando sus mediciones se aproximan al verdadero valor de la magnitud que se mide. Comúnmente para determinar la exactitud, se calcula el promedio de distintas mediciones y se compara con algún parámetro establecido.

Precisión.

Un instrumento es preciso si al realizar un determinado número de veces una misma medición, se obtiene igual resultado o muy cercano. Se representa mediante el valor más pequeño de una magnitud que se puede medir con dicho instrumento.

2.2 INSTRUMENTOS DOCUMENTALES

Los instrumentos documentales miden variables subjetivas en la investigación científica y deben pasar por la prueba de validez y confiabilidad, estos son: cuestionarios, escalas de actitudes e inventarios. Se preparan para medir constructos estructurados en preguntas.

2.2.1 Cuestionario

Es instrumento documental, contiene preguntas, ítems o indicadores, diseñado para obtener información específica de manera sistemática y ordenada de temas variados, en forma individual o colectiva. Ejemplo: los exámenes para evaluar rendimiento académico. Los cuestionarios deben:

1. Ser formulados en lenguaje, claro, simple y directo, usando frases simples.
2. Contener sólo una frase lógica, en lo posible con menos de 20 palabras
3. Contener ítems relevantes con respecto a lo que se quiere medir.
4. No ser extenso, un cuestionario largo es cansado sus preguntas finales se responden por cumplir.
5. Tener presente el sistema de codificación.
6. Tomar en cuenta el modelo de procesamiento de los datos.

Los cuestionarios se clasifican de acuerdo a las respuestas de sus preguntas en:

a. **Cuestionario Cerrado.**

Es aquel en el que se solicita respuestas breves, específicas y delimitadas. Es fácil de codificar y contestar, el informante contesta las preguntas, que pueden ser dicotómicas o politómicas, con las alternativas que se ofrecen.

- Pregunta dicotómica. Es aquella que tiene dos alternativas de respuesta que se excluyen mutuamente, para representar lo que se quiere evaluar se necesita un gran número de preguntas de este tipo.

Ejemplos.

- ¿En la actualidad, usted trabaja? Sí ( ) o No ( ).
- ¿La mediana es una medida de dispersión? Falso (F) o Verdadero (V).
- ¿Cuál es tu factor sanguíneo? Positivo (+) o negativo (-).

- Pregunta politómica. Es aquella que tiene más de dos alternativas y pueden elegirse más de una. Permite obtener más información.

Ejemplos.

- ¿Cuál es su grado de instrucción?

  A. Inicial.

  B. Primaria.

  C. Secundaria.

  D. Superior.

- ¿Con qué frecuencia utiliza el servicio de carga?

  a) Diario

  b) Semanal

  c) Mensual

b. Cuestionario Abierto.

Es aquél en el que la respuesta no tiene sugerencia alguna, se puede responder como se quiera. Las preguntas de esta naturaleza proporcionan respuestas de mayor profundidad, requieren más tiempo para ser respondidas, se usa en estudios pilotos para obtener información sobre la variable que se investiga.

Ejemplos.

¿Por qué te gusta Estadística?

……

¿Cuáles son los indicadores para determinar el índice de desarrollo humano?

……

c. Cuestionario Mixto.

Es una combinación de los dos tipos de cuestionarios anteriores, mayormente las preguntas abiertas se orientan a justificar las respuestas de las preguntas cerradas.

Ejemplos.

- ¿Está de acuerdo con la decisión del jurado?
  ( ) Sí
  ( ) No
  ¿Por qué? ……..……

- ¿Se considera lo suficientemente informado sobre los distintos aspectos que conforman este curso?
  ( )Sí
  ( ) No
  ¿Por qué? …….

En el ámbito educativo, los tipos de cuestionarios más usados son las listas de cotejo y las pruebas de aprovechamiento.

a. Lista de cotejo.

Es un cuestionario de preguntas de contenido sobre los aspectos que se desea evaluar, la medición puede ser cualitativa o cuantitativa. La

elaboración demanda conocimiento de la temática en extensión y profundidad para que el cotejo disminuya la subjetividad al momento de cotejar.

Ejemplo.

Tabla 2.1
Indicadores para evaluación de proyectos

1: Hecho ,2: Pendiente ,3: No realizado	*1*	*2*	*3*
¿Los resultados del proyecto fueron presentados y socializados?			
¿Las evaluaciones del desarrollo fueron incluidas en la evaluación final?			
¿Prepararon la evaluación de acuerdo a lo planteado en el proyecto?			
¿Se registró información para la evaluación durante la ejecución del proyecto?			
Se recolectaron todos los registros, trabajos, informes, etc., para la evaluación final.			
Los objetivos y criterios de la evaluación fueron conocidos desde el inicio.			
Todos los involucrados en el proyecto analizaron los resultados de la evaluación.			

b. Pruebas de aprovechamiento.

Se usan para evaluar el aprendizaje en las diferentes áreas del conocimiento, estos instrumentos deben ser diseñados de acuerdo a las competencias que se quieren logran en los procesos de enseñanza y aprendizaje y tomar en cuenta su validez de contenido fundamentalmente.

2.2.2 Escalas de actitudes.

Las escalas de actitudes son constructos teóricos**,** para medir aspectos o cualidades de personas o grupos a quienes se les solicita que marquen o señalen valorando su respuesta a una serie de preguntas.

Las actitudes, intereses y valores se miden con escalas, cada una de estas características se denominan dimensiones» o factores.

Las actitudes no se observan directamente, se infieren a partir de las expresiones que vierten o las conductas que manifiestan.

Toda actitud tiene tres características:

La dirección que puede ser positiva o negativa.

La magnitud o grado de valoración de la respuesta.

La intensidad o puntuación de la respuesta.

La suma total de las valoraciones indica la dirección e intensidad de la actitud. su construcción.

Las **escalas múltiples o baterías de reactivos**, se utilizan para establecer o comprobar un perfil de actitudes, entre estas escalas tenemos:

La escala de Osgood o de diferencial semántico, la escala de Stapel o unipolar, la escala de Likert o aditiva, la escala de Gutmann o de Escalograma y la escala de Thurstone o de Intervalos Iguales; de estas trataremos las tres primeras por ser las más usadas en la investigación y particularmente la escala de Likert.

a. Escala de Charles Osgood o de Diferencial Semántico.

La escala de Osgood se fundamenta en el diferencial semántico, se usa para medir la actitud hacia ciertos objetos, hechos, situaciones o personas, con expresiones directas del sentido denotativo (real) y connotativo (figurado) de la palabra. Por ejemplo, la expresión: "aquella ventana está limpia" tiene significado denotativo, mientras que "tus labios de rubí" posee significado connotativo, simbólica o figurada de la palabra.

Se usan adjetivos calificativos y sus antónimos (diferencial semántico), con grados (posibilidades de respuesta) impares (3,5 o 7) o pares (2,4,6 u 8).

Las actitudes que se expresan en este modelo pueden ser valorativas como bueno o malo; de potencia como fuerte o débil, también pude ser de actividad como rápido o lento.

Ejemplo 1.

Tabla 2.2
Valoraciones opuestas en diferencial semántico

Bueno	+3	+2	+1	0	-1	-2	-3	Malo
Limpio	+3	+2	+1	0	-1	-2	-3	Sucio
Dulce	+3	+2	+1	0	-1	-2	-3	Amargo
Fuerte	+3	+2	+1	0	-1	-2	-3	Débil
Grande	+3	+2	+1	0	-1	-2	-3	Pequeño
Pesado	+3	+2	+1	0	-1	-2	-3	Ligero
Activo	+3	+2	+1	0	-1	-2	-3	Pasivo
Rápido	+3	+2	+1	0	-1	-2	-3	Lento
Caliente	+3	+2	+1	0	-1	-2	-3	Frío

Ejemplo 2.

Marca con X la posición política cultural de la universidad peruana:

Tabla 2.3
Valoraciones positivas en diferencial semántico

Pasiva	1	2	3	4	5	6	7	Activa
Progresiva	1	2	3	4	5	6	7	Conservadora
Superficial	1	2	3	4	5	6	7	Profunda
Dogmática	1	2	3	4	5	6	7	Crítica

b. Escala de Alexander Stapel o Unipolar.

Se ha elaborado para medir al mismo tiempo la dirección y la intensidad de las actitudes, es una escala de clasificación de 10 puntos, positivos y negativos (+ 5 y -5) donde se responde con un solo adjetivo o frase que describa el objeto a evaluar, se analiza igual que el diferencial semántico.

Ejemplo.

Evalúa la precisión con que cada atributo describe mejor el café con leche en el desayuno. Si en tu opinión, el atributo lo describe totalmente, entonces marcarás + 5, y si no lo describe para nada marcarás -5.

Tabla 2.4
Valoración de atributos del objeto de investigación

+5	+5	+5	+5	+5
+4	+4	+4	+4	+4
+3	+3	+3	+3	+3
+2	+2	+2	+2	+2
+1	+1	+1	+1	+1
Aroma	**Color**	**Efecto vigorizante**	**Efecto energizante**	**Sabor**
-1	-1	-1	-1	-1
-2	-2	-2	-2	-2
-3	-3	-3	-3	-3
-4	-4	-4	-4	-4
-5	-5	-5	-5	-5

c. Escala de Rensis Likert o Aditiva.

Para elaborar la escala de Likert se debe definir primero el concepto que se va a investigar, esto implica la validez de contenido (estado del arte) incluyendo el juicio de expertos, para ello se sigue los siguientes pasos:

1. Proponer un número de preguntas abiertas sobre la variable a investigar tomando en cuenta sus dimensiones.

2. Aplicar el cuestionario abierto a una muestra de la población a estudiar.

3. Confirmar, a partir de las respuestas, la dirección de la actitud hacia lo que se pregunta en cada ítem.

4. Reformular las preguntas para que puedan ser respondidas usando una escala valorativa que varía entre actitud desfavorable y favorable. Los autores del libro prefieren una escala con una valoración de cuatro alternativas para evitar la respuesta de estar indeciso. En los enunciados desfavorables se invierte las alternativas de respuesta.

5. Solicitar la evaluación interdisciplinaria por expertos de los ítems en el cuestionario, sobre: claridad, relevancia, coherencia, relación lógica con la dimensión conceptual y suficiencia, para obtener la medición de la variable.

6. Realizar un muestreo piloto para aplicar el cuestionario con escala de Likert y determinar su confiabilidad.

7. Seleccionar solamente los ítems que tengan mayor desviación estándar para mejorar el índice discriminante.

Para evaluar la escala de Likert:

1. Se suman las puntuaciones obtenidas en cada ítem por cada sujeto, la suma total indica la actitud hacia la variable en estudio.
2. Se consideran con la puntuación de un ítem con respuesta neutra, en caso que un ítem no sea respondido.

Tabla 2.5
Valoraciones con puntuaciones en una escala de Likert

Alternativa A		**Alternativa B**		**Alternativa C**		**Alternativa D**		**Alternativa E**	
Nada de acuerdo	1	Muy de acuerdo	5	Totalmente de acuerdo	5	Definitivamente sí	5	Completamente verdadero	5
Algo en desacuerdo	2	De acuerdo	4	De acuerdo	4	Probablemente sí	4	Verdadero	4
Ni acuerdo ni desacuerdo	3	Ni de acuerdo, ni desacuerdo	3	Neutral	3	Indeciso	3	Ni falso, ni verdadero	3
Algo de acuerdo	4	En desacuerdo	2	En desacuerdo	2	Probablemente no	2	Falso	2
Completo desacuerdo	5	Muy en desacuerdo	1	Totalmente en desacuerdo	1	Definitivamente no	1	Completamente falso	1

Tabla 2.6
*Actitud de los trabajadores de una empresa. Escala de Likert**

DP	Proposiciones	Valoraciones				
		DA	D	I	A	AA
		1	2	3	4	5
+	1.Los trabajadores que tienen una personalidad definida son responsables de sus acciones					X
+	2.Los trabajadores que han cometido inconductas no deben ser considerados en la lista de ascensos.				X	
+	3. Los trabajadores que permanentemente llegan tarde no muestran identificación con la empresa.			X		
+	4.La empresa debe tener un manual de méritos por las actividades que realizan sus trabajadores.					X
+	5.La probabilidad de que la empresa de nuevas oportunidades al trabajador conflictivo es muy baja.				X	
+	6.Es imposible confiar en un trabajador que tiene signos de ser adicto a las drogas					X
_-	7.Las inconductas pueden ser asumidas por los trabajadores	5	4 X	3	2	1
_-	8. Se gasta demasiado dinero en capacitación para los trabajadores.	5	4 X	3	2	1
	PUNTAJE TOTAL= 34		8	3	8	15
D P: Dirección de las proposiciones						
AA: Acuerdo Absoluto, A: Acuerdo, I: Indiferente, D: Desacuerdo, DA: Desacuerdo Absoluto,						
*elaborado en base a la tabla de Raúl Pino Gotuzzo en Metodología de la Investigación						

2.2.3 Inventarios.

Están elaborados por preguntas a las que la persona responde indicando su posición ante ellas sin que esta reciba la calificación de correcta o incorrecta.
Antes de ser aplicados en forma definitiva es recomendable realizar un ensayo piloto a fin de develar algunos errores, determinar el tiempo de respuestas, instrucciones, claridad de las preguntas, etc.

Es frecuente en la práctica educativa, sobre todo a nivel inicial, la aplicación de inventarios con el propósito de conocer características de la personalidad, psicológicas, lingüísticas, hábitos y aprestamientos para iniciar la educación de los niños. Además, pueden usarse en niveles de estudios superiores para describir las características de la personalidad, la inteligencia, el rendimiento, actitudes, repitencia y deserción de los estudiantes.

Un ejemplo de inventarios es el test de inteligencias múltiples que indica: inteligencia lingüística, lógico matemática, musical, naturalista, etc.

Inventario de la personalidad

Un inventario de personalidad consta de un conjunto de ítems agrupados en escalas destinadas a evaluar rasgos de la personalidad de una persona, su perfil de pensamientos actitudes y comportamientos; se

administra de manera grupal o individual, antes de aplicar el inventario, el investigador debe conocer su validez y confiabilidad.

Ejemplo

Tabla.2.7

Inventario de personalidad

Marca el número que indique, que tan cierto es en relación con usted cada afirmación.

	1	2	3	4	5
1. Hago amigos con facilidad					
2. Tiendo a ser tímido					
3. Me gusta estar con otros					
4. Me gusta ser independiente de la gente					
5. Por lo general prefiero hacer las cosas sólo					
6. Siempre estoy en movimiento					
7. Me gusta salir y correr tan pronto me despierto en la mañana					
8. Me gusta mantenerme ocupado todo el tiempo					
9. Tengo mucha energía					
10. Prefiero los pasatiempos tranquilos e inactivos a los activos					
11. Tiendo a llorar con facilidad					
12. Me asusto con facilidad					
13. Tiendo a ser algo emocional					
14. Me enfado con facilidad					
15. Tiendo a irritarme con facilidad					

Tomado de Test Psicológicos y Evaluación, por Lewis R. Aiken. Undécima edición.

Antes de procesar los datos se debe analizar las respuestas de los ítems en base al contenido que se estudia, para este ejemplo la valoración de las respuestas de los ítems 2, 4, 5 y 10 debe invertirse para ser sumados.

Las preguntas del 1 al 5 miden el índice de sociabilidad, del 6 al 10 el índice de actividad y del 11 al 15 la emocionalidad.

CAPÍTULO III

VALIDEZ Y CONFIABILIDAD DE INSTRUMENTOS PARA MEDICIÓN DE VARIABLES

Los instrumentos para la medición de variables pueden ser documentales y físicos, para los documentales es necesario determinar su validez y confiabilidad y para los físicos su exactitud y precisión.

3.1 VALIDEZ DE UN INSTRUMENTO

La validez es el grado de coherencia lógica y metodológica de los ítems de un cuestionario para la obtención de información de la variable a

investigar; tiene tres dimensiones: la validez de contenido, la validez de constructo y validez de criterio.

3.1.1 Validez de contenido.

Todo instrumento tiene validez de contenido cuando las preguntas, para conocer la variable que deseamos medir, se fundamentan científicamente en la base teórica de la variable o relación de las variables del objeto de investigación.

La validez de contenido puede conocerse mediante tres métodos: el método del juicio de expertos, el método Delphy y el análisis factorial.

a) Método Juicio de Expertos.

En base al juicio emitido por uno, tres o cinco personas versadas en la materia que se investiga se pueden identificar errores en la construcción del instrumento. En la validación los expertos analizan independientemente: la relevancia, coherencia, suficiencia y claridad con la que están redactados los ítems del instrumento ajustándose a la base teórica.

Cada experto recibe un instrumento con información suficiente sobre los objetivos, el universo de contenido, la tabla de operacionalización de las variables, con los instrumentos de validación. Las preguntas en las que coinciden favorablemente los expertos, se toman en cuenta para el instrumento, en las que mantienen coincidencia parcial se reformulan o se cambian y

pueden nuevamente someterse a validación, y en aquellas que exista coincidencia desfavorable se eliminan.

Ejemplo.

Juicio de experto sobre validez de un instrumento

A-Datos generales

1. Instrumento de validación
2. Autores de la investigación
3. Grado académico del investigador
4. Institución donde estudia
5. Institución donde labora
6. Título de la investigación

B. Criterios para la validación

Tabla.3.1
Criterios de evaluación del instrumento para medición de variables

Criterio	**Valoración del indicador**	**1**	**2**	**3**	**4**	**5**
Claridad	Está formulado con un lenguaje claro					
Objetividad	Está expresado en conductas observables					
Actualidad	Responde al avance científico tecnológico					
Organización	Está estructurado en forma lógica					
Suficiencia	Presenta la cantidad adecuada de reactivos					
Intencionalidad	Está diseñado según los propósitos del estudio					
Consistencia	Presenta sustento científico					
Coherencia	Reactivos coherentes con la operacionalización de variables.					
Metodología	Responde al tipo y diseño de la investigación.					
Oportunidad	Está diseñado para aplicarse en el momento preciso.					
	Puntaje =					
1: deficiente	2: regular 3: buena 4: muy buena 5: excelente					

C. Opinión sobre la aplicabilidad.

1. Se recomienda su aplicabilidad ( )
2. Se recomienda mejorarlo antes de aplicar ( )

Fecha…….

…….
Firma del experto

D.N.I.

b) Método Delphy

Con respecto a este método, Hurtado (2002) sostiene que:

> Este procedimiento utiliza un grupo de expertos para el análisis que se mantienen aislados con objeto de minimizar el efecto de presión social y otros aspectos del comportamiento de pequeños grupos. Los expertos pueden ser especialistas internos o externos […]. Este método no requiere que se llegue a un consenso. El objetivo es más bien obtener un número de opiniones que se haya reducido por la aplicación del método […]. Como investigación es un proceso sistemático, formal y profundo para obtener y probar las hipótesis sobre el tema en cuestión […]. Este método nos permite consultar un conjunto de expertos para validar nuestra propuesta sustentado en sus conocimientos, investigaciones, experiencia, estudios bibliográficos, etc. Da la posibilidad a los expertos de analizar el tema con tiempo sobre todo si no hay posibilidades de que lo hagan de manera conjunta […]. Esta vía se caracteriza por permitir el análisis de un problema complejo […]. Siempre se [comienza] este proceso enviando un modelo a los posibles expertos con una explicación breve sobre los objetivos del trabajo y los resultados que se desean obtener. La secuencia establecida es la siguiente:

1. Se establece contacto con los expertos conocedores y se les pide que participen en panel.
2. Se manda un cuestionario a los miembros del panel y se les pide que den su opinión en los temas de interés.
3. Se analizan las respuestas y se identifican las áreas en que están de acuerdo y en las que difieren.
4. Se manda al análisis resumido de todas las respuestas a los miembros del panel, se les pide que llenen de nuevo el cuestionario y den sus razones respecto a las opiniones en que difieren.

También, es adecuado contar con una autovaloración sobre el grado de conocimiento o información que tienen sobre el tema a estudiar, indicando en una escala de 1 al 10, cuál es el nivel de conocimientos, así como, la experiencia valorándose como alta, media o baja, eligiendo el experto más competente.

c) Análisis Factorial.

"El análisis factorial es una técnica de reducción de datos que sirve para encontrar grupos homogéneos de variables […] que correlacionan mucho entre sí y procurando, inicialmente, que unos grupos sean independientes de otros" (de la Fuente, 2011, p.1). Para determinar los datos que se pueden reducir se parte de la varianza común y no común, de la Fuente (2011) explica estos dos conceptos del modo siguiente:

> Sean unos ítems de una escala de actitudes, donde la puntuación de cada sujeto encuestado es la suma de las respuestas a todos los ítems, según la clave de corrección diseñada:
> 1º Me lo paso muy bien en mi casa, con mis padres
> Muy de acuerdo = 5 De acuerdo = 4
> 2º Algunas veces me gustaría marcharme de mi casa
> Muy de acuerdo = 1 De acuerdo = 2
> […] La varianza de cada ítem puede ser compartida con la varianza de otros ítems: Algunos individuos encuestados están muy bien en su casa con sus padres (ítem 1) y nunca piensan irse de su casa (ítem 2). […] Esta relación viene expresada por el coeficiente de correlación 'r' de Pearson, donde r^2 expresa la proporción de varianza común o de variación conjunta. Es decir, si la correlación entre estos dos ítems es de 0,90, significa que tienen un 81% de varianza común (variación en las respuestas). El resto de la varianza (19%) [es no común]. (pp. 1-2)

3.1.2 Validez de constructo.

Un constructo es un objeto abstracto producto de la comprensión teórica sobre algo, no es susceptible de medición directamente si no a partir de los aspectos observables,

"La validez de constructo es el grado de correspondencia o congruencia que existe entre los resultados de una prueba y los conceptos teóricos

en los que se basan los temas que se pretenden medir" (Mejía Mejía 2005, p.35).

Paul Meehl y Lee Cronbach (1956) proponen tres pasos para evaluar esta validez. Vásquez Sánchez et al. (2021) lo formulan de la siguiente manera:

1. Verificar la relación teórica entre los conceptos que contiene el instrumento y la variable de estudio.
2. Comprobar que los métodos y técnicas para medir los constructos sean adecuados.
3. Probar empíricamente las relaciones planteadas en el instrumento (p.48).

