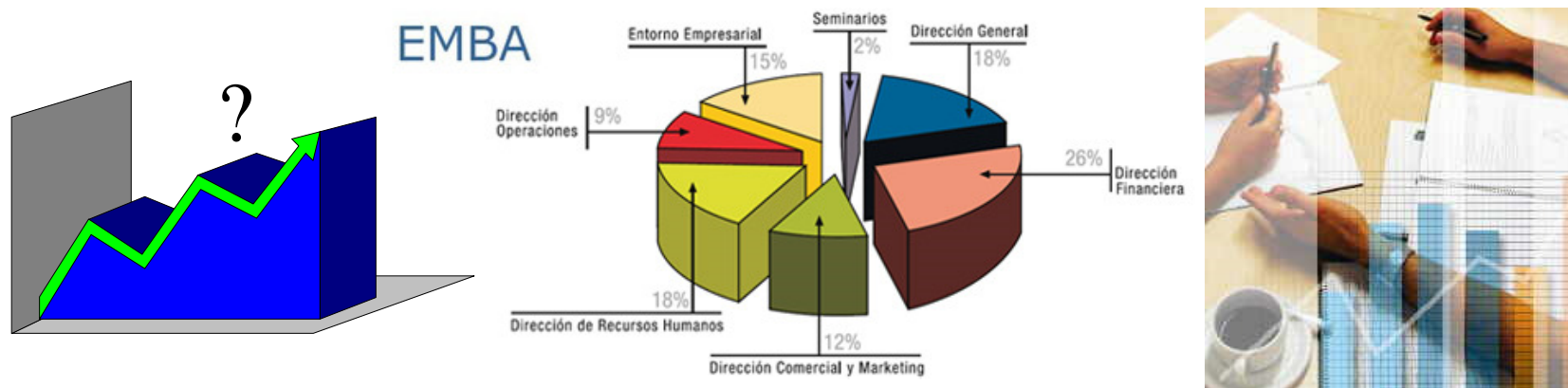


Titulaciones empresariales

PROFESOR: Gerardo Silva Serrano



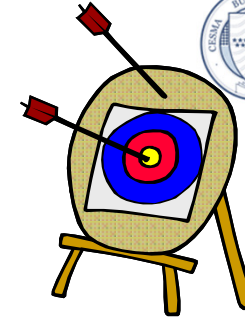
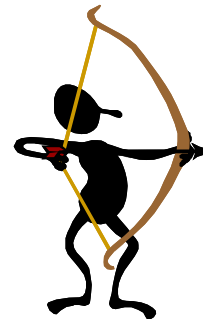
Mi historia profesional



Marketing Staff



Objetivos del curso

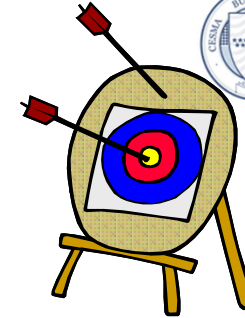
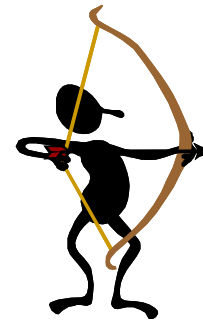


CESMA
BUSINESS SCHOOL

Al terminar el curso, el alumno deberá:

- Extraer una primaria pero valiosa información a partir de una gran cantidad de datos mediante el estudio de su representación gráfica.
- Manejar las medidas estadísticas unidimensionales básicas que le permitan evaluar los resultados estadísticos con el rigor necesario.
- Saber analizar la relación lineal de dos variables y, en su caso, predecir una en función de la otra estudiando la fiabilidad de la predicción realizada.

Objetivos del curso

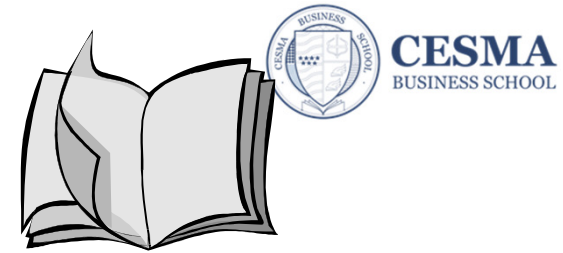


CESMA
BUSINESS SCHOOL

Al terminar el curso, el alumno deberá:

- Comprender y valorar la importancia de las muestras en el análisis de poblaciones.
- Conocer los distintos tipos de muestreo así como los tipos de errores que pueden cometerse y la posibilidad de aproximarnos a su medición.
- Utilizar las técnicas de muestreo aleatorio simple para inferir conclusiones sobre toda una población a partir de la información proporcionada por una parte de ella.

Contenido del curso



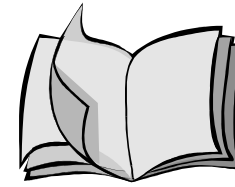
I. INTRODUCCIÓN Y CONCEPTOS BÁSICOS

- 1.1. Introducción.
- 1.2. Para qué sirve la estadística.
- 1.3. Conceptos básicos: población, muestra y variables.

2. PRESENTACIÓN DE LA INFORMACIÓN ESTADÍSTICA

- 2.1. Tipos de datos.
- 2.2. Tablas de frecuencias.
- 2.3. Representaciones gráficas.

Contenido del curso



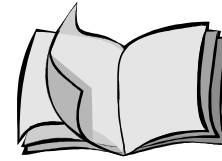
3. MEDIDAS RESUMEN DE LA INFORMACIÓN ESTADÍSTICA

- 3.1. Medidas de posición.
- 3.2. Medidas de dispersión.
- 3.3. Medidas de asimetría y apuntamiento.

4. ESTADÍSTICA BIDIMENSIONAL

- 4.1. Tabla de frecuencia conjunta.
- 4.2. Descripción numérica.
- 4.3. Representaciones gráficas.
- 4.4. relación entre dos variables: El modelo de regresión lineal.

Contenido del curso



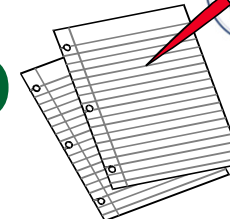
5. INTRODUCCIÓN A LA INFERENCIA ESTADÍSTICA

- 5.1. Población y muestra.
- 5.2. Herramientas de la inferencia: probabilidad y variables aleatorias.
- 5.3. Los errores de muestreo.

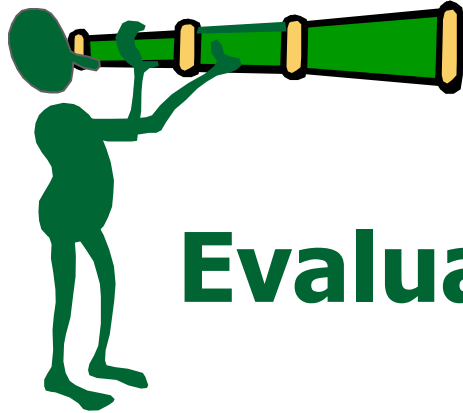
6. MUESTREO ALEATORIO SIMPLE

- 6.1. Parámetros poblacionales
- 6.2. estimadores puntuales.
- 6.3. Estimadores por intervalos de confianza.

Metodología del curso



- ✓ Exposiciones teóricas por parte del profesor y manejo de ejemplos prácticos de realidades empresariales.
- ✓ Realización por parte de los alumnos de ejercicios y resolución de casos de manera individual, donde se aplicara todo lo aprendido en las clases.



Evaluación de la asignatura

Distribución de la calificación:

Asistencia	20%
Participación, casos y ejercicios	20%
Examen Final	60%

I. INTRODUCCIÓN Y CONCEPTOS BÁSICOS



Introducción a la Estadística

La gestión diaria de las empresas genera una cantidad enorme de información en sus diferentes departamentos, que debe ser clasificada, analizada y luego utilizada para la toma de decisiones estratégicas y operativas de la dirección de la Dirección y de los diferentes departamentos que la componen.

Es aquí donde la gestión eficiente de los Sistemas de Información y el uso de la Estadística, juegan un papel fundamental en la gestión exitosa de la empresa.

No se puede gestionar lo que no se mide. Las mediciones son la clave. Lo que, no se mide no se puede controlar. Si no se puede controlar, no se puede gestionar. Si no se puede gestionar, no se puede mejorarlo.

La falta sistemática o ausencia estructural de estadísticas en las organizaciones impide una administración científica de las mismas.

Dirigir sólo en base a datos financieros del pasado, realizar predicciones basadas más en la intuición o en simples extrapolaciones, y tomar decisiones desconociendo las probabilidades de éxito u ocurrencia, son sólo algunos de los problemas o inconvenientes más comunes hallados en las empresas.

I. INTRODUCCIÓN Y CONCEPTOS BÁSICOS



Introducción a la estadística

Carecer de datos estadísticos en cuanto a lo que acontece tanto interna como externamente, impide decidir sobre bases racionales, y adoptar las medidas preventivas y correctivas con el suficiente tiempo para evitar daños, en muchos casos irreparables, para la organización.

Realizar mediciones y análisis estadísticos de datos en las empresas resulta hoy en día una labor mucho menos complicada que en el pasado, ya que, se dispone actualmente de ordenadores de alta velocidad y programas de software que hacen posible la realización rápida y eficiente de cualquier tipo de análisis.

Las empresas que no hagan uso de estas nuevas potencialidades y afronten debidamente éstas nuevas exigencias, no sólo perderán capacidad competitiva, sino que quedarán rezagadas ante los continuos cambios del entorno, poniendo en serio riesgo su propia continuidad.

En esta época de turbulencia, el empresario necesita predecir a tiempo los niveles de demanda de sus productos, necesita reconocer a tiempo los cambios de tendencia, debe no sólo saber en que se gastó, sino como se gastó en el tiempo y en qué conceptos.

I. INTRODUCCIÓN Y CONCEPTOS BÁSICOS



Introducción a la estadística

Para negociar, para tomar decisiones, para corregir problemas de calidad, para aumentar la productividad, para fijar precios, para mejorar el mantenimiento y disponibilidad de las máquinas e instalaciones, para mejorar la concesión y cobranza de los créditos se requiere sí o sí contar con datos estadísticos.

Toda decisión, todo análisis, todo presupuesto, está prácticamente en el aire si no se cuenta con datos estadísticos suficientes y fiables.

No sólo a nivel empresa, sino también a nivel país, los que más han avanzado han sido aquellos que hicieron de las estadísticas una herramienta fundamental.

Japón es un claro ejemplo, las técnicas estadísticas enseñadas por **Deming** fue lo que hizo que este país pasara de ser un fabricante de imitaciones baratas a líder internacional en productos de primera calidad.

Sin estadísticas una empresa carece de capacidad para reconocer que actividades o productos le generan utilidades, y cuales sólo pérdidas.

No contar con datos e interpretarlos correctamente es para los administradores gestionar al azar. Contar con datos, les permite ver lo que está aconteciendo y en consecuencia tomar las medidas más apropiadas.

Introducción a la estadística

Ejemplos de información que nos proporciona el análisis estadístico de datos para tomar decisiones:

De la Dirección de la empresa:

- ¿Cuántos son, como son, donde están, cuales son las cuotas de mercado, etc., de mis competidores?
- ¿Qué productos nuevos lanzar al mercado?
- ¿Qué nuevos mercados abrir y que mercados actuales potenciar?
- ¿Cuánto cuestan los nuevos lanzamientos y cómo y cuando recuperar las inversiones?

I. INTRODUCCIÓN Y CONCEPTOS BÁSICOS



Introducción a la estadística

Ejemplos de información que nos proporciona el análisis estadístico de datos para tomar decisiones:

De Marketing y Ventas:

- ¿Qué productos se venden más? ¿Cuáles generan más beneficios? ¿Cuáles son las tendencias decrecimiento?
- ¿Qué clientes generan los mayores beneficios?
- ¿Qué zonas o regiones son las que generan mayores ventas en unidades monetarias y volúmenes? (en total y por producto)
- En que etapa del ciclo de vida se encuentra cada uno de sus productos o servicios ofrecidos?
- ¿Cuál es el nivel de satisfacción de los clientes?

I. INTRODUCCIÓN Y CONCEPTOS BÁSICOS



Introducción a la estadística

Ejemplos de información que nos proporciona el análisis estadístico de datos para tomar decisiones:

De Producción/Operaciones:

- ¿Cuáles son las reparaciones que más se han producido en el último trimestre?
- ¿En que día de trabajo de cada mes se logra llegar al punto de equilibrio?
- ¿Qué tipo de reparaciones han generado mayores gastos?
- ¿Cuáles son las capacidades de los diferentes procesos en materia de costes, productividad y calidad?
- ¿Cuál es su nivel en sigma de cada una de las actividades?

Introducción a la estadística

Ejemplos de información que nos proporciona el análisis estadístico de datos para tomar decisiones:

De Finanzas y Contabilidad:

- ¿Cuál es el gasto promedio en dietas y desplazamientos de los ejecutivos de ventas de la empresa?
- ¿A cuánto ascienden los ingresos por ventas antes de impuestos?
- ¿Cuál es el presupuesto asignado para I&D para el siguiente año y cuál ha sido su incremento con respecto al año anterior?

Introducción a la estadística

Ejemplos de información que nos proporciona el análisis estadístico de datos para tomar decisiones:

De Recursos Humanos:

- ¿Cuál es la media de horas extras trabajadas por el personal de producción en el último mes?
- ¿Cuántas horas de capacitación recibió el personal de la empresa?
- ¿Cuál es la media de rotación de empleados?
- ¿Quiénes son los trabajadores con más ausencias laborales?

Introducción a la estadística

Sin esta información las empresas no pueden:

- Adoptar a tiempo las medidas correctivas;
- Confeccionar un presupuesto viable y efectivo;
- Administrar eficazmente su flujo de fondos;
- Evitar excesos de stock y obsolescencia de inventarios;
- Gestionar la mejora de los diversos procesos;
- Saber cuando está mejorando la productividad;
- Negociar un incremento de precios;
- Detectar la causa de un problema y solucionarlo.

Introducción a la estadística

¿Para qué la estadística?

Además de lo dicho anteriormente para:

- El control de calidad.
- El nivel de averías y sus frecuencias.
- Los tiempos para cambios o preparación de herramientas.
- Los niveles de productividad de distintos procesos, actividades y productos.
- Los costes correspondientes a distintos tipos de conceptos y actividades.
- La gestión de créditos y cobranzas.

Introducción a la estadística

¿Para qué la estadística?

- El seguimiento del flujo de fondos.
- Los niveles de satisfacción de los clientes y usuarios.
- Los tipos de accidentes y sus frecuencias.
- El análisis paretiano de defectos, costes, rentabilidades, ventas.
- Ventas por clientes, vendedores, zonas y productos.
- Predicciones de ventas por zonas, productos, servicios o sucursales.
- Capacidad de los procesos en cuanto a generación de niveles de costes, calidad y productividad.

Introducción a la estadística

¿Para qué la estadística?

- Evolución de los distintos ratios económicos – financieros y patrimoniales a lo largo del tiempo.
- Estudios e investigación de mercado.
- Tiempos de máquinas y personas por actividad.
- Cantidad y representación porcentual de distintos problemas y sus efectos económicos en la organización.
- Tasa de polivalencia del personal.
- Productos más demandados, a nivel global, por zona y por canal de comercialización.

Introducción a la estadística

¿Para qué la estadística?

- Tiempos totales de ciclos productivos.
- Tiempos de respuestas.
- Gestión de inventarios.
- Cumplimiento de aprovisionamiento por parte de los proveedores.
- Predicción de ventas por canales de comercialización.
- Proyectos de inversión.
- Probabilidades para la construcción del "Árbol para la Toma de Decisiones".

Introducción a la estadística

¿Para qué la estadística?

- Porcentajes de actividades generadoras de valor agregado para los clientes finales, de valor agregado para la empresa y carentes de valor agregado.
- Tiempos promedios, máximos y mínimos de reparaciones por tipo de averías.
- Cálculos de costes y en especial para el Costeo Basado en Actividades.
- Para los cálculos de productividades.
- Coeficientes de correlación.
- Estadística del personal (directivos y empleados).

Introducción a la estadística

¿Por qué se aplica tan poco la estadística?

- Por la cultura de los empresarios, pero en mayor medida a la falta de preparación de los profesionales, en materia estadística.
- Lo antes descrito es menos frecuente en los países anglosajones, los cuales tienen una fuerte cultura e inclinación por las estadísticas y las probabilidades.
- Otro tanto se da en Japón o Corea, países que dan a la educación de las estadísticas y matemáticas una fuerte preponderancia en sus planes de estudios y luego en la aplicación práctica.

Introducción a la estadística

Qué hacer para aplicar la estadística?

- Concientizar, para luego pasar a capacitar y luego implementar.
- Diagnosticar para saber que datos necesita la organización es un paso fundamental, pues a partir de allí se diseñarán los software más apropiados a las actividades, procesos y requerimientos específicos de cada empresa.
- Si bien la intuición nunca dejará de perder importancia, el tener el respaldo de datos confiables permitirá al directivo poder adoptar decisiones sobre una base más apropiada.

Conceptos básicos

- **La estadística** se ocupa de los métodos y procedimientos para recoger, resumir y analizar los datos, así como de realizar inferencias a partir de ellos, con la finalidad de ayudar a la toma de decisiones y en caso formular predicciones.
- Una parte del trabajo es por tanto **descriptiva** y otra parte es **inferencial**.
- **La estadística descriptiva** trata de describir, analizar y representar un conjunto de datos, utilizando métodos numéricos y gráficos que resumen y presentan la información contenida en ellos.

Conceptos básicos

- Con la **estadística inferencial**, a partir de los datos muestrales y apoyándose en el cálculo de probabilidades, se efectuarán estimaciones, decisiones, predicciones u otras generalizaciones sobre un conjunto mayor de datos.

Etapas de un estudio estadístico. Ejemplo: Calculo del IPC

1. Recogida de datos: Para calcular el Índice de Precios al Consumo (**IPC**), el Instituto Nacional de Estadística (**INE**) mensualmente, obtiene de la encuesta continua de presupuesto familiares (ECPF), cuánto y en qué gastan su dinero los hogares españoles. Esta medición es trimestral.

Incluye 484 bienes. Alimentación, transporte, hostelería y vivienda son las principales partidas del gasto familiar.

Conceptos básicos

Etapas de un estudio estadístico. Ejemplo: Calculo del IPC

Cabe destacar que la rúbrica "vivienda" no incluye los precios de compra ya que esto se considera una inversión, no un consumo, sólo los de alquiler y otros como suministros (gas, agua, electricidad) y reparaciones.

El INE pasa mensualmente una encuesta a 141 establecimientos comerciales de alimentación y 97 para el resto de los bienes, controlando más de 180.000 precios. En esta encuesta se procura establecer pormenorizadamente las características de los artículos pero teniendo en cuenta las peculiaridades regionales.

El IPC es un índice "base 100", es decir, tomando como referencia un dato inicial igual a 100 se van expresando el resto de la serie temporal. El año base del actual índice es 2001, esto significa que el índice publicado, en enero de 2004, que es de 107,4, indica que la cesta de la compra costaba ese año 7,4% más que en 2001.

Conceptos básicos

Etapas de un estudio estadístico. Ejemplo: Calculo del IPC

2. Estimación: El IPC se usa para estimar la tasa de inflación subyacente en España. Lo más interesante no es el valor del IPC en sí; es conocer su variación, es decir, si se ha encarecido el coste de la vida y cuánto. Para ello, se utilizan la tasa de variación mensual (sobre el mes anterior) y, especialmente, la tasa de variación sobre el mismo mes del año anterior.

La primera viene muy influida por factores estacionales. En enero de 2004, el IPC se redujo ocho décimas respecto a diciembre (pasó de 108,2 a 107,4) pero esto se ha debido al efecto de las rebajas, por lo que el dato no es significativo.

Por el contrario, la variación sobre el mismo mes del año anterior compara datos homogéneos evitando así estas distorsiones estacionales. En enero, el dato fue del 2,3%, tres décimas menos que en enero del año anterior, por lo que puede afirmarse que la tasa de inflación en España descendió en 2004.

Conceptos básicos

Etapas de un estudio estadístico. Ejemplo: Cálculo del IPC

3. Predicción: Observando la serie mensual del IPC (por ejemplo, desde 2001 hasta 2012), se predice la tasa de inflación en futuros meses (año 2013).