Tabla 3.2
Cuadro de análisis de constructos para elaborar instrumentos.

El instrumento parece realmente medir el constructo "A".	Teoría: Se encontraron investigaciones donde A se relaciona directamente con B, C y D y negativamente con E.	Si el instrumento mide "A", entonces sus resultados deben relacionarse con los resultados de las mediciones de B, C y D y negativamente con las de E.	El instrumento mide el constructo "A".

3.1.3 Validez de criterio

"Este tipo de validez denota la relación existente entre las puntuaciones de un instrumento de medición y una variable independiente externa (criterio), que mide directamente el comportamiento o las características en cuestión" (Ary et al., 1979/1982, pp. 205-206). Un

instrumento tiene validez de criterio si es:

Atingente. Que se corresponda directamente con el rasgo característico que se está midiendo.

Insesgado. Con preguntas cuya varianza sea mínima.

Estable. Que muestra confiabilidad para su replicabilidad.

Accesible. Que posee disponibilidad y aplicación práctica.

Los factores que tienden a distorsionar la validez de un instrumento son:

1. Contenido del instrumento.

- Instrumento sin información de la unidad de estudio.
- Instrucciones imprecisas.
- Preguntas sin coherencia lógica.
- Falta de estructuras semántica y sintáctica correctas de las preguntas.
- Preguntas con diferentes interpretaciones.
- Mal uso de conectores lógicos en las preguntas.
- Preguntas sin contenido de la variable en estudio.
- Falta de orden en las preguntas.
- Preguntas con patrón de respuestas identificable.

2. Administración del instrumento.

- Espacio geográfico inadecuado.
- Tiempo insuficiente para su aplicación y desarrollo.

- Respuestas influenciadas por terceros
- Factores incontrolables al normal desarrollo.
- Distractores evidentes en el desarrollo.
- Elección de grupos de aplicación inadecuados.

3. Calificación del instrumento.

- Sin plan de análisis de datos con relación a los objetivos.
- Respuestas al azar de las preguntas.
- Respuestas atípicas del instrumento.
- Métodos estadísticos inadecuados para el análisis.

3.2 CONFIABILIDAD DE UN INSTRUMENTO

a) Método Test-Re Test.

"Consiste en repetir la aplicación del test al mismo grupo y correlacionar las puntuaciones obtenidas [, es muy adecuado] para la medición de aptitudes físicas y atléticas, test de personalidad y motoras" (Corral, 2009, p. 239).

Se usa la fórmula de correlación de **Pearson** siempre que:

1. La correlación sea lineal entre las variables.
2. Las variables sean continuas y en escala de intervalo o de

razón con distribución normal.

3. Exista variación homogénea entre las variables (homocedasticidad).
4. La variable X sea la primera aplicación y la variable Y sea la segunda.

$$r = \frac{n\sum \mathrm{XY} - \sum \mathrm{X} \sum \mathrm{Y}}{\sqrt{[n \sum \mathrm{X}^2 - (\sum \mathrm{X})^2][n \sum \mathrm{Y}^2 - (\sum \mathrm{Y})^2]}}$$

b) Método de división por mitades o hemi-test.

Para este caso el test se divide en dos mitades equivalentes y de igual longitud, y se utiliza la fórmula de Pearson para correlacionar los puntajes de la primera mitad llamada X con los de la segunda mitad llamada Y. Al *r* calculado se aplica la corrección de Spearman-Brown para longitud doble:

$$R = \frac{2r}{1 + \mathrm{r}}$$

De este modo, la correlación *R* estima el valor como si cada mitad tuviera el doble de ítems. Como el énfasis se pone al puntaje de los sujetos, *R* brinda la consistencia interna del cuestionario.

c) El coeficiente de correlación de Spearman.

Se utiliza cuando las variables se miden con escala ordinal y es posible establecer rangos a sus valores; cada rango de las variables se le asigna un número. En la siguiente fórmula, X' y Y' son los valores numéricos asignados, $D = X' - Y'$ y n es el número de pares de datos.

$$r = 1 - \frac{6\sum D^2}{n(n^2 - 1)}$$

Ejemplo:

Tabla 3.3
Ejemplo de cálculos para la determinación del coeficiente de Spearman

Datos		Ordenación de puestos				D	D^2
X	Y	X'		Y'		X'-Y'	
7	4	2	7 puesto 2	1	4 puesto 1	2-1	1
5	7	1	5 puesto 1	2	7 puesto 2	1-2	1
8	9	3	8 puesto 3	4	9 puesto 4	3-4	1
9	8	4	9 puesto 4	3	8 puesto 3	4-3	1

$$r = 1 - \frac{6\sum D^2}{n(n^2 - 1)}$$

Reemplazando valores:

$$r = 1 - \frac{6(4)}{4(4^2 - 1)} = 1 - 0.4 = 0.6$$

d) Método de División por Mitades de Rulon.

En este método el cuestionario se aplica una sola vez, haciendo luego la separación del cuestionario en dos mitades, de manera

preferente una mitad estará constituida por los ítems pares haciendo de variable X y la otra mitad por los impares haciendo de variable Y, en forma paralela, dado que muchas veces los ítems tienen incremento gradual de la dificultad de respuesta, no es recomendable tomar la mitad del instrumento directamente.

Esta covariación o correlación entre las mitades mide la *consistencia interna* del instrumento, Rulon no exige homocedasticidad, o varianzas iguales en las mitades, considera que la diferencia entre las dos mitades es de carácter aleatorio.

$$r = \frac{1 - S_d^2}{S_t^2}$$

Dónde:

r = Coeficiente de correlación.

S_d^2= Varianza de la diferencia entre las puntuaciones de las mitades.

S_t^2= Varianza total de las puntuaciones del test.

Los pasos para el cálculo de *r* son:

1. Aplicar el cuestionario a la muestra determinada una sola vez.
2. Dividir el cuestionario en dos partes con el mismo número de ítems.
3. Calcular los puntajes en cada mitad.

4. Calcular la diferencia entre las puntuaciones de cada participante. $d = X\text{-}Y$ (*X* ítem impar, *Y* ítem par).
5. Calcular la varianza del total y la varianza de las diferencias, d.

"La consistencia interna de un cuestionario [que es una] característica fundamental de la confiabilidad, […] mide el grado de homogeneidad total y covariación que tienen las preguntas entre sí, requisito en la contrastación empírica de las hipótesis" (Vásquez Sánchez et al, 2021, p. 54). Para la estimación de la consistencia interna se usan los siguientes coeficientes:

a) Coeficiente Alfa De Cronbach.

Cuando las preguntas del cuestionario tienen respuestas politómicas, dicotómicas, de escala tipo Likert, si las pruebas son de ensayo o si los cuestionarios de encuestas son de opinión de opinión; si el valor del cálculo del coeficiente es mayor que 0.8 se considera que el instrumento es adecuado para investigar. Su fórmula es la siguiente:

$$\alpha = \frac{K}{(K-1)}\left[1 - \frac{\sum S_i^2}{S_T^2}\right]$$

Donde:

$K =$ Número de ítems del instrumento.

$S_i^2 =$ Varianza de cada ítem.

$S_T^2 =$ Varianza total de los ítems.

Ejemplo.

En la Tabla 3.4 se muestran los datos de la aplicación de un cuestionario a 9 personas, para el que se obtiene un alfa de Cronbach de 0,83, lo que significa que el instrumento es confiable para la investigación.

Tabla 3.4

Respuesta a 9 ítems por 9 personas en la escala de Likert

Encuestados	0: Totalmente en desacuerdo,1: En desacuerdo, 2: De acuerdo, 3: Totalmente de acuerdo									Suma de respuestas C/ Ítem
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Total**
1	3	3	2	3	2	3	3	3	3	25
2	3	3	2	3	2	3	3	3	3	25
3	3	3	3	3	2	2	3	2	3	24
4	1	2	3	3	1	2	3	3	3	21
5	2	3	3	3	3	2	3	3	3	25
6	3	3	3	3	3	3	3	3	2	26
7	3	2	3	2	2	2	1	2	2	19
8	3	3	3	2	2	2	2	2	3	22
9	2	2	0	1	0	2	2	2	2	13
Varianza, S^2	0.53	0.25	1.03	0.53	0.86	0.25	0.52	0.28	0.25	17.19

$$\alpha = \frac{9}{8}\left[1 - \frac{4{,}5}{17{,}19}\right] = 0{,}83$$

b) Kuder- Richarson 20.

La estimación de la confiabilidad de un instrumento con la

fórmula 20 de Kuder-Richardson se usa si el instrumento se aplica a la muestra de la investigación una sola vez y tiene alternativa de respuesta dicotómica o de la forma correcta e incorrecta con diferentes índices de dificultad.

$$\mathrm{KR}_{20} = \frac{K}{K-1}\left[\frac{S_T^2 - \sum_{i=1}^{K} p_i q_i}{S_T^2}\right]$$

Donde:
$K =$ Número de ítems del instrumento.

$S_T^2 =$ Varianza total de los ítems.

$p_i =$ Proporción de respuestas correctas del i-ésimo ítem.

$q_i =$ Proporción de respuestas incorrectas del i-ésimo ítem.

c) Kuder-Richarson 21

Esta fórmula se utiliza en casos similares que KR_{20}, pero con la condición de que todos los ítems poseen igual varianza e índice de dificultad.

$$\mathrm{KR}_{21} = \frac{K}{K-1}\left[1 - \frac{\overline{Y}(K-\overline{Y})}{KS_T^2}\right]$$

Donde:

$K =$ Número de ítems del instrumento.

$S_T^2 =$ Varianza total de los ítems.

$\bar{Y} =$ Media aritmética de los puntajes totales.

Respecto a la confiabilidad de instrumentos Vásquez et al (2021) acotan:

> Existen instrumentos para recolección de datos que por su naturaleza no es posible determinar su confiabilidad mediante fórmulas, tales como: guías de observación, hojas de registros, listas de cotejo, rúbricas, entrevistas, inventarios, algunos tipos de encuestas, entre otros. En estos casos el juicio de los expertos juega un rol muy importante.
>
> También existen instrumentos que son válidos y confiables desde el momento en que se consignan los datos como es el caso de las historias clínicas, dado que son los expertos, los especialistas que brindan la información. (p. 57)

Generalmente la estructura formal de un instrumento para la toma de datos es la siguiente:

Encabezado

-Institución que realiza la investigación.

-Nombre del instrumento

-Objetivo de la investigación

Cuerpo

-Datos de la entidad o persona investigada.

-Instrucciones para responder las preguntas.

-Cuestionario de preguntas sobre la variable o variables del objeto de investigación.

Observaciones. Aclaraciones sobre el recojo de los datos o la medición de las variables.

Fecha de aplicación.

CAPÍTULO IV

TÉCNICAS PARA LA RECOLECCIÓN DE DATOS

Elaborado el cuestionario en base a los indicadores, elementos teóricos de las variables en estudio, y determinado el diseño de contrastación de la hipótesis, es necesario precisar la manera como se van a obtener los datos, esto es, elegir las técnicas de recolección que pueden ser, la documentación, la observación científica, la entrevista, la encuesta y las técnicas psicosociales.

4.1 DOCUMENTACIÓN

Esta técnica consiste en la búsqueda de información en documentos para construir el nuevo conocimiento, orientada por los objetivos de la investigación. Su fundamento está en el análisis de contenido de la

documentación que se revisa, pues constituye el marco teórico del objeto de la investigación.

Los libros, monografías, tesis, artículos (expuestos en conferencias y congresos) e informes técnicos constituyen las fuentes primarias, mientras que las fuentes secundarias son los resúmenes, enciclopedias, manuales, entre otros. La información de estas fuentes puede ser argumentativa, si trata de probar algo nuevo, o expositiva, cuando simplemente muestra el conocimiento establecido.

4.2 OBSERVACIÓN CIENTÍFICA

"Es un modo refinado de aprehender el mundo perceptible y de poner a prueba nuestras ideas sobre el mismo: está influenciada por el conocimiento científico" (Bunge, 1969/1973, p.729).

La observación científica es la técnica más usada en la investigación básica, donde inicia toda comprensión empírica de la realidad, posee los siguientes elementos:

> El *objeto* de la observación, el *sujeto* u observador (incluyendo, como es natural, sus percepciones), las circunstancias de la observación (o medio ambiente del objeto y el sujeto), los *medios* de observación (sentidos, instrumentos auxiliares y procedimientos), y el cuerpo de *conocimiento* en el cual se

encuentran relacionados los anteriores elementos. (Bunge, 1969/1973, p.729)

Según Bunge (1969/1973) tiene las siguientes características:

- **Intencionada:** porque se hace con un objetivo determinado.
- **Ilustrada:** porque va guiada de algún modo por un cuerpo de conocimiento.
- **Selectiva:** porque permite discriminar el objeto de la observación.
- **Interpretativa:** porque describe y explica el objeto de la observación (p.727).

La observación científica es de dos tipos: directa, cuando el objeto es perceptible, e indirecta, cuando surge de hipótesis a partir de una observación directa. Por ejemplo, en la investigación-acción, para tener una observación directa más objetiva, el investigador se involucra en la actividad desde el interior del grupo, esta técnica también es recomendada en la investigación etnográfica.

4.3 ENTREVISTA

Kerlinger (1973/1975) define a la entrevista como "una confrontación interpersonal, en la cual una persona (el entrevistador) formula a otra

(el respondiente) preguntas cuyo fin es conseguir contestaciones relacionadas con el problema de investigación" (p.338).

Entrevistas según el tipo de preguntas.

Las **entrevistas cerradas,** o estructuradas, cuentan con un sistema de preguntas preestablecido por el entrevistador acerca de la información que desea conocer.

Las **entrevistas abiertas,** o no estructuradas, si las preguntas se realizan libremente, acá el entrevistado aborda la respuesta de la forma que cree conveniente.

Las **entrevistas semiabiertas, s**e realizan combinando el modo abierto y cerrado.

La entrevista a profundidad.

En una entrevisto a profundidad, tanto el entrevistador como los entrevistados deben tener conocimiento amplio de lo que se investiga y manejar expresiones lingüísticas de contenidos que se comprendan pues sigue el modelo de conversación. Entre ellas se encuentran:

a) **La historia de vida:** Se realiza generalmente a personas que han pasado alguna experiencia académica socioeconómica o cultural especial importante que se desea investigar lo más objetiva posible.

b) **Las entrevistas pedagógicas:** Se realizan sobre acontecimientos y actividades que no se pudieron observar directamente

c) **Grupo focal:** Es una entrevista en grupo guiada sobre una temática preestablecida. El conductor debe ser conocedor de la temática a fin de que se produzca la información lo más objetiva y actualizada. Se usa especialmente en investigación educativa, sociológica y económica.

4.4 ENCUESTA

Es una técnica para obtener información con procedimientos estandarizados de interrogación que se realiza a las personas de manera individual o colectiva para conocer sus actitudes, comportamientos o valoraciones acerca de algo que se quiera investigar. El cuestionario de preguntas que se usa en la encuesta puede ser autoadministrado o aplicado por un encuestador, que no necesariamente es el investigador o del equipo de investigación.

Las preguntas que brindan mayor información son las abiertas, porque habría variedad de respuestas. En las preguntas de respuestas cerradas la opción "otros" es una salida cuando no se tiene respuesta anticipada.

La técnica de la encuesta ha sido desarrollada por psicólogos, sociólogos, economistas, politólogos y estadísticos. Por lo general, se

investiga con encuestas: factores demográficos, socioeconómicos y culturales, como hábitos, costumbres, calidad de vida, etc.

Tipos de encuestas.

Las **encuestas para investigación exploratoria,** se usan cuando existe muy poca información sobre algún hecho, suceso o fenómeno, para determinar cómo desarrollar hipótesis, establecer relaciones y plantear problemas adecuadamente en posteriores investigaciones.

Las **encuestas para investigación descriptiva,** se usan para describir un fenómeno, caracterizarlo o diferenciarlo de otro. Estas encuestas permiten la descripción de las relaciones entre variables en una determinada situación.

Las **encuestas para investigación explicativa,** se usan para develar probables relaciones causales entre variables del objeto de investigación, además que permite descubrir otras variables explicativas (independientes) y variables extrañas para poder controlarlas a priori con el diseño o después en el análisis de datos.

La aplicación de encuestas se realiza también con carácter evaluativo y predictivo, con fines políticos, sociales, comerciales, etc.

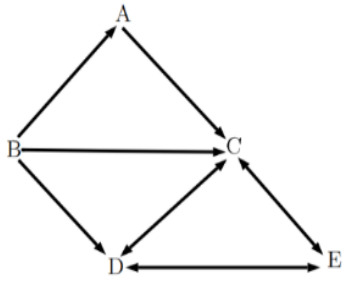
4.5 TÉCNICAS SOCIOMÉTRICAS Y PSICOMÉTRICAS

Las técnicas sociométricas se usan para estudiar la organización y coherencia de un grupo social, para esto sus miembros indican con quienes les agradaría trabajar o realizar actividades específicas, pudiendo ser por orden de prioridad.

Ejemplo.

A 5 personas (A, B, C, D y E) de un grupo social se les solicita elegir con quien o quienes les gustaría evaluar un proyecto de investigación. Luego, con la información obtenida, se construye el sociograma de la Figura 4.1, en el cual se aprecia que C es el más preferido, D, C y E se escogen entre sí, A escoge solamente a C, B escoge a A, C y D, por último, nadie escoge a B.

Figura. 4.1
Sociograma de un grupo.



Las técnicas psicométricas se usan para estudiar el comportamiento de las personas mediante preguntas o ítems validados en un test, que pueden ser de distribución normal o de distribución criterial que siguen modelos teóricos de referencia.

CAPÍTULO V

ANÁLISIS ESTADÍSTICO DE DATOS

La recopilación de datos, en la investigación científica, es un proceso organizado y sistemático en el que se aplica métodos y técnicas estadísticas teniendo como soportes la base conceptual y operacional de las variables y los instrumentos.

Por otro lado, el análisis de datos es el procesamiento de los mismos con el propósito de obtener medidas estadísticas que nos permitan realizar la descripción científica de las variables en estudio, de acuerdo a los objetivos que se persigue en la investigación.

5.1 ETAPAS DEL ANÁLISIS ESTADÍSTICO DE DATOS

5.1.1 Codificación y procesamiento de los datos.

Realizada la medición de las variables del objeto de investigación, se tienen los datos, éstos se codifican asignándoles números, letras o símbolos que permitan su procesamiento. Luego de la codificación se preparan las tablas y figuras para presentar los resultados que dependen de la definición conceptual y operacional de las variables, de las que se calculan estadísticas descriptivas e inferenciales. Este procesamiento puede realizarse de manera manual si el volumen de datos es pequeño, como generalmente se tiene grandes volúmenes se usan softwares estadísticos adecuados como Excel o SPSS.

a) Codificación de datos.

Las variables se codifican para posibilitar el análisis e interpretación de los resultados. Las respuestas a las preguntas cerradas se codifican al diseñarlas, mientras que las respuestas a las preguntas abiertas o semiabiertas se codifican después de recoger los datos

Ejemplo.

Marque con aspa

¿Cuál es su sexo?

Masculino ( )

Femenino ( )

¿Cuál es su instrucción?,

Primaria ( )

Secundaria ( )

Superior ( )

Codificación: Primero se categoriza las variables, luego se codifican las alternativas de respuesta como se muestra en la tabla 5.1.

Tabla 5.1
Codificación de la variable sexo y grado de instrucción.

Variable	Categoría	Código
Sexo	Masculino	01
	Femenino	02
Grado de instrucción	Primaria	01
	Secundaria	02
	Superior	03

Ejemplo.

¿Por qué no comprará electrodomésticos el próximo año?

Contestaron así:

a) No tengo dinero para comprarlo.

b) Son muy caros.

c) Espero que bajen de precio.

d) Son grandes para mi cocina.

e) Se malogran muy rápido.

Codificación: Las respuestas a, b y c pueden entrar en una categoría denominada “falta de dinero” y asignarles el código 1; las respuestas d y e entran en la categoría “inconformidad con el producto” a la que se le asigna el código 2.

b) Procesamiento de datos.

Abarca desde la organización de los datos codificados hasta la presentación de los resultados, que contienen las variables del objeto de investigación o sus dimensiones, en tablas y figuras. Si estas presentan valores puntuales, en ocasiones es preferible escribirlos en un párrafo.

La **tabulación** es la preparación de tablas para el análisis de datos de la variable o la relación de variables de acuerdo a lo propuesto en la hipótesis y objetivos de la investigación. En esta parte se realiza el conteo de respuestas y la obtención de los resultados según los modelos o pruebas estadísticas que se requieran. Las categorías de clasificación para el análisis están condicionadas al concepto o definición y tipo de las variables que se investigan. Generalmente de éstas se obtienen frecuencias absolutas y porcentuales, promedios, desviaciones estándar y pruebas de significación en el caso de contrastación de hipótesis.

En SPSS, la base de datos de una investigación se consigue introduciendo estos de la siguiente forma: los individuos o unidades de estudio por columnas y las variables por filas, para esto se configura su codificación respectiva en la ventana "Vista de variables".

Elementos de una tabla de resultados de investigación.

El título.

Debe contener la variable o relación de variables que se investiga. Si el lugar y el tiempo condicionan en alguna medida los resultados de la investigación, también deben ser considerados.

El cuerpo.

Llamado también campo de la tabla, se compone de filas y columnas, y contiene las categorías, dimensiones o medidas estadísticas de las variables en estudio. Los totales marginales o generales se consideran sólo si el tipo de tabla y el objetivo planteado lo ameritan.

Las notas.

Se ubican en la parte inferior de la tabla y se utilizan para brindar información extra que ayude a la completa comprensión de los contenidos del cuerpo. En ocasiones son llamadas desde el mismo cuerpo a través de super índices utilizando números, letras o símbolos.

La fuente.

Indica el origen de los datos, se ajusta a la norma general de la cita bibliográfica y se ubica en la parte inferior de la tabla. Si la tabla es de elaboración propia no es necesario escribir la fuente.

5.1.2 Análisis e interpretación,

El análisis se corresponde con la descripción natural de los resultados, presentados en tablas o figuras que han sido diseñadas de acuerdo a los objetivos de la investigación, a partir de éstas y haciendo uso de los antecedentes y la base teórica se llega a una correcta interpretación de los datos.