4. Inferencia: El gobierno quiere saber (contrastar) si la tasa de inflación anual es menos de, por ejemplo, el 3%, para ver si se cumplen los requisitos europeos.

5. Ayuda en la toma de decisiones: Basado en la inflación predicha, se debe decidir si cambiar o no las tasas de interés del Banco de España.

Conceptos básicos

Población o universo (N): Conjunto de elementos (personas, animales o cosas) de las que tratamos de obtener cierta información. Puede ser finita (<100.000 elementos). Infinita (>100.000 elementos).

Muestra (n): Número de elementos de la población que se seleccionan para el estudio.

Variable: Fenómeno o característica de la población que se quiere estudiar. Las variables a estudiar pueden ser **discretas o continuas**.

Variable continua: Es aquella que toma cualquier valor dentro de un rango determinado. **Ejemplos**, La estatura de una persona, puede ser 1,70 mts. ó 1,75 mts., pero en potencia al menos podría tomar cualquier valor intermedio como 1,73 mts.

El peso de una persona puede ser 80 kilos o 75,2 kilos.

Conceptos básicos

Un atributo esencial de una variable continua es que, a diferencia de lo que ocurre con una variable discreta, nunca se la puede medir exactamente. Con una variable continua debe haber inevitablemente un error de medida.

Variable discreta: Aquella que solo toma valores enteros y no se puede fraccionar. Ejemplos: El número de personas que acuden a un estadio, el número de caras que tiene un dado, el número de hijos que puede tener una mujer.

Variable discreta dicotómica: Aquella que solo toma dos valores. Ejemplo: resultados que se obtienen al lanzar una moneda. Cara o cruz; personas que conocen o no un producto.

Estadístico: Es una medida resumida de una característica de la muestra.

Parámetro: Es una medida resumida de una característica de la población.

I. INTRODUCCIÓN Y CONCEPTOS BÁSICOS



Conceptos básicos

Ejemplos que ilustran estos conceptos: Un constructor de viviendas, quiere saber el grado de satisfacción de sus clientes 6 meses después de haberla entregado a sus clientes.

Población o universo (N): Todas las persona que han adquirido una vivienda de la constructora. $N = 15.550$ en total.

Muestra (n): 300 compradores seleccionados al azar para el estudio, de los diferentes planes construidos.

Variable: Grado de satisfacción global, Grado de satisfacción con los materiales de construcción, grado de satisfacción con los acabados, etc.

Estadísticos: Calificación promedio obtenida de la muestra para las tres variables en una escala de calificación de 1 a 10. Las desviaciones típicas de estas variables.

Parámetro: Las calificaciones promedio y sus correspondientes desviaciones típicas calculadas para toda la población.

Conceptos básicos

Otro ejemplo: Probar la aceptación de un nuevo chocolate, entre amas de casa de Madrid, captadas por detección en centros comerciales:

Población: Todas las amas de casa de 21 a 50 años de edad, de todas las clases sociales residentes en Madrid.

Muestra: 400 amas de casa entrevistadas a la entrada de centros comerciales.

Variables: Sabor del chocolate, interés por el chocolate, gasto en consumo de chocolate , etc.

Estadísticos: Calificación promedio del sabor, % de personas que definitivamente lo comprarían, el gasto medio en chocolates, etc., obtenidos de las encuestas a las 400 amas de casa.

Parámetros: Estos mismos valores referidos a todas las amas de casa de Madrid (datos inferidos).

Conceptos básicos

Operadores suma producto

En estadística se requiere la suma de grandes cantidades de datos y es pertinente tener una notación simplificada para indicar la suma de estos datos. Así, si una variable se puede denotar por X , entonces las observaciones sucesivas de esta variable se escriben:

$$X_1 + X_2 + X_3 + \dots + X_n$$

En general, la i -ésima observación se escribe X_i ; $i=1, \dots, n$.

La letra griega sigma mayúscula (Σ) se emplea para indicar la suma de estas n observaciones.

$$\sum_{i=1}^n X_i = X_1 + X_2 + X_3 + \dots + X_n$$

Conceptos básicos

Operadores suma producto

La notación: $\sum_{i=1}^n X_i$ se lee la sumatoria de X sub-i donde i va desde 1 hasta n.

La letra debajo del operador Σ se llama índice de la suma; en la expresión:

$$\sum_{i=1}^n X_i \quad \text{Note que el índice de la suma es } i.$$

Las sumatorias se pueden representar bajo dos tipos de notaciones:

Notación suma abierta:

Esta notación va de una representación de sumatoria a cada uno de los elementos que la componen, por ejemplo:

Conceptos básicos

Operadores suma producto

Notación suma abierta:

$$\sum_{i=1}^3 X_i = X_1 + X_2 + X_3$$

Notación suma pertinente:

Esta notación es al contrario de la suma abierta, va de la representación de cada uno de los elementos de una sumatoria a su representación matemática resumida, por ejemplo:

$$X_1 + X_2 + X_3 = \sum_{i=1}^3 X_i$$

Conceptos básicos

Operadores suma producto

Ejemplo 1:

Si $X_1=3$ $X_2=9$ $X_3=11$ Encontrar: $\sum_{i=1}^n X_i$

Solución:

$$\sum_{i=1}^3 X_i = X_1 + X_2 + X_3 = 3 + 9 + 11 = 23$$

Ejemplo 2:

Si $X_1=1$ $X_2=2$ $X_3=-1$ Encontrar: $\sum_{i=1}^n X_i^2$

$$\sum_{i=1}^3 X_i^2 = X_1^2 + X_2^2 + X_3^2 = (1)^2 + (2)^2 + (-1)^2 = 1 + 4 + 1 = 6$$

Conceptos básicos

Operadores suma producto

Ejemplo 3:

Si $X_1=9$ $X_2=6$ $X_3=5$ $X_4=8$ $X_5=12$ Encontrar:

$$\begin{aligned}\sum_{i=1}^n (X_i - 5)^2 &= (X_1 - 5)^2 + (X_2 - 5)^2 + (X_3 - 5)^2 + (X_4 - 5)^2 + (X_5 - 5)^2 \\ &= (9 - 5)^2 + (6 - 5)^2 + (5 - 5)^2 + (8 - 5)^2 + (12 - 5)^2 \\ &= (4)^2 + (1)^2 + (0)^2 + (3)^2 + (7)^2 \\ &= 16 + 1 + 0 + 9 + 49 = 75\end{aligned}$$

Ahora bien, cuando se trabajan estas expresiones en forma algebraica se necesita identificar variables y constantes, así si X es una variable, a y b son dos constantes, podemos probar que:

Conceptos básicos

Operadores suma producto

$$\sum_{i=1}^n (aX_i + b) = a \sum_{i=1}^n X_i + nb$$

De lo anterior es evidente que la suma de una expresión que es la suma de dos ó más términos es igual a la suma de las sumas de los términos por separado.

Por ejemplo:

$$\sum_{i=1}^n (X_i + Y_i + Z_i) = \sum_{i=1}^n X_i + \sum_{i=1}^n Y_i + \sum_{i=1}^n Z_i$$

La suma de una constante multiplicada por una variable es lo mismo que la constante multiplicada por la suma de la variable, esto es:

$$\sum_{i=1}^n ax_i = a \sum_{i=1}^n x_i$$

I. INTRODUCCIÓN Y CONCEPTOS BÁSICOS



Conceptos básicos

Operadores suma producto

La suma de una constante, es igual a n veces la constante, esto es:

$$\sum_{i=1}^n a = na$$

Ejercicio Práctico 1.1

Describe la aplicación de la estadística y defina cuales serían: La variable a estimar, la población, la muestra, dos parámetros, dos estadísticos y decir las decisiones que tomarías con base en la información obtenida para las siguientes empresas:

1. El hospital Ramón y Cajal
2. Los procesos de envasado de Coca Cola
3. El grupo Leche Pascual
4. SEAT
5. La agencia de Publicidad Ogilvy

Conceptos básicos

Ejercicio Práctico 1.2

Escriba en forma abierta cada una de las sumas que se indican:

a. $\sum_{i=1}^6 X_i$

b. $\sum_{i=6}^8 X_i^2$

c. $\sum_{i=1}^3 (X_i - 2)^2$

d. $\sum_{i=1}^5 CX_i$

e. $\sum_{j=3}^7 (Y_j + 7)$

Conceptos básicos

Ejercicio Práctico 1.3

Si $X_1 = X_2 = 3$; $X_3 = -2$; $X_4 = 0$; $X_5 = 4$ encontrar:

a. $\sum_{i=1}^5 (X_i - 1)$

b. $\sum_{i=1}^3 (X_i - 7) X_i$

c. $\sum_{i=1}^5 (X_i - 2)^2$

d. $\sum_{i=1}^5 CX_i$

e. $\sum_{i=3}^5 (X_i + 7)$

Medición de variables

Objetivo de la Estadística Descriptiva:

Aprender cómo resumir las características más importantes de una muestra de datos.

Analizar la población es muy complicado y costoso, por ello, se selecciona una muestra.

Tipos de datos

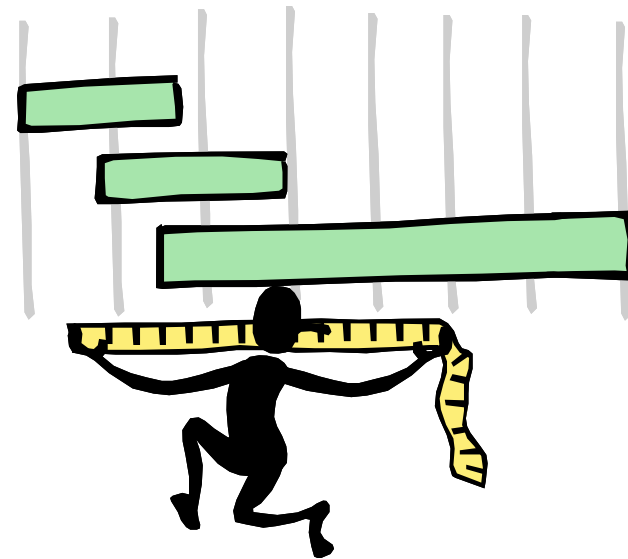
Es posible clasificar distintos tipos de variables. En primer lugar, se distinguen entre:

- ❑ Variables de naturaleza categórica (variable cualitativa)
- ❑ Variables de naturaleza numérica (variable cuantitativa)

Medición de variables

CONCEPTO DE MEDICIÓN

Significa asignar números u otros símbolos a las características de los objetos, de acuerdo a ciertas reglas especificadas con anterioridad. No medimos a los consumidores sino sólo sus percepciones, actitudes, preferencias u otras características relevantes.



Medición de variables

ESCALAS DE MEDICIÓN

CONCEPTO:

Las escalas son una extensión de la medición. Comprenden la creación de un continuo sobre el que se localizan los objetos medidos.

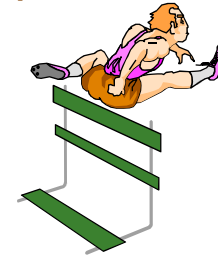
TIPOS DE ESCALAS DE MEDICIÓN

- | | | |
|---------------------------|---|--------------------|
| 1. Escala nominal | } | Escala no métricas |
| 2. Escala ordinal | | |
| 3. Escala de intervalos y | } | Escala métricas |
| 4. Escala de razón. | | |

Tema 2. Medición de variables y distribuciones de frecuencias

ESCALA NOMINAL

Escala cuyos números sirven sólo como etiquetas o señales para identificar y clasificar los objetos, con una correspondencia estricta de uno a uno entre los números y los objetos.



¿PARA QUE SE UTILIZA?

Identificar a los entrevistados, marcas, atributos, tiendas, territorios de ventas, mujeres que trabajan o que no trabajan, etc.

Los números solo sirven como etiquetas para las clases o categorías. Cumple con la primera característica del sistema numérico.

I. INTRODUCCIÓN Y CONCEPTOS BÁSICOS



Medición de variables

Escala nominal. Ejemplos:

¿Cuál es su estado Civil?

- ☐ Soltero (1)
- ☐ Casado (2)
- ☐ Divorciado (3)
- ☐ Viudo (4)

¿Que pasta de dientes utiliza normalmente?

- ☐ Signal (1)
- ☐ Colgate (2)
- ☐ Licor del polo (3)
- ☐ Otra marca (4)

Medición de variables

Escala nominal. Operaciones estadísticas permitidas:

- Conteos de frecuencias: Absolutas y relativas
- Moda
- Prueba binomial
- Prueba de chi cuadrada

I. INTRODUCCIÓN Y CONCEPTOS BÁSICOS

Medición de variables

Escala ordinal

Es una escala de clasificación, en la que los números se asignan a los objetos para indicar el grado relativo en el que los objetos poseen cierta **característica**. Comprende la característica del sistema numérico del orden de los numerales.

Mide la relación ordenada de los objetos (orden de preferencia), pero no la magnitud.



Medición de variables

Escala ordinal. Ejemplo

Clasificar por orden de preferencia los supermercados que te presentamos en la siguiente tarjeta. Poner 1 al que más prefiera y un 5 al que menos prefiera:

- ☐ ALCAMPO
- ☐ CAPRABO
- ☐ DÍA
- ☐ HIPERCOR
- ☐ CARREFOUR

I. INTRODUCCIÓN Y CONCEPTOS BÁSICOS



Medición de variables

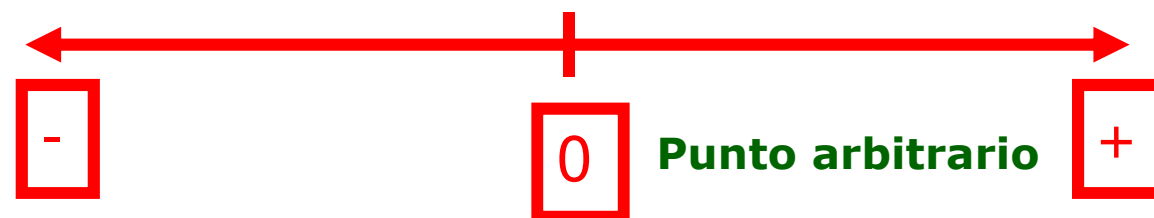
Escala ordinal. Operaciones estadísticas permitidas

- Las mismas que para las escalas nominales más,
- Las estadísticas con base en los centiles (deciles, cuartiles, etc.),
- La mediana y
- La correlación de rango-orden.

Medición de variables

Escala de intervalo

Es una escala en la cuál los números se utilizan para clasificar los objetos, de manera que las distancias numéricamente iguales en la escala representan distancias iguales en la característica que mide. Es una escala métrica. Cumple con la tercera característica del sistema numérico.



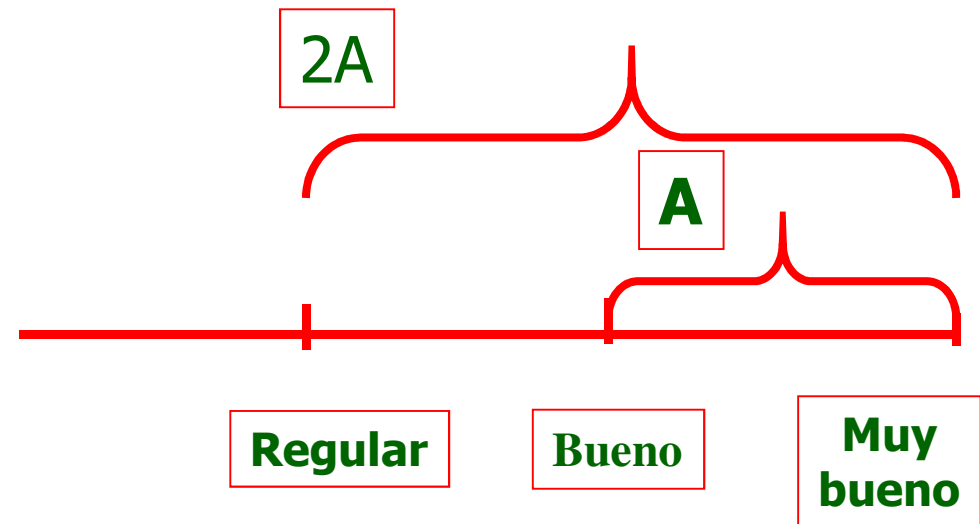
I. INTRODUCCIÓN Y CONCEPTOS BÁSICOS

Medición de variables

Escala de intervalo. Ejemplo:

En una escala de 5 puntos donde 1 es muy malo y cinco es muy bueno que calificación daría Usted, a Caprabo como supermercado? (MOSTRAR TARJETA)

- ☐ MUY MALO (1)
- ☐ MALO (2)
- ☐ REGULAR (3)
- ☐ BUENO (4)
- ☐ MUY BUENO (5)



Medición de variables

Escala de intervalo. Operaciones estadísticas permitidas

- Las mismas que se aplican a escalas nominales y ordinales, además:
- La media aritmética.
- El rango.
- La varianza y la desviación típica.
- Las correlaciones producto momento.

I. INTRODUCCIÓN Y CONCEPTOS BÁSICOS



Medición de variables

Escala de razón

Es la escala más alta de medición. Posee todas las características de las escalas nominal, ordinal y de intervalos y, además un punto cero absoluto.

EJEMPLO:

Que cantidad de dinero se gastó usted, la última semana en:

ALCAMPO	_____	Euros
CAPRABO	_____	Euros
DIA	_____	Euros
HIPERCOR	_____	Euros
CARREFOUR	_____	Euros.

Medición de variables

Escala de razón. Operaciones estadísticas permitidas

- Todas las técnicas estadísticas pueden aplicarse a los datos de razón. Estas incluyen estadísticas especializadas como la media geométrica, la media armónica y el coeficiente de variación.

Medición de variables

Ejercicio Práctico 1.4

Una compañía aérea te solicita que realices un estudio para conocer, los hábitos de viajes de sus clientes, con el propósito de crear un club de clientes y así poderles otorgar premios por su fidelidad.

Principales objetivos de la investigación

1. Conocer los hábitos de viaje: ¿A dónde viajan? ¿Cuándo viajan? ¿Con qué frecuencia viajan? ¿Con quién viajan? ¿Cuáles son los motivos por los que viajan?
2. Gasto aproximado en viajes en euros.
3. Compañías aéreas por las que viajan habitualmente.
4. Características demográficas, socioeconómicas y de estilos de vida de los entrevistados.

Elaborar las preguntas con sus respectivas escalas de medición que sirvan para responder a estos objetivos de investigación.

Tablas de frecuencias

Cómo resumir una muestra de datos

Dada una muestra de datos, se quiere extraer la información pertinente. Mostrando a alguien la muestra entera, no van a ser capaces de ver los rasgos importantes. Para resumir los datos, en primer lugar, es conveniente hacer una tabla de frecuencias.

La Tabla de frecuencias

Para una variable con k categorías, $x_1, \dots, x_k$ sea n_k el número de observaciones en cada clase o categoría, entonces se construye la tabla de frecuencias:

2. PRESENTACIÓN DE LA INFORMACIÓN ESTADÍSTICA

Tablas de frecuencias

La tabla de frecuencias

Categoría	Frecuencia absoluta	Frecuencia relativa
X_1	n_1	$f_1 = n_1/n$
X_2	n_2	$f_2 = n_2/n$
...	...	...
X_k	n_k	$f_k = n_k/n$

Donde $n = n_1 + n_2 + \dots + n_k$ es el tamaño de la muestra.

2. PRESENTACIÓN DE LA INFORMACIÓN ESTADÍSTICA



Tablas de frecuencias

Ejemplo de distribución de frecuencias:

El número de equipos informáticos (X) de 30 sucursales de una institución financiera se distribuye de la siguiente forma:

Nº de equipos	Nº de sucursales
10	9
20	6
30	9
40	3
50	3

2. PRESENTACIÓN DE LA INFORMACIÓN ESTADÍSTICA

Tablas de frecuencias

Ejemplo de distribución de frecuencias:

La distribución de frecuencias de esta variable es:

x_1	n_1	f_1	N_1	F_1
10	9	$9/30 = 0.3$	9	0.3
20	6	$6/30 = 0.2$	$9+6 = 15$	$0.3+0.2 = 0.5$
30	9	$9/30 = 0.3$	$15+9 = 24$	$0.5+0.3 = 0.8$
40	3	$3/30 = 0.1$	$24+3 = 27$	$0.8+0.1 = 0.9$
50	3	$3/30 = 0.1$	$27+3 = 30$	$0.9+0.1 = 1$

Donde:

f_1 es la frecuencia relativa, N_i es la frecuencia absoluta acumulada y F_i es la frecuencia relativa acumulada de la clase i -ésima

Tablas de frecuencias

La Tabla de frecuencias para datos de tipo continuo

Para una variable continua X que toma valores en un rango continuo, es necesario agrupar la variable en intervalos para analizar mejor los datos.

Al construir la tabla de distribución de frecuencias debemos prestar atención a 3 aspectos:

1. Seleccionar un número adecuado de clases o categorías para la tabla.
2. Obtener el intervalo de clase o amplitud apropiado para cada clase.
3. Establecer los límites de cada clase para evitar traslapes.

2. PRESENTACIÓN DE LA INFORMACIÓN ESTADÍSTICA



Tablas de frecuencias

La Tabla de frecuencias para datos de tipo continuo

1. Selección del número de clases:

El número de clases que se utilizan depende de la cantidad de datos. Es decir, cuanto mayor sea el número de observaciones, mayor será el número de clases que se requieren. Por regla general una distribución de frecuencias debe tener al menos 5 clases y no más de 15.

2. Obtención de los intervalos de clase:

- Ordenamos los datos, $X_1, X_2, \dots, X_n$, de menor a mayor.
- Se calcula el tamaño del intervalo

$$\text{Tamaño del intervalo} = \frac{\text{Rango}}{\text{Número de clases que se desean}}$$

2. PRESENTACIÓN DE LA INFORMACIÓN ESTADÍSTICA



Tablas de frecuencias

La Tabla de frecuencias para datos de tipo continuo

2. Obtención de los intervalos de clase:

- **Rango o recorrido de la variable:** $\max(x_i) - \min(x_i)$; $i=1, \dots, n$ (n es el tamaño de la muestra)
- Definimos el primer intervalo como $[l_1, l_2)$, siendo l_1 el límite inferior y l_2 el límite superior.
- El segundo intervalo estaría definido por $[l_2, l_3)$, y así sucesivamente hasta llegar al último intervalo, $[l_{k-1}, l_k]$.

2. PRESENTACIÓN DE LA INFORMACIÓN ESTADÍSTICA



Tablas de frecuencias

La Tabla de frecuencias para datos de tipo continuo

Definimos:

Amplitud del primer intervalo: $a_1 = l_2 - l_1$ (ó longitud del intervalo)

Marca de clase: $m_1 = (l_2 - l_1)/2$ (es decir, el punto medio del intervalo)
De forma similar se define para el resto de los intervalos

Frecuencia absoluta de cada intervalo intervalo: $n_1, \dots, n_k$ contando las observaciones que caen en cada uno de ellos

Frecuencia relativa de cada intervalo: $f_k = n_k/N$

2. PRESENTACIÓN DE LA INFORMACIÓN ESTADÍSTICA

Tablas de frecuencias

La Tabla de frecuencias para datos de tipo continuo

Ejemplo:

Los gastos de consumo de electricidad de una muestra aleatoria de 50 apartamentos del centro de Madrid de 3 habitaciones son:

Datos originales de la factura por servicios en euros

96	171	202	178	147	102	153	197	127	82
157	185	90	116	172	111	148	213	130	165
141	149	206	175	123	128	144	168	109	167
95	163	150	154	130	143	187	166	139	149
108	119	183	151	114	135	191	137	129	158

2. PRESENTACIÓN DE LA INFORMACIÓN ESTADÍSTICA



Tablas de frecuencias

La Tabla de frecuencias para datos de tipo continuo

1. Selección del número de clases:

Decidimos agrupar los datos en 5 categorías o clases

2. Obtención de los intervalos de clase:

- **Rango del consumo:** $\max(x_i) - \min(x_i)$; $i=1, \dots, n$ (n es el tamaño de la muestra)

$$\text{Rango} = 213 - 82 = 131$$

- Se calcula el tamaño del intervalo

$$\text{Tamaño del intervalo} = \frac{131}{5} = 26,2 \text{ se aproxima a } 30$$

2. PRESENTACIÓN DE LA INFORMACIÓN ESTADÍSTICA



Tablas de frecuencias

La Tabla de frecuencias para datos de tipo continuo

Definimos:

La Amplitud de cada intervalo (a), las marca de clase (m_k) de cada intervalo y las frecuencia absolutas de la variable en cada intervalo (n_k). Definimos las frecuencias relativas (f_k), absolutas acumuladas (N_k) y las frecuencias relativas acumuladas (F_k) para cada intervalo.

La distribución de frecuencias de esta variable es:

$[l_k, l_{k+1})$	m_k	n_k	f_k	N_k	F_k
82 a < 112	97	8	$8/50 = 0.16$	8	0.16
112 a < 142	127	15	$15/50 = 0.30$	$8 + 15 = 23$	$0.16 + 0.30 = 0.46$
142 a < 172	157	18	$18/50 = 0.36$	$23 + 18 = 41$	$0.46 + 0.36 = 0.82$
172 a < 202	187	6	$6/50 = 0.12$	$41 + 6 = 47$	$0.82 + 0.12 = 0.94$
202 a < 232	217	3	$3/50 = 0.06$	$47 + 3 = 50$	$0.94 + 0.06 = 1$

Donde N_k es la frecuencia absoluta acumulada y F_k es la frecuencia relativa acumulada del intervalo k -ésimo (en este ejemplo tenemos 5 clases ó intervalos)

2. PRESENTACIÓN DE LA INFORMACIÓN ESTADÍSTICA



Tablas de frecuencias Ejercicio práctico 2.1

En diversas agencias de publicidad de España se selecciona una muestra aleatoria de 50 vicepresidentes ejecutivos y se obtienen los sueldos anuales de estos funcionarios. Los salarios varían de 52,000 a 137,000 euros al año. Establecer los límites de clase para una distribución de frecuencias

- a) Si se desean 5 intervalos de clase.
- b) Si se desean 6 intervalos de clase.

Ejercicio Práctico 2.2

Los siguientes datos son los valores en libros (es decir el capital social, dividido entre el número de acciones en circulación de una muestra aleatoria de 50 acciones de la bolsa de valores de Madrid

7	9	8	6	12	6	9	15	9	16
8	5	14	8	7	6	10	8	11	4
10	6	16	5	10	12	7	10	15	7
10	8	8	10	18	8	10	11	7	10
7	8	15	23	13	9	8	9	9	13

Tema 2. Medición de variables y distribuciones de frecuencias

Ejercicio 2.3

Se te pide:

- Agrupar los datos en 5 clases.
- Agrupar los datos en 7 clases.
- ¿Cuál de estas dos agrupaciones es mejor para el conjunto de datos? Explícalo.
- Calcular la marca de clase para los dos agrupaciones de datos
- Calcular las frecuencias absolutas, relativas, absolutas acumuladas y relativas acumuladas para las dos agrupaciones de datos.

Ejercicio 2.4

Los datos que aparecen ordenados en la siguiente tabla corresponden a la vida útil en horas de una muestra de 40 focos de 100 wats fabricados por Luminotec:

Tema 2. Medición de variables y distribuciones de frecuencias

Ejercicio 2.5

684	697	720	773	821
831	835	848	852	852
859	860	868	870	876
893	899	905	909	911
922	924	926	926	938
939	943	946	954	971
972	977	984	1005	1014
1016	1041	1052	1080	1093

Se te pide:

- Agrupar los datos en 5 clases.
- Calcular la marca de clase para esta agrupación de datos.
- Calcular las frecuencias absolutas, relativas, absolutas acumuladas y relativas acumuladas.

Representaciones gráficas de los datos

La gente prefiere ver imágenes en lugar de tablas de números, y entonces es útil usar gráficos para mostrar los datos. El gráfico más importante para variables cualitativas y para datos de tipo discreto es el diagrama de barras.

Diagrama de barras

Si la variable cualitativa es ordinal, es natural ordenar las barras en el orden de las categorías. Si es nominal, el orden de las barras no importa.

Cada barra representa la frecuencia absoluta de cada clase.

También es posible construir un diagrama de barras usando frecuencias relativas en lugar de frecuencias absolutas.

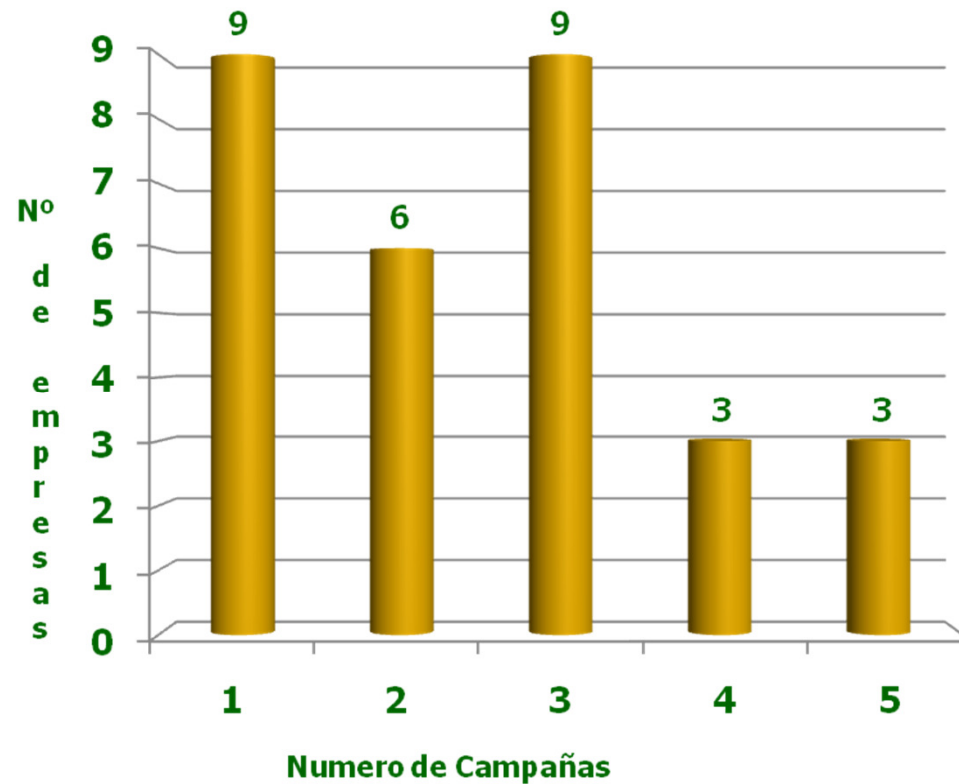
2. PRESENTACIÓN DE LA INFORMACIÓN ESTADÍSTICA

Representación gráfica de los datos

Ejemplo representación grafica diagrama de barras:

El número campañas publicitarias (X) de 30 empresas se distribuyen según la tabla:

Nº de campañas	Nº empresas
1	9
2	6
3	9
4	3
5	3



2. PRESENTACIÓN DE LA INFORMACIÓN ESTADÍSTICA

Representación gráfica de los datos

Otro ejemplo representación gráfica de diagrama de barras

Facturación en España de los institutos de investigación de mercados, según tipo de estudio.



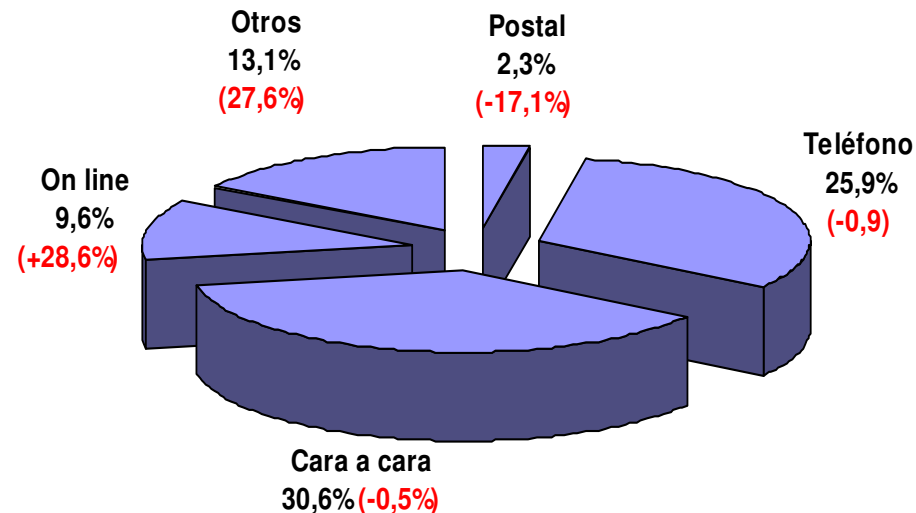
2. PRESENTACIÓN DE LA INFORMACIÓN ESTADÍSTICA

Representación gráfica de los datos

Diagramas de sectores o de Pastel

Es otra manera de representación gráfica para variables cualitativas. Se divide un círculo en sectores donde el área de un sector es proporcional al número de datos en una categoría.

% de facturación de los Institutos de Investigación en España
desglosado por tipo de estudios cuantitativos



2. PRESENTACIÓN DE LA INFORMACIÓN ESTADÍSTICA



Representación gráfica de los datos

Representación gráfica de variables cuantitativas

Los Histogramas:

Son gráficos de barras verticales en los que se construyen las barras rectangulares en los límites de cada clase.

Cuando se traza el histograma la variable a estudiar se grafica en el eje horizontal; el eje vertical representa el número o porcentaje de observaciones de cada intervalo de clase.

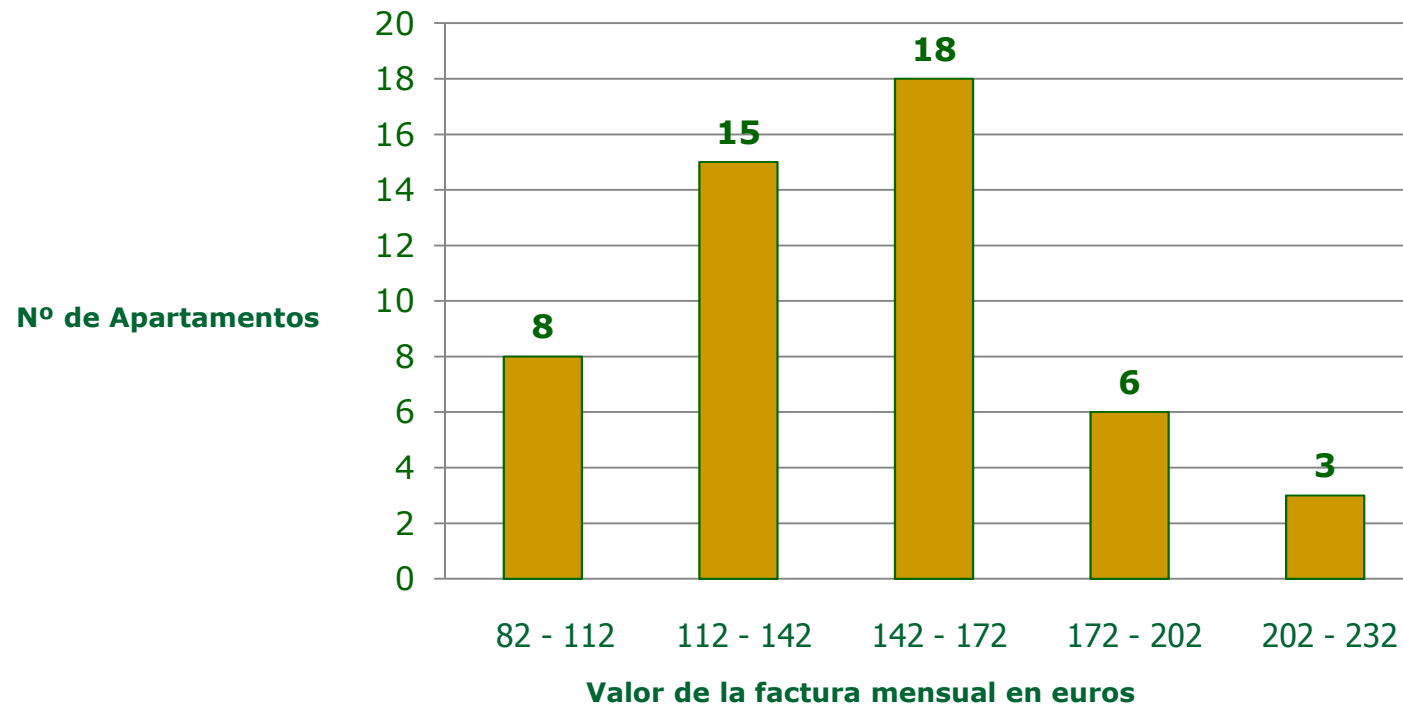
Por lo que se pueden trazar histogramas de frecuencias absolutas e histogramas de frecuencias relativas

2. PRESENTACIÓN DE LA INFORMACIÓN ESTADÍSTICA

Representación gráfica de los datos

Representación gráfica de variables cuantitativas

Histogramas de frecuencias absolutas del consumo de electricidad de 50 apartamentos de 3 habitaciones en Madrid.



2. PRESENTACIÓN DE LA INFORMACIÓN ESTADÍSTICA



Representación gráfica de los datos

Representación gráfica de variables cuantitativas

Los polígonos:

Cuando se construye el polígono se grafica la variable a estudiar sobre el eje horizontal; el eje vertical representa el número o porcentaje de observaciones de cada intervalo de clase.

Los polígonos se forman uniendo por líneas las marcas de clase de todas las categorías o clases que forman la tabla de frecuencias.

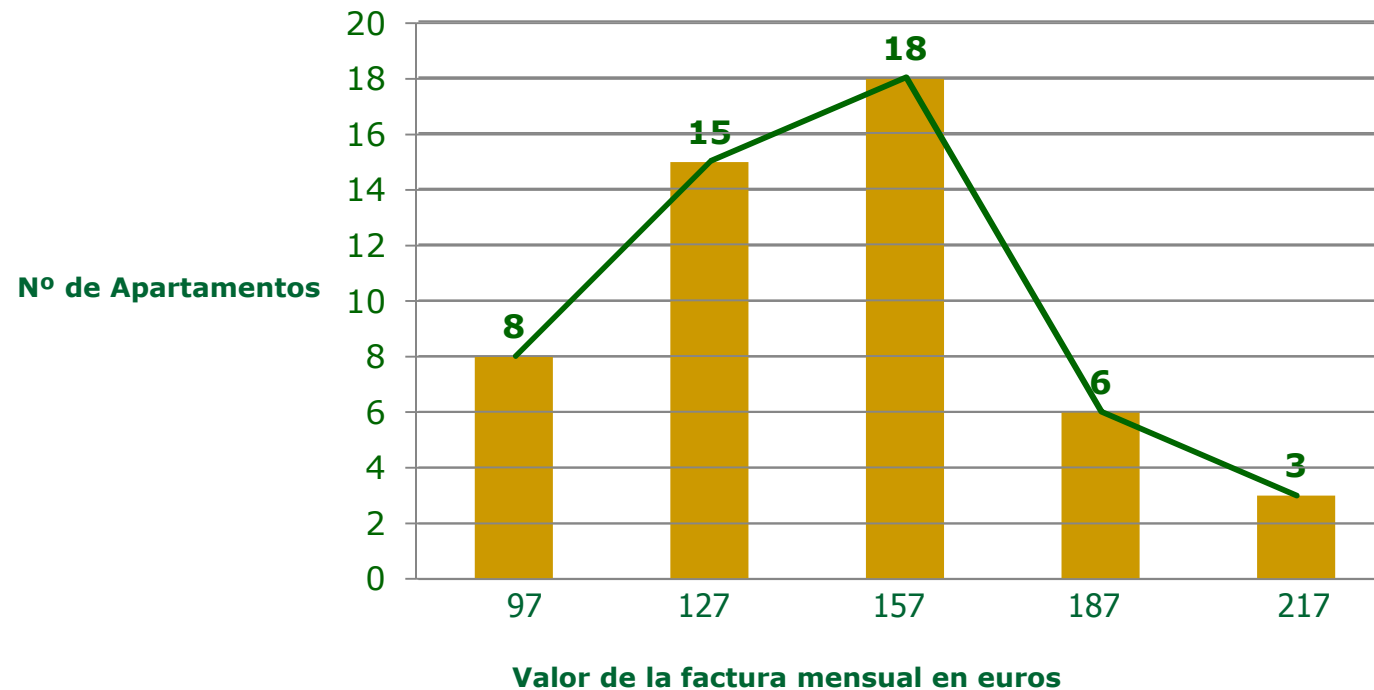
Se pueden trazar polígonos de frecuencias absolutas y polígonos de frecuencias relativas.