5.2 ANÁLISIS ESTADÍSTICO PARA UNA VARIABLE

5.2.1 Elaboración de tablas de distribución de frecuencias.

La distribución de frecuencias revela el comportamiento de la variable en estudio. Para obtener esta distribución es necesario: las categorías o clases para las cualitativas, o intervalos para las cuantitativas, según los fundamentos teóricos y estadísticos, y los objetivos de la investigación.

La amplitud de los intervalos dependerá siempre de la definición conceptual y operacional de la variable, pudiendo tener igual o diferente amplitud e incluso amplitud cero.

Para intervalos de igual amplitud interválica se procede del modo siguiente:

1. **Se halla el rango R.**

Llamado también amplitud total se obtiene por diferencia del máximo valor con el mínimo:

$$R = máx - mín\ .$$

2. **Se calcula el número de intervalos de clase.**

Para esto hacemos uso de:

$$m = 2.5\ \sqrt[4]{n}\ ,$$

o según la regla de Sturges:

$$m\ =\ 1\ +\ 3.3 \log n\ .$$

Donde n es el tamaño de muestra y m es el número de intervalos. Si la variable es discreta se usa intervalos cerrados: $[Y'_{j-1}, Y'_j]$, y si es continua será semiabierto, es decir un intervalo abierto por la izquierda y cerrado por la derecha: $(Y'_{j-1}, Y'_j]$, o viceversa: $[Y'_{j-1}, Y'_j)$.

3. **Se determina la amplitud y se construyen los intervalos.**

La amplitud es la distancia entre el límite superior y el inferior de un intervalo, ésta se calcula por:

$$c\ =\ \frac{R}{m}\ ;$$

con este valor de c se forman los intervalos fijando el mínimo valor de los datos y agregando la amplitud interválica hasta completar el rango.

Ejemplo de construcción de tabla de frecuencias.

Supongamos que los datos que siguen corresponden a cocientes de inteligencia de 30 alumnos de la facultad de ciencias físicas y matemáticas de una universidad.

90, 70, 78, 81, 101, 113, 120, 88, 100, 112, 97, 91, 83, 130, 93, 122, 93, 101, 111, 102, 111, 103, 109, 105, 101, 103, 85, 98, 98, 94.

Rango: $máx - mín = 130 - 70 = 60$.

Número de intervalos.

a) $m = 2.5\sqrt[4]{n} = 2.5\sqrt[4]{30} = 5.85 \approx 6$.

b) $m = 1 + 3.31 \log 30 = 5.89 \approx 6$.

Amplitud de intervalos.

$$c = \frac{R}{m} = \frac{60}{6} = 10 .$$

Determinación de los intervalos $(Y'_{j-1}, \ Y'_j]$:

Sean Y'_{j-1} y Y'_j los límites inferior y superior del intervalo j respectivamente, entonces Y'_0 es el mínimo valor del rango total por lo

que sí estará incluido en el primer intervalo. Haciendo los cálculos respectivos se obtiene: $Y_0' = 70$, $Y_1' = Y_0' + 10 = 80$, $Y_2' = Y_1' + 10 = 90$ y así sucesivamente hasta encontrar lo límites de todos lo intervalos.

La marca de clase Y_i:

Se obtiene calculando la semisuma de los límites del intervalo y sirve para graficar el polígono de frecuencias. Por ejemplo $Y_1 = \frac{70 + 80}{2} = 75$.

La frecuencia absoluta n_i:

Se obtiene contando los valores de la variable en el í-ésimo intervalo incluyendo el valor igual al límite superior. La sumatoria de las frecuencias absolutas da el tamaño de muestra: $\Sigma n_i = 30 = n$.

La frecuencia n_2 = 5 significa que 5 alumnos tienen un cociente intelectual entre 80 y 90 inclusive, para este conteo no se toma en cuenta 80 porque ya se consideró en el intervalo anterior; es decir se tomaron los valores inmediatamente mayores a 80.

Para el primer intervalo sí se incluye el límite inferior, por lo que n_1= 2 y estos valores son 70 y 78, como se aprecia no hay 80 para incluir en este intervalo.

La frecuencia relativa h_i :

Se obtiene por división de $\frac{n_i}{n}$ así, por ejemplo:

$$h_3 = \frac{n_3}{n} = \frac{8}{30} = 0.26666 \, .$$

$\sum h_i = 1$ (o aproximadamente 1) al graficar estas frecuencias se obtiene la distribución de probabilidad o distribución del tanto por uno. Si se suma iterativamente estas frecuencias se obtiene la frecuencia relativa acumulada H_i.

La frecuencia acumulada N_i :

Se obtiene sumando iterativamente las frecuencias absolutas.

$$N_1 = n_1 = 2 \, ,$$
$$N_2 = N_1 + n_2 = 2 + 5 = 7 \, ,$$
$$\vdots$$
$$N_6 = N_5 + n_6 = 28 + 2 = 30 \, .$$

La última frecuencia acumulada es igual al tamaño de la muestra, las frecuencias acumuladas se usan para construir la ojiva o polígono acumulativo.

La frecuencia porcentual h_i (100) ó %

Resulta de multiplicar la frecuencia relativa h_i por 100 así:

$$h_3(100) = 0.2666(100) = 26.66 \, .$$

El uso de porcentajes permite la comparación de frecuencias en igual o diferente tamaño de muestra, con igual cantidad de intervalos.

Tabla 5.2
Distribución de frecuencias del cociente intelectual de 30 alumnos de una universidad.

$(Y'_{i-1}, Y'_i]$	Y_i	n_i	h_i	N_i	%
70-80	75	2	0.0666	2	6.66
80-90	85	5	0.1666	7	16.66
90-100	85	8	0.2666	15	26.66
100-110	105	8	0.2666	23	26.66
110-120	115	5	0.1666	28	26.66
120-130	125	2	0.0666	30	6.66
Total		**30**	**0.9996**		**99.96**

5.2.2 Gráficas de distribuciones de frecuencias.

Las gráficas constituyen la representación objetiva de los resultados de una investigación y permiten descubrir la tendencia de la variable en estudio.

Histograma.

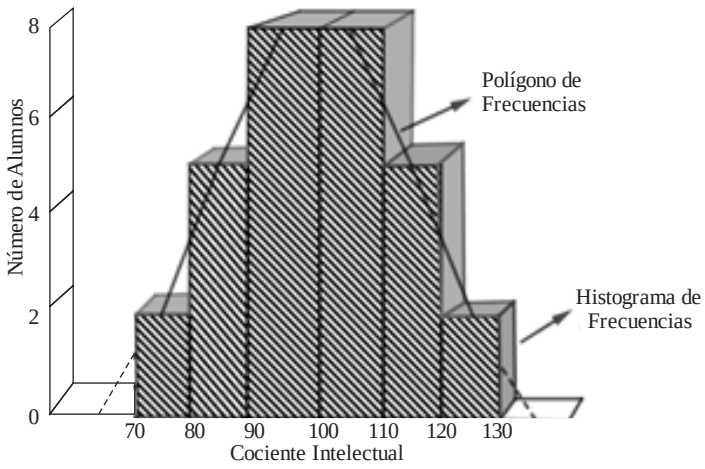
Es una gráfica de barras que representa las frecuencias de cada categoría o intervalo de las variables en estudio. Las barras se levantan una a continuación de otras si la variable es ordinal o continua y separadas si la variable es nominal o discreta, tomando en cuenta la amplitud interválica en la construcción de sus bases.

El polígono de frecuencias.

Muestra la variación de la variable en estudio y se construye uniendo los puntos formados por las marcas de clase de los intervalos y sus frecuencias absolutas; el primer y último punto se unen a la base mediante línea punteada para configurar el polígono. Para comparar distribuciones se puede trazar más de un polígono de frecuencias sobre la misma base.

Figura 5.1

Histograma y polígono de frecuencias del cociente intelectual de 30 alumnos de una universidad (Tabla 5.2).

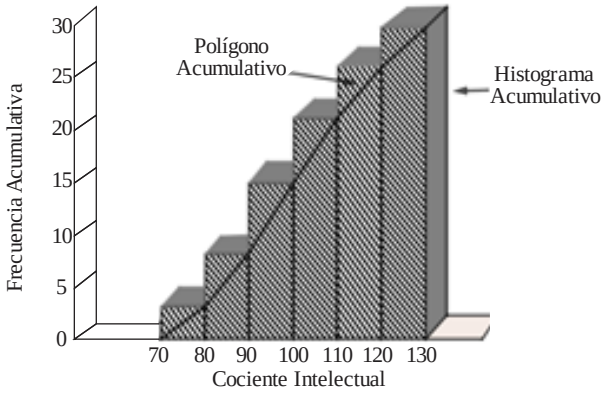


Histograma acumulativo y polígono acumulativo

El histograma acumulativo se forma con las frecuencias acumuladas, y el polígono acumulativo uniendo los puntos formados por el límite superior de cada intervalo y su frecuencia acumulada, tomando como

inicio el límite inferior del primer intervalo. Su uso es adecuado cuando las variables son discretas y continuas.

Figura 5.2
Histograma acumulativo y polígono acumulativo del cociente intelectual de 30 alumnos de una universidad (Tabla 5.2).



5.2.3 Estadísticas de centralización de datos.

Son valores que centralizan los datos y tienen por objetivo representarlos por medio de un número y darles una característica global, interpretándose de acuerdo a la naturaleza de la variable y el hecho o fenómeno que se estudia. Estas son:

La media aritmética.

La media aritmética o promedio, se calcula sumando los datos y dividiendo el resultado entre el tamaño de muestra. Cuando los datos **no están agrupados** el promedio se obtiene mediante la fórmula:

$$\bar{X} = \frac{\sum X_i}{n},$$

donde X_i es el i-ésimo dato, n es el tamaño de la muestra o número total de datos y $\bar{X}$ es el promedio aritmético. Si los datos fueran 0, 1, 2, 1, el promedio sería:

$$\sum X_i = 0 + 1 + 2 + 1 = 4\,, \quad n = 4\,,$$
$$\bar{X} = \frac{4}{4} = 1\,.$$

Para datos **agrupados en intervalos** el promedio se obtiene por:

$$\bar{Y} = \frac{\sum Y_i n_i}{n},$$

Y_i es la marca de clase del i-ésimo intervalo; n_i es la frecuencia o el número de datos en cada intervalo y n es el tamaño de muestra.

Supongamos la distribución:

Tabla 5.3
Distribución de frecuencias para el cálculo del promedio de datos agrupados.

$(Y'_{i-1}\ ,\ Y'_i\ ]$	Y_i	n_i	$Y_i n_i$
00 – 10	5. 0	4	20.0
10 – 13	11. 5	8	92.0
13 – 16	14. 5	16	232.0
16 – 18	17. 0	8	136.0
18 – 20	19. 0	4	76.0
T o t a l		40	556.0

$$\bar{Y} = \frac{\sum Y_i n_i}{n} = \frac{556}{40} = 13.9\,.$$

Como solamente se conocen los límites reales de los intervalos, puede que se dé algún error por agrupamiento. Un caso particular de datos agrupados en el que la amplitud interválica es 0 tiene la distribución:

Tabla 5.4
Ejemplo de distribución de frecuencias para el cálculo del promedio de datos agrupados con amplitud interválica cero.

X_i	0	1	2	3
n_i	3	2	1	0

Cada punto de recorrido de la variable X_i se repite un determinado número de veces. El promedio será:

$$\bar{X} = \frac{\sum X_i n_i}{n} = \frac{4}{6} = 0.67 .$$

La media ponderada.

En muchas situaciones los datos a promediar tienen ponderación (peso) diferente, en este caso los datos se multiplican por su ponderación y se divide entre la suma de tales, si se tienen los datos $X_1, X_2, \dots, X_n$ con sus pesos respectivos $W_1, W_2, \dots, W_n$, la media ponderada será:

$$\bar{X}_w = \frac{\sum X_i W_i}{\sum W_i} .$$

Por ejemplo, para la nota final de un curso, el promedio de pruebas escritas tiene peso 2 y un trabajo de aplicación tiene peso 3, si el promedio de pruebas es 10 y el trabajo tiene nota 11 su media ponderada será:

$$\bar{X}_w = \frac{10(2)+11(3)}{5} = 10.6 .$$

La media armónica.

Representada por $\bar{X}_h$, se calcula mediante la formula:

$$\bar{X}_h = \frac{n}{\sum \frac{1}{X_i}} .$$

La media armónica es utilizada:

- Cuando los datos siguen una progresión armónica.
- Para promediar velocidades; cuando las distancias recorridas a cada velocidad son iguales.
- Para promediar razones de precio por cantidad, cuando se ofrece cantidades variables de mercadería por la misma cantidad de dinero.
- Para promediar resistencias en paralelo.

Así, si deseamos determinar la velocidad media de un automóvil que ha hecho los primeros 10 km de un viaje a razón de 30 km/h y los segundos 10 km a 60 km/h.

$$\bar{X}_h = \frac{2}{\frac{1}{30}+\frac{1}{60}} = 40 \text{ km/h} .$$

La media geométrica.

Esta media exige que todos los datos sean mayores que cero, se designa por $\bar{X}_g$ y se calcula mediante:

$$\bar{X}_g = \sqrt[n]{X_1 . X_2 \dots X_n} .$$

La media geométrica se aplica:

-En caso en que los datos sigan una proporción geométrica.

-Para promediar, aumentos o disminuciones porcentuales de una cantidad en el tiempo.

Se conoce que en un año el precio de un litro de leche aumentó el 15% y el año siguiente el 8%, entonces el aumento promedio en los dos años se calcula de la siguiente manera:

$$\bar{X}_g = \sqrt{115 \times 108} = 111.45 \text{ ,}$$

obteniendo que el aumento promedio es del 11.45%.

La relación existente entre promedios es:

$$\bar{X}_h \leq \bar{X}_g \leq \bar{X} \text{ .}$$

La moda (Mo).

Es el valor que más se repite en un conjunto de datos, por ejemplo: si los datos son 0, 1, 2, 3, 3 la moda sería 3 ($Mo = 3$); si los datos son 0, 0, 1, 2, 3, 3 habría dos modas $Mo_1 = 0$ y $Mo_2 = 3$. En ciertos casos no existirá moda y se conocerá como amodal, por ejemplo: 0, 1, 2, 3, 4 ó 2, 2, 2, 2 en este último caso a pesar de que existe un valor que más se repite, éste es constante por lo que no se le considera moda.

Cuando los **datos están agrupados,** la moda se define como el valor alrededor del cual las observaciones tienden a concentrarse más y se calcula mediante la siguiente fórmula:

$$Mo = Y'_{j-1} + C_j \frac{n_j - n_{j-1}}{(n_j - n_{j-1}) + (n_j - n_{j+1})} \text{ ,}$$

siendo:

Y'_{j-1}: límite inferior del intervalo con mayor frecuencia (intervalo modal).

C_j: amplitud interválica modal.

n_j: máxima frecuencia.

n_{j-1}: frecuencia anterior a n_j.

n_{j+1}: frecuencia posterior a n_j.

Ejemplo:

Se desea calcular la moda de la siguiente distribución:

Tabla 5.5
Distribución de frecuencias para el cálculo de la moda de datos agrupados.

$(Y'_{i-1}$, $Y'_i]$	Y_i	n_i	
0 10	5.0	4	
10 13	11.5	8	n_{j-1}
13 16	14.5	16	n_j
16 18	17.0	8	n_{j+1}
18 20	19.0	4	

Para esto, se sustituyen los datos que se muestran en la tabla, y se obtiene:

$$Mo = 13 + 3\ \frac{8}{8+8} = 14.5\ .$$

Cuando existe más de un intervalo modal, se suman todas las frecuencias absolutas a partir del centro del polígono de frecuencias

hacia el extremo izquierdo y el extremo derecho. Luego se comparan ambas cantidades y se elige el intervalo ubicado en el lado con mayor sumatoria, procurando que esté lo más cerca al centro.

La mediana (Me).

Es el valor situado al centro de los datos ordenados de menor a mayor o viceversa. Si **n es impar**, la mediana es la observación que se ubica exactamente en el centro y si **n es par**, es la semisuma de las observaciones que ocupan el centro de la distribución. Por ejemplo, si los datos son 1, 2, 3, 4, 5, 6 la mediana será:

$$Me = \frac{3+4}{2} = 3.5\ .$$

Cuando los datos están distribuidos en intervalos de clase, el intervalo mediano es aquel que posee la menor frecuencia acumulada mayor a la mitad de la muestra, y la mediana se calcula por:

$$Me = Y'_{j-1} + C_j \frac{\frac{n}{2} - N_{j-1}}{N_j - N_{j-1}},$$

donde:

$\frac{n}{2}$: mitad de la muestra.

N_j: menor frecuencia acumulada mayor a $\frac{n}{2}$.

N_{j-1}: frecuencia acumulada anterior a N_j.

Y'_{j-1}: límite inferior del intervalo mediano.

C_j: amplitud interválica.

En la distribución de la tabla 5.6,

Tabla 5.6
Distribución de frecuencias para el cálculo de la mediana de datos agrupados.

$(Y'_{i-1}\ ,\ Y'_i]$	Y_i	n_i	N_i
0-10	5.0	4	4
10-13	1.5	8	12 (N_{j-1})
13-16	4.5	16	28 (N_j)
16-18	7.0	8	36
18-20	9.0	4	40
T o t a l		40	

la mediana es:

$$Me = 13 + 3\ \frac{\frac{40}{2}-12}{28-12} = 14.5\ .$$

Ejemplo de cálculo de medidas de centralización para una variable discreta.

El número de faltas ortográficas cometidas por 15 alumnos en un dictado de literatura es como sigue:

0, 6, 8, 8, 11, 10, 12, 13, 14, 13, 10, 21, 15, 17, 18.

-Cuando los datos no están agrupados.

Para el **promedio** se suman todos los datos y se divide entre el total, así:

$$\bar{X} = \frac{\sum X_i}{15} = \frac{176}{15} = 11.73 \approx 12\ .$$

Para calcular la **mediana**, se ordenan los datos en forma creciente (o decreciente) y se ubica el dato en la mitad del recorrido, de la siguiente manera:

0, 6, 8, 8, 10, 10, 11, **12**, 13, 13, 14, 15, 17, 18, 21.

$$Me = 12 \ .$$

Por último, para calcular la **moda**, se ubican los datos que más se repiten, en este caso se tiene tres modas que se repiten dos veces:

$Mo_1 = 8, Mo_2 = 10, Mo_3 = 13$.

-Cuando los datos se agrupan.

La distribución de frecuencias se hace de la siguiente manera:

El rango:

$$R = 21 - 0 = 21 \ .$$

Los intervalos:

$$m = 2.5 \sqrt[4]{15} = 4.92 \ .$$

La amplitud interválica:

$$C = \frac{R}{m} = \frac{21}{4.92} = 4.27 \approx 4 \ .$$

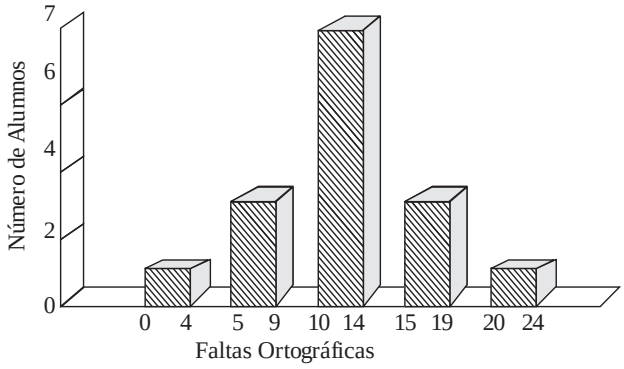
Como la variable faltas ortográficas es discreta, los intervalos de la distribución son cerrados. Por lo que la tabla de frecuencias queda de la siguiente manera:

Tabla 5.7
Distribución de frecuencias de faltas ortográficas de 15 alumnos.

$[Y'_{i-1}$, $Y'_i]$	Y_i	n_i	N_i	$Y_i n_i$
0 4	2	1	1	2
5 9	7	3	4	21
10 14	12	7	11	84
15 19	17	3	14	51
20 24	22	1	15	22
T o t a l		15		180

Figura 5.3

Histograma de frecuencias de faltas ortográficas de 15 alumnos (Tabla 5.7).



Para sus estadísticas de centralización se usan las fórmulas de datos agrupados.

Promedio:

$$\bar{Y} = \frac{\Sigma Y_i n_i}{n} = \frac{180}{15} = 12 \text{ faltas.}$$

Mediana:

$$Me = Y'_{j-1} + C_j \frac{\frac{n}{2} - N_{j-1}}{N_j - N_{j-1}} = 10 + 4\frac{7.5-4}{11-4} = 12 \text{ faltas.}$$

Moda:

$$Mo = Y'_{j-1} + C_j \frac{n_j - n_{j-1}}{(n_j - n_{j-1}) + (n_j - n_{j+1})} = 10 + 4\ \frac{4}{4+4} = 12 \text{ faltas.}$$

Si se compara las medidas de centralización halladas antes de agrupar los datos con las calculadas después, se puede observar que no siempre coinciden. Esto se debe a que pueden existir errores producidos por la agrupación que se realiza, el tamaño de muestra y los redondeos por la naturaleza de la variable.

La media y la mediana permiten analizar el sesgo de las distribuciones; si la distribución es normal, el promedio, la mediana y moda tienen valor muy parecido; si éstos coinciden la distribución es simétrica.

Para el análisis de los datos, si la variable se distribuye con mucha asimetría, evitar la media, si los datos son muy escasos en la parte central, evitar la mediana, y si carecen de un punto principal de concentración, evitar la moda.

5.2.4. Estadísticas de posición de datos.

Los cuartiles.

Son valores de la variable en estudio que dividen a la distribución ordenada de los datos, de menor a mayor, en cuatro partes de igual cantidad.

Si los datos **no están agrupados**, el cuartil dos (Q_2) es igual a la mediana, mientras que el cuartil uno (Q_1) y el cuartil tres (Q_3) son las medianas respectivas de las dos partes a ambos lados de Q_2.
En cambio, si los **datos se agrupan en intervalos**, los cuartiles se determinan con la siguiente fórmula.

$$Q_r = Y'_{j-1} + C_j \frac{\frac{rn}{4} - N_{j-1}}{N_j - N_{j-1}},$$

donde:

r: número de cuartil que se calcula.

n: tamaño de muestra.

N_j: menor frecuencia acumulada mayor a $\frac{rn}{4}$.

N_{j-1}: frecuencia acumulada anterior a N_j.

Y'_{j-1}: límite inferior del intervalo cuartil.

C_j: amplitud del intervalo.

En el ejemplo de faltas ortográficas de 15 alumnos en un dictado de literatura (Tabla 5.7), el cuartil 2 será:

$$Q_2 = 10 + 4\,\frac{\frac{2(15)}{4} - 4}{11-4} = 10 + 4\,\frac{3.5}{7} = 12\,.$$

Esto quiere decir que el 50% de los alumnos tienen 12 faltas ortográficas o menos.

Los deciles.

Son valores de la variable en estudio que dividen a la distribución ordenada de los datos, de menor a mayor, en diez partes de igual cantidad. Existen nueve deciles, de los cuales el decil cinco (D_5) es igual a la mediana.

Cuando están agrupados en intervalos su cálculo se determina por la siguiente fórmula:

$$D_r = Y'_{j-1} + C_j\,\frac{\frac{rn}{10} - N_{j-1}}{N_j - N_{j-1}}\,,$$

donde:

r: número del decil que se calcula.

n: tamaño de muestra.

N_j: menor frecuencia acumulada mayor a $\frac{rn}{10}$.

N_{j-1}: frecuencia acumulada anterior a N_j.

Y'_{j-1}: límite inferior del intervalo decil.

C_j: amplitud del intervalo.

En la distribución del número de faltas ortográficas (Tabla 5.7), el decil 6 será:

$$D_6 = 10 + 4\ \frac{\frac{6(15)}{10} - 4}{11-4} = 10 + 4\frac{5}{7} = 12.857 \approx 13\ .$$

Debido a que la variable es discreta, para elaborar una conclusión el valor del decil se redondea a 13. Por lo tanto, el 60% de los alumnos tienen 13 faltas ortográficas o menos.

Los percentiles.

Son valores de la variable en estudio que dividen a la distribución ordenada de los datos, de menor a mayor, en cien partes de igual cantidad. Existen 99 percentiles, de los cuales el percentil 50 (P_{50}) es igual a la mediana. En datos agrupados por intervalos se usa la siguiente fórmula:

$$P_r = Y'_{j-1} + C_j\ \frac{\frac{rn}{100} - N_{j-1}}{N_j - N_{j-1}}\ ,$$

donde:

r: número del percentil que se calcula.

n: tamaño de muestra.

N_j: menor frecuencia acumulada mayor a $\frac{rn}{100}$.

N_{j-1}: frecuencia acumulada anterior a N_j.

Y'_{j-1}: límite inferior del intervalo percentil.

C_j: amplitud del intervalo.

En el ejemplo del número de faltas ortográficas (Tabla 5.7) el percentil 75 será:

$$P_{75} = 15 + 4\,\frac{\frac{75(15)}{100} - 11}{14-11} = 15 + 4\frac{0.25}{3} = 15.33 \approx 15\ .$$

Debido a que la variable es discreta, el resultado se redondea, y de esta forma se concluye que el 75% de estudiantes tienen 15 faltas ortográficas o menos.

Cálculo de cuartil, decil y percentil para datos agrupados sin intervalos según su frecuencia

Para este caso, los cuartiles, deciles y percentiles se calculan tomando en cuenta la frecuencia acumulada inmediata mayor o igual a: $\frac{rn}{4}$, $\frac{rn}{10}$, $\frac{rn}{100}$ respectivamente.

200 universitarios fueron evaluados en manejo de ordenadores, los puntajes obtenidos se muestran en la tabla siguiente:

Tabla 5.8
Puntajes obtenidos en una evaluación sobre manejo de ordenadores por 200 universitarios.

X **(puntajes obtenidos)**	***n*** **(frecuencia de puntajes)**	***N*** **(frecuencia acumulada)**
18	1	1
19	6	7
20	12	19
21	18	37
22	22	59
23	37	96
24	45	141
25	32	173
26	21	194
27	4	198
28	2	200
Total	**200**	

Se desea calcular el cuartil tres (Q_3), el decil uno (D_1) y el percentil 48 (P_{48}).

Q_3: Se inicia calculando el valor de referencia $\frac{rn}{4}$:

$$\frac{3(200)}{4} = 150 \text{ ,}$$

luego de esto, se busca la frecuencia acumulada mayor o igual a 150 que es 173 y se corresponde a 25 de puntaje obtenido. Por lo tanto, $Q_3 = 25$, que se interpreta como: las tres primeras cuartas partes de universitarios obtuvieron un puntaje menor o igual a 25.

D_1: Se inicia calculando el valor de referencia $\frac{rn}{10}$:

$$\frac{1(200)}{10} = 20 \ ,$$

luego de esto, se busca la frecuencia acumulada mayor o igual a 20 que es 37 y se corresponde a 21 de puntaje obtenido. Por lo tanto, $D_1 = 21$, que se interpreta como: la primera décima parte de universitarios obtuvieron un puntaje menor o igual a 21.

P_{48}: Se inicia calculando el valor de referencia $\frac{rn}{100}$:

$$\frac{48(200)}{100} = 96 \ ,$$

luego de esto, se busca la frecuencia acumulada mayor o igual a 96 que es 96 exactamente y se corresponde a 23 de puntaje obtenido. Por lo tanto, $P_{48} = 23$, que puede interpretarse como: un universitario que tiene 23 de puntaje obtenido, supera al 48% de estudiantes, 52% de estudiantes tienen mejor nota.

Si bien existen distintos métodos para calcular los cuartiles, deciles y percentiles; los métodos mostrados en este libro fueron elegidos por considerarse los más intuitivos y prácticos.

Diagramas de caja y bigotes.

Estos diagramas sirven para dividir una población o muestra ordenada en cuatro partes de igual cantidad equivalente al 25 %.

Para su construcción, el 50% de los valores centrales estará contenido en una caja, indicándose dentro de ella la mediana. El intervalo de caja

se llama rango intercuartil (RIQ), pues sus valores están entre Q_1 y Q_3. Los valores externos a la caja se representan con segmentos que son los bigotes, el extremo de cada bigote estará separado máximo 1.5 veces la diferencia de $Q_3 - Q_1$; si un dato queda fuera se le conoce como atípico y se representa con un asterisco.

Ejemplo.

Representar con un diagrama de caja y bigotes la siguiente distribución.

Tabla 5.9
Distribución de frecuencias para el diagrama de caja o bigotes

X	n	N
0	13	13
1	12	25
2	26	51
3	34	85
4	10	95
5	4	99
6	1	100
Total	**100**	

Q_1: Se calcula el valor de referencia $\frac{rn}{4}$:

$$\frac{1(100)}{4} = 25\ ,$$

luego de esto, se busca la frecuencia acumulada mayor o igual a 25 que es exactamente 25, que se corresponde al valor 1 de la variable, por lo tanto $Q_1 = 1$.

Q_3: Se calcula el valor de referencia $\frac{rn}{4}$:

$$\frac{3(100)}{4} = 75\ ,$$

luego de esto, se busca la frecuencia acumulada mayor o igual a 75 que es 85, que se corresponde al valor 3 de la variable, por lo tanto $Q_3 = 3$.

$\boldsymbol{M_e}$**:** Como hay 100 valores, la mediana será la semisuma de los valores en las posiciones 50 y 51. Observando la tabla se aprecia que ambos son 2, en consecuencia $M_e = 2$.

$\boldsymbol{RIQ}$**:**

$$Q_3 - Q_1 = 3 - 1 = 2$$

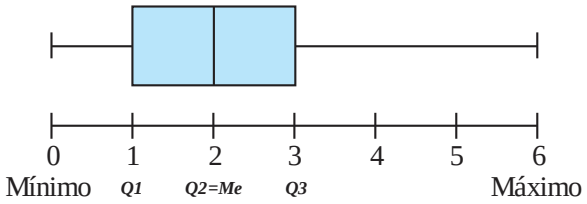
Los bigotes se alejaran a lo máximo 1.5 veces el rango intercuartil, para nuestro caso $1.5(2) = 3$.

Bigotes:

El bigote derecho llegará hasta máximo $3 + 3 = 6$, el bigote izquierdo llegará mínimo a $1 - 3 = -2$, como el mínimo es cero el bigote llega solamente hasta cero.

Figura 5.4

Digrama de cajas y bigotes elaborado con las frecuencias de los datos de la tabla 5.9



5.2.5 Medidas de dispersión de datos.

Las medidas de dispersión indican el grado de variabilidad de las mediciones, entre éstas están el rango, la desviación estándar y el coeficiente de variación, las dos últimas dependen del promedio

El rango.

El rango es la diferencia entre el máximo y el mínimo valor de las mediciones de la variable que se investiga. También se le conoce como amplitud total, campo de variación o recorrido, se usa mucho en control de calidad y en el análisis de muestras pequeñas de tamaño no superior a 12.

Ejemplo.

Se desea calcular el rango de la estatura de los alumnos de la escuela profesional de estadística sabiendo que la máxima es 180 cm. y la mínima 156 cm.

$$R = \text{Máx} - \text{Mín} = 180\ \text{cm} - 156\ \text{cm} = 24\ \text{cm}$$

La desviación estándar.

Indica el grado de dispersión de los datos respecto del promedio. Los datos serán más homogéneos si su valor es bajo, y más disperso si su valor es alto. Se calcula a partir de la varianza:

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

donde S es desviación estándar y S^2 es la varianza.

Si los **datos no están agrupados** la varianza se calcula por:

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

o también mediante la fórmula equivalente:

$$S^2 = \frac{n\sum X_i^2 - (\sum X_i)^2}{n(n-1)}$$

Ejemplo.

Supongamos las notas de exámenes de 5 alumnos: 12, 10, 11, 15, 08.

$$\bar{X} = \frac{\sum X_i}{n} = \frac{56}{5} = 11.2$$

$$S^2 = \frac{(12-11.2)^2+(10-11.2)^2+(11-11.2)^2+(15-11.2)^2+(8-11.2)^2}{5-1}$$

$$S^2 = \frac{26.8}{4} = 6.7$$

por lo tanto, la desviación estándar es:

$$S = \sqrt{6.7} = 2.59$$

Si los **datos están agrupados** se calcula con:

$$S^2 = \frac{\sum(Y_i - \bar{Y})^2 n_i}{n-1},$$

donde:

Y_i: marca de clase.

$\bar{Y}$: promedio de datos agrupados.

n_i: frecuencia de clase.

n: tamaño de muestra.

O por la fórmula equivalente:

$$S^2 = \frac{n\sum Y_i^2 ni - (\sum Y_i n_i)^2}{n(n-1)}$$

Ejemplo.

Se tiene la siguiente tabla de distribución de frecuencias:

Tabla 5.10
Distribución de notas de 40 alumnos

$(Y'_{j-1}$, $Y'_j]$	Y_i	n_i	$Y_i^2 n_i$	$Y_i n_i$
0 10	5.0	4	100	20
10 13	11.5	8	1058	92
13 16	14.5	16	3364	232
16 18	17.0	8	2312	136
18 20	19.0	4	1444	76
T o t a l		**40**	**8278**	**556**

$$S^2 = \frac{40(8278) - (556)^2}{40(40-1)} = 14.90$$

$$S = \sqrt{14.09} = 3.75$$

Ejemplo.

Se tiene la siguiente distribución:

Tabla 5.11
Ejemplo de distribución de frecuencias para calcular la desviación estándar

Y_i	n_i	$Y_i^2 n_i$	$Y_i n_i$
0	3	0	0
1	2	2	2
2	1	4	2
Total	**6**	**6**	**4**

Se puede observar que los **intervalos de clase tienen amplitud cero**, en este caso los Y_i son los puntos de recorrido de la variable, calculando la varianza y desviación estándar de la siguiente manera:

$$S^2 = \frac{6(6)-4^2}{6(6-1)} = \frac{36-16}{30} = 0.66$$

$$S = \sqrt{0.66} = 0.81$$

Propiedades de la desviación estándar:

1. Es siempre un valor positivo.

2. Es la medida menos influenciada por cambios de muestra de una sola población.

3. Es influenciada por los valores extremos de la variable.

4. Se puede estimar conociendo solamente el rango de la variable usando la fórmula: $S \approx \frac{R}{6}$. Siempre que la distribución concentre la mayoría de sus valores cerca el promedio y sus valores extremos no estén alejados de éste.

5. Para muestras de poblaciones normales o aproximadamente normales se tiene:

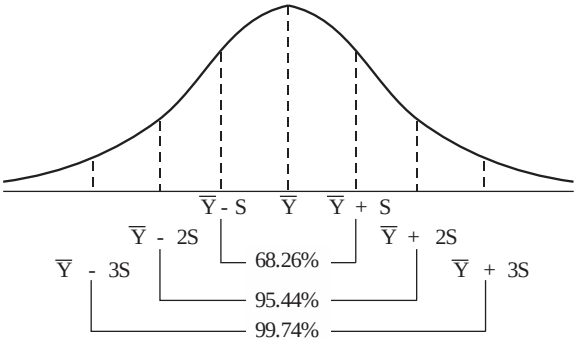
a) El 68.26% de las observaciones se hallan comprendidas entre: $\bar{Y} - S$ y $\bar{Y} + S$.

b) El 95.45% entre $\bar{Y} - 2S$ y $\bar{Y} + 2S$.

c) El 99.74% entre $\bar{Y} - 3S$ y $\bar{Y} + 3S$.

Figura 5.5

Concentración de datos en una muestra de una población normal o aproximadamente normal.



Coeficiente de variación C.V.

Es una medida de variabilidad relativa, a diferencia de la desviación estándar que es absoluta. Se calcula dividiendo la desviación estándar entre el promedio:

$$C.V. = \frac{S}{\bar{Y}}.$$

También puede expresarse como porcentaje:

$$C.V. = \frac{S}{\bar{Y}} 100\% .$$

Ejemplo.

Para el caso de las notas de los 40 alumnos (Tabla 5.10).

$$C.V. = \frac{3.75}{13.9} 100\% = 26.98\%$$

dado que

$$\bar{Y} = \frac{\sum y_i n_i}{n} = \frac{556}{40} = 13.9 .$$

El coeficiente de variación permite la comparación adecuada entre grupos de datos, pues analizas su variabilidad independientemente de: la población a la que pertenecen, sus escalas de medición o las unidades que utilizan. Cabe resaltar que, si la media está próxima a cero, el coeficiente de variación merece poca confianza ya que su valor tendería al infinito.

5.2.6 Normalidad de una distribución.

Una distribución de frecuencias de una variable continua es llamada normal si está definida por la siguiente fórmula:

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

donde:

X: variable en estudio.

$f(X)$: frecuencia absoluta de los datos con infinitos intervalos.

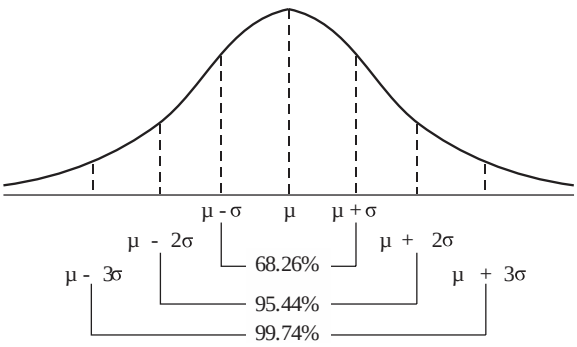
μ: media de la población de datos.

σ: desviación estándar de la población de datos.

$\pi \approx 3.1416$.

$e \approx 2.7183$.

Figura 5.6
Concentración de datos en una población normal con respecto al promedio y desviación estándar.



Esta distribución muestra un comportamiento centralizado alrededor del promedio; su media, mediana y moda coinciden. Si $\mu = 0$ y $\sigma = 1$, la distribución se conoce como normal estandarizada.

Las distribuciones de variables continuas se aproximan a la distribución normal, cuanto más elevado sea el número de datos recogidos y su gráfica tome forma de campana. Por esto, es conveniente analizar si una distribución puede considerarse "normal", dichas pruebas son la asimetría y curtosis.

Asimetría.

Una distribución es simétrica sí coinciden su promedio aritmético, mediana y moda. Como esto no ocurre frecuentemente en la realidad, la mayoría de distribuciones son asimétricas, pudiendo ser asimétricas positivas, si se cumple que $Mo < Me < \bar{Y}$, o asimétricas negativas, si $\bar{Y} < Me < Mo$ (Figura 5.10).

-Coeficientes de asimetría de Pearson

Existen 2 coeficientes de asimetría de Pearson, uno que depende de la moda y el otro que depende de la mediana.

$$AS_1 = \frac{\bar{Y} - Mo}{S},$$

$$AS_2 = \frac{3(\bar{Y} - Me)}{S}.$$

AS_2 proviene de AS_1 debido a que en distribuciones moderadamente asimétricas se cumple aproximadamente la relación:

$$3(\bar{Y} - Me) \approx (\bar{Y} - Mo).$$

El valor de los coeficientes de asimetría dado por Pearson indica el tipo de distribución.

Simétrica: $AS = 0$.

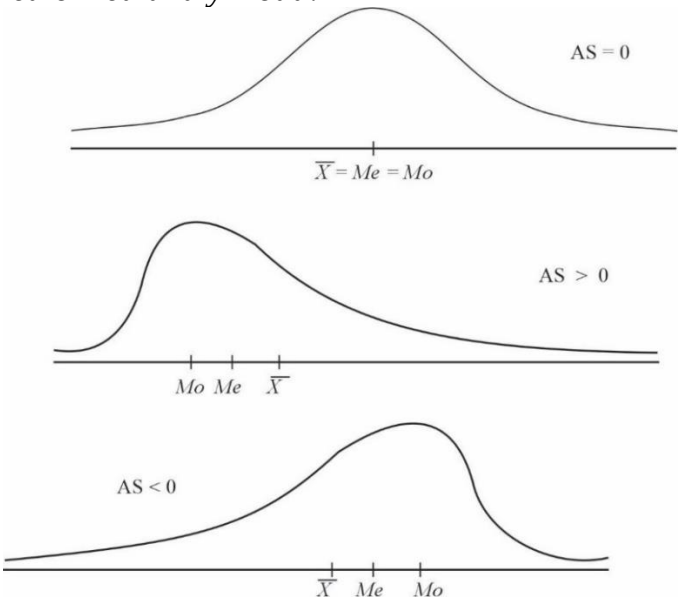
Asimétrica positiva: $AS > 0$.

Asimétrica negativa: $AS < 0$.

Si el valor está entre -1 y 1, la distribución puede considerarse cercana a la normal.

Figura 5.7

Representación de la simetría de las distribuciones de frecuencias según promedio mediana y moda.



En la distribución de notas de 40 alumnos (Tabla 5.10) la asimetría es:

$$AS_2 = \frac{3(13.9-14.5)}{3.75} = -\frac{1.8}{3.75} = -0.48 \ .$$

Por lo tanto, la distribución tiene asimetría negativa, pero puede considerarse cercana a la normal.

Para el caso del número de faltas ortográficas (Tabla 5.7), la distribución será simétrica:

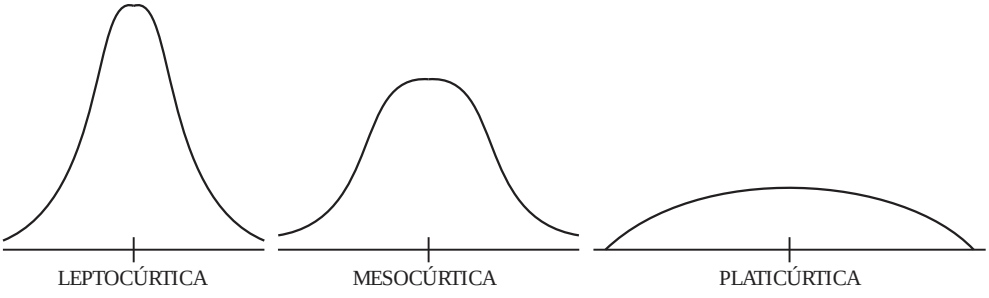
$$AS_1 = \frac{12-12}{4.52} = 0 \ .$$

Curtosis.

Es una medida de forma de una distribución de datos que está referida a su altura. Según esto las distribuciones pueden ser: leptocúrtica, mesocúrtica o platicúrtica (Fig. 5.8).

Figura 5.8

Formas de curvas que generan la distribución de frecuencias.



Existen diferentes formas para el cálculo de la curtosis, aquella basada en los percentiles y cuartiles es la siguiente:

$$Cu = \frac{q_3 - q_1}{2(P_{90} - P_{10})}.$$

De acuerdo a este valor una distribución puede ser:

Leptocúrtica: $Cu > 0.263$.

Mesocúrtica o normal: $Cu = 0.263$.

Platicúrtica: $Cu < 0.263$.

Una distribución con valor de curtosis cercano a 0.263, puede considerarse que tenga un comportamiento similar a la normal.

Se desea determinar la curtosis de la siguiente distribución:

Tabla 5.12
Distribución de horas de taller en tres días de 40 alumnos de ingeniería de una universidad

$(Y'_{i-1}\ ,\ Y'_i]$	Y_i	n_i	N_i
0-4	2	2	2
4-8	6	6	8 (N_{i-1})
8-12	10	8	16 (N_i)
12-16	14	16	32
16-20	18	8	40
T o t a l		40	

Los cuartiles son:

$$q_3 = 12 + 4\ \frac{\frac{3(40)}{4} - 16}{32 - 16} = 15.5$$

$$q_1 = 8 + 4\ \frac{\frac{1(40)}{4} - 8}{16 - 8} = 9.0$$

Los percentiles son:

$$P_{90} = 16 + 4\,\frac{\frac{90(40)}{100} - 32}{40 - 32} = 18$$

$$P_{10} = 4 + 4\,\frac{\frac{10(40)}{100} - 2}{8 - 2} = 5.3$$

Sustituyendo los valores en la fórmula:

$$Cu = \frac{15.5-9}{2(18-5.3)} = 0.255\ .$$

Como la curtosis es 0.255 se trata de una distribución **platicúrtica**, sin embargo, al estar muy cerca a 0.263, podemos asumir que tiene un comportamiento similar a la normal.