2. PRESENTACIÓN DE LA INFORMACIÓN ESTADÍSTICA

Representación gráfica de los datos

Representación gráfica de variables cuantitativas

Polígono de frecuencias absolutas del consumo de electricidad de 50 apartamentos de 3 habitaciones en Madrid.



2. PRESENTACIÓN DE LA INFORMACIÓN ESTADÍSTICA



Representación gráfica de los datos

Representación gráfica de variables cuantitativas

Los polígonos acumulados:

Para las distribuciones acumuladas absolutas y relativas se pueden construir gráficos de representación llamados polígonos acumulados u ojivas.

Cuando se construye la ojiva la variable a estudiar se grafica en el eje horizontal; el eje vertical representa el número o porcentaje de observaciones acumuladas en cada intervalo de clase.

Si fuéramos a construir una tabla de distribución relativa o porcentual acumulada registraríamos en primer lugar los límites inferiores de cada una de las marcas de clase, según aparecen en la distribución y, después, añadiendo un límite extra al final.

2. PRESENTACIÓN DE LA INFORMACIÓN ESTADÍSTICA



Representación gráfica de los datos

Representación gráfica de variables cuantitativas

Ejercicio práctico 2.6

Con los datos de la distribución de frecuencias calculada en el ejercicio 2,3 donde se utilizaron 5 categorías o clases, graficar:

- El histograma de frecuencias absolutas de la tabla. ¿Qué significa el valor observado en la categoría 3?
- El polígono de frecuencias relativas de la tabla. ¿Qué significa el valor observado en la categoría 3?
- El polígono de frecuencias relativas acumuladas de la tabla. ¿Qué significa el valor cero? ¿Qué significa el valor observado en el límite inferior de la categoría 4?

2. PRESENTACIÓN DE LA INFORMACIÓN ESTADÍSTICA



Representación gráfica de los datos

Representación gráfica de variables cuantitativas

Ejercicio práctico 2.7

Con los datos de la distribución de frecuencias calculada en el ejercicio 2,4 donde se utilizaron 5 categorías o clases, graficar:

- El histograma de frecuencias absolutas de la tabla. ¿Qué significa el valor observado en la categorías 1 y 5?
- El polígono de frecuencias relativas de la tabla. ¿Qué significa el valor observado en la categorías 5 y 6?
- El polígono de frecuencias relativas acumuladas de la tabla. ¿Qué significa el valor cero? ¿Qué significa el valor observado en el límite inferior de la categoría 3?

Descripción y Resumen numéricos de datos

Cuando disponemos de una muestra de observaciones asociadas a cierta variable estadística, además, de representarse gráficamente se pueden resumir por unas medidas que nos den una idea global de cómo están distribuidos los valores de la muestra.

Propiedades de los datos Cuantitativos

Las propiedades principales que describen un conjunto de datos numéricos son:

1. Tendencia central de posición
2. Dispersión
3. Forma

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de posición o tendencia central

La mayor parte de los conjuntos de datos tienden a agruparse alrededor de un punto central, por lo que es posible elegir algún valor promedio que describa todo el conjunto de datos. A estas Medidas se denominan medidas de tendencia central o de "posición".

Las medidas de tendencia central son: La media aritmética, la mediana, la moda y el rango medio.

La media aritmética

La medida de posición más usada. Se calcula sumando todas las observaciones de un conjunto de datos, dividiendo después ese total entre el número total de elementos involucrados.

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de posición o tendencia central

La fórmula matemática de la media aritmética:

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

Donde:

- $\bar{X}$ = media aritmética de la muestra
- n = tamaño de la muestra
- X_i = i-ésima observación de la variable x

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA

Medidas de posición o tendencia central

Cálculo de la media aritmética

Ejemplo:

Los precios de un cuarto sencillo en un día entre semana de una muestra de 12 hoteles de Madrid fue:

145	210	125	110	135	90
120	105	130	124	118	122

$$\bar{X} = \frac{\sum_{i=1}^{12} X_i}{12} = \frac{145+210+125+110+135+90+120+105+130+124+118+122}{12} = 127,83$$

El precio medio de un cuarto en los 12 hoteles de Madrid es de **127,83** euros.

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de posición o tendencia central

Los problemas con la media aritmética

Como la media se calcula analizando todas las observaciones, si existen en la muestra datos extremos estos afectan su resultado.

Supongamos que tenemos una muestra de 99 datos iguales a 1 y otro dato igual a 101.

Si calculamos la media:

$$\bar{X} = \frac{(99 \times 1) + (101 \times 1)}{100} = \frac{200}{100} = 2$$

La media es 2, pero esta medida no es una buena medida de localización de la mayoría de los datos.

Como la media es el punto que esta exactamente en la mitad de los datos siempre se cumple que:

$$\sum_{i=1}^n (x_i - \bar{X}) = 0$$

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de posición o tendencia central

La Mediana

Es el valor que se encuentra en el centro de una secuencia ordenada de datos. La mediana no se ve afectada por observaciones extremas en un conjunto de datos.

Cómo calcular la mediana:

Se ordenan los datos de menor a mayor. Esto se denomina arreglo ordenado. Después se utiliza la fórmula **de posicionamiento del punto**

$$\frac{n + 1}{2}$$

Para localizar el lugar que ocupa el valor de la mediana en el arreglo ordenado.

Regla 1 Si el tamaño de la muestra es un número impar, la mediana esta representada por El valor numérico $(n+1)/2$ de las observaciones ordenadas.

Regla 2 Si el tamaño de la muestra es un número par, la mediana será el promedio de los valores numéricos correspondientes a esas dos observaciones centrales.

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA

Medidas de posición o tendencia central

Ejemplo del calculo de la mediana

Los precios de un cuarto sencillo en un día entre semana de una muestra de 12 hoteles de Madrid fue:

145	210	125	110	135	90
120	105	130	124	118	122

Ordenamos los datos de menor a mayor

90	105	110	118	120	122	124	125	130	135	145	210
1	2	3	4	5	6	7	8	9	10	11	12

Observación ordenada

Mediana = **123**

Para estos datos la posición mediana es $(n+1)/2 = (12 + 1)/2 = 6,5$. Por tanto la mediana se obtiene promediando las observaciones ordenadas que ocupan los lugares 6º y 7º:

$$\frac{122 + 124}{2} = \mathbf{123}$$

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA

Medidas de posición o tendencia central

Calculo de la mediana con número impar de datos

Supongamos que en el ejemplo anterior la muestra es de solo 11 hoteles y eliminamos el hotel ordenado en 12º lugar cuyo valor es de 210 euros:
La mediana esta representada por el valor numérico de la observación que ocupa la posición $(11+1)/2 = 6$

90	105	110	118	120	122	124	125	130	135	145
1	2	3	4	5	6	7	8	9	10	11

Observación ordenada

↓
Mediana = 122

Empates en los datos

Cuando se calcula la mediana se ignora el hecho de que pueda haber valores iguales en los datos.

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de posición o tendencia central

Cálculo de la mediana cuando hay empates en los datos

Supongamos que en el ejemplo anterior, en la muestra de 11 hoteles, el valor ordenado correspondiente al hotel N° 5 no es de 120 si no, de 122.

La mediana sigue siendo 122 que es el valor correspondiente a la observación observada en 6º lugar, que es el valor de la mediana, aunque el 5º sea igual.

90	105	110	118	122	122	124	125	130	135	145
1	2	3	4	5	6	7	8	9	10	11

Observación ordenada

↓

Mediana = 122

Características de la mediana

El cálculo del valor de la mediana se ve afectada por la cantidad de observaciones y no por la magnitud de las observaciones.

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de posición o tendencia central

La Moda

Es el valor de un conjunto de datos que se repite con más frecuencia. Se Obtiene fácilmente a partir de un arreglo ordenado.

Ejemplo del calculo de la moda:

Los precios de un cuarto sencillo en un día entre semana de una muestra de 12 hoteles de Madrid fue:

90	105	110	118	120	122	124	125	130	135	145	210
1	2	3	4	5	6	7	8	9	10	11	12

Observación ordenada

En este ejemplo se observa que no existe ninguna moda.

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de posición o tendencia central

Ejemplo del calculo de la moda:

Supongamos que los resultados de la muestra de Los precios de un cuarto sencillo en un día entre semana de una muestra de 12 hoteles de Madrid fue:

90	105	110	120	120	120	124	124	130	135	145	210
1	2	3	4	5	6	7	8	9	10	11	12

Observación ordenada

En este ejemplo la moda sería 120 euros.

Un conjunto de datos puede tener más de una moda.

Supongamos que en el ejemplo anterior la distribución de los precios de un cuarto sencillo en un hotel de Madrid un día entre semana de la muestra aleatoria de 12 hoteles se distribuye como se ve en la siguiente tabla:

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA

Medidas de posición o tendencia central

Ejemplo del cálculo de la moda:

90	105	110	120	120	120	124	124	124	135	145	210
1	2	3	4	5	6	7	8	9	10	11	12

Observación ordenada

En este ejemplo la moda sería 120 euros y 124 euros. Estos datos se describen como bimodales.

A diferencia de la media aritmética, la moda no se ve afectada por valores extremos. Sin embargo, solo se utiliza para propósitos descriptivos por que es más variable, para distintas muestras, que las demás medidas de tendencia central.

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de posición o tendencia central

El rango medio

Es el promedio de las observaciones mayor y menor de un conjunto de datos. Se expresa de la siguiente manera:

$$\text{Rango medio} = \frac{X_{\text{menor}} + X_{\text{mayor}}}{2}$$

Cálculo del rango medio

Utilizando el arreglo ordenado de la muestra aleatoria de los precios de los cuartos de 12 hoteles en Madrid, el rango medio sería:

90	105	110	118	120	122	124	125	130	135	145	210
1	2	3	4	5	6	7	8	9	10	11	12

Observación ordenada

$$\frac{90 + 210}{2} = 150$$

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de posición o tendencia central

El rango medio

A pesar de su sencillez, el rango medio se debe utilizar con cautela, ya que solo implica las observaciones mayor y menor de un conjunto de datos, si hay observaciones extremas se distorsiona como medida de tendencia central.

No obstante, con frecuencia se utiliza en forma exitosa el rango medio como parámetro de medición tanto para análisis financieros como para informes sobre el clima, puesto que ofrece un valor adecuado "rápido y sencillo" para resumir todo un conjunto de datos, ya sea una serie de precios diarios de cierre de una acción para todo un año o un conjunto de temperaturas registradas cada hora durante todo un día. En tales casos, no es probable que ocurra un valor extremo.

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de posición o tendencia central

Ejercicio práctico 3.1

Con los siguientes conjuntos de datos, ambos con un tamaño de muestra $n=7$:

Lote 1	10	2	3	2	4	2	5
Lote 2	20	12	13	12	14	12	15

- Calcule la media aritmética, la mediana, la moda y el rango medio para ambos conjuntos de datos.
- Compare los resultados y resuma sus conclusiones.
- Compare el primer elemento muestreado de cada conjunto; compare el segundo elemento muestreado y así sucesivamente. Describa brevemente las conclusiones en este caso, a la luz del resumen hecho en la parte b).

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de posición o tendencia central

Ejercicio práctico 3.2

El director de Recursos humanos de la Brigada de Bomberos de Madrid debe elegir cual de dos aspirantes contratar en función a los resultados obtenidos en 5 pruebas de 100 metros lisos en un periodo de una hora con descansos de 5 minutos. Los resultados registrados fueron:

Aspirante 1	12.1	12.0	12.0	16.8	12.1
Aspirante 2	12.3	12.4	12.4	12.5	12.4

- Con base en los resultados anteriores ¿A cuál de los dos aspirantes debe elegir el director de recursos humanos?
- ¿Debería ser diferente la selección si el director de recursos humanos sabe que el aspirante 1 se cayó al principio de la cuarta carrera? ¿Por qué?

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de posición o tendencia central

Ejercicio práctico 3.3

Un fabricante de pilas para transistores tomó una muestra de 13 piezas de la producción de un día y las utilizó de forma continua hasta cuando empezaron a fallar. El resultado en horas de funcionamiento fue:

342	426	317	545	264	451	1049
631	512	266	492	562	298	

- Calcule la media, la mediana, la moda y el rango medio. Observando la distribución de los tiempos ¿Qué medidas descriptivas parecen ser las mejores y cuáles las peores? ¿Por qué?
- ¿De qué forma le puede ser útil la información al fabricante?

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de dispersión

Una segunda propiedad muy importante que describe un conjunto de datos es la dispersión. Dispersión es el grado de variabilidad o diseminación de los datos. Así podemos decir, que dos conjuntos de datos pueden diferir tanto en su tendencia central como en dispersión.

Las principales medidas de dispersión son: El rango, la varianza, la desviación típica o estándar y el coeficiente de variación.

El rango

Es la diferencia entre las observaciones mayor y menor del conjunto de datos:

$$\text{Rango} = X_{\text{mayor}} - X_{\text{menor}}$$

El rango mide la dispersión total del conjunto de datos. Su debilidad radica en que no toma en cuenta como se distribuyen los datos.

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de dispersión

Ejemplo del cálculo del rango

Los precios de un cuarto sencillo en un día entre semana de una muestra aleatoria de 12 hoteles de Madrid fue:

145	210	125	110	135	90
120	105	130	124	118	122

$$\text{Rango} = X_{\text{mayor}} - X_{\text{menor}}$$

$$\text{Rango} = 210 - 90 = 120$$

Podemos decir que la diferencia entre la habitación más cara y la más barata de los hoteles muestreados en Madrid es de 120 euros. Sin embargo, no podemos decir nada acerca de si del total de hoteles analizados hay habitaciones más caras o más baratas.

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de dispersión

La varianza y la desviación típica

Son las medidas de dispersión usadas con más frecuencia. Toman en consideración la forma en que se distribuyen los datos. Miden la forma en que los valores fluctúan alrededor de la media.

Definición de la varianza muestral

Es el promedio de los cuadrados de las diferencias entre cada una de las observaciones de un conjunto de datos y la media. Así, para una muestra que contiene n observaciones $X_1, X_2, \dots, X_n$, la varianza muestral sería:

$$S^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_n - \bar{X})^2}{n - 1}$$

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de dispersión

La varianza y la desviación típica

Utilizando la notación sumatoria se puede plantear la fórmula de forma mas simple de la siguiente manera:

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

$\bar{X}$ = media aritmética de la muestra

n = Tamaño de la muestra

X_i = i-ésimo valor de la variable aleatoria X

$\sum_{i=1}^n (X_i - \bar{X})^2$ = Sumatoria de todas las diferencias al cuadrado entre los valores de X_i y $\bar{X}$

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de dispersión

La desviación típica

La desviación típica o estándar, es simplemente la raíz cuadrada de la varianza muestral. Se simboliza como S

Desviación estándar: $S = \sqrt{S^2}$

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

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de dispersión

Cálculo de S y S^2

Para calcular la varianza:

1. Calcule las diferencias entre cada una de las observaciones y la media.
2. Eleve al cuadrado cada una de las diferencias.
3. Sume estas diferencias elevadas al cuadrado.
4. Divida esa suma entre $n - 1$.

Para obtener la desviación típica simplemente saque la raíz cuadrada de la Varianza.

Ejemplo:

Para la muestra de los precios de los cuartos sencillos un día entre semana de la muestra aleatoria de 12 hoteles de Madrid los datos originales fueron:

145	210	125	110	135	90
120	105	130	124	118	122

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de dispersión

Cálculo de S y S²

La varianza se calcula de la siguiente manera:

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

$$S^2 = \frac{(145-127,83)^2 + (210-127,83)^2 + \dots + (122-127,83)^2}{12 - 1} = 871,61$$

$$S = \sqrt{871,61} = 29,52$$

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de dispersión

Qué indican la varianza y la desviación típica

Miden la dispersión promedio en torno a la media. Es decir, como fluctúan las observaciones mayores por encima de la media y como se distribuyen las observaciones menores por debajo de ella.

La varianza al ser calculada nos da unidades al cuadrado. En el ejemplo anterior la varianza del precio de las habitaciones individuales de los 12 hoteles muestreados en Madrid es de **871,61** euros cuadrados. Por ello, en la práctica la principal medida de dispersión que se utiliza es la desviación típica cuyo valor esta dado en unidades originales.

En nuestro ejemplo diríamos entonces que los precios de las habitaciones individuales un día entre semana de los 12 hoteles muestreados en Madrid tienen una desviación típica de **29,52** euros. Es decir que la mayor parte de los precios de las habitaciones individuales de los hoteles muestreados en Madrid, varía entre **98,31** y **157,35** euros. Es decir, $(127,83 \pm 29,52)$

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de dispersión

¿Por qué se elevan las desviaciones al cuadrado?

Como la media es el centro de gravedad del conjunto de datos, es decir, el punto que esta exactamente en la mitad. Las fórmulas de la varianza y desviación típica no podrían calcularse tan solo con:

$$\sum_{i=0}^n (x_i - \bar{x}) \text{ por que este resultado es } = 0$$

La suma de las desviaciones al cuadrado permite estudiar la dispersión de los datos. Por ello se utiliza:

$$\sum_{i=0}^n (x_i - \bar{x})^2$$

En el proceso de elevar al cuadrado, las observaciones que se encuentran más alejadas de la media reciben más peso que las observaciones que están más cerca.

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de dispersión

Generalizaciones acerca de la varianza y la desviación típica

1. Cuanto más “separados” o dispersos estén los datos. Tanto mayores serán el rango, la varianza y la desviación típica.
2. Si los datos están mas “concentrados” o son homogéneos, menores serán el rango, la varianza y la desviación típica.
3. Si todas las observaciones son iguales, de manera que no haya variación en los datos, el rango, la varianza y la desviación típica serán iguales a cero.

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de dispersión

Fórmulas abreviadas de cálculo de la varianza

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

Donde:

$\sum_{i=1}^n X_i^2$ = sumatoria de los cuadrados de cada observación

$(\sum_{i=1}^n X_i)^2$ = cuadrado de la sumatoria total

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA

Medidas de dispersión

Fórmula abreviada para el cálculo de la desviación típica

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

Donde:

- $\sum_{i=1}^n X_i^2$ = sumatoria de los cuadrados de cada observación
- $(\sum_{i=1}^n X_i)^2$ = cuadrado de la sumatoria total

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de dispersión

Cálculo de S^2 y S con la fórmula abreviada

En el ejemplo que venimos manejando:

$$S^2 = \frac{(145^2 + 210^2 + \dots + 122^2) - \frac{(145 + 210 + \dots + 122)^2}{12}}{12 - 1}$$
$$S^2 = \frac{205684 - (2353156/12)}{12 - 1} = \frac{205684 - 196096,33}{11} = 871,61$$
$$S = \sqrt{871,61} = 29,52$$

Como vemos los resultados son exactamente iguales que con la fórmula anterior.

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de dispersión

Coeficiente de Variación

Es una medida relativa de la dispersión. Se expresa como porcentaje y no en términos de las unidades de los datos manejados.

Es útil cuando se compara la variabilidad de dos o mas conjuntos de datos expresados en diferentes unidades de medición.

Se expresa de la siguiente manera:

$$CV = (S/\text{media}) 100\%$$

En nuestro ejemplo el coeficiente de variación para los precios de las habitaciones sencillas de los 12 hoteles mostrados en Madrid sería:

$$CV = (29,52/127,83) 100\% = \mathbf{23,1\%}$$

Es decir, para esta muestra de 12 hoteles la dispersión promedio del precio de la habitación de un cuarto sencillo es del 23,1%

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de dispersión

Ejercicio Práctico 3.4

Con los datos del ejercicio práctico 3.1 calcular:

- El rango, la varianza y la desviación típica para los dos conjuntos de datos.
- Compare los resultados y resuma tus conclusiones.
- Calcular el coeficiente de variación para los dos conjuntos de datos. Compare los resultados y resuma tus conclusiones.