5.3 ANÁLISIS ESTADÍSTICO PARA DOS VARIABLES

5.3.1 Diseño de tablas y gráficos para distribuciones bidimensionales.

En la investigación científica se busca determinar la relación entre variables, a partir de los datos que provienen de cada unidad de análisis. Estas relaciones se muestran en tablas y figuras que se diseñan de

acuerdo a la definición conceptual y operacional de las variables en estudio. Las tablas de doble entrada pueden formarse entre variables de cualquier tipo.

Ejemplo.

Para estudiar la edad de la madre y el número de hijos en edad escolar se toma una muestra de 30 madres.

Tabla 5.13
Datos de la edad y el número de hijos de 30 madres de una institución educativa

Número de madre	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Edad de la madre	25	22	24	30	31	35	33	30	32	34	40	48	50	45	35
n° de hijos	1	1	1	2	2	2	2	3	1	2	2	3	3	2	2
Número de madre	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Edad de la madre	25	28	23	22	20	24	27	26	28	30	32	32	40	42	40
n° de hijos	1	1	2	1	1	1	2	2	3	2	2	1	3	2	2

La variable edad la consideramos continua pues, a pesar de que la respuesta esté en años cumplidos, su naturaleza está en los números reales, por lo que se agrupará en intervalos. Por otro lado, el número de hijos es una variable discreta ya que se define con números naturales.

Tabla 5.14
Edad y número de hijos de 30 madres de una institución educativa

Edad de la madre $(Y'_{i-1}, Y'_i]$	Número de hijos						Total	
	1		**2**		**3**			
	n	%	n	%	n	%	n	%
20 25	7	23	1	3	0	0	8	26
25 35	2	7	11	37	2	7	15	50
35 45	0	0	4	13	1	3	5	17
45 50	0	0	0	0	2	7	2	7
Total	**9**	**30**	**16**	**53**	**5**	**17**	**30**	**100**

En las dos últimas columnas de la tabla 5.14 se tiene las frecuencias marginales, absolutas y porcentuales, de la edad. De modo similar, en la última fila se tiene las frecuencias marginales absolutas y porcentuales, del número de hijos.

Las frecuencias conjuntas se obtienen relacionando los valores de ambas variables y sus porcentajes se obtienen con respecto al tamaño de muestra total.

La distribución de las frecuencias muestra que:

-El 50% de las madres tienen una edad comprendida entre 25 a 35 años.

-El 53% de las madres tienen 2 hijos.

-El 37% de las madres tienen 2 hijos y su edad está entre 25 a 35 años.

Las distribuciones bidimensionales se representan gráficamente mediante un **estereograma** que se construye con los valores de las variables y las frecuencias conjuntas.

Ejemplo de distribución bidimensional con una variable cuantitativa y una cualitativa.

Con los datos del rendimiento académico y el tipo de colegio en la tabla 5.15, se construye el estereograma de la figura 5.9.

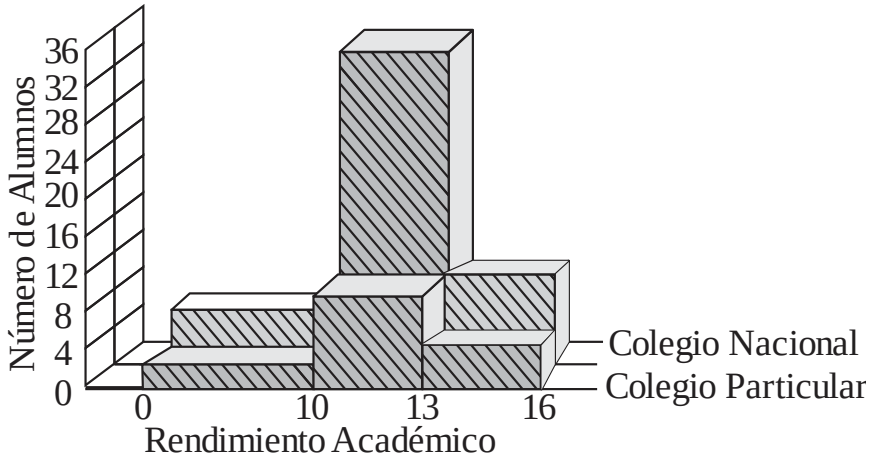
Tabla 5.15

Rendimiento académico según el tipo de colegio de 72 alumnos.

Rendimiento Académico	Tipo de colegio				Total	
	Nacional		**Particular**			
	n	%	n	%	n	%
0 10	8	14	2	13	10	34
10 13	36	64	10	62	46	62
13 16	12	22	4	25	16	4
T o t a l	**56**	**100**	**16**	**100**	**108**	**100**

Figura 5.9

Estereograma de frecuencias, rendimiento académico según el tipo de colegio de 72 alumnos.



Ejemplo de distribución bidimensional con dos variables cuantitativas.

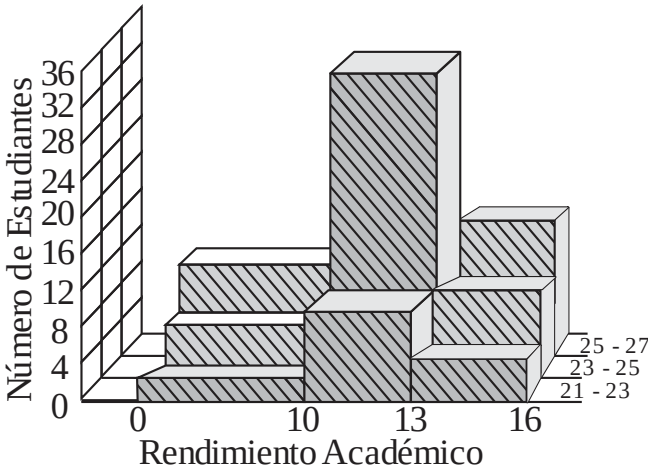
Con los datos del rendimiento académico y la edad de los estudiantes en la tabla 5.16, se construye el estereograma de la figura 5.10.

Tabla 5.16

Rendimiento académico y edad en los estudiantes de una universidad.

Rendimiento académico.	Edad						total	
	21 - 23		23 - 25		25 - 27			
	n	%	n	%	n	%	n	%
0 - 10	2	12	8	14	14	36	**24**	**22**
10 - 13	10	63	36	64	4	11	**50**	**45**
13 – 16	4	25	12	22	20	53	**36**	**33**
Total	**16**	**100**	**56**	**100**	**38**	**100**	**110**	**100**

Figura. 5.10
Estereograma de frecuencias del rendimiento académico y edad en los estudiantes de una universidad.



5.3.2 Correlación y regresión.

En una investigación científica, en la que se trabaja con más de una variable, se busca determinar el grado y tipo de relación estadística, conocida como **correlación**, que existe entre ellas.

Para conocer el grado de correlación, se inicia por hacer un análisis de la dispersión de los datos y así develar empíricamente el modelo al que se ajustarían los datos de las variables. El proceso de ajustar la dispersión de los datos a un modelo que permita la descripción, la explicación y la predicción del comportamiento con nuevos datos de las variables en estudio se denomina **regresión**.

El análisis e interpretación adecuada de la correlación y regresión necesita de las medidas de centralización, dispersión, posición, asimetría y curtosis que explican el comportamiento de las variables de manera independiente.

El análisis de regresión y correlación se explicará con más detalle en el capítulo siguiente.

AUTOEVALUACIÓN I

1. Elija al azar dos grupos de alumnos, uno de 30 y otro de 32 y analice las siguientes características:

Grupo A	**Grupo B**
Edad	Edad
Peso	Peso
Sexo	Sexo
N° de hermanos	N° de hermanos
Estatura	Estatura

a. Define las variables.

b. Precise las escalas con que se miden las variables.

c. Elabore las tablas de distribución de las frecuencias.

d. Grafique el histograma de frecuencias absolutas, acumuladas, así como sus respectivos polígonos de frecuencia.

2. Clasifique las variables en cualitativas, discretas y continuas.

- Edad de los estudiantes de un instituto.
- Procedencia de los trabajadores de una empresa.
- Grupo sanguíneo de donantes de órganos.
- Temperatura ambiental de la zona costera.
- Cociente intelectual de los alumnos de una universidad.

3. Grafique la distribución de los datos que se presentan y calcule: Las frecuencias relativas y las frecuencias acumuladas.

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

4. Calcule la asimetría y la curtosis de la pregunta 3.

5. Para los datos que siguen: 0, 0, 4, 4, 2, 3, 8, 2, 2, 3, 3, 3, 1, 1, 1, 1. Calcule su promedio, mediana, moda, la varianza, desviación estándar, coeficiente de variación y la asimetría de la distribución.

6. Grafique la siguiente distribución y Determine su promedio aritmético, su mediana, su moda, desviación estándar, su coeficiente de variación y su asimetría.

X	0	1	3	5	7
n	4	3	2	1	0

7. Las notas de 40 niños en una prueba de lenguaje se distribuyen como sigue:

$(Y'_{j-1}$, $Y'_j]$	n_j
0 10	5
10 13	8
13 16	14
16 18	8
18 20	5

a. Grafique el histograma y los polígonos de frecuencias absolutas y acumuladas.

b. Calcule las medidas de centralización y dispersión.

c. ¿Cuál es la asimetría que presenta la distribución?

d. ¿Cuántos niños tienen nota entre $\bar{Y} \pm 2S$?

e. ¿Cuántos niños tienen nota mayor o igual que $\bar{Y} + S$?

f. ¿Cuál es la probabilidad de encontrar niños con nota mayor de 16?

g. Calcule e interprete Q_3, D_6, P_{90}.

h. Pruebe con la curtosis si la distribución se aproxima a una normal.

CAPÍTULO VI

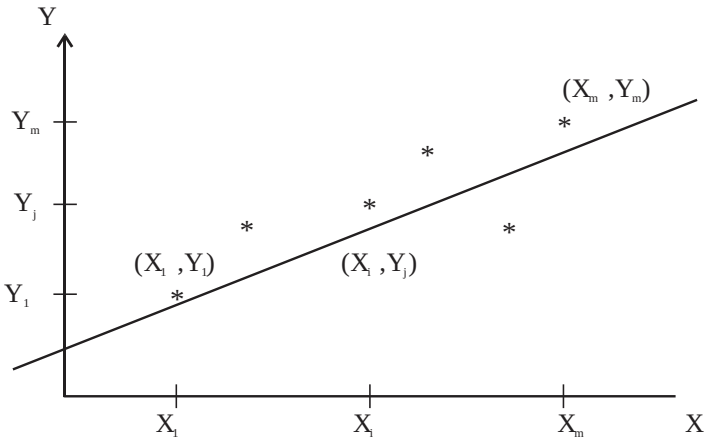
ANÁLISIS DE REGRESIÓN Y CORRELACIÓN

6.1 REGRESIÓN

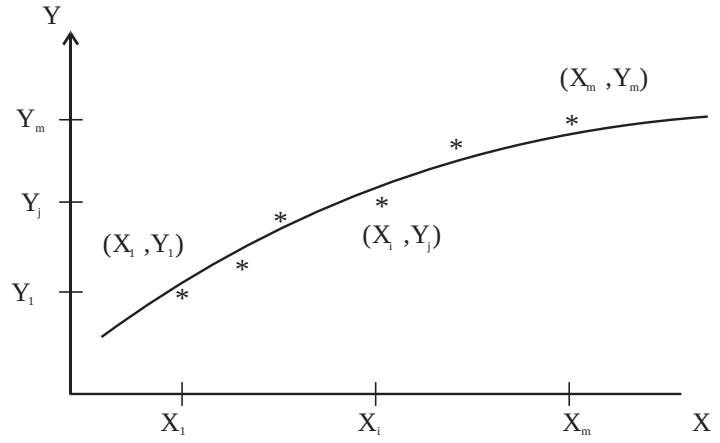
Para determinar la dependencia de una variable respecto a otra es necesario conocer la ecuación a la que mejor se ajustan los datos. Dadas las variables X y Y, un diagrama de dispersión muestra la localización de los puntos (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n), en un sistema de coordenadas rectangular; con este diagrama es posible representar el lugar geométrico de la distribución de los datos. Si el diagrama se aproxima a una recta, entonces existe una relación lineal, si mantienen

relación y no siguen una línea recta entonces existe una relación no lineal. Las ecuaciones de aproximación que mejor se ajusten a la distribución de datos se llaman **curvas de ajuste**.

Figura 6.1
Diagramas de dispersión de datos de dos pares de variables con sus respectivas curvas de ajuste.



Relación Lineal entre X y Y



Relación No Lineal entre X y Y

6.1.1 Regresión lineal

Si la forma del diagrama de dispersión de puntos se aproxima a una recta ($Y = a + bX$), ésta se escogerá de modo que sea la que mejor se ajuste a los datos. Sabiendo que:

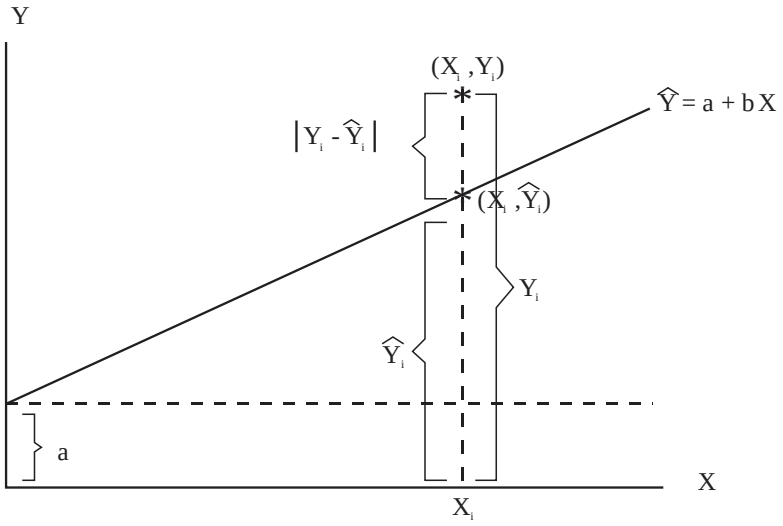
X: variable independiente, estímulo o causa.
Y: variable dependiente, respuesta, o efecto.
b: coeficiente de regresión, pendiente o proporción de cambio.
a: coeficiente autónomo.

Entonces $\hat{Y} = a + bX$ (a y b conocidos), rendirá un valor de $\hat{Y}$ para cada valor de *X*, que será un estimado de *Y*.

a es el punto donde la recta $\hat{Y}$ corta al eje de la *Y* (intercepto), si ésta pasa por el origen entonces $a = 0$.

b es la pendiente de la recta $\hat{Y}$ y establece su velocidad de cambio con respecto a *X*. Cuando b es positiva ambas variables aumentan o disminuyen juntas; cuando b es negativa, al aumentar una variable disminuye la otra.

Figura 6.2
Distancia de un valor de Y a la recta de ajuste.



Como se puede observar en la Figura 6.2, para un valor de X, $\hat{Y}$ puede estar distante $d = |Y - \hat{Y}|$ del valor observado Y. Se usa valor absoluto porque pueden existir valores observados ubicados debajo de la recta.
Una forma de lograr la línea de mejor ajuste es hallar aquella para la cual la suma de las distancias $|Y - \hat{Y}|$ (errores), para todos los valores dados de X sea lo más pequeño posible, es decir que $\sum |Y - \hat{Y}|$ sea mínimo.

El problema del valor absoluto es que no posee derivada continua, éste se elimina al usar el método de los mínimos cuadrados que está dado por:

$$\min_{\hat{Y}} \sum (Y - \hat{Y})^2$$

Para el cálculo de a y b, en la expresión $\Sigma(Y-\hat{Y})^2$ sustituimos $\hat{Y}$ por su valor; obteniendo:

$$\sum(Y-a-bX)^2$$

Diferenciando esta expresión respecto a los coeficientes a y b e igualando a cero, se obtienen las ecuaciones normales:

$$\sum Y = na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

Por lo tanto:

$$a = \frac{\sum Y\sum X^2 - \sum X\sum XY}{n\sum X^2 - (\sum X)^2}$$

$$b = \frac{n\sum XY - \sum X\sum Y}{n\sum X^2 - (\sum X)^2}$$

En función de los promedios de las variables X, Y y el coeficiente b; a está dado por:

$$a = \bar{Y} - b\bar{X}$$

Así, conocidos a y b arribamos al modelo $\boldsymbol{\hat{Y} = a + bX}$, llamado de estimación, predicción o regresión de Y/X. Como una comprobación puede verificarse que $\sum Y = \sum \hat{Y}$.

Para graficar la recta de ajuste $\hat{Y}$ se unen los puntos correspondientes a los valores extremos de X con sus $\hat{Y}$ respectivos.

La ecuación $\hat{Y} = a + bX$ puede utilizarse tanto para interpolar valores de $\hat{Y}$ (dentro del recorrido de X) como para extrapolar valores de $\hat{Y}$ (fuera del recorrido de X) en este caso se deben utilizar valores próximos a los extremos de X, para brindar mayor confiabilidad en la estimación.

Ejemplo didáctico.

Supóngase que tenemos una muestra aleatoria de 5 estudiantes con sus respectivas horas de estudio y nota obtenida en determinada asignatura. Se desea hallar el modelo de regresión de los puntajes sobre las horas dedicadas al estudio.

Alumnos	Horas de estudio (X)	Nota (Y)
A	2	6
B	3	8
C	4	10
D	6	12
E	7	15

Solución.

A partir de los datos se procede a construir la tabla 6.1

Tabla 6.1
Cálculos para el modelo de regresión

Alumno	X	Y	XY	X^2	$\hat{Y}$	$(Y - \hat{Y})^2$
A	2	6	12	4	6.216	0.046656
B	3	8	24	9	7.876	0.015376
C	4	10	40	16	9.536	0.215296
D	6	12	72	36	12.856	0.732736
E	7	15	105	49	14.516	0.234256
Total	**22**	**51**	**253**	**114**	**51.000**	**1.244320**

Reemplazando lo obtenido en la ecuación normal de b:

$$b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} = \frac{5(253) - 22(51)}{5(114) - 22^2} = 1.66$$

$b = 1.66$ puntos por hora de estudio.

Con la fórmula de promedios se calcula a:

$$a = \bar{Y} - b\bar{X}$$

$$a = \frac{51}{5} - 1.66\frac{22}{5}$$

$a = 2.896$ puntos.

Luego el modelo de regresión sería:

$$\hat{Y} = 2.896 + 1.66X$$

A partir de esta fórmula se puede calcular las estimaciones de Y

$$\hat{Y}_1 = 2.896 + 1.66\ (2) = 6.216$$

$$\hat{Y}_2 = 2.896 + 1.66\ (3) = 7.876$$

$$\hat{Y}_3 = 2.896 + 1.66\ (4) = 9.536$$

$$\hat{Y}_4 = 2.896 + 1.66\ (6) = 12.856$$

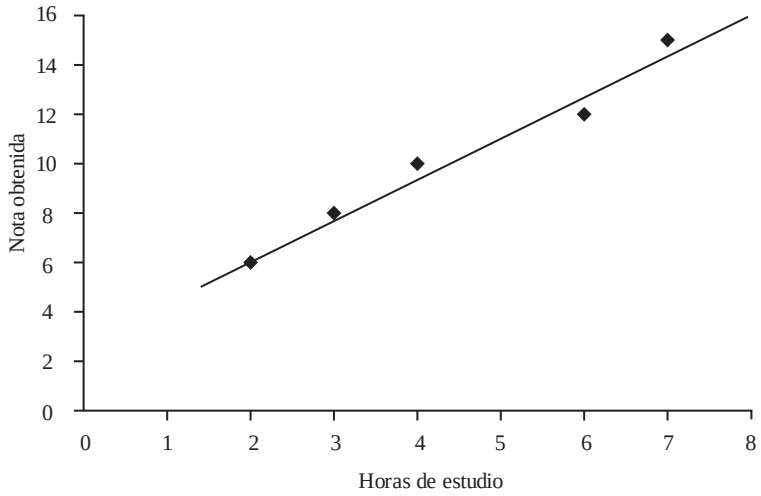
$$\hat{Y}_5 = 2.896 + 1.66\ (7) = 14.516$$

$$\sum \hat{Y} = 51\ .$$

Como se puede observar $\sum Y = \sum \hat{Y}$.

Figura 6.3

Diagrama de dispersión y recta de ajuste con los datos de la tabla 6.1.



A partir del modelo de regresión se puede estimar la nota que se obtendría con 5 horas de estudio (interpolación):

$\hat{Y} = 2.896 + 1.66(5) = 11.196$

Igualmente se puede estimar la nota con 8 horas de estudio (extrapolación o predicción):

$\hat{Y} = 2.896 + 1.66(8) = 16.176.$

Error típico de estimación.

Es la dispersión de puntos alrededor de la recta de estimación, se calcula por:

$$S_{Y/X} = \sqrt{\frac{\sum(Y - \hat{Y})^2}{n - 2}}$$

donde n es el número de pares de datos (muestra), reemplazando valores el error de estimación será:

$$S_{Y/X} = \sqrt{\frac{1.24432}{5 - 2}}$$

$$S_{Y/X} = 0.644 \ .$$

Las propiedades de $S_{Y/X}$ son análogas a la desviación estándar de una sola variable, en donde las desviaciones de las observaciones se miden respecto a su promedio, mientras que en la regresión las desviaciones de las observaciones Y_i son respecto a $\hat{Y}$.

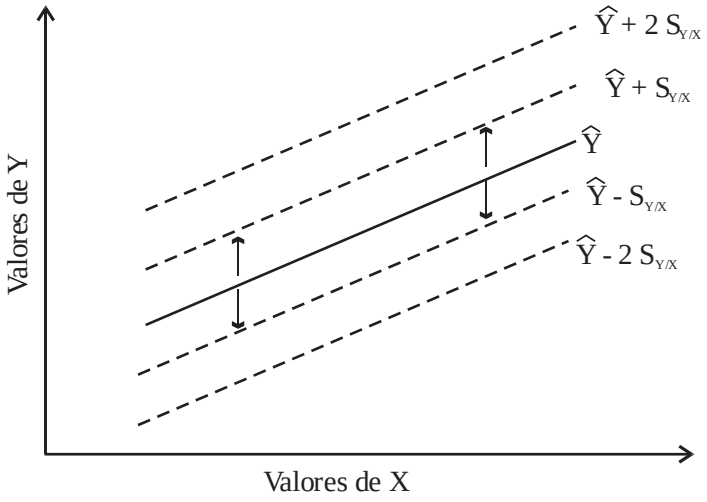
Si construimos las siguientes paralelas a la recta de regresión:

$$\hat{Y} \pm S_{Y/X} \ ; \ \hat{Y} \pm 2S_{Y/X} \ ; \ \hat{Y} \pm 3S_{Y/X} \ ,$$

entre estas líneas quedarían incluidos aproximadamente el 68.27%, 95.45% y 99.73% de los datos de la muestra, determinándose así franjas o cinturones de confiabilidad. Entonces al 68.27% de confianza, para 8 horas de estudio la nota que se espera estará entre $\hat{Y} - S_{Y/X}$ y $\hat{Y} + S_{Y/X}$, es decir entre $16.176 - 0.644 = 15.532$ y $16.176 + 0.644 = 16.820$.

Figura 6.4

Franjas de confiabilidad para un modelo de regresión lineal.



Análisis de regresión en series de tiempo.

Si una de las variables es el tiempo y la otra evoluciona o experimenta cambios a través de ésta, entonces la recta o curva de regresión es exclusivamente de *Y* en *X* y toma el nombre de recta o curva de tendencia. En este caso, la relación entre las variables no es estrictamente de tipo causal y se utiliza frecuentemente para fines de estimación, predicción o pronóstico.

Con el objeto de abreviar el cálculo de los coeficientes, a la variable tiempo se le asigna nuevos valores tratando de que su sumatoria sea igual a cero.

Ejemplo cuando la muestra es impar.

En la tabla 6.2 se indica la compra de seguros de vida (en millones de soles) en una ciudad entre 2015 y 2019. Se puede observar que al valor central de la variable año se le asignó 0, a los valores que le preceden -1 y -2 y a los que le siguen 1 y 2.

Tabla 6.2

Compra de seguros de vida (en millones de soles) en una ciudad entre 2015 y 2019.

Año	X	Y	XY	X^2	Y^2	$\hat{Y}$
2015	-2	51.7	-103.4	4	2672.89	50.32
2016	-1	52.9	- 52.9	1	2798.41	53.25
2017	0	55.0	0.0	0	3025.00	56.18
2018	1	57.0	57.0	1	3249.00	59.11
2019	2	64.3	128.6	4	4134.49	62.04

Debido a que $\sum X = 0$ y consecuentemente $\bar{X} = 0$, los coeficientes de la regresión lineal se caculan de la siguiente forma:

$$b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum \mathrm{X})^2} = \frac{\sum XY}{\sum X^2}$$

$$b = \frac{29.3}{10} = 2.93$$

$$a = \bar{Y} - b\bar{X} = \bar{Y} = 56.18 \ .$$

El modelo de estimación queda como:

$$\hat{Y} = 56.18 + 2.93X \ .$$

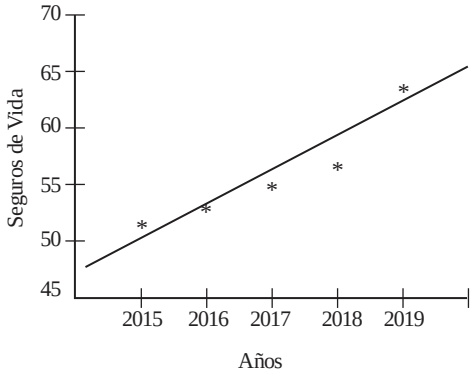
Para predecir la compra de seguros de vida para el año 2020, X tomaría el valor de 3:

$$\hat{Y}_{2020} = 56.18 + 2.93(3) = 64.97$$

entonces la compra de seguros de vida en el año 2020 sería de 64.97 millones de soles.

Figura 6.5

Diagrama de dispersión y recta de ajuste con los datos de la tabla 6.2.



Ejemplo cuando la muestra es par.

Tabla 6.3

Postulantes a la Universidad nacional Pedro Ruiz Gallo 2014-2019.

Año	X	Y	XY	X^2	$\hat{Y}$
2014	-5	4000	-20000	25	3990.48
2015	-3	4400	-13200	9	4407.62
2016	-1	4500	4500	1	4824.76
2017	1	5800	5800	1	5241.90
2018	3	5500	16500	9	5659.04
2019	5	6000	30000	25	6076.18
Total	**0**	**30200**	**14600**	**70**	**30199.98**

Como en este caso hay dos valores centrales, a éstos se les asigna -1 y 1 respectivamente. Debido a que la diferencia entre -1 y 1 es 2, los

valores que están por debajo de -1, serán -3 y -5, y los que están por encima 3 y 5. Calculándose el modelo de estimación así:

$$b = \frac{\sum XY}{\sum X^2} = \frac{14600}{70} = 208.57$$

$$a = \bar{Y} = 5033.33$$

$$\hat{Y} = 5033.33 + 208.57X \ .$$

La estimación de postulantes para el 2020 se calcula con $X = 7$:

$$\hat{Y}_{2020} = 5033.33 + 208.57(7)$$

$$\hat{Y} = 6493.32$$

como el número de postulantes debe ser un valor entero, se redondea a 6493.

6.1.2 Regresión no lineal.

En este tipo de regresión la forma del diagrama de dispersión de puntos sigue una función que no es lineal, pudiendo ser cuadrática, cúbica, exponencial, entre otras. El método más común para determinar los coeficientes de la curva a la que más se ajusten los datos es el de mínimos cuadrados.

Mínimos cuadrados para regresión no lineal cuadrática.

El modelo de regresión cuadrática se expresa de la siguiente manera:

$$\hat{Y} = a + bX + cX^2 \ .$$

Aplicando mínimos cuadrados:

$$\min_{\hat{Y}} \sum (Y - \hat{Y})^2$$

$$\min_{a,b,c} \sum (Y - (a + bX + cX^2))^2 \ ,$$

donde a, b y c se determinan resolviendo el sistema de ecuaciones normales que se obtienen al derivar la función anterior respecto a los coeficientes:

$$\sum Y = na + b\sum X + c\sum X^2$$

$$\sum XY = a\sum X + b\sum X^2 + c\sum X^3$$

$$\sum X^2 Y = a\sum X^2 + b\sum X^3 + c\sum X^4 \ .$$

En la práctica, este sistema de ecuaciones se obtiene multiplicando por 1, X y X^2 respectivamente a la ecuación original ($Y = a + bX + cX^2$) y aplicando sumatoria a ambos miembros de las ecuaciones. Por este método se pueden obtener las ecuaciones normales de mínimos cuadrados para curvas cuadráticas, cúbicas, etc.

Ejemplo didáctico.

Ilustremos la regresión no lineal cuadrática con los datos de horas de estudio y notas de 5 alumnos de la tabla 6.1 y a partir de estos se procede a construir la tabla 6.4.

Tabla 6.4
Cálculos para el modelo de regresión cuadrática

X	Y	XY	X^2	X^2Y	X^3	X^4	$\hat{Y}$
2	6	12	4	24	8	16	6.22398
3	8	24	9	72	27	81	7.86683
4	10	40	16	160	64	256	9.51942
6	12	72	36	432	216	1296	12.85382
7	15	105	49	735	343	2401	14.53563
22	**51**	**253**	**114**	**1428**	**658**	**4050**	**50.99968**

Reemplazando en las ecuaciones normales se tiene:

$$51 = 5a + 22b + 114c$$

$$253 = 22a + 114b + 658c$$

$$1423 = 114a + 658b + 4050c\ .$$

Resolviendo el sistema por el método de Cramer:

$$\Delta = \begin{vmatrix} 5 & 22 & 114 \\ 22 & 114 & 658 \\ 114 & 658 & 4050 \end{vmatrix} = 2464$$

$$a = \frac{\begin{vmatrix} 51 & 22 & 114 \\ 253 & 114 & 658 \\ 1423 & 658 & 4050 \end{vmatrix}}{\Delta} = \frac{7312}{2464} = 2.9675$$

$$b = \frac{\begin{vmatrix} 5 & 51 & 114 \\ 22 & 253 & 658 \\ 114 & 1423 & 4050 \end{vmatrix}}{\Delta} = \frac{3988}{2464} = 1.6185$$

$$c = \frac{\begin{vmatrix} 5 & 22 & 51 \\ 22 & 114 & 253 \\ 114 & 658 & 1423 \end{vmatrix}}{\Delta} = \frac{12}{2464} = 0.00487.$$

Luego la ecuación del modelo de regresión cuadrática es:

$$\hat{Y} = 2.9675 + 1.6185X + 0.00487X^2 \ .$$

Como se puede observar la parte lineal: $2.9675 + 1.6185X$, está muy cerca de la recta calculada por medio de regresión lineal: $2.896 + 1.66X$. A partir del modelo se obtienen las siguientes estimaciones:

$$\hat{Y}_1 = 2.9675 + 1.6185(2) + 0.00487(2)^2 = 6.22398$$
$$\hat{Y}_2 = 2.9675 + 1.6185(3) + 0.00487(3)^2 = 7.86683$$
$$\hat{Y}_3 = 2.9675 + 1.6185(4) + 0.00487(4)^2 = 9.51942$$
$$\hat{Y}_4 = 2.9675 + 1.6185(6) + 0.00487(6)^2 = 12.85382$$
$$\hat{Y}_5 = 2.9675 + 1.6185(7) + 0.00487(7)^2 = 14.53563$$
$$\sum \hat{Y} = 50.99968 \ .$$

Extrapolando para 8 horas de estudio, la nota esperada sería:

$$\hat{Y}_{8h} = 2.9675 + 1.6185(8) + 0.00487(8)^2 = 16.22718 \ ,$$

estimación que muy semejante a la que se consigue con el modelo lineal, esto es debido a que en el coeficiente de X^2 es muy pequeño.
Para cuantificar la dispersión alrededor de la parábola de ajuste calculamos el error típico de estimación:

$$S_{Y/X} = \sqrt{\frac{\sum(Y - \hat{Y})^2}{n - 3}}$$

donde $\hat{Y} = a + bX + cX^2$, en este caso se le resta 3 al tamaño de muestra pues se tiene 3 coeficientes: a, b y c.

Tabla 6.5
Cálculos para el error típico de la regresión cuadrática.

X	Y	$\hat{Y}$	$(Y - \hat{Y})^2$
2	6	6.22398	0.05016704
3	8	7.86683	0.01773424
4	10	9.51942	0.23095713
6	12	12.85382	0.72900859
7	15	14.53563	0.21563949
			1.24350649

$$S_{Y/X} = \sqrt{\frac{1.24350649}{5 - 3}} = 0.7885$$

Con este resultado, podemos concluir que este modelo tiene más error que el lineal que es de 0.644.

Análisis de regresión en series de tiempo.

Es similar al caso lineal y de igual manera, a la variable tiempo se le asigna nuevos valores para que su sumatoria sea cero. Acá también se debe tomar en cuenta que $\sum X^3 = 0$, $\sum X^5 = 0$, $\sum X^7 = 0$ y así sucesivamente. A continuación, trataremos el análisis de regresión

en series de tiempo de curvas con tendencia parabólica, hiperbólica y exponencial.

A) Tendencia parabólica.

Las ecuaciones normales debido al cambio que se hace a la variable tiempo se reducen a:

$$\sum Y = na + c\sum X^2$$

$$\sum XY = b\sum X^2$$

$$\sum X^2 Y = a\sum X^2 + c\sum X^4$$

despejando a, b y c se tiene:

$$a = \frac{\sum Y \sum X^4 - \sum X^2 Y \sum X^2}{n\sum X^4 - (\sum X^2)^2}$$

$$b = \frac{\sum XY}{n\sum X^2}$$

$$c = \frac{n\sum X^2 Y - \sum Y^2 \sum X^2}{n\sum X^4 - (\sum X^2)^2}$$

Ejemplo:

Tabla 6.6

Líneas telefónicas instaladas en el centro poblado Mimave:2009-2019

Año	X	Y	XY	X^2	X^2Y	X^4	$\hat{Y}$
2009	- 5	5	- 25	25	125	625	4.5430
2010	- 4	15	- 60	16	240	256	13.9075
2011	- 3	25	- 75	9	225	81	27.2112
2012	- 2	45	- 90	4	180	16	44.4541
2013	- 1	65	- 65	1	65	1	65.6362
2014	0	90	0	0	0	0	90.7575
2015	1	120	120	1	120	1	119.8180
2016	2	135	310	4	620	16	152.8177
2017	3	190	570	9	1710	81	189.7566
2018	4	230	920	16	3680	256	230.6347
2019	5	275	1375	25	6875	625	275.4520
Total	**0**	**1215**	**2980**	**110**	**13840**	**1958**	**214.9885**

Sustituyendo en las fórmulas:

$$a = \frac{1215(1958) - 13840(110)}{11(1958) - 110^2} = 90.7575$$

$$b = \frac{2980}{110} = 27.0909$$

$$c = \frac{11(13840) - 1215(110)}{11(1958) - 110^2} = 1.9696$$

y el modelo de estimación será:

$$\hat{Y} = 90.7575 + 27.0909X + 1.9696X^2$$

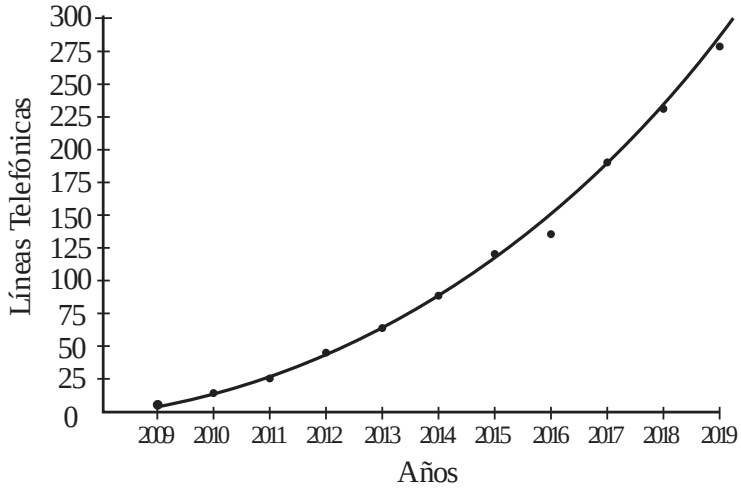
Para estimar el número de líneas telefónicas instaladas en el centro poblado de Mimave para el 2020 se usa $X = 6$:

$$\hat{Y}_{2020} = 90.7575 + 27.0909(6) + 1.9696\ (6)^2$$

$$\hat{Y} = 324.2085$$

como el número de líneas debe ser un valor entero, se redondea a 324.

Figura 6.6
Diagrama de dispersión y recta de ajuste con los datos de la tabla 6.6.



B) Tendencia hiperbólica.

$$\hat{Y} = a + bX^{-1}$$

Esta ecuación representa la rama positiva de una hipérbola equilátera, para obtener sus coeficientes se resuelven las siguientes ecuaciones normales halladas al aplicar mínimos cuadrados:

$$\sum Y = na + b\sum X^{-1}$$
$$\sum X^{-1}Y = a\sum X^{-1} + b\sum X^{-2}$$

C) Tendencia exponencial

$$\hat{Y} = ab^{X}$$

Esta ecuación debe linealizarse antes de aplicar mínimos cuadrados, para esto se calcula su logaritmo, obteniendo:

$$\log \hat{Y} = \log a + X \log b \ .$$

Con esto sus ecuaciones normales serán:

$$\sum \log Y = n \log a + \log b \sum X$$
$$\sum X \log Y = \log a \sum X + \log b \sum X^2 \ .$$

6.2 CORRELACIÓN

Es la relación que existe entre variables y permite conocer el grado de dispersión que tendrían con respecto a alguna curva o recta de ajuste. El valor del coeficiente de correlación (r) no depende del modelo de regresión utilizado y puede ser positivo (correlación directa) si ambas variables incrementan o disminuyen al mismo tiempo, o negativo (correlación inversa) cuando una aumenta si la otra disminuye o viceversa. Si las variables son independientes $r \approx 0$.

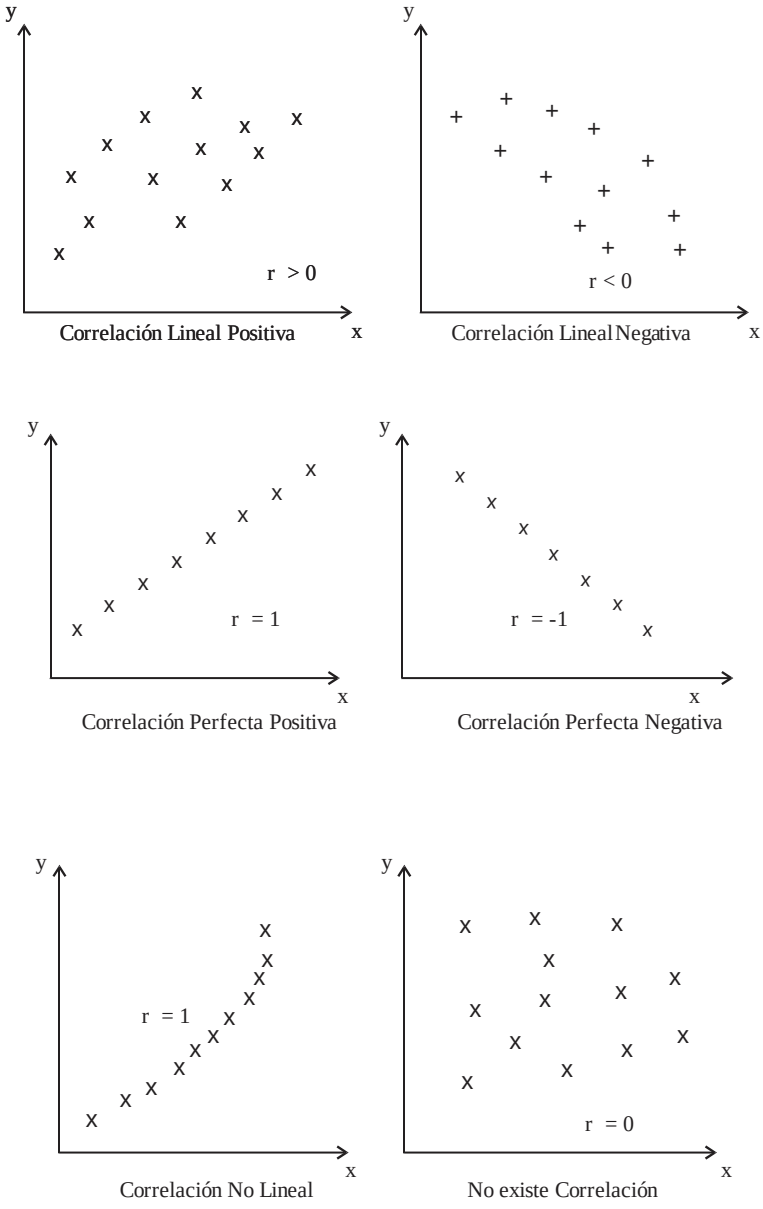
En la tabla 6.7 se presenta un referente teórico del grado de interrelación entre variables a partir del valor de r.

Tabla 6.7
Grado de interdependencia entre variables establecidas por "r"

Coeficiente de correlación "*r*"	**Grado de Interrelación**
0.00 ± 0.20	Correlación nula
± 0.20 ± 0.40	Correlación débil
± 0.40 ± 0.60	Correlación moderada
± 0.60 ± 0.80	Correlación fuerte
± 0.80 ± 1.00	Correlación muy fuerte

Figura 6.7

Diagramas de dispersión según coeficiente de correlación de dos variables.



6.2.1 Correlación de Pearson para datos no agrupados.

Se calcula mediante la fórmula:

$$r = \frac{n\sum XY - \sum X \sum Y}{\{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]\}^{\frac{1}{2}}}$$

Ejemplo:

Calcular r para las horas de estudio y la nota obtenida en una muestra aleatoria de 5 alumnos.

Tabla 6. 8
Horas de estudio y nota obtenida en una muestra de 5 alumnos.

Alumnos	X (Horas de estudio)	Y (Nota)	XY	X^2	Y^2
A	2	6	12	4	36
B	3	8	24	9	64
C	4	10	40	16	100
D	6	12	72	36	144
E	7	15	105	49	225
Total	**22**	**51**	**253**	**114**	**569**

Sustituyendo valores en la fórmula:

$$r = \frac{5(253) - 22(51)}{\{\,[5(114) - 22^2][5(569) - (51)^2]\}^{\frac{1}{2}}}$$

$$r = \frac{143}{\{(86)(244)\}^{\frac{1}{2}}} = \frac{143}{144.856} = 0.987$$

Como se observa la relación es directa casi perfecta entre horas de estudio y la nota obtenida.

6.2.2 Correlación de Pearson para datos agrupados.

Cuando la muestra es grande se hace necesario agrupar los datos en intervalos, con amplitudes no necesariamente iguales, éstos se determinan según la definición conceptual y operacional de las variables del objeto de investigación. A partir de estos intervalos, y sus respectivas frecuencias marginales y conjuntas, se puede calcular el coeficiente de correlación, dado por:

$$r = \frac{n\sum Z_x^{''} Z_y^{''} n_{xy} - (\sum Z_x^{''} n_x)(\sum Z_y^{''} n_y)}{\{[n\sum Z_x^{''2} n_x - (\sum Z_x^{''} n_x)^2][n\sum Z_y^{''2} n_y - (\sum Z_y^{''} n_y)^2]\}^{\frac{1}{2}}}$$

donde:

n_x: frecuencia marginal de la variable x.

n_y: frecuencia marginal de la variable y.

n_{xy}: frecuencia conjunta.

n: tamaño de la muestra.

Para calcular las expresiones $Z_x^{''}$ y $Z_y^{''}$ primero se hallan:

$$Z'_{x_i} = X_i - O_{tx},$$
$$Z'_{y_i} = Y_i - O_{ty}.$$

X_i y Y_i son marcas de clase, O_{tx} y O_{ty} se les denomina orígenes de trabajo y son puntos de referencia que se eligen entre las marcas de

clase, preferentemente aquellas que ocupen el lugar central de las distribuciones.