Ejercicio práctico 3.5

Con los datos del ejercicio 3.2 calcular:

- El rango, la varianza y la desviación típica para los dos aspirantes.
- Con los resultados obtenidos. ¿Refuerzan estos datos la decisión tomada con base en los resultados obtenidos de las medidas de tendencia central?
- En términos porcentuales, ¿qué aspirante tiene la variación promedio más alta en sus tiempos? Justifique su respuesta.

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de dispersión

Ejercicio Práctico 3.6

Con los datos del ejercicio práctico 3.3 calcular:

- El rango, la varianza y la desviación típica.
- ¿Qué significan estos valores en este caso en particular?
- Calcular el coeficiente de variación. Interprete lo que significa este valor en este caso. Saque tus propias conclusiones.

Medidas de forma

La tercera propiedad que describe un conjunto de datos es su forma; la manera en que se distribuyen los datos. Una distribución de datos puede ser **simétrica** o no. Las distribuciones no simétricas se denominan asimétricas o **sesgadas**.

Para saber la forma del conjunto de datos comparamos la media y la mediana. Es decir:

Media > mediana: sesgo positivo o hacia la derecha

Media = mediana: simétrica o sesgo cero

Media < mediana: sesgo negativo o hacia la izquierda.

Veamos algunos ejemplos de la forma de varios conjuntos de datos.

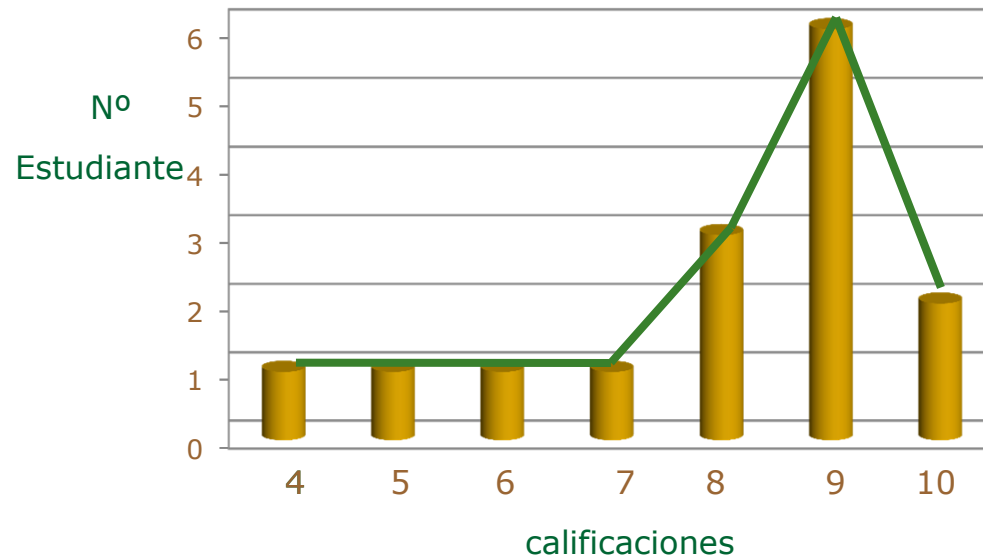
3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA

Medidas de forma

Sesgo a la izquierda

La siguiente tabla muestra las calificaciones de 15 alumnos en un examen:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Nº estudiante
4	5	6	7	8	8	8	9	9	9	9	9	9	10	10	Calificación



Media = 8

Mediana = 9

Sesgo negativo o a la izquierda. La media es menor que la mediana ya que hay pocas calificaciones bajas

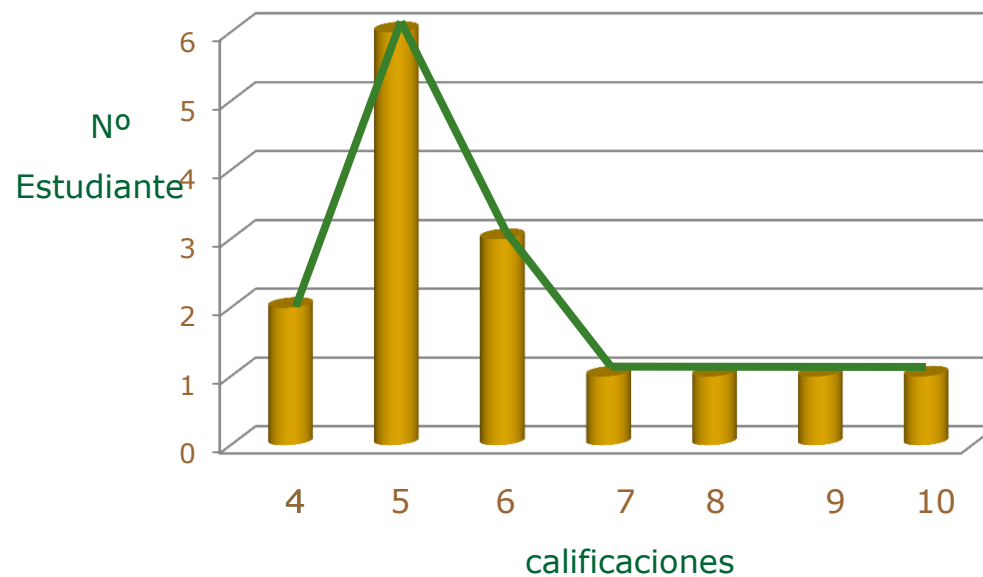
3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA

Medidas de forma

Sesgo a la derecha

La siguiente tabla muestra las calificaciones de 15 alumnos en un examen:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Nº estudiante
4	4	5	5	5	5	5	5	6	6	6	7	8	9	10	Calificación



Media = 6

Mediana = 5

Sesgo positivo a la derecha.
Muestra el mal desempeño de los estudiantes en el examen.

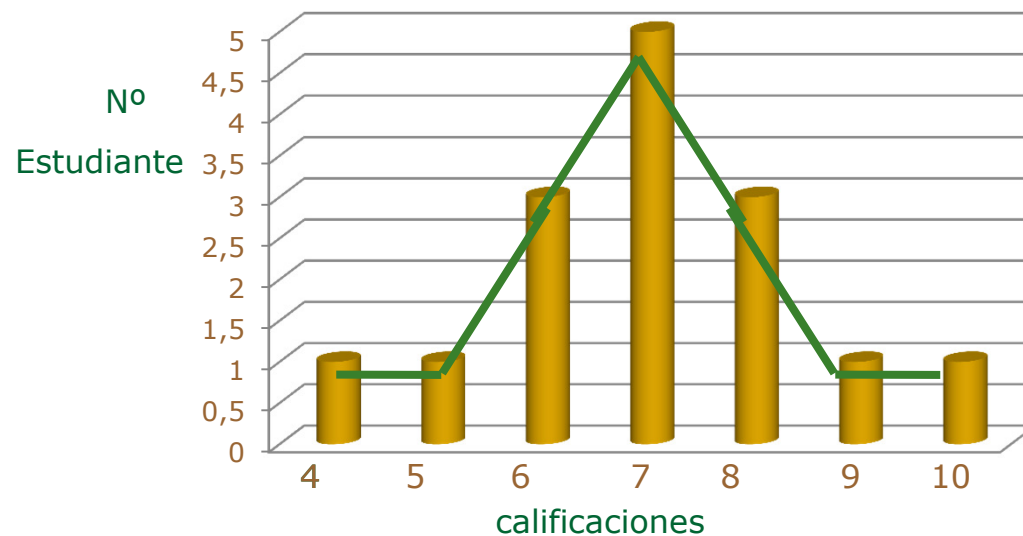
3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA

Medidas de forma

Datos simétricos

La siguiente tabla muestra las calificaciones de 15 alumnos en un examen:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Nº estudiante
4	5	6	6	6	7	7	7	7	7	8	8	8	9	10	Calificación



Media = 7

Mediana = 7

Distribución simétrica. También son iguales la moda y el rango medio.

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de forma

¿Cuál es la forma de los datos del ejemplo del precio de un cuarto sencillo en un día entre semana de 12 hoteles de Madrid?

145	210	125	110	135	90
120	105	130	124	118	122

Media = **127,83** euros

Mediana = **123** Euros

Como la media es $>$ que la mediana los precios tienen sesgo positivo o están sesgados a la derecha. Es decir, hay más habitaciones sencillas caras que baratas en la muestra de hoteles madrileños.

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de forma

Ejercicio Práctico 3.7

Con los datos del ejercicio práctico 3.1:

- Describa la forma de los datos.
- Compare los resultados y comenta tus conclusiones.

Ejercicio práctico 3.8

Con los datos del ejercicio 3.2

- Describir la forma de los datos (los tiempos para los dos aspirantes).
- Compare los resultados y comenta tus conclusiones.

Ejercicio Práctico 3.9

Con los datos del ejercicio 3.3

- Describir la forma de los datos de la vida útil de las pilas para transistores.
- Compare los resultados y comenta tus conclusiones.

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de forma

Los cuantiles

Además de las medidas de tendencia central, dispersión y forma, también hay algunas medidas útiles de posición no central que suelen utilizarse al resumir o descubrir las propiedades de un conjunto de datos cuantitativos. A estas medidas se le denominan cuantiles. Algunos de los cuantiles más utilizados son **cuartiles** que dividen los datos en cuartos; los **deciles** (que dividen los datos ordenados en décimos) y los **percentiles** (los que los dividen en centésimos). Revisaremos los cuartiles.

Los cuartiles

La mediana divide los datos en un arreglo ordenado en la mitad (50% de las observaciones son menores y el 50% de las observaciones son mayores), los cuartiles son medidas descriptivas que dividen los datos ordenados en cuatro cuartos.

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Tema 3. Descripción y Resumen numéricos de datos

Los cuartiles

Primer cuartil, Q1, es el valor según el cual el 25% de las observaciones son menores que él y el 75% son mayores.

El segundo cuartil, Q2, es la mediana: el 50% de las observaciones son menores y el 50% mayores.

El tercer cuartil, Q3, es el valor según el cual el 75 de las observaciones son menores y el 25% son mayores.

Para aproximar los cuartiles de una población que contiene N observaciones se utilizan las fórmulas de posicionamiento.

Q1 = Valor correspondiente a la $\frac{N + 1}{4}$ observación ordenada

Q2 = Mediana, el valor correspondiente a la $\frac{2(N + 1)}{4}$ observación ordenada

Q3 = Valor correspondiente a la $\frac{3(N + 1)}{4}$ Observación ordenada

Medidas de forma

Los cuartiles

Las siguientes reglas se usan para obtener los valores de los cuartiles:

1. Si la posición resultante es un entero, la observación numérica correspondiente a esa posición es la que constituye el cuartil.
2. Si la posición resultante se encuentra a mitad entre dos valores enteros, se elige el promedio de los dos valores correspondientes.
3. Si la posición que se obtiene no es ni un entero, ni un valor intermedio entre dos valores enteros, se sigue una regla sencilla para aproximar al cuartil específico, la cual consiste en redondear a la posición entera más cercana y elegir el valor numérico de la observación correspondiente.

Ejemplo:

Las calificaciones obtenidas por los 18 estudiantes del curso de estadística de la carrera de Administración de Negocios de una universidad fueron:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	No. estudiante
10	3	4	8	7	8	9	6	5	6	6	5	6	8	9	6	4	5	Calificación

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA

Medidas de forma

Ejemplo:

Ordenamos las calificaciones de menor a mayor

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	No. estudiante
3	4	4	4	5	5	5	6	6	6	6	6	7	8	8	9	9	10	Calificación
				↑				└─┘					↑					
				Q1				Q2					Q3					

$$Q1 = \frac{18 + 1}{4} = 4,75 \approx 5 \text{ posición ordenada. Por tanto: } Q1 = 5$$

$$Q2 = \frac{2(18 + 1)}{4} = \frac{N + 1}{2} = 9,5 \text{ posición ordenada. Por tanto: } Q2 = 6 + 6/2$$
$$Q2 = 6$$

$$Q3 = \frac{3(18 + 1)}{4} = 14,25 \approx 14 \text{ posición ordenada. Por tanto } Q3 = 8$$

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de forma

Interpretación de lo que significan los cuartiles en el ejemplo:

El 25% de los estudiantes de estadística de la carrera de Administración de negocios, obtuvieron una calificación inferior a 5.

El 50% de los estudiantes de estadística de la carrera de Administración de negocios, obtuvieron una calificación inferior a 6.

El 75% de los estudiantes de estadística de la carrera de Administración de negocios, obtuvieron una calificación inferior a 8.

Medidas descriptivas de resumen utilizando cuartiles

Los cuartiles son útiles no como medidas de posición no central; también sirven para elaborar otras medidas importantes de tendencia central y dispersión.

El eje medio una medida de tendencia central

$$\text{Eje medio} = \frac{Q1 + Q3}{2}$$

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de forma

El eje medio una medida de tendencia central

En nuestro ejemplo:

$$\text{Eje medio} = \frac{5 + 8}{2} = 6,5$$

Se utiliza para superar algunos problemas potenciales presentados con valores extremos en los datos.

El rango intercuartilico es una medida de dispersión

Es la diferencia entre los cuartiles 1 y 3.

$$\text{Rango intercuartilico} = Q3 - Q1$$

En nuestro ejemplo Rango intercuartilico = $8 - 5 = 3$

Representa la dispersión en el 50% medio de los datos. Por tanto, no se ve influenciada por valores extremos

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de forma

Ejercicio práctico 3.10

Los siguientes datos son los valores en libros (es decir el capital social, dividido entre el número de acciones en circulación) del total de 55 empresas manufactureras españolas que cotizan en la bolsa de valores de Madrid

7	9	8	6	12	6	9	15	9	16
8	5	14	8	7	6	10	8	11	4
10	6	16	5	10	12	7	10	15	7
10	8	8	10	18	8	10	11	7	10
7	8	15	23	13	9	8	9	9	13
14	8	12	18	15					

- Cualcular los cuartiles 1, 2 y 3.
- Interprete lo que significa Q1 y Q3 en este ejercicio.
- Calcular el eje medio y el rango intercuartilico.
- Interprete lo que significan el eje medio y el rango intercuartilico en el ejercicio.

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de forma

Ejercicio práctico 3.11

Con los datos del ejercicio 2.4

- a. Calcular los cuartiles 1, 2 y 3.
- b. Interprete lo que significa $Q1$ y $Q3$ en este ejercicio.
- c. Calcular el eje medio y el rango intercuartilico.
- d. Interprete lo que significan el eje medio y el rango intercuartilico en el ejercicio.

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de forma

Utilización de cinco números de resumen para el diagrama de bloques y líneas, una técnica gráfica

Cinco números de resumen

Para identificar y describir las principales características de los datos, el método de "análisis exploratorio de datos" utiliza medidas de tendencia central y de dispersión que tienen la propiedad de resistencia; es decir, estadísticos que son relativamente insensibles a valores extremos de algunos de los datos. La mediana, el eje medio y el rango intercuartilico son tres estadísticos resistentes de uso común. Si se combinan estas medidas resistentes con información referente a los extremos, se logra entonces una mejor idea de la forma de la distribución. Cinco números de resumen son:

X menor Q1 Mediana Q3 Xmayor

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de forma

Para el ejemplo de las calificaciones obtenidas por 18 estudiantes del curso de estadística de los estudiantes de Administración de Negocios de una universidad estos números son:

$$X_{\text{menor}} = 3 \quad Q1 = 5 \quad \text{Mediana} = 6 \quad Q3 = 8 \quad X_{\text{mayor}} = 10$$

Ahora se pueden utilizar los cinco números para estudiar la forma de la distribución. Si los datos fueran perfectamente simétricos sería cierto qué:

1. La distancia de $Q1$ a la mediana sería igual que la distancia de la mediana a $Q3$.
2. La distancia de X_{menor} a $Q1$ sería igual a la distancia de $Q3$ a X_{mayor} .
3. La mediana, el eje medio y el rango medio serían todos iguales. (Además estas medidas serían iguales a la media de los datos.

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de forma

Por otro lado, para distribuciones no simétricas sería cierto que:

1. En las distribuciones con sesgo hacia la derecha, la distancia entre $Q3$ a X_{mayor} excede en gran medida a la distancia de X_{menor} a $Q1$.
2. En las distribuciones con sesgo a la derecha

$\text{Mediana} < \text{eje medio} < \text{rango medio}$

3. En las distribuciones con sesgo hacia la izquierda, la distancia de X_{menor} a $Q1$ excede en gran medida la distancia de $Q3$ a X_{mayor} .
4. En las distribuciones con sesgo hacia la izquierda

$\text{Rango medio} < \text{eje medio} < \text{mediana}$

Según estas reglas la distribución de las calificaciones de los 18 Estudiantes del curso de estadística de la carrera de Administración de Negocios tienen sesgo hacia la derecha: $6 < 6,5 < 6,5$

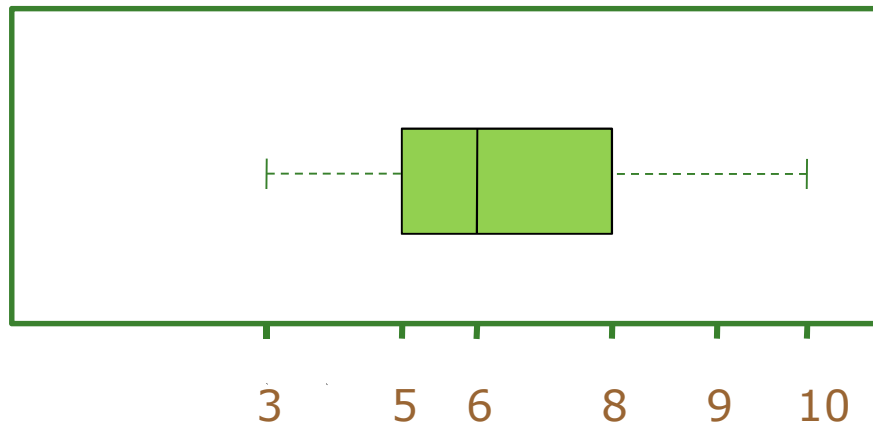
3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA

Medidas de forma

Diagrama de bloques y líneas

Es la representación gráfica de los datos a través de los cinco números de Resumen.

Diagrama de bloques y líneas de las calificaciones de los 18 estudiantes del curso de estadística de La carrera de Administración de negocios:



Distribución sesgada a la derecha

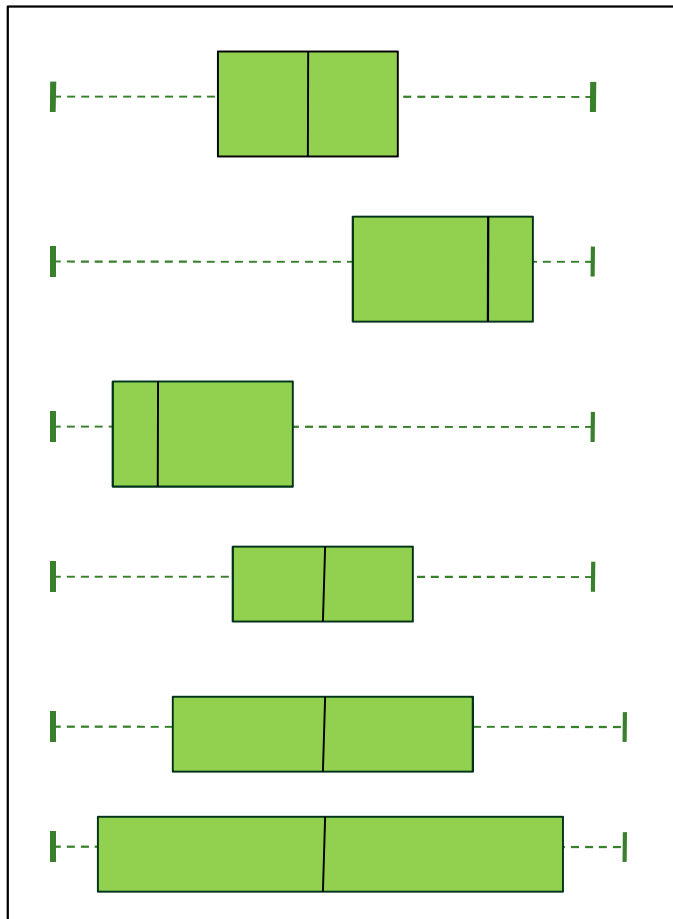
La línea vertical trazada dentro del bloque representa el valor de la mediana de los datos. Nótese que la línea vertical del lado izquierdo del bloque representa $Q1$ y la línea vertical del lado derecho del bloque representa la ubicación de $Q3$.

Se puede ver que la caja contiene el 50% medio de las observaciones de la distribución. El 25% inferior de los datos se representa mediante la línea punteada que conecta el lado izquierdo del bloque con la ubicación del valor, X_{menor} . De manera similar, el 25% superior de los datos se representa mediante una línea punteada que conecta el lado derecho del bloque con X_{mayor} .

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA

Medidas de forma

Diagrama de bloques y líneas



Curva en forma de campana

Distribución sesgada hacia la izquierda

Distribución sesgada hacia la derecha

Curva en forma triangular

Curva en forma rectangular

Curva en forma de U

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Medidas de forma

Ejercicio Práctico 3.12

Con los datos del ejercicio práctico 3.10

- Calcule los cinco números de resumen.
- Trace el diagrama de bloques y líneas y describa su forma.

Ejercicio práctico 3.13

Con los datos del ejercicio 2.4

- Calcule los cinco números de resumen.
- Trace el diagrama de bloques y líneas y describa su forma.

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Calculo de las medidas descriptivas para una población

Para una población de tamaño N la forma de calcular las medidas descriptivas de la población es la siguiente:

$$\text{Media: } \mu_x = \frac{\sum_{i=1}^N x_i}{N}$$

$$\text{Rango medio} = \frac{X_{\text{menor}} + X_{\text{mayor}}}{2}$$

$$\text{Rango} = X_{\text{mayor}} - X_{\text{menor}}$$

$$\text{Varianza: } \sigma_x^2 = \frac{\sum_{i=1}^N (x_i - \mu_x)^2}{N}$$

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu_x)^2}{N}}$$

Donde N = tamaño de la población x_i = i-ésimo valor de la variable X

$\sum_{i=1}^N x_i$ "Sumatoria de" todos los valores x_i de la población

Sumatoria de todas las
Diferencias al cuadrado entre
 x_i y μ_x

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Calculo de las medidas descriptivas para una población

Ejercicio práctico 3.14

Los siguientes datos corresponden a los precios en céntimos de las 20 marcas de agua mineral que se venden en los supermercados de la provincia de Almería:

65	70	68	75	72
67	80	95	90	88
65	70	75	85	80
70	75	92	90	80

Calcular:

- La media, la mediana, la moda, el rango medio, para esta población.
- El Q1, Q3, el eje medio, el rango, el rango intercuartilico, la varianza, la desviación típica y el coeficiente de variación para esta población.
- Decir cual es la forma de los datos de la población.

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Calculo de las medidas descriptivas para una población

Ejercicio práctico 3.15

Supongamos que de la población de marcas de agua mineral se tomo una muestra aleatoria de 7 marcas para analizar los precios y los resultados fueron:

65	70	95	75	72	80	88
----	----	----	----	----	----	----

Calcular:

- Las mismas medidas (estadísticos) de esta muestra de datos que los calculados para la población (parámetros)
- Decir la forma de los datos.
- Comparar los parámetros poblacionales con los estadísticos muestrales y saque tus propias conclusiones.

Obtención de medidas descriptivas de resumen a partir de datos agrupados (distribución de frecuencias)

La media

Cuando se agrupan los datos en una distribución de frecuencias, se utiliza el punto medio de cada clase o la marca de clase m_k para calcular la media. El proceso se realiza de la siguiente manera:

1. Multiplicar cada una de las marcas de clase m_k por el número de observaciones n_k que representa.
2. Obtener la sumatoria de estos productos.
3. Dividir ese total entre el número de observaciones.

Es decir, para una muestra de n observaciones:

$$\bar{X} \approx \frac{\sum_{i=1}^k m_k n_k}{n}$$

Donde $\bar{X}$ = media de la muestra

n = número de observaciones de la muestra

K = número de clases de la distribución de frecuencias

m_k = marca de clase o punto medio de la k -ésima clase

n_k = número de observaciones clasificadas en la k -ésima clase

Obtención de medidas descriptivas de resumen a partir de datos agrupados (distribución de frecuencias)

Ejemplo calculo de la media:

Para calcular la media de los gastos de consumo de electricidad de una muestra aleatoria de 50 apartamentos del centro de Madrid que se agrupo en 5 clases como se ve en la tabla. Procederíamos así:

$[l_k, l_{k+1})$	m_k	n_k	$m_k \times n_k$
82 a < 112	97	8	776
112 a < 142	127	15	1905
142 a < 172	157	18	2826
172 a < 202	187	6	1122
202 a < 232	217	3	651
		50	7280

$$\bar{X} \approx \frac{\sum_{i=1}^k m_k n_k}{n} \approx \frac{7280}{50} = 145,6$$

El valor medio de la factura de consumo de electricidad de la muestra de 50 apartamentos de Madrid es de aproximadamente 145,6 euros.

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Obtención de medidas descriptivas de resumen a partir de datos agrupados (distribución de frecuencias)

Calculo de la media para la población:

$$\mu_{\bar{x}} \approx \frac{\sum_{i=1}^k m_k n_k}{N}$$

Donde $\mu_{\bar{x}}$ = media de la población

N = número total de observaciones de la población

K = número de clases de la distribución de frecuencias

m_k = marca de clase o punto medio de la k-ésima clase

n_k = número de observaciones clasificadas en la k-ésima clase

Obtención de medidas descriptivas de resumen a partir de datos agrupados (distribución de frecuencias)

Aproximación de otras medidas de tendencia central

La mediana se aproxima a partir de graficar la ojiva de frecuencias relativas acumuladas.

La moda se puede aproximar a partir de la marca de clase que contiene el mayor número de observaciones.

Se puede aproximar el rango medio promediando el límite superior de la última clase y el límite inferior de la primera clase.

En el ejemplo anterior: La Moda $\approx$ 157.

$$\text{El rango Medio} = \frac{232 + 82}{2} = 157$$

Obtención de medidas descriptivas de resumen a partir de datos agrupados (distribución de frecuencias)

Aproximación de la varianza y la desviación típica

1. Se obtienen los cuadrados de las diferencias entre cada marca de clase y la media.
2. Se multiplica o se “pondera” cada una de las diferencias al cuadrado por el correspondiente número de observaciones de la clase.
3. Se suman estos productos.
4. Se divide este total entre $n-1$ en el caso de muestras y entre N en el caso de poblaciones.

$$S^2 \approx \frac{\sum_{i=1}^k (m_k - \bar{X})^2 n_k}{n - 1}$$

Donde $\bar{X}$ = media aproximada de la muestra

S^2 = varianza de la muestra

n = número de observaciones de la muestra

K = número de clases de la distribución de frecuencias

m_k = marca de clase o punto medio de la k -ésima clase

n_k = número de observaciones clasificadas en la k -ésima clase

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Obtención de medidas descriptivas de resumen a partir de datos agrupados (distribución de frecuencias)

Aproximación de la varianza y la desviación típica

En tanto para la población:

$$\sigma_x^2 \approx \frac{\sum_{i=1}^k (m_k - \mu_x)^2 n_k}{N}$$

La desviación típica es simplemente la raíz cuadrada de la varianza.
Para la muestra:

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

Para la población:

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

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Obtención de medidas descriptivas de resumen a partir de datos agrupados (distribución de frecuencias)

Ejemplo calculo de la varianza y la desviación típica:

Para calcular la varianza y la desviación típica de los gastos de consumo de electricidad de una muestra aleatoria de 50 apartamentos del centro de Madrid que se agrupó en 5 clases como se ve en la tabla. $\bar{x}$ Procederíamos así:

$[l_k, l_{k+1})$	m_k	n_k	$m_k - \bar{X}$	$(m_k - \bar{X})^2$	$(m_k - \bar{X})^2 n_k$
82 a < 112	97	8	(97-145,6)	2361,96	18895,68
112 a < 142	127	15	(127-145,6)	345,96	5189,4
142 a < 172	157	18	(157-145,6)	129,96	2339,28
172 a < 202	187	6	(187-145,6)	1713,96	10283,76
202 a < 232	217	3	(217-145,6)	5097,96	15293,88
Totales		50			52002

Obtención de medidas descriptivas de resumen a partir de datos agrupados (distribución de frecuencias)

Ejemplo calculo de la varianza y la desviación típica:

$$S^2 \approx \frac{52002}{49} = 1061,27$$

$$S^2 \approx 1061,27$$

$$S \approx \sqrt{1061,27} = 32,58$$

Los gastos de consumo eléctrico de la muestra de apartamentos de Madrid tienen una variabilidad promedio de 32,58 euros

Obtención de medidas descriptivas de resumen a partir de datos agrupados (distribución de frecuencias)

Aproximación de otras medidas de dispersión:

El rango se aproxima en datos agrupados mediante la diferencia entre el límite superior de la última clase y el límite inferior de la primera clase.

$$\text{Rango} = 232 - 82 = 150$$

El coeficiente de variación, estén los datos agrupados o no se sigue definiendo como una medida de dispersión relativa con respecto a la media y se expresa como el cociente de la desviación típica y la media.

$$C_v \text{ muestra} = (S/\bar{X}) \times 100\% \quad C_v = (32,58/145,6) \times 100\% = 22,38\%$$

3. MEDIDAS DE RESUMEN DE LA INFORMACIÓN ESTADÍSTICA



Obtención de medidas descriptivas de resumen a partir de datos agrupados (distribución de frecuencias)

Ejercicio práctico 3.16

Con la tabla de frecuencias del ejercicio, 2.3. Calcular aproximando:

- La media, la mediana, la moda y el rango medio.
- La varianza, la desviación típica, el rango y el coeficiente de variación.
- Calcule todas las medidas obtenidas en los apartados *a* y *b* con los datos no agrupados.
- Compare las mediciones y saque tus propias conclusiones.

Ejercicio práctico 3.17

Con la tabla de frecuencias del ejercicio, 2.4. Calcular aproximando:

- La media, la mediana, la moda y el rango medio.
- La varianza, la desviación típica, el rango y el coeficiente de variación.
- Calcule todas las medidas obtenidas en los apartados *a* y *b* con los datos no agrupados.
- Compare las mediciones y saque tus propias conclusiones.

Tema 4. Estadística Bidimensional

Estadística Bidimensional



Hasta ahora hemos estudiado métodos numéricos y gráficos para estudiar el comportamiento de una variable que podía ser cualitativa o cuantitativa (discreta o continua):

- Tabla de frecuencias
- Medidas de tendencia central: Media, mediana, moda, rango medio, etc.
- Medidas de dispersión: Varianza, desviación típica, rango, etc.
- Forma: Simetría y sesgo o asimetría.

En estadística, muchas veces nos interesa describir el comportamiento conjunto de 2 o más variables.

Veamos algunos ejemplos de relacionar variables:

¿Cómo afecta una promoción de ventas las ventas de un producto?

¿Los beneficios de una empresa dependen de sus ingresos?

¿La clase social de un colectivo afecta las ventas de una determinada marca de coche?

Cuando se analizan conjuntamente dos variables hablamos de **Estadística Bidimensional**.

Tema 4. Estadística Bidimensional



Estadística Bidimensional

Ejemplo:

Un centro de idiomas de una universidad tiene el objetivo de fomentar el interés de los alumnos universitarios por los idiomas y coordinar todas las tareas relativas a su enseñanza. Se realiza una encuesta para planificar la oferta de cursos de 5 idiomas: Inglés, Alemán, Francés, Italiano y Japonés.

Los resultados de una muestra de 40 alumnos de Administración de Empresas (A), Economía (E), Ingeniería (I) y Otras Titulaciones (O), a los que se les preguntó cuál era el idioma que deseaban estudiar los resultados fueron:

(O,ALE) (I, ING) (I, ING) (I, ING) (A, FRA) (A, ING) (I, JAP) (A, FRA)
(E, ING) (I, ING) (A, FRA) (O, ING) (E, ING) (I, ING) (O, ING) (E, ING)
(I, ING) (O, ING) (A, FRA) (I, ING) (E, ITA) (O, JAP) (O, ING) (I, ING)
(I, ING) (I, ING) (I, ING) (A, FRA) (O, ING) (E, ING) (E, ITA) (A, ING)
(I, ING) (I, ITA) (I, JAP) (I, JAP) (I, ING) (A, FRA) (I, ING) (E, ING)

Tema 4. Estadística Bidimensional

Tabla de doble entrada: Frecuencias absolutas

Resumimos los datos de la encuesta anterior en una tabla. Como analizamos dos variables a la vez: **Carrera del estudiante** e **idioma** que quiere estudiar. Esta tabla se llama de doble entrada.

Carrera Idioma	Administración de Empresas	Economía	Ingeniería	Otras titulaciones	Totales
Inglés	2	5	14	5	26
Francés	6	0	0	0	6
Alemán	0	0	0	1	1
Italiano	0	2	1	0	3
Japonés	0	0	3	1	4
Totales	8	7	18	7	40

En esta tabla podemos ver que 14 estudiantes de la carrera de ingeniería estudiarán inglés; 6 estudiantes de administración de empresas estudiarán francés; ningún estudiante de economía estudiará alemán, etc.

Tema 4. Estadística Bidimensional

Tabla de doble entrada: Distribuciones marginales absolutas

Se llaman distribuciones marginales absolutas a los totales de las filas y las columnas de la tabla. En nuestro ejemplo las distribuciones marginales absolutas del idioma y la carrera serían:

Carrera Idioma	Administración de Empresas	Economía	Ingeniería	Otras titulaciones	Totales
Inglés					26
Francés					6
Alemán					1
Italiano					3
Japonés					4
Totales	8	7	18	7	40

Del idioma: Del total de 40 estudiantes encuestados, 26 estudiarán inglés; 6 francés, 1 alemán, 3 italiano y 4 japonés.

De la carrera: Del total de 40 estudiantes encuestados, 8 estudian ADE, 7 Economía, 18 Ingeniería y 7 otras titulaciones.

Tema 4. Estadística Bidimensional

Tabla de doble entrada: Distribuciones relativas

Se obtienen dividiendo la frecuencia absoluta de cada celda entre el total.

Carrera Idioma	Administración de Empresas	Economía	Ingeniería	Otras titulaciones	Totales
Inglés	0.05	0.125	0.35	0.125	0.65
Francés	0.15	0	0	0	0.15
Alemán	0	0	0	0.025	0.025
Italiano	0	0.05	0.025	0	0.075
Japonés	0	0	0.075	0.025	0.10
Totales	0.20	0.175	0.45	0.175	1

En nuestro ejemplo dividimos los valores de cada celda entre 40 (total de alumnos encuestados). En esta tabla podemos ver que el 5% de los estudiantes de Administración de empresas escogieron estudiar inglés. El 7,5% de los estudiantes de Ingeniería escogieron estudiar japonés y así sucesivamente interpretaríamos los valores de cada celda.

Tema 4. Estadística Bidimensional



Tabla de doble entrada: Distribuciones marginales relativas

Se llaman distribuciones marginales relativas a los totales en (%) de las filas y las columnas de la tabla. En nuestro ejemplo serían:

Carrera Idioma	Administración de Empresas	Economía	Ingeniería	Otras titulaciones	Totales
Inglés					0.65
Francés					0.15
Alemán					0.025
Italiano					0.075
Japonés					0.10
Totales	0.20	0.175	0.45	0.175	1

Distribución marginal del idioma: El 65% de los estudiantes estudiarán inglés, el 15% francés; el 2,5% alemán; el 7,5% italiano y el 10% japonés.

Distribución Marginal de la carrera: El 20% de los estudiantes encuestados son de ADE; el 17,5% de economía; el 45% de Ingeniería y el 17,5% de otras carreras.

Tema 4. Estadística Bidimensional

Tabla de doble entrada: Distribución condicionada

Se llama distribución condicionada al análisis bien sea de una columna o de una fila ya sean de frecuencias absolutas o frecuencias relativas. En nuestro ejemplo, supongamos que solo nos interesa analizar el idioma que estudiarían los estudiantes de ingeniería (columnas):

Carrera Idioma	Ingeniería
Inglés	14
Francés	0
Alemán	0
Italiano	1
Japonés	3
Totales	18

Carrera Idioma	Ingeniería
Inglés	$14/18=0.78$
Francés	0
Alemán	0
Italiano	$1/18=0.06$
Japonés	$3/18=0.17$
Totales	1

14 de los 18 estudiantes encuestados de ingeniería estudiarán inglés. O, El 78% de los estudiantes de ingeniería estudiarán inglés.

Tema 4. Estadística Bidimensional



Tabla de doble entrada: Distribución condicionada

Supongamos que ahora nos interesa analizar los alumnos de las diferentes carreras que se apuntarían a inglés (fila):

Carrera Idioma	Administración de Empresas	Economía	Ingeniería	Otras titulaciones	Totales
Inglés	2	5	14	5	26

Carrera Idioma	Administración de Empresas	Economía	Ingeniería	Otras titulaciones	Totales
Inglés	$2/26 = 0.08$	$5/26 = 0.19$	$14/26 = 0.54$	$5/26 = 0.19$	26

De los 26 alumnos que estudiarán inglés, 5 son de la carrera de Economía.
El 19% de los alumnos que estudiarán inglés son estudiantes de Economía.

Tema 4. Estadística Bidimensional



Tabla de doble entrada: Distribución condicionada a partir de las frecuencias relativas

La información anterior se puede obtener a partir de las frecuencias relativas, dividiendo todas las celdas de la fila o la columna, por el total relativo:

Carrera Idioma	Administración de Empresas	Economía	Ingeniería	Otras titulaciones	Totales
Inglés	0.05	0.125	0.35	0.125	0.65

Carrera Idioma	Administración de Empresas	Economía	Ingeniería	Otras titulaciones	Totales
Inglés	$0.05/0.65=0.08$	$0.125/0.65=0.19$	$0.35/0.65=0.54$	$0.125/0.65=0.19$	0.65

Del 65% de los alumnos que estudiarán inglés, el 54% son de la carrera de Ingeniería.

Tema 4. Estadística Bidimensional

Formulación general:

Supongamos que observamos los valores de dos variables **X** e **Y** con **I** y **J** categorías o valores respectivamente ($X_1, X_2, X_3, \dots, X_I$; $Y_1, Y_2, Y_3, \dots, Y_J$) en una muestra de **n** elementos, **La tabla de doble entrada** de frecuencias absolutas es la siguiente:

		Y					
X	Y \ X	Y ₁	Y ₂	Y ₃		Y _J	
	X ₁	n _{1,1}	n _{1,2}	n _{1,3}		n _{1,j}	n_{1.}
	X ₂	n _{2,1}	n _{2,2}	n _{2,3}		n _{2,j}	n_{2.}
	X ₃	n _{3,1}	n _{3,2}	n _{3,3}		n _{3,j}	n_{3.}
							
	X _I	n _{i,1}	n _{i,2}	n _{i,3}		n _{i,j}	n_{i.}
		n.₁	n.₂	n.₃		n._J	n

Donde:

$n_{i,j}$ es el número de veces que aparece la pareja (X_i, Y_j) $i = 1, \dots, I$; $j = 1, \dots, J$.

$n_{i.}$ son las frecuencias marginales de X_i ; $i = 1, \dots, I$

$n_{.j}$ son las frecuencias marginales de Y_j $j = 1, \dots, J$

Tema 4. Estadística Bidimensional

Formulación general

		Y					
		Y ₁	Y ₂	Y ₃		Y _J	
X	X ₁	f _{1,1}	f _{1,2}	f _{1,3}		f _{1,j}	f_{1.}
	X ₂	f _{2,1}	f _{2,2}	f _{2,3}		f _{2,j}	f_{2.}
	X ₃	f _{3,1}	f _{3,2}	f _{3,3}		f _{3,j}	f_{3.}
	X _I	f _{i,1}	f _{i,2}	f _{i,3}		f _{i,j}	f_{i.}
		f_{.1}	f_{.2}	f_{.3}		f_{.J}	1

La distribución condicionada de Y dado X = X_i, es:

$$f(Y = y_1) / (X = x_i) = \frac{f_{i,1}}{f_i}$$

$$f(Y = y_2) / (X = x_i) = \frac{f_{i,2}}{f_i}$$

.....

$$F(Y = y_J) / (X = X_i) = \frac{f_{i,J}}{f_i}$$

Tema 4. Estadística Bidimensional

Formulación General

Tabla de doble entrada de frecuencias relativas:

La distribución condicionada de X dado Y = y_j, es:

$$f(X = x_1) / (Y = y_j) = \frac{f_{1,j}}{f_j}$$

$$f(X = x_2) / (Y = y_j) = \frac{f_{2,j}}{f_j}$$

.....

$$f(X = x_I) / (Y = y_j) = \frac{f_{i,j}}{f_j}$$

Tema 4. Estadística Bidimensional



Ejercicio práctico 4.1

El departamento de Investigación de Mercados de una fábrica de chocolates realizó un estudio con 424 consumidores para medir la valoración que estos daban a un nuevo chocolate. Se utilizó una escala de intervalos para medir la preferencia, como la que se muestra a continuación.