La variable $Z^{''} = \frac{Z'}{C}$ define Z' respecto a la amplitud interválica promedio (c) obteniendo:

$$Z^{''}_{X_i} = \frac{X_i - O_{tx}}{C_x}$$

$$Z^{''}_{Y_i} = \frac{Y_i - O_{ty}}{C_y}$$

Ejemplo

Se desea determinar el coeficiente de correlación entre el neuroticismo y la extroversión de 100 estudiantes de sociología de una universidad, a partir de los resultados de la aplicación del test de Eysenck que se presentan en la tabla 6.9.

Tabla 6.9
Neuroticismo y extroversión de100 estudiantes de una universidad

Neuroticismo			**Extroversión**											
		X	**0 - 10**		**10 – 13**		**13 - 16**		**16– 24**					
		X_i	5		11.5		14.5(Ot)		20.0					
		$Z^{''}_{X_i}$	*- 1.9*		*-0.6*		*0.00*		*1.1*					
Y	Y_i	$Z^{''}_{Y_i}$	n_{xy}	$Z^{''}_x Z^{''}_y n_{xy}$	n_{xy}	$Z^{''}_x Z^{''}_y n_{xy}$	n_{xy}	$Z^{''}_x Z^{''}_y n_{xy}$	n_{xy}	$Z^{''}_x Z^{''}_y n_{xy}$	n_y	$Z^{''}_y n_y$	$Z^{''2}_y n_y$	$\sum Z^{''}_y Z^{''}_x n_{xy}$
0 - 7	3.5	***-1.6***	**2**	**6.1**	**7**	**6,7**	**8**	**0**	**9**	**-15.8**	26	-41.6	66.6	-3.0
7-11	9	***-0.7***	**5**	**6.7**	**15**	**6.3**	**14**	**0**	**2**	**-1.5**	36	-25,2	17,6	11.5
11-15	13(Ot)	***0.0***	**9**	**0.0**	**6**	**0.0**	**9**	**0**	**1**	**0.0**	25	0.0	0.0	0.0
15- 24	19.5	***1.0***	**7**	**-13.3**	**3**	**-1.8**	**2**	**0**	**1**	**1.1**	13	13.0	13.0	-14.0
	n_x		23		31		33		13		100	-53,8	97.2	-5.5
	$Z^{''}_x n_x$		-43.7		-18.6		0		14.3		-48.9			
	$Z^{''2}_x n_x$		83.03		11.16		0		15.73		109.9			
	$\sum Z^{''}_x Z^{''}_y n_{xy}$		-0.4		11.2		0		-16.2		-5.5			

Las marcas de clase X_i y Y_i son las semisumas de los límites de cada intervalo, por ejemplo $Y_1 = \frac{0+7}{2} = 3.5$. De las marcas de clase se elige el origen de trabajo, o punto de referencia, en este caso se eligieron: $O_{tx} = X_3 = 14.5$ y $O_{ty} = Y_3 = 13$. Como las amplitudes de los intervalos no son iguales, se obtienen sus promedios: $c_x = 5$ y $c_y = 6$. Con estos datos ya podemos calcular los $Z_x^{''}$ y $Z_y^{''}$, por ejemplo $Z_{x_1}^{''}$ y $Z_{Y_1}^{''}$ quedarían calculados de la siguiente manera:

$$Z_{X_1}^{''} = \frac{X_1 - O_{tx}}{c_x} = \frac{5-14.5}{5} = -1.9,$$
$$Z_{Y_1}^{''} = \frac{Y_1 - O_{ty}}{c_y} = \frac{3.5-13}{6} = -1.6\ .$$

Finalmente se calcularán los $Z_x^{''} Z_y^{''} n_{xy}$, con la primera frecuencia conjunta se tendría:

$$Z_{X_1}^{''} Z_{Y_1}^{''} n_{11} = (-1.9)(-1.6)(2) = 6.08 \approx 6.1.$$

Por lo tanto, reemplazando todos los valores, el coeficiente de correlación sería:

$$r = \frac{100(-6.6) - (-48.9)(-53.8)}{\{[100(109.9) - (-48.9^2][100(97.2) - (-53.8)^2]\}^{\frac{1}{2}}}$$

Sustituyendo

$$r = \frac{-3180.82}{7661.04.} = -0.42$$

El signo indica que la correlación es inversa, esto significa que, si se incrementa el valor de una variable disminuye el valor de la otra y viceversa. Tomando en cuenta la tabla 6.9, para nuestro caso la correlación sería moderada.

6.2.3. Coeficiente de determinación r^2 .

Mide la proporción de variación existente en Y que es explicada por la variable de X, se expresa en términos de porcentaje y se puede calcular mediante dos fórmulas:

- Elevando al cuadrado el coeficiente de correlación r.
- Usando el hecho de que en el análisis de regresión la variación total de Y, $\sum(Y-\bar{Y})^2$, está dado por la suma de la variación explicada y la variación no explicada.

  $$\sum(Y-\bar{Y})^2 = \sum(\hat{Y}-\bar{Y})^2 + \sum(Y-\hat{Y})^2$$

  Por lo que el coeficiente de determinación sería:

$$r^2 = \frac{\sum(\hat{Y}-\bar{Y})^2}{\sum(Y-\bar{Y})^2}$$

Si $r^2 = 0$, X no explica nada de la variabilidad de Y, pues $\sum(\hat{Y}-\bar{Y})^2 = 0$. Si $r^2 = 1$, X explica toda la variabilidad de Y, pues $\sum(Y-\hat{Y})^2 = 0$.

Ejemplo didáctico.

Una investigación tiene como objetivo determinar la relación entre el porcentaje de graduados de escuelas superiores y el sueldo medio en miles de dólares de una ciudad. Se toma una muestra aleatoria de 5 ciudades y se determina por los datos del censo la información de la tabla 6.10.

Tabla 6. 10
Porcentaje de graduados y sueldo en una región geográfica.

% Graduados (X)	Sueldo (Y)	XY	X^2	Y^2
7.2	4.2	30.24	51.84	17.64
6.7	4.9	32.83	44.89	24.01
10.2	6.2	63.24	104.04	38.44
6.3	3.8	23.94	39.69	14.44
6.0	4.4	26.40	36.00	19.36
36.4	**23.5**	**176.75**	**276.46**	**113.89**

$\hat{Y}$	$(Y-\bar{Y})^2$	$(\hat{Y}-\bar{Y})^2$	$(Y-\hat{Y})^2$
4.66	0.25	0.0016	0.2116
4.42	0.04	0.0784	0.2304
6.13	0.25	2.0449	0.0049
4.22	0.81	0.2304	0.1764
4.07	0.09	0.3969	0.1089
23.50	**3.44**	**2.7522**	**0.7322**

Los datos que dependen del modelo de regresión lineal utilizado, se han calculado de la siguiente manera:

$$\bar{X} = 7.28 \qquad \bar{Y} = 4.7$$

$$b = \frac{n\Sigma XY - \Sigma X\,\Sigma Y}{n\Sigma X^2 - (\Sigma X)^2} = \frac{5(176.68) - 36.4(23.5)}{5(276.46) - (36.4)^2}$$

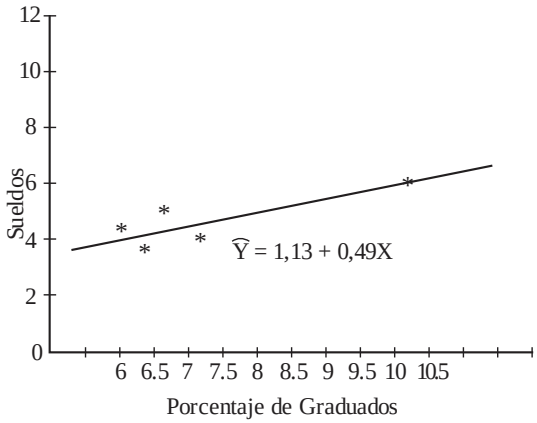
$$b = \frac{27.85}{57.34} = 0.49$$

$$a = \bar{Y} - b\bar{X} = 4.7 - 0.49(7.28) = 1.3$$

$$\hat{Y} = 1.13 + 0.49X$$

$$\hat{Y}_1 = 1.13 + 0.49(7.2) = 4.66$$
$$\hat{Y}_2 = 1.13 + 0.49(6.7) = 4.42$$
$$\hat{Y}_3 = 1.13 + 0.49(10.2) = 6.13$$
$$\hat{Y}_4 = 1.13 + 0.49(6.3) = 4.22$$
$$\hat{Y}_5 = 1.13 + 0.49(6) = 4.07$$
$$\sum \hat{Y} = 23.50 \ .$$

Figura 6.8
Diagrama de dispersión y recta de ajuste con los datos de la tabla 6.10.



Sustituyendo:

$$r^2 = \frac{2.7522}{3.44} = 0.80,$$

lo que indica que el 80% de la variación del ingreso es debido al nivel de instrucción. Además: $r = (0.8)^{\frac{1}{2}} = 0.89$, lo que ilustra la alta relación entre las variables, esto se puede comprobar usando la fórmula original de Pearson:

$$r = \frac{5(176.65) - 36.4(23.5)}{\{[5(276.46) - (36.4)^2][5(113.89) - (23.5)^2]\}^{\frac{1}{2}}}$$

$$r = \frac{27.85}{\{(57.34)(17.2)^2\}^{\frac{1}{2}}}$$

$$r = \frac{27.85}{\sqrt{986.25}} = 0.89$$

También se puede calcular el error típico de estimación o dispersión alrededor del modelo $S_{Y/X}$, que confirma el grado de correlación

$$S_{Y/X} = \sqrt{\frac{\sum(Y - \hat{Y})^2}{n - 2}} = \sqrt{\frac{0.7322}{3}} = 0.49$$

6.2.4 Coeficiente de correlación de Spearman.

Se usa para calcular la relación entre variables cualitativas ordinales, con la siguiente fórmula:

$$r = 1 - \frac{6.\sum d^2}{n(n^2 - 1)}$$

donde d es la diferencia entre los rangos de las variables luego de ser ordenados, n es el tamño de muestra y r es el estimador de ρ.

Ejemplo.

La evaluación a cinco directivos en autoritarismo y prestigio social es como sigue:

Personas	Autoritarismo	Prestigio social
A	50	9
B	30	1
C	60	5
D	20	6
E	80	4

Se desea determinar el grado de correlación de estas variables (asociación por ser cualitativas). Al ordenar los datos por rangos y estableciendo las diferencias entre ellos, se obtiene la tabla 6.11.

Tabla 6.11
Evaluación a cinco directivos en autoritarismo y prestigio social.

Directivos	Autoritarismo	Rango	Prestigio social	Rango	d	d^2
A	50	3	9	1	2	4
B	30	4	1	5	-1	1
C	60	2	5	3	-1	1
D	20	5	6	2	3	9
E	80	1	4	4	-3	9
						24

Reemplazando.

$$r = 1 - \frac{6(24)}{5(25-1)} = 1 - \frac{144}{120} = 1 - 1.20 = -0.20\ .$$

Este valor indica que el grado de correlación (asociación) es insignificante e inversa, siendo el coeficiente de determinación sólo el 4%.

Si más de una observación tuvieran igual rango, se saca el promedio y ese promedio se asigna a cada rango.

AUTOEVALUACIÓN II

1. Los datos que siguen corresponden a una prueba de inteligencia y de adaptación social de 10 personas.

C. I.	80	75	71	80	50	64	46	70	64	74
A. Soc.	146	90	114	77	143	26	88	105	78	44

   Muestre el diagrama de dispersión y grafique la recta que podría ajustarlos datos.

   a. Calcule los coeficientes a y b del modelo.
   b. Si la adaptación social está relacionada a la inteligencia, ¿Cuál es su valor para una persona que tiene 85 en la prueba de inteligencia?
   c. Determine la correlación entre el coeficiente intelectual y la adaptación social.
   d. Calcule la variación en la adaptación social que está explicada por la inteligencia.
   e. Calcule el error típico del modelo.
   f. Estime con 95% de confiabilidad el valor de la adaptación social, para una persona que tiene 90 en la prueba de inteligencia.
   g. Cuál es la probabilidad de encontrar personas con una adaptación social entre $\hat{Y} \pm 3S_{X/Y}$

2. Realice la medición de la estatura y el peso a 30 alumnos y analice la relación entre estas variables, elaborando una distribución

bidimensional de frecuencias. Calcule el coeficiente correlación para datos agrupados.

3. El rendimiento académico en los niños de 5 a 8 años de las escuelas rurales y urbanas de determinada región respectivamente son como siguen:

   Escuela rural: $\bar{X} = 15 \quad S = 5$

   Escuela urbana: $\bar{X} = 14 \quad S = 3$

   ¿Son los niños de las escuelas rurales mejores que los niños de las escuelas urbanas? ¿Por qué?

4. $\hat{Y}_A = 14 + 0.8X \qquad S_{Y/X} = 1.5$

   $\hat{Y}_B = 14 + 0.5X \qquad S_{Y/X} = 0.5$

   Elija el modelo que explique mejor la relación de las variables y grafique cada modelo con una franja confidencial del 95%.

5. De 5 mujeres entre 35 y 40 años y se registró su ansiedad y frecuencia cardiaca como sigue:

Ansiedad	41	45	41	42	45
Frecuencia cardiaca	55	60	65	70	65

   Realice un análisis de regresión completo con los datos del ejemplo anterior.

CAPÍTULO VII

NÚMEROS ÍNDICES

7.1 ÍNDICES GENERALES

Los índices son medidas de resumen que expresan en un solo dato el estado de un fenómeno o suceso; se obtienen al operar las magnitudes o valores de los indicadores para realizar comparaciones.

Ejemplos:

- Índice de Precios al Consumidor. (Alimentación, Vivienda, Salud, Vestido, Educación, Transporte, Cultura, Recreación)
- Índice de Desempleo. (Número de desempleados, Población Activa)

- Índice de Desempeño Docente. (Conocimientos, Motivación, Habilidades Lógica, Habilidades de enseñanza)
- Índice de Costos Educacionales. (Presupuesto asignado a la unidad, Total de matrículas registradas por año)
- Índice de Producción de Graduados. (Número de graduados por año, Población de alumnos de la Escuela Profesional por año)

El índice adecuadamente calculado, garantiza exactitud potencialidad y reproductibilidad.

La *exactitud* es el grado en que refleja lo que se mide, tiene límites definidos de variabilidad, según los indicadores que lo forman.
La *potencialidad* es la capacidad de información sobre el concepto que mide.
La *reproductibilidad*, capacidad de reconstrucción del concepto que mide a partir de la comprensión del propio índice.

7.1.1 Razón

Una razón es el cociente de dos números en el que el numerador y el denominador se refieren a cosas distintas, sin que ninguna se contenga en la otra. Tienen las mismas unidades y el rango es de 0 a infinito.

Ejemplo:

En Perú en el año 1995 hubo 7 342 casos de leishmaniasis, de los cuales 1 122 en Cuzco, 649 en Madre de Dios, 233 en Ucayali.

La razón $= \frac{\text{N° de casos en Cuzco}}{\text{N° de casos en Madre de Dios}} = \frac{1122}{649} = 1.73;$

se interpreta como por cada caso de Leishmaniosis en Madre de Dios, en Cuzco hubieron 1,73. La razón $= \frac{\text{N° de casos en Cuzco}}{\text{N° de casos en Ucayali}} = \frac{1122}{233} =$ 4.81; se interpreta como por cada caso de Leishmaniosis en Ucayali, en Cuzco hubieron 4 casi 5 casos.

7.1.2 Proporción

Es el cociente en el que el denominador contiene a los elementos del numerador, por esto su valor se encuentra entre 0 y 1. Puede ser expresada como porcentaje y se usa para estimar la probabilidad de un suceso.

Ejemplos:

Personas con leishmaniasis en Cuzco con respecto a todos los casos en Perú: $\frac{1122}{7343} = 0.15$; se interpreta como que el 15% de las personas con leishmaniasis en Perú fueron detectadas en Cuzco.

Personas con leishmaniasis en Madre de Dios con respecto a todos los casos en Perú: $\frac{649}{7343} = 0.09$; se interpreta como que el 9% de las personas con leishmaniasis en Perú fueron detectadas en Madre de Dios.

Suponiendo que en cierta región el número de concepciones fue de 1752 y el número de muerte fetales 332, la proporción de muertes fetales $= \frac{\text{nº demuertes fetales}}{\text{nº de concepciones}} = 0{,}19$, esto se interpreta como que el 19% de concepciones se vieron interrumpidas por muerte.

7.1.3 Tasa

"Es la relación del número de casos, frecuencias o eventos de una categoría entre el número total de observaciones, multiplicadas por un múltiplo de 10, generalmente 100 ó 1000" (INEI, 2006, p. 58).

$$\text{Tasa} = \frac{\text{Número de eventos durante un periódo t}}{\text{Número total de observaciones en el periodo t}} \times 10n$$

Algunas tasas presentadas por el INEI (2006) en su glosario de términos estadísticos son:

- **Tasa bruta de mortalidad para el año z (TBM_z)**

$$\text{TBM}_z = \frac{\text{Defunciones ocurridas durante el año z}}{\text{Población al 30 de junio del año z}} \times 100$$

- **Tasa de analfabetismo para el año z (TA_z)**

$$\text{TA}_z = \frac{\text{Analfabetos de 15 y más años en el año z}}{\text{Población de 15 y más años en el año z}} \times 100$$

- **Tasa de inflación (TI)**

$$\text{TI} = \frac{\text{IPC año actual} - \text{IPC año base}}{\text{IPC año base}} \times 100$$

Donde IPC es el índice de precios al consumidor.

- **Tasa de mortalidad infantil para el año z ($\mathbf{TMI_z}$)**

$$\mathrm{TMI}_z = \frac{\text{Defunciones de menores de 1 año en el año z}}{\text{Nacidos vivos en el año z}} \times 1000$$

Ejemplo 1:

En el año 1995 se encontraban censados en Cuzco 1 055 401 personas, en Ucayali 321 819, en Madre de Dios 66 684 y en Perú 23 029 603.

La tasa de leishmaniasis en Cuzco en el año 1995 = $\frac{1125}{1055401} = 1.06 \times 10^{-3}$; esto significa que, de 1 000 personas, 1,06 se contagiaron de leishmaniasis en Cuzco.

La tasa de leishmaniasis en Madre de Dios en el año 1995= $\frac{649}{66684} = 9.73 \times 10^{-3}$; esto significa que, de 1 000 personas, 9,73 se contagiaron de leishmaniasis en Madre de Dios.

La tasa de leishmaniasis en Perú en el año 1995= $\frac{7343}{23029603} = 3.18 \times 10^{-4}$ esto significa que, de 10 000 personas, 3,18 se contagiaron de leishmaniasis en Perú.

Ejemplo 2.

La población en Perú el año de 2012 era 30 135 875 habitantes, de los cuales 1 229 260 en Lambayeque, 1 006 953 en Loreto, 417 508 en Amazonas y 1 799 607 en Piura. Los casos de dengue declarados

durante ese año en Perú fueron 29 964 de los cuales 732 en Lambayeque, 4 588 en Loreto, 676 en Amazonas y 1 355 en Piura.

En ese año, la tasa del dengue en Lambayeque $= \frac{732}{1229260} = 5.95 \times 10^{-4}$; o sea 5,95 personas por cada 10 000 tuvieron dengue en Lambayeque.

La tasa del dengue en Loreto $= \frac{4588}{1006953} = 4.56 \times 10^{-3}$; o sea 4,56 personas por cada 1 000 tuvieron dengue en Loreto. (45 personas por cada 10 000)

La tasa del dengue en Amazonas $= \frac{676}{1006953} = 1{,}62 \times 10^{-3}$; o sea 1,62 personas por cada 1 000 tuvieron dengue en Amazonas. (16 personas por cada 10 000)

La tasa del dengue en Piura $= \frac{1355}{1006953} = 7{,}53 \times 10^{-4}$; o sea 7,53 personas por cada 10 000 tuvieron dengue en Piura.

7.1.4 Incidencia.

En epidemiología es la ocurrencia de nuevos casos de enfermedad, lesión, u otra condición médica durante un periodo de tiempo específico, típicamente calculada como una proporción o tasa (Mcnutt y Krug, 2016).

Cuando es calculada como proporción se le denomina incidencia acumulada (ia) y se usa la siguiente fórmula:

$$\text{ia} = \frac{\text{Nuevos casos en cierto periodo de tiempo}}{\text{Población total en riesgo durante ese periodo de tiempo}}$$

Ejemplo 1.

En un viaje de promoción que duró 10 días, de 43 alumnos 4 se contagiaron de Covid-19, en este caso la incidencia acumulada sería: $\text{ia} = \frac{4}{43} = 0.093$, que significa un 9.3% de incidencia acumulada durante el periodo de 10 días.

Por otra parte, la tasa de incidencia usa lo que se denominará sujeto-tiempo, que es una medida que toma en consideración la cantidad de sujetos y el tiempo que cada uno de ellos ha sido evaluado; por ejemplo, si se evalúa a 100 personas durante un año se tendrá 100 personas-año, si a 10 personas se les evalúa durante 10 años también se tendrá 100 personas-año, por último si en un grupo de 10 personas, a 4 de ellas se las evalúa durante 5 años, a 3 durante 5 años y a otras 3 durante 6 años, entonces se tendrá $(4 \times 5) + (3 \times 5) + (3 \times 6) = 53$ personas-año.
El cálculo de la tasa se obtiene con la fórmula:

$$\text{ti} = \frac{\text{Nuevos casos en cierto periodo de tiempo}}{\text{Total de sujetos-tiempo}}$$

Ejemplo 2.

Entre los años 2020 y 2022 se evaluaron a 1000 personas, de las cuales 140 se contagiaron de varicela. De las 1000 personas, a 600 se la evaluó durante 3 años, a 300 durante 2 años y a 100 durante un año, la tasa de incidencia sería

$$\text{ti} = \frac{140}{(600\times3)+(300\times2)+(100\times1)} = 0.056,$$

lo cual equivale a decir que hay una incidencia de 56 casos de varicela por cada 1000 personas-año.

7.1.5 Prevalencia.

Es la proporción de una población con una enfermedad o una condición particular en un punto específico de tiempo (prevalencia puntual) o durante un periodo de tiempo (prevalencia de periodo). (Mcnutt y Krug, 2013)

$$\text{Pr} = \frac{\text{Casos existentes en un tiempo determinado}}{\text{Población total en riesgo en un tiempo determinado}}$$

Ejemplo.