¿Después de haber probado el chocolate, ¿Cómo califica el sabor de este?

- 5 Muy bueno
- 4 Bueno
- 3 Regular
- 2 Malo
- 1 Muy malo

Se consideraron para el análisis tres variables:

Variable 1, tamaño de la población: Hábitat urbanos (ciudades > 50000 habitantes); hábitat rurales (poblaciones con < 50000 habitantes).

Variable 2, tamaño del hogar: uno o dos miembros; tres o más miembros.

Variable 3, edad de la ama de casa: 34 años o menos; 35 años o más.

Tema 4. Estadística Bidimensional

Ejercicio práctico 4.1

Los resultados del muestreo se muestran en las siguientes tablas:

Tabla 1

Hábitat Calific.		Urbano	Rural
Muy bueno	5	44	112
Bueno	4	33	53
Regular	3	19	14
Malo	2	25	62
Muy malo	1	18	44

Tabla 2

Hogar Calific.		1 o 2	3 o +
Muy bueno	5	126	30
Bueno	4	37	49
Regular	3	18	15
Malo	2	45	42
Muy malo	1	41	21

Tema 4. Estadística Bidimensional

Ejercicio práctico 4.1

Los resultados del muestreo se muestran en las siguientes tablas:

Tabla 3

Calific. \ Edad		34 o menos	35 o más
Muy bueno	5	44	112
Bueno	4	32	54
Regular	3	13	20
Malo	2	21	66
Muy malo	1	23	39

Tema 4. Estadística Bidimensional



Ejercicio práctico 4.1

Se te pide:

- Calcular las frecuencias marginales absolutas para cada una de las tablas. Interpretar los resultados.
- Calcular las frecuencias marginales relativas para cada una de las tablas. Interpretar los resultados.
- Calcular la frecuencias condicionadas de las columnas de cada una de las tablas. Interpretar los resultados.
- Analizando conjuntamente todos los cálculos anteriores, ¿Qué conclusiones podemos obtener del estudio?
- ¿Qué otras medidas o características de este conjunto de datos podríamos obtener para un más profundo análisis?

Tema 4. Estadística Bidimensional



Variables discretas

Con dos variables discretas, de igual modo, se pueden construir las distribuciones conjuntas, marginales y condicionadas.

Ejemplo:

La siguiente tabla muestra el número de veces (X) que una muestra de 50 estudiantes de Economía han tenido que repetir la asignatura de Estadística hasta aprobarla y el número de años (Y) que han tardado en acabar la carrera.

$\begin{smallmatrix} Y \\ \diagdown \\ X \end{smallmatrix}$	4	5	6	7
0	15	5	3	2
1	4	8	2	1
2	0	2	1	3
3	0	0	0	4

15 estudiantes de economía que aprobaron a la primera el curso de estadística terminaron la carrera en 4 años.

Tema 4. Estadística Bidimensional

Variables discretas

Distribución de frecuencias absolutas (conjuntas y marginales)

X \ Y	4	5	6	7	Total
0	15	5	3	2	25
1	4	8	2	1	15
2	0	2	1	3	6
3	0	0	0	4	4
Total	19	15	6	10	50

25 estudiantes aprobaron el curso de estadística a la primera; 15 la repitieron una vez, 6 dos veces y cuatro tres veces.

19 estudiantes terminaron la carrera en 4 años; 15 en cinco años; 6 en seis años y 10 en siete años.

Tema 4. Estadística Bidimensional

Variables discretas

Distribución de frecuencias relativas (conjuntas y marginales)

$\begin{matrix} Y \\ X \end{matrix}$	4	5	6	7	Total
0	0.3	0.1	0.06	0.04	0.5
1	0.08	0.16	0.04	0.02	0.3
2	0	0.04	0.02	0.06	0.12
3	0	0	0	0.08	0.08
Total	0.38	0.3	0.12	0.2	1

El 50% de los estudiantes aprobaron el curso de estadística a la primera; el 30% la repitieron una vez; el 12% dos veces y el 8% tres veces. El 38% de los estudiantes terminaron la carrera en 4 años; el 30% en cinco años; el 12% en seis años y el 20% en siete años.

Tema 4. Estadística Bidimensional



Variables discretas

Distribución de frecuencias relativas (conjuntas y marginales)

¿Cuál es la distribución del número de años que tarda en terminar la licenciatura suponiendo que un estudiante tiene que repetir dos veces la asignatura de Estadística?

	4	5	6	7
Tenemos que calcular: $f(Y/X = 2)$:	-----			
	0	0.33	0.17	0.5

Podemos calcular la **media** condicionada de $Y|X=2$ a través de esta tabla:

$$(4 \times 0) + (5 \times 0.33) + (6 \times 0.17) + (7 \times 0.5) = 6.17$$

La **varianza** condicionada es:

$$(4^2 \times 0 + 5^2 \times 0.33 + 6^2 \times 0.17 + 7^2 \times 0.5) - (6.17)^2 = 0.806$$

Y la **desviación típica** condicionada: $\sqrt{0.806} \approx 0.898$

Tema 4. Estadística Bidimensional



Variables discretas

Distribución de frecuencias relativas (conjuntas y marginales)

Si comparamos la distribución de $Y|X=2$ con la distribución marginal de Y ,

Y	4	5	6	7
$f(Y X=2)$	0	0.33	0.17	0.5

Marginal de Y :

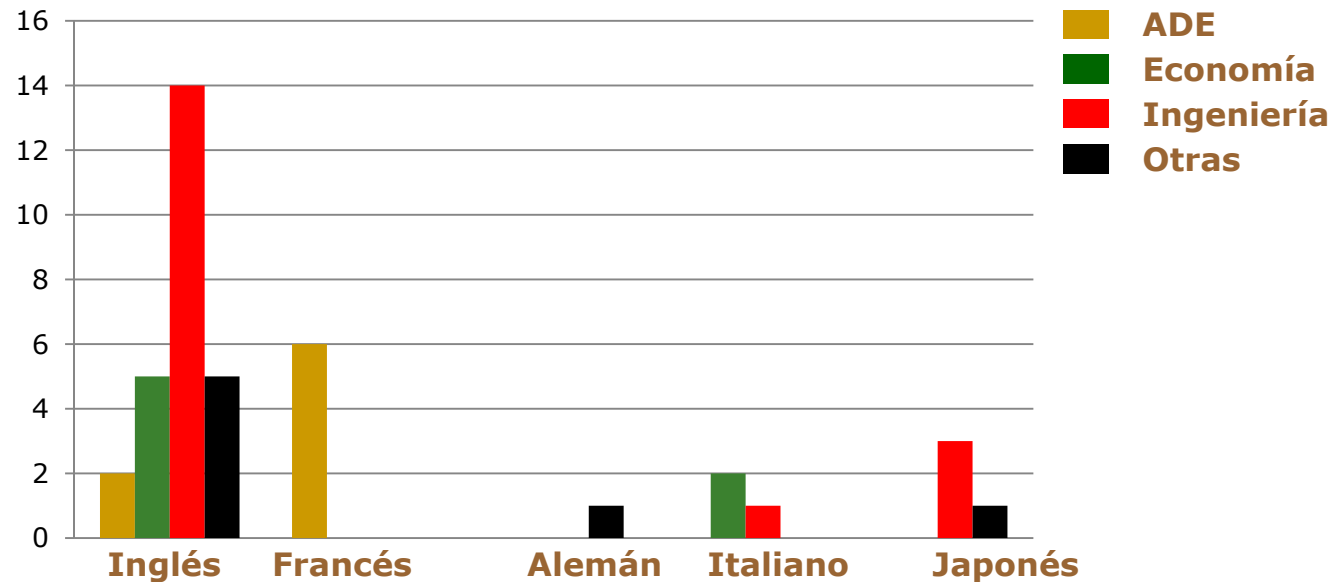
Y	4	5	6	7
$f_{.j}$	0.38	0.3	0.12	0.2

Entonces, vemos que dependiendo del valor de la variable X , así será la distribución de la variable Y .

Esto nos da una idea de **dependencia** entre las variables X e Y .

Tema 4. Estadística Bidimensional

Diagrama de barras bidimensional para el ejemplo carreras e idiomas que elegirían los estudiantes



Relación entre variables

Cuando estudiamos dos variables de forma simultánea en los negocios, podemos ver si existe o no alguna relación entre dichas variables.

Ejemplos:

¿Existe alguna relación entre la experiencia profesional de un trabajador y el sueldo que cobra?

¿Existe alguna relación entre el número de facings de un producto en un lineal y las ventas del mismo?

¿Existe alguna relación entre la producción agrícola y la cantidad de fertilizantes utilizados? Etc.

En este análisis de dependencia definiremos la variable Dependiente o efecto como **Y**. Las variables independiente o causa como **X**.

Tema 4. Estadística Bidimensional



Relación entre variables

Esta dependencia a la que se hace referencia es relacional matemática y no necesariamente de causalidad. Así, para un mismo número de unidades producidas, pueden existir niveles de coste, que varían empresa a empresa.

Si se da ese tipo de relaciones, se suele recurrir a los estudios de regresión en los cuales se obtiene una nueva relación pero de un tipo especial denominado función, en la cual la variable independiente se asocia con un indicador de tendencia central de la variable dependiente. Cabe recordar que en términos generales, una función es un tipo de relación en la cual para cada valor de la variable independiente le corresponde uno y sólo un valor de la variable dependiente.

Regresión simple y Correlación. Aspectos teóricos

El análisis de Regresión: Se utiliza para fines de predicción.

El análisis correlación: Se utiliza para medir la fuerza de asociación entre las dos variables.

Tema 4. Estadística Bidimensional



Relación entre variables

Regresión simple y Correlación. Aspectos teóricos

Muchos estudios se basan en la creencia de que es posible identificar y cuantificar alguna Relación Funcional entre dos o más variables, donde una variable depende de la otra variable.

Se puede decir que **Y** depende de **X**, en donde Y y X son dos variables cualquiera en un modelo de Regresión Simple.

"Y es una función de X"

$$Y = f(X)$$

Como Y depende de X,
Y es la variable **dependiente**, y
X es la variable **independiente**.

En el Modelo de Regresión es muy importante identificar cuál es la variable dependiente y cuál es la variable independiente.

Tema 4. Estadística Bidimensional



Relación entre variables

Regresión simple y Correlación. Aspectos teóricos

En el Modelo de Regresión Simple se establece que Y es una función de sólo una variable independiente, razón por la cual se le denomina también Regresión Bivariada porque sólo hay dos variables, una dependiente y otra independiente y se representa así:

$$Y = f(X)$$

"Y esta en función de X" o dicho de otra manera los valores que toma Y depende de los valores que tome X

La variable dependiente es la variable que se desea explicar, predecir. También se le llama **VARIABLE EFECTO**.

La variable Independiente X se le denomina **VARIABLE EXPLICATIVA** ó y se le utiliza para EXPLICAR Y.

Tema 4. Estadística Bidimensional

Relación entre variables



Regresión simple y Correlación. Diagrama de dispersión

En el análisis de regresión que incluya a una variable independiente X y una variable dependiente Y , los valores individuales se trazan en una grafica de dos dimensiones llamada diagrama de dispersión. Cada valor se grafica en sus coordenadas particulares X y Y .

Ejemplo:

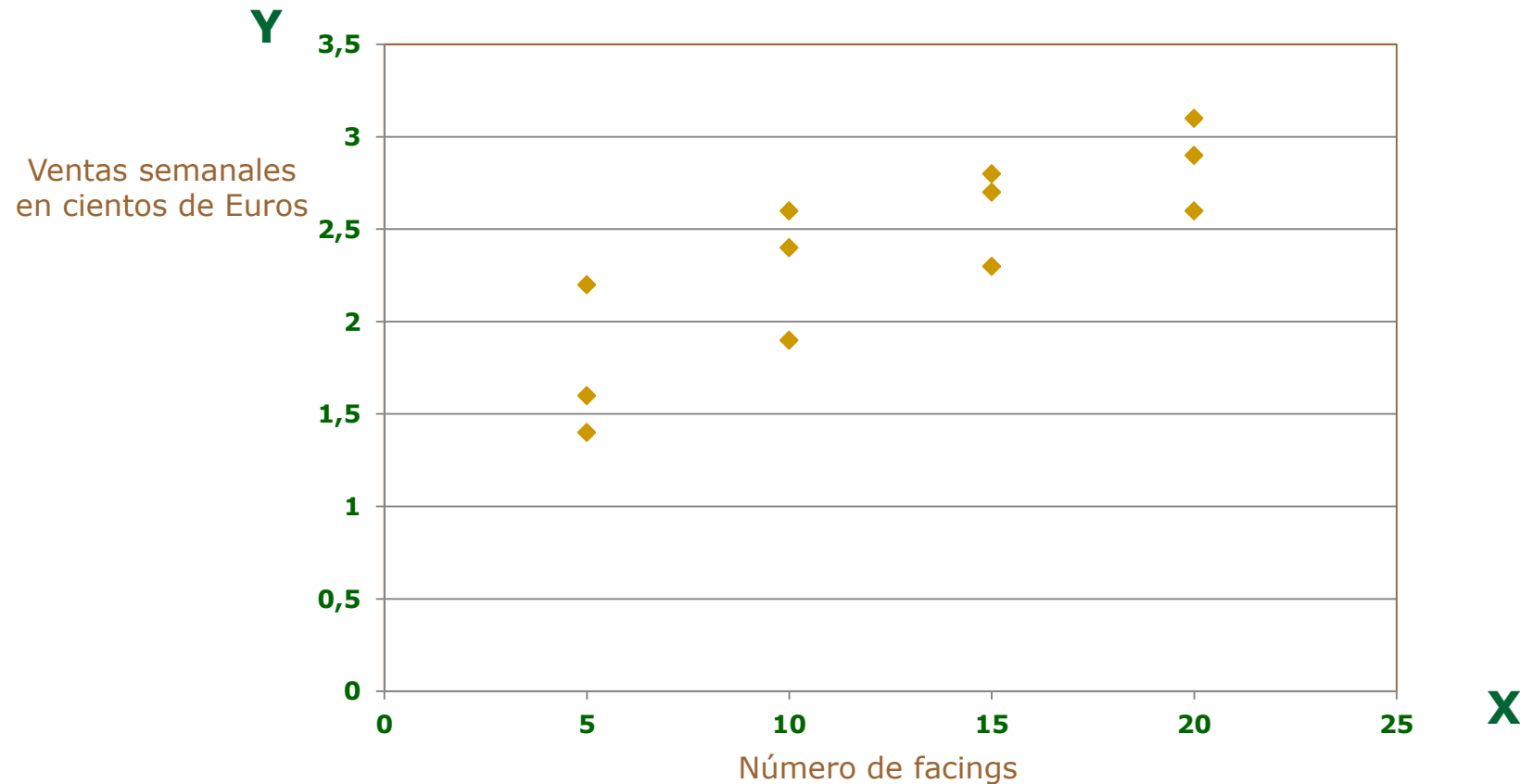
El Brand Manager de una cadena de supermercados quiere determinar si el número de facings de bricks de leche influye en las ventas del producto en de los supermercados. Se seleccionó una muestra de 12 tiendas de igual tamaño y los resultados de la ventas semanales en cientos (00) de euros, fueron los siguientes:

Tienda	1	2	3	4	5	6	7	8	9	10	11	12
Facings X	5	5	5	10	10	10	15	15	15	20	20	20
Ventas (00) Euros Y	1.6	2.2	1.4	1.9	2.4	2.6	2.3	2.7	2.8	2.6	2.9	3.1

Tema 4. Estadística Bidimensional

Relación entre variables

Regresión simple y Correlación. Diagrama de dispersión



Tema 4. Estadística Bidimensional



Relación entre variables

Regresión simple. Determinación de la ecuación de regresión.

Método de los mínimos cuadrados

$$\hat{Y}_i = b_0 + b_1 X_i$$

Donde:

$\hat{Y}_i$: Es el valor predicho de **Y** para la observación **i** y **X_i** es el valor de **X** para la observación **i**.

Esta ecuación requiere la determinación de dos coeficientes **b₀** (la ordenada en el origen) y **b₁** (la pendiente) con el fin de predecir los valores de **Y**.

Una vez obtenidos **b₀** y **b₁** se conoce la línea recta y se traza en el diagrama de dispersión.

Visualmente podemos observar que tanto se ajuste esta línea recta a los datos originales.

Tema 4. Estadística Bidimensional

Relación entre variables



Regresión simple. Determinación de la ecuación de regresión.

El análisis de regresión lineal simple tiene por objeto encontrar la línea recta que mejor se ajuste a los datos. Es decir:

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Donde: Y_i = Valor real de Y para la observación i

$\hat{Y}_i$ = Valor predicho de Y para la observación i

Puesto que $\hat{Y}_i = b_0 + b_1x_i$, se simplifica en

$$\sum_{i=1}^n [Y_i - (b_0 + b_1x_i)]^2$$

Esta ecuación tiene dos incógnitas b_0 y b_1 . La técnica matemática que determina los valores b_0 y b_1 que mejor se ajusten a los datos observados se conoce como el **método de mínimos cuadrados**.

Tema 4. Estadística Bidimensional

Relación entre variables



Regresión simple. Determinación de la ecuación de regresión.

Cualquier valor para b_0 y b_1 que no sean los determinados mediante el método de mínimos cuadrados dará como resultado una suma mayor de las diferencias al cuadrado entre el valor real de Y y el valor predicho de Y .

Con el método de mínimos cuadrados se obtienen las dos ecuaciones siguientes, llamadas ecuaciones normales:

$$\text{I.} \quad \sum_{i=1}^n Y_i = nb_0 + b_1 \sum_{i=1}^n X_i$$

$$\text{II.} \quad \sum_{i=1}^n X_i Y_i = b_0 \sum_{i=1}^n X_i + b_1 \sum_{i=1}^n X_i^2$$

Como hay dos ecuaciones con dos incógnitas se solucionan simultáneamente para b_0 y b_1 en la forma siguiente:

Tema 4. Estadística Bidimensional

Relación entre variables



Regresión simple. Determinación de la ecuación de regresión.

$$b_1 = \frac{n \sum_{i=1}^n X_i Y_i - \left(\sum_{i=1}^n X_i \right) \left(\sum_{i=1}^n Y_i \right)}{n \sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i \right)^2}$$

$$b_0 = \bar{Y} - b_1 \bar{X}$$

Donde:

$$\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n} \quad y \quad \bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

Tema 4. Estadística Bidimensional

Relación entre variables



Regresión simple. Determinación de la ecuación de regresión. O, bien:

$$b_1 = \frac{\sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n X_i \right) \left(\sum_{i=1}^n Y_i \right)}{n}}{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i \right)^2}{n}} \quad y$$

$$b_0 = \frac{\sum_{i=1}^n Y_i}{n} - b_1 \frac{\sum_{i=1}^n X_i}{n}$$

Tema 4. Estadística Bidimensional

Relación entre variables



Regresión simple. Determinación de la ecuación de regresión.