En un centro geriátrico masculino se detectó que en 2 años hubo 20 casos de cáncer de próstata dentro de una población total de riesgo de 500, por lo que su prevalencia es $\Pr = \frac{20}{500} = 0.04$, lo que signifca que hubo una prevalencia de cáncer de próstata del 4% en 2 años.

7.2 ÍNDICES ECONÓMICOS

Son indicadores estadísticos que facilitan el análisis y la interpretación de la situación económica pasada, presente o futura. Entre estos se tiene:

Índice de precios. "Compara los cambios en el precio entre dos periodos. El índice de precios al consumidor mide los cambios globales de varios bienes de consumos y también de los servicios, y se utiliza para definir el costo de vida" (Levin, 1987/1988, p.774).

Índice de cantidad. "Mide cuanto cambia con el tiempo el número o cantidad de una variable" (Levin, 1987/1988, p.774).

Índice de valor. "Mide los cambios en el valor monetario de una variable [combinando] precio y cantidad para presentar un índice más informativo" (Levin, 1987/1988, p.774).

Ejemplo:

Tabla 7.1
Precios y cantidad de productos para cálculo de números índices

Producto	2019 (**Periodo Base**)		2023 (**Periodo Actual**)	
	Precio (P_0)	Cantidad (Q_0)	Precio (P_1)	Cantidad (Q_1)
Palta	15	400	20	300
Papaya	30	100	25	200
Manzana	10	250	15	100

Índice de precio simple de la palta

$$I_P = \frac{P_1}{P_0} \times 100$$

$$I_P = \frac{20}{15} \times 100$$

$$I_P = 133.33 \rightarrow 133.33 - 100 = 33.33$$

El precio de la plata ha tenido un aumento del 33.33%.

Índice de cantidad simple de la papaya

$$I_Q = \frac{Q_1}{Q_0} \times 100$$

$$I_Q = \frac{200}{100} \times 100$$

$$I_Q = 200 \rightarrow 200 - 100 = 100$$

La cantidad de papaya ha tenido un aumento del 100%.

Índice de valor simple de la manzana

$$I_V = \frac{P_1 Q_1}{P_0 Q_0} \times 100$$

$$I_V = \frac{(15 \times 100)}{(10 \times 250)} \times 100$$

$$I_V = 60 \rightarrow 60 - 100 = -40$$

El valor de la manzana ha tenido una disminución del 40%.

Además de los índices simples, existen los **índices ponderados**:

Índice de Laysperes

$$I_L = \frac{\sum P_1 Q_0}{\sum P_0 Q_0} \times 100$$

$$I_L = \frac{(20 \times 400) + (25 \times 100) + (15 \times 250)}{(15 \times 400) + (30 \times 100) + (10 \times 250)} \times 100$$

$$I_L = \frac{14250}{11500} \times 100$$

$$I_L = 123.91 \rightarrow 123.91 - 100 = 23.91$$

El precio de los productos ha tenido un crecimiento del 23.91%.

Índice de Paasche

$$I_P = \frac{\sum P_1 Q_1}{\sum P_0 Q_1}$$

$$I_P = \frac{(20 \times 300) + (25 \times 200) + (15 \times 100)}{(15 \times 300) + (30 \times 200) + (10 \times 100)} \times 100$$

$$I_P = \frac{12500}{11500} \times 100$$

$$I_P = 108.70 \rightarrow 108.70 - 100 = 8.70$$

El precio de los productos ha tenido un crecimiento del 8.70%.

Índice de Fisher.

$$I_F = \sqrt{I_L I_P}$$

$$I_F = \sqrt{\left(\frac{14250}{11500} \times 100\right) \times \left(\frac{12500}{11500} \times 100\right)}$$

$$I_F = 116.06 \rightarrow 116.06 - 100 = 16.06$$

El precio de los productos ha tenido un crecimiento del 16.06%.

REFERENCIAS

Abreu, J. L. (2012). Constructos, Variables, Dimensiones, Indicadores y Congruencia. *Daena: International Journal of Good Conscience*, 7(3), 123-130. http://www.spentamexico.org/v7-n3/7(3)123-130.pdf

Ary, D., Jacobs, L. C. y Razavieh, A. (1982). *Introducción a la investigación Pedagógica* (Trad. J. Salazar y J. Pecina). Nueva editorial interamericana. (Trabajo original publicado en 1979).

Bunge, M. (1973). *La investigación científica* (Trad. M. Sacristán). Editorial Ariel. (Trabajo original publicado en 1969).

Corral, Y. (2009). Validez y confiabilidad de los instrumentos para la recolección de datos. *Revista ciencias de la educación*, (33), 228-247.

Cronbach, L. J. y Meehl, P. E. (1956). Construct validity in psychological tests. *Minnesota studies in the philosophy of science*, *1*, 174-204.

de la Fuente, S. (2011). *Análisis Factorial*. Universidad Autónoma de Madrid. https://www.fuenterrebollo.com/Economicas/ECONOMETRIA/MULTIVARIANTE/FACTORIAL/analisis-factorial.pdf

Hurtado, S. (2002). *Criterio de expertos. Su Procesamiento a través del método Delphy*. Histodidáctica. http://www.ub.edu/histodidactica/index.php?option=com_content&view=article&id=21:criterio-de-expertos-su-procesamiento-a-traves-del-metodo-delphy&catid=11:metodologia-y-epistemologia&Itemid=103

Instituto Nacional de Estadística e Informática. (2006). *Glosario básico de términos estadísticos.* https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib0900/Libro.pdf

Kerlinger, F. N. (1975). *Investigación del comportamiento* (Trad. J. Blengio y J. Pecina). Nueva editorial interamericana. (Trabajo original publicado en 1973).

Levin, R. I. (1988). *Estadística para administradores* (Trad. M. Efrén Alatorre). PRENTICE-HALL HISPONAMERICANA. (Trabajo original publicado en 1987).

Mcnutt, L. y Krug, A. (14 de abril de 2016). *Incidence. Encyclopedya Britannica.* [incidencia. Enciclopedia Britannica]. https://www.britannica.com/science/incidence-epidemiology

Mcnutt, L. y Krug, A. (16 de diciembre de 2013). *Prevalence. Encyclopedya Britannica.* [prevalencia. Enciclopedia Britannica]. https://www.britannica.com/science/prevalence

Mejía Mejía, E. (2005). *Técnicas e instrumentos de investigación*. Centro de Producción Editorial e Imprenta de la Universidad Nacional Mayor de San Marcos.

Rojas, A. (31 de agosto de 2007). *Investigación científica: 24 lecciones para conocer y usar la ciencia*. Filosofía y lenguaje de la ciencia. http://lenguajecientifico.blogspot.com/2007/08/leccin-1-introduccin-la-epistemologa-de.html

Stevens, S. S. (1946). On the theory of scales of measurement [Sobre la teoría de las escalas de medición]. *Science*, *103*(2684), 677-680.

Vásquez Sánchez, E. (1996). *Elementos de estadística.* Autoedición.

Vásquez Sánchez, E., Rodríguez Alayo, N. M., Ortiz Basauri, G. M. y Vásquez Ortiz, E. A. (2021). *El proyecto de investigación.* Editorial Universitaria. UNPRG. Chiclayo-Perú.

BIBLIOGRAFÍA

Alarcón, R. (1991). *Métodos y diseños de investigación del comportamiento.* Editorial universitaria. Universidad Cayetano Heredia.

Benza, J. C. (1966). *Estadística general con énfasis en muestreo*. Editorial Jurídica.

Dixon, J. W. y Massey Jr, F. J. (1951) *Introduction to statistical analysis.* McGraw-Hill.

Merril, W. C., Fox, K.A. y Kitaigorodzki, M. (1972). *Introducción a la estadística económica*. Amorrortucentro regional de ayuda técnica.

Morice, E. (1975). *Diccionario de estadística.* Editorial Cesca.

Neter, J. y Wasserman, W. (1973). *Fundamentos de estadística.* Editorial Continental.

Yamane, T. (1974). *Estadística*. Editorial Harla.

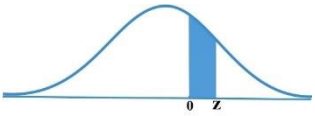
Wayne, W. A. (1981). *Estadística con aplicaciones a las ciencias sociales y educación*. McGraw-Hill.

ANEXOS

Tabla de números aleatorios

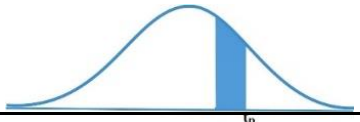
51772	**74640**	**42331**	**29044**	**46621**	**62898**	**93582**
24033	**23491**	**83587**	**6568**	**21960**	**21387**	**76105**
45939	**60173**	**52078**	**25424**	**11645**	**55870**	**56974**
30586	**2133**	**75797**	**45406**	**31041**	**86707**	**12973**
3585	**79353**	**81938**	**82322**	**96799**	**85659**	**36081**
64937	**3355**	**95863**	**20790**	**65304**	**55189**	**745**
15630	**64759**	**51135**	**98527**	**62586**	**41889**	**25439**
9448	**56301**	**57683**	**30277**	**94623**	**85418**	**68829**
21531	**91157**	**77331**	**60710**	**52290**	**16835**	**48653**
91097	**17480**	**29414**	**6829**	**87843**	**28195**	**27279**
50532	**25496**	**95652**	**42457**	**73547**	**76552**	**50020**
7136	**40876**	**79971**	**54195**	**25708**	**51817**	**36732**
27989	**64728**	**10744**	**8396**	**56242**	**90985**	**28868**
85184	**73949**	**36601**	**46253**	**477**	**25234**	**9908**
54398	**21154**	**97810**	**36764**	**32869**	**11785**	**55261**
65544	**34371**	**9591**	**7839**	**58892**	**92843**	**72828**
8263	**65952**	**85762**	**64236**	**39238**	**18776**	**84303**
39817	**67906**	**48236**	**16057**	**81812**	**15815**	**63700**
62257	**4077**	**79443**	**95203**	**2479**	**30763**	**92486**
53298	**90276**	**62545**	**21944**	**16530**	**3878**	**7516**
4186	**19640**	**87056**	**73533**	**92520**	**51759**	**32783**
10863	**97453**	**90581**	**52850**	**13932**	**38045**	**58107**
37428	**93507**	**94271**	**34654**	**61261**	**51658**	**70502**
17169	**88116**	**42187**	**62230**	**64164**	**74266**	**93657**
50884	**14070**	**74950**	**98839**	**12848**	**14316**	**89167**
65253	**11822**	**15804**	**41783**	**60065**	**8274**	**38953**
88036	**24034**	**67283**	**63127**	**47593**	**97941**	**65191**
6652	**41982**	**49159**	**56605**	**13466**	**65194**	**28614**
71590	**16159**	**14676**	**32646**	**48427**	**38276**	**43042**
47152	**35683**	**47280**	**48625**	**61842**	**40851**	**22811**
24819	**52984**	**76168**	**25174**	**5647**	**7041**	**48805**
72484	**94923**	**75936**	**9517**	**78124**	**91188**	**69171**
99431	**50995**	**20507**	**25990**	**17246**	**70539**	**82473**
36574	**72139**	**70185**	**93088**	**18509**	**35479**	**36801**
59009	**38714**	**38723**	**35259**	**95078**	**4961**	**98140**
91341	**84821**	**63886**	**34878**	**62890**	**41462**	**8471**
99247	**46149**	**3229**	**6225**	**1709**	**13377**	**75116**
85915	**19219**	**45943**	**32946**	**77203**	**38975**	**10365**
54083	**23631**	**5825**	**98529**	**72586**	**53115**	**65749**
95715	**2526**	**33537**	**40366**	**9702**	**39856**	**55330**

AREAS BAJO LA CURVA NORMAL TIPIFICADA DE 0 a z



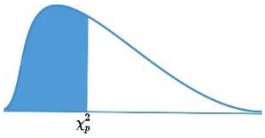
z	0	1	2	3	4	5	6	7	8	9
0	0,0000	0,0040	0,0080	0,0120	0,016	0,0199	0,0239	0,0279	0,0319	0,0359
0.1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0754
0.2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0.3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0.4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0.5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0.6	0,2258	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2518	0,2549
0.7	0,2580	0,2512	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0.8	0,2381	0,2910	0,2939	0,2967	0,2996	0,3023	0,3051	0,3078	0,3106	0,3133
0.9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1.1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1.2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1.3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1.4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1.5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1.6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1.7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1.8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1.9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2.1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2.2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2.3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2.4	0,4918	0,4920	0,4922	0,4925	0.4927	0,4929	0,4931	0,4932	0,4934	0,4936
2.5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2.6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2.7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2.8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2.9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3.1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993
3.2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4994	0,4995	0,4995	0,4995
3.3	0,4995	0,4995	0,4995	0,4996	0,4996	0,4996	0,4996	0,4996	0,4996	0,4997
3.4	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998
3.5	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998
3.6	0,4998	0,4998	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3.7	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3.8	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3.9	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000

DISTRIBUCIÓN t STUDENT CON v GRADOS DE LIBERTAD (ÁREA SOMBREADA=p)



v	$t_{0,995}$	$t_{0,99}$	$t_{0,975}$	$t_{0,95}$	$t_{0,90}$	$t_{0,80}$	$t_{0,75}$	$t_{0,70}$	$t_{0,60}$	$t_{0,65}$
1	63,66	31,82	12,71	6,31	3,08	1,376	1,000	0,727	0,325	0,158
2	9,92	6,96	4,30	2,92	1,89	1,061	0,816	0,617	0,289	0,142
3	5,84	4,54	3,18	2,35	1,64	0,978	0,765	0,584	0,277	0,137
4	4,60	3,75	2,78	2,13	1,53	0,941	0,741	0,569	0,271	0,134
5	4,03	3,36	2,57	2,02	1,48	0,920	0,727	0,559	0,267	0,132
6	3,71	3,14	2,45	1,94	1,44	0,906	0,718	0,553	0,265	0,131
7	3,50	3,00	2,36	1,90	1,42	0,896	0,711	0,549	0,263	0,130
8	3,36	2,90	2,31	1,86	1,40	0,889	0,706	0,546	0,262	0,130
9	3,25	2,82	2,26	1,83	1,38	0,883	0,703	0,543	0,261	0,129
10	3,17	2,76	2,23	1,81	1,37	0,879	0,700	0,542	0,260	0,129
11	3,11	2,72	2,20	1,80	1,36	0,876	0,697	0,540	0,260	0,129
12	3,06	2,68	2,18	1,78	1,36	0,873	0,695	0,539	0,259	0,128
13	3,01	2,65	2,16	1,77	1,35	0,870	0,694	0,538	0,259	0,128
14	2,98	2,62	2,14	1,76	1,34	0,868	0,692	0,537	0,258	0,128
15	2,95	2,60	2,13	1,75	1,34	0,866	0,691	0,536	0,258	0,128
16	2,92	2,58	2,12	1,75	1,34	0,865	0,690	0,535	0,258	0,128
17	2,90	2,57	2,11	1,74	1,33	0,863	0,689	0,534	0,257	0,128
18	2,88	2,55	2,10	1,73	1,33	0,862	0,688	0,534	0,257	0,127
19	2,86	2,54	2,09	1,73	1,33	0,861	0,688	0,533	0,257	0,127
20	2,84	2,53	2,09	1,72	1,32	0,860	0,687	0,533	0,257	0,127
21	2,83	2,52	2,08	1,72	1,32	0,859	0,686	0,532	0,257	0,127
22	2,82	2,51	2,07	1,72	1,32	0,858	0,686	0,532	0,256	0,127
23	2,81	2,50	2,07	1,71	1,32	0,858	0,685	0,532	0,256	0,127
24	2,80	2,49	2,06	1,71	1,32	0,857	0,685	0,531	0,256	0,127
25	2,79	2,48	2,06	1,71	1,32	0,856	0,684	0,531	0,256	0,127
26	2,78	2,48	2,06	1,71	1,32	0,856	0,684	0,531	0,256	0,127
27	2,77	2,47	2,05	1,70	1,31	0,855	0,684	0,531	0,256	0,127
28	2,76	2,47	2,05	1,70	1,31	0,855	0,683	0,530	0,256	0,127
29	2,76	2,46	2,04	1,70	1,31	0,854	0,683	0,530	0,256	0,127
30	2,75	2,46	2,04	1,70	1,31	0,854	0,683	0,530	0,256	0,127
40	2,70	2,42	2,02	1,68	1,30	0,851	0,681	0,529	0,255	0,126
60	2,66	2,39	2,00	1,67	1,30	0,848	0,679	0,527	0,254	0,126
100	2,62	2,36	1,98	1,66	1,29	0,845	0,677	0,526	0,254	0,126
∞	2,58	2,33	1,96	1,645	1,28	0,842	0,674	0,524	0,253	0,126

DISTRIBUCIÓN χ^2_p
CON v GRADOS DE LIBERTAD
(AREA SOMBREADA = p)



v	$\chi^2_{0,995}$	$\chi^2_{0,99}$	$\chi^2_{0,975}$	$\chi^2_{0,95}$	$\chi^2_{0,90}$	$\chi^2_{0,75}$	$\chi^2_{0,50}$	$\chi^2_{0,25}$	$\chi^2_{0,10}$	$\chi^2_{0,05}$	$\chi^2_{0,025}$	$\chi^2_{0,01}$	$\chi^2_{0,005}$
1	7,88	6,63	5,02	3,94	2,71	1,32	0,455	0,102	0,0158	0,0009	0,0010	0,0002	0,0000
2	10,6	9,21	7,38	5,99	4,61	2,71	1,39	0,575	0,211	0,103	0,0506	0,0201	0,0100
3	12,8	11,3	9,35	7,81	6,25	4,11	2,37	1,21	0,584	0,352	0,216	0,115	0,072
4	14,9	13,3	11,1	9,49	7,78	5,39	3,36	1,92	1,06	0,711	0,484	0,297	0,207
5	16,7	15,1	12,8	11,1	9,24	6,33	4,35	2,67	1,61	1,15	0,831	0,554	0,412
6	18,5	16,8	14,4	12,6	10,6	7,84	5,35	3,45	2,20	1,64	1,24	0,872	0,676
7	20,3	18,5	16,0	14,1	12,0	9,04	6,35	4,25	2,83	2,17	1,69	1,24	0,969
8	22,0	20,1	17,5	15,5	13,4	10,2	7,34	5,07	3,49	2,73	2,18	1,65	1,34
9	23,6	21,7	19,0	16,9	14,7	11,4	8,34	5,90	4,17	3,33	2,70	2,09	1,73
10	25,2	23,2	20,5	18,3	16,0	12,5	9,34	6,74	4,87	3,94	3,25	2,56	2,16
11	26,8	24,7	21,9	19,7	17,3	13,7	10,3	7,58	5,58	4,57	3,82	3,05	2,60
12	28.3	26,2	23,3	21,0	18,5	14,8	11,3	8,44	6,30	5,23	4,40	3,57	3,07
13	29,8	27,7	24,7	22,4	19,8	16,0	12,3	9,30	7,04	5,89	5,01	4,11	3,57
14	31,3	29,1	26,1	23,7	21,1	17,1	13,3	10,2	7,79	6,57	5,63	4,66	4,07
15	32,8	30,6	27,5	25,0	22,3	18,2	14,3	11,0	8,55	7,26	6,26	5,23	4,60
16	34, 3	32,0	28,8	26,3	23,5	19,4	15,3	11,9	9,31	7,96	6,91	5,81	5,14
17	35,7	33,4	30,2	27,6	24,8	20,5	16,3	12,8	10,1	8,67	7,56	6,41	5,70
18	37,2	34,8	31,5	28,9	26,0	21,6	17,3	13,7	10,9	9,39	8,23	7,01	6,26
19	38,6	36,2	32,9	30,1	27,2	22,7	18,3	14,6	11,7	10,1	8,91	7,63	6,84
20	40,0	37,6	34,2	31,4	28,4	23,8	19,3	15,5	12,4	10,9	9,59	8,26	7,43
21	41,4	38,9	35,5	32,7	29,6	24,9	20,3	16,3	13,2	11,6	10,3	8,90	8,03
22	42,8	40,3	36,8	33,9	39,8	26,0	21,3	17,2	14,0	12,3	11,0	9,54	8,64
23	44,2	41,6	38,1	35,2	32,0	27,1	22,3	18,1	14,8	13,1	11,7	10,2	9,26
24	45,6	43,0	39,4	36,4	33,2	28,2	23,3	19,0	15,7	13,8	12,4	10,9	9,89
25	46,9	44,3	40,6	37,7	34,4	29,3	24,3	19,9	16,5	14,6	13,1	11,5	10,5
26	48,3	45,6	41,9	38,9	35,6	30,4	25,3	20,8	17,3	15,4	13,8	12,2	11,2
27	49,6	47,0	43,2	40,1	36,7	31,5	26,3	21,7	18,1	16,2	14,6	12,9	11,8
28	51,0	48,3	44,5	41,3	37,9	32,6	27,3	22,7	18,9	16,9	15,3	13,6	12,5
29	52,3	49,6	45,7	42,6	39,1	33,7	28,3	33,6	19,8	17,7	16,0	14,3	13,1
30	53,7	50,9	47,0	43,8	40,3	34,8	29,3	24,5	20,6	18,5	16,8	15,0	13,8
40	66,8	63,7	59,3	55,8	51,8	45,6	39,3	33,7	29,1	26,5	24,4	22,2	20,7
50	79,5	76,2	71,4	67,5	63,2	56,3	49,3	42,9	37,7	34,8	32,4	29,7	28,0
60	92,0	88,4	83,3	79,1	74,4	67,0	59,3	52,3	46,5	43,2	40,5	37,5	35,5
70	104,2	100,4	95,0	90,5	85,5	77,6	69,3	61,7	55,3	51,7	48,8	45,4	43,3
80	166,3	112,3	106,6	101,9	96,6	88,1	79,3	71,1	64,3	60,4	57,2	53,5	51,2
90	128,3	124,1	118,1	113,1	107,6	98,6	89,3	80,6	73,3	69,1	65,6	61,8	59,2
100	140,2	135,8	129,6	124,3	118,5	109,1	99,3	90,1	82,4	77,9	74,2	70,1	67,3

La estadistica descriptiva
permite ver lo que los datos nos dicen,
no lo que queremos escuchar.

John Tukey

ISBN: 978-9972-55-035-5