Al examinar las ecuaciones anteriores observamos que tenemos que calcular cinco cantidades para calcular b_0 y b_1 . Estas son:
 n , tamaño de la muestra;

$\sum_{i=1}^n X_i$ La suma de los valores de X

$\sum_{i=1}^n Y_i$ La suma de los valores de Y

$\sum_{i=1}^n X_i^2$ La suma del cuadrado de los valores de X y

$\sum_{i=1}^n X_i Y_i$ La suma del producto cruzado de X y Y

Tema 4. Estadística Bidimensional

Relación entre variables

En nuestro ejemplo, los cálculos a realizar serían:

	Facings X	Ventas sem. (00) Euros Y	X ²	Y ²	X*Y
	5	1,6	25	2,56	8
	5	2,2	25	4,84	11
	5	1,4	25	1,96	7
	10	1,9	100	3,61	19
	10	2,4	100	5,76	24
	10	2,6	100	6,76	26
	15	2,3	225	5,29	34,5
	15	2,7	225	7,29	40,5
	15	2,8	225	7,84	42
	20	2,6	400	6,76	52
	20	2,9	400	8,41	58
	20	3,1	400	9,61	62
Sumatorias Σ	150	28,5	2250	70,69	384
Medias	12,50	2,38			

Tema 4. Estadística Bidimensional

Relación entre variables



Regresión simple. Determinación de la ecuación de regresión en el ejemplo:

$$b_1 = \frac{n \sum_{i=1}^n X_i Y_i - \left(\sum_{i=1}^n X_i \right) \left(\sum_{i=1}^n Y_i \right)}{n \sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i \right)^2}$$

$$b_1 = \frac{12 (384) - 150 (28,5)}{12 (2,250) - (150)^2} = \mathbf{0.074}$$

$$b_0 = \bar{Y} - b_1 \bar{X}$$

$$b_0 = 2,38 - 0.074 (12,5) = \mathbf{1,455}$$

Tema 4. Estadística Bidimensional

Relación entre variables



Ecuación de regresión del ejemplo

$$\hat{Y}_i = 1,455 + 0.074x_1$$

La pendiente **b₁** se calculó como **+0.074**. Esto significa que por cada aumento de un facing (un brick) en el lineal, las ventas de leche aumentan en **7,4** Euros.

La ordenada en el origen **b₀** se calculó como **+1,455**. Esto representa el valor de Y cuando X = 0. Esto significa que si la leche no tuviera ningún facing en el lineal las ventas de la misma sería de **145,5** euros.

Tema 4. Estadística Bidimensional

Relación entre variables

Cálculo de la recta de regresión:

Facings X	Ventas sem.Y	Y [^]
5	1,6	1,83
5	2,2	1,83
5	1,4	1,83
10	1,9	2,20
10	2,4	2,20
10	2,6	2,20
15	2,3	2,57
15	2,7	2,57
15	2,8	2,57
20	2,6	2,94
20	2,9	2,94
20	3,1	2,94

$$\hat{Y}_i = 1,455 + 0.074x_1$$

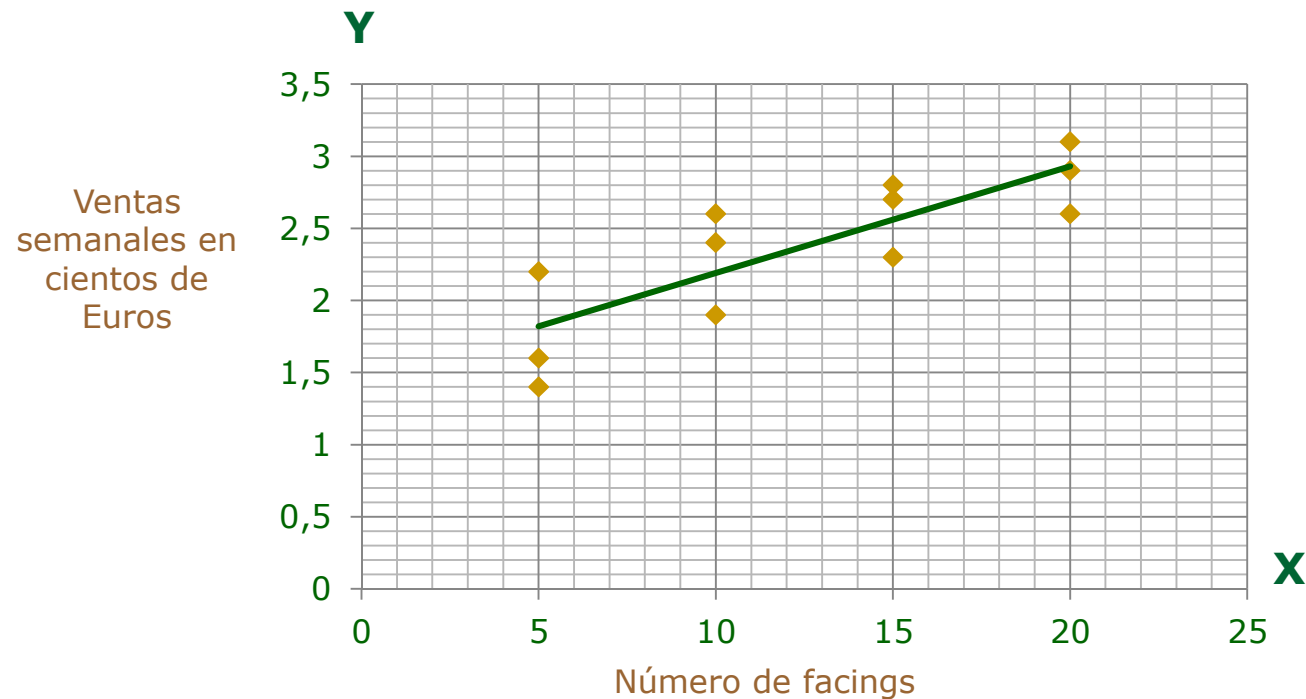


Recta de regresión

Tema 4. Estadística Bidimensional

Relación entre variables

En nuestro ejemplo, los cálculos a realizar serían:



Tema 4. Estadística Bidimensional

Relación entre variables



Predicciones en el análisis de regresión del ejemplo

¿A cuánto ascenderían las ventas de leche si se tienen 8 facings del producto en el lineal?

$$\hat{Y}_i = 1,455 + 0.074 (8) = 2,047$$

Si se tienen 8 facings en el lineal del supermercado las ventas de leche ascenderían a 204,7 euros.

Al usar un modelo de regresión para fines de predicción, solo podemos tomar en cuenta el rango pertinente de la variable independiente. Este rango incluye todos los valores de X desde el más pequeño hasta el más grande. En nuestro ejemplo desde 5 facings hasta 20 facings. Por consiguiente cuando se predice un valor de Y para un determinado valor de X, se puede interpolar dentro de este rango pertinente de los valores de X, pero no se puede extrapolar más allá del rango de los valores de X. En nuestro ejemplo no podríamos predecir las ventas en los supermercados que tuvieran más de 20 facings o menos de 5 facings.

Tema 4. Estadística Bidimensional

Relación entre variables

Errores o residuos de la predicción

Los errores de predicción e_i se definen como:

$$e_i = Y_i - \hat{Y}_i \quad i = 1, \dots, n$$

Podemos comprobar que la media de los residuos siempre es igual a 0

X_i	Y_i	$\hat{Y}_i$	e_i
5	1,6	1,83	-0,23
5	2,2	1,83	0,38
5	1,4	1,83	-0,43
10	1,9	2,20	-0,30
10	2,4	2,20	0,21
10	2,6	2,20	0,41
15	2,3	2,57	-0,27
15	2,7	2,57	0,14
15	2,8	2,57	0,24
20	2,6	2,94	-0,34
20	2,9	2,94	-0,04
20	3,1	2,94	0,17

Tema 4. Estadística Bidimensional

Relación entre variables



Ejercicio práctico 4.2

El gerente de personal de una empresa intuye que quizás haya relación entre el ausentismo y la edad y quería tomar la edad de un trabajador para desarrollar un modelo de predicción de días de ausencia durante un año laboral. Se seleccionó una muestra aleatoria de 10 trabajadores con los resultados que se presentan a continuación

Trabajador	Edad en años X	Días de Ausentismo Y
1	27	15
2	61	6
3	37	10
4	23	18
5	46	9
6	58	7
7	29	14
8	36	11
9	64	5
10	40	8

Se te pide:

- Elaborar el diagrama de dispersión.
- Observando el diagrama. A simple vista ¿Se puede observar alguna relación entre la edad y los días de ausentismo de los trabajadores de la empresa?
- Utilice el método de los mínimos cuadrados para calcular los coeficientes de regresión b_0 y b_1 .
- Interprete el significado de la pendiente b_1 en este problema.
- Prediga los días de ausencia de dos trabajadores: uno de 28 años de edad y otro de 59 años de edad.
- Calcule los valores de la recta de regresión y graficar sobre el diagrama de dispersión.

Tema 4. Estadística Bidimensional

Relación entre variables

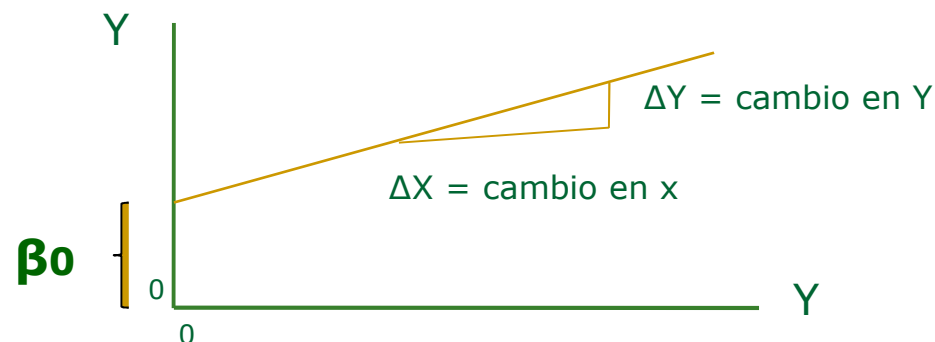
Tipos de modelos de regresión en la población

En la población la naturaleza de la relación puede tomar muchas formas, que van desde funciones matemáticas sencillas hasta otras más complicadas. La relación más sencilla es la relación en línea recta o lineal. El modelo lineal de regresión en la población se representa como:

$$Y_i = \beta_0 + \beta_1 X_1 + \epsilon_i$$

Donde:

- β_0 = Ordenada en el origen de la población
- β_1 = Pendiente real de la población
- ϵ_i = Error aleatorio en Y para la observación i

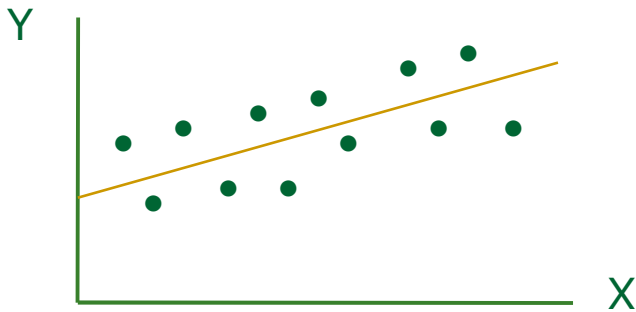


Representación gráfica de una relación lineal positiva

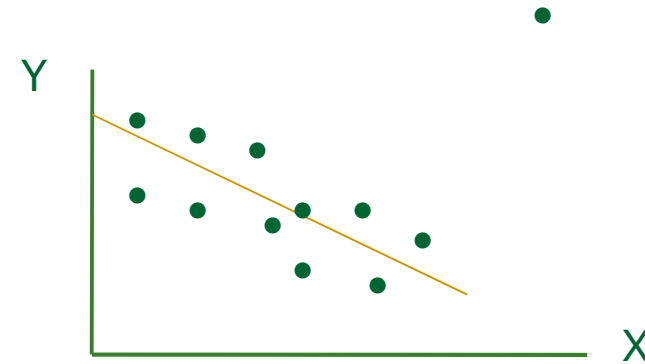
Tema 4. Estadística Bidimensional

Relación entre variables

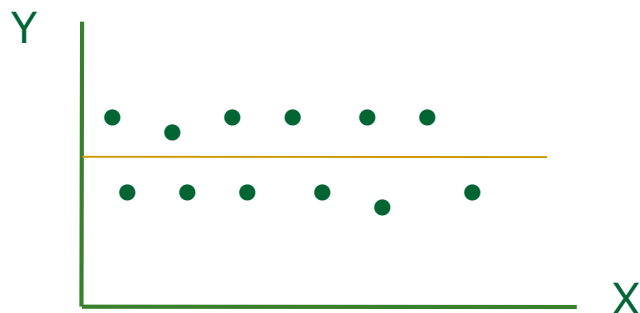
Tipos de modelos de regresión en la población



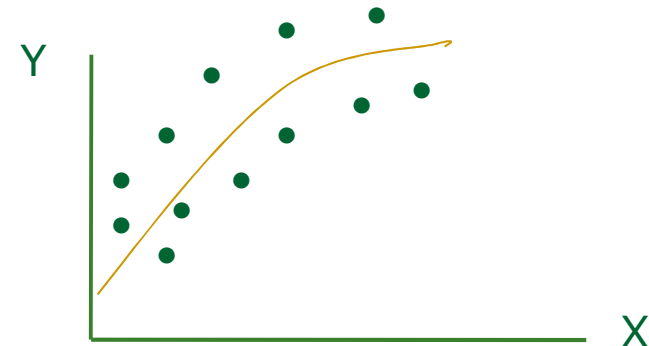
Relación lineal positiva



Relación lineal negativa



Ninguna relación entre X y Y

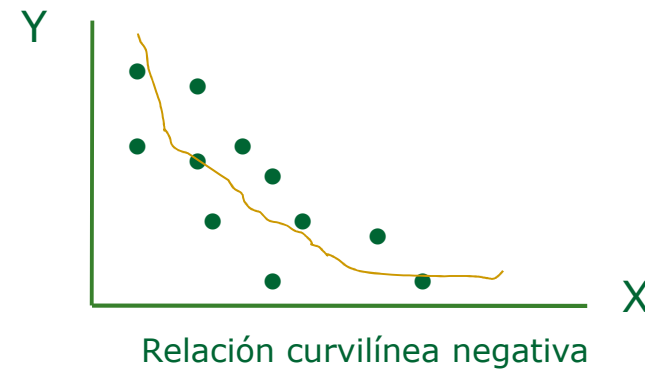
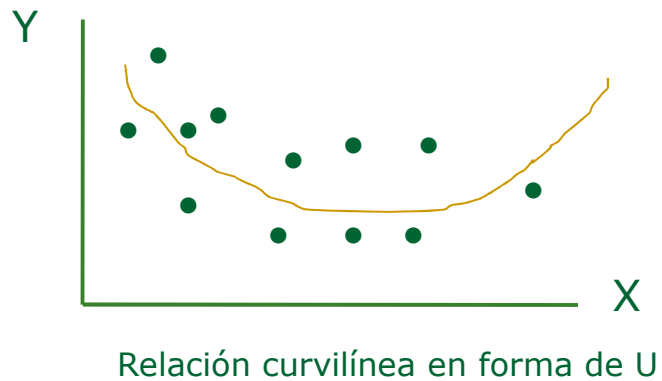


Relación curvilínea positiva

Tema 4. Estadística Bidimensional

Relación entre variables

Tipos de modelos de regresión en la población



Tema 4. Estadística Bidimensional

Relación entre variables



La covarianza y el coeficiente de correlación

En el ejemplo que venimos manejando se ve gráficamente que existe una relación positiva (ó creciente) y aproximadamente lineal entre las ventas de bricks de leche y el número de facings de los lineales de los supermercados.

La covarianza es una medida de la fuerza de la relación lineal entre dos variables cuantitativas. Si X y Y son independientes, la covarianza es 0.

Definición de covarianza:

Para una muestra de n datos bivariantes, (X_1, Y_1) , (X_2, Y_2) ,..., (X_n, Y_n) , la covarianza entre las variables X y Y es:

$$S_{xy} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}) (Y_i - \bar{Y})$$

Tema 4. Estadística Bidimensional

Relación entre variables



La covarianza

Que también puede expresarse como:

$$S_{xy} = \frac{1}{n} \left(\sum_{i=1}^n X_i Y_i \right) - \bar{X} \bar{Y}$$

Correlación

Si, en nuestro ejemplo las unidades de la variable X son Bricks que miden 5 centímetros cada uno y las unidades de la variable Y son euros, entonces, si cambiamos la escala de las unidades, cambia la covarianza. Esto hace que el valor de la covarianza sea difícil de interpretar. ¿Cómo decir si la covarianza es grande o pequeña? Para evitar que la covarianza dependa de sus unidades de medida, se calcula una medida alternativa, la **correlación**.

Tema 4. Estadística Bidimensional



Relación entre variables

Definición de correlación

Para una muestra de n datos bivariantes, (X_1, Y_1) , (X_2, Y_2) ,..., (X_n, Y_n) , la correlación entre las variables X y Y es:

$$r_{xy} = \frac{S_{xy}}{S_x S_y}$$

donde S_x e S_y son las desviaciones típicas de X y Y , respectivamente.

Tema 4. Estadística Bidimensional

Relación entre variables



Cálculo de la covarianza y el coeficiente de correlación

Volviendo al ejemplo que venimos manejando, calculamos la covarianza:

$$S_{xy} = \frac{1}{n} \left(\sum_{i=1}^n X_i Y_i \right) - \bar{X} \bar{Y}$$

$$S_{xy} = \frac{1}{12} (384) - (12,5 \times 2,38) = \mathbf{2,25}$$

Calculamos la varianza S^2 y la desviación típica S tanto para X como para y .

$$S_x^2 = 34,091 \quad S_x = \mathbf{5,84}$$

$$S_y^2 = 0,2729 \quad S_y = \mathbf{0,522}$$

Tema 4. Estadística Bidimensional



Relación entre variables

Cálculo de la covarianza y el coeficiente de correlación:

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{2,25}{5,84 \times 0,522} = \mathbf{0.7381 \text{ o } 73,81\%}$$

Como observamos a simple vista en el diagrama de dispersión y ahora con el cálculo del coeficiente de correlación, vemos que si existe una correlación positiva lineal bastante alta.

Interpretación del coeficiente de correlación:

Indica el porcentaje de variación en la variable Y que es explicado por un cambio unitario en la variable independiente X.

En nuestro ejemplo:

El 73,81% del aumento de las ventas de leche en los supermercados se debe al aumento de una unidad de facing en el lineal. Esto indica una correlación bastante alta ya que solo en el 21,69% de los casos de aumento de las ventas se debe a otros factores.

Tema 4. Estadística Bidimensional



Relación entre variables

Propiedades de la correlación:

- No depende de las unidades de medida.
- $-1 \leq r_{XY} \leq 1$
- Si $r_{XY} = 1$, existe una relación lineal positiva perfecta entre X y Y.
- Si $r_{XY} = -1$, existe una relación lineal negativa perfecta entre X y Y.
- Si no existe ninguna relación entre X y Y, $r_{XY} = 0$.
- Si $r_{XY} = 0$, no existe relación lineal entre X y Y, pero puede existir entre ellas otro tipo de relación entre ellas (cuadrática, cúbica, etc..)
- Si la correlación está cerca de -1 ó 1, entonces hay una relación (negativa ó positiva) entre ambas variables aproximadamente lineal.

Tema 4. Estadística Bidimensional



Relación entre variables

Ejercicio 4.3

Con los datos del ejercicio 4.2, calcular:

- a. La covarianza.
- b. El coeficiente de correlación
- c. ¿Qué significado tiene el coeficiente de correlación en este caso?

Tema 4. Estadística Bidimensional



Relación entre variables

Ejercicio 4.4

EL departamento de producción de una fábrica de colchones, quiere probar la elasticidad de un tipo de muelle utilizado en la fabricación de una marca concreta. Con este objetivo, se sometió a un muelle a varios niveles de fuerza (en Newtons) y se midió la extensión total del muelle (en mm) en cada caso. Los resultados fueron los siguientes:

Fuerza X	0.1	0.1	0.2	0.2	0.3	0.3	0.4	0.4	0.5	0.5
Extensión Y	18	11	25	22	35	50	54	45	52	68

Se te pide:

- Elaborar el diagrama de dispersión.
- Observando el diagrama. A simple vista ¿Se puede observar alguna relación entre la fuerza y la extensión alcanzada por el muelle?
- Utilice el método de los mínimos cuadrados para calcular los coeficientes de regresión b_0 y b_1 .
- Interprete el significado de la pendiente b_1 en este problema.
- Prediga la extensión de un muelle sometido a una resistencia de 0.35 y otro de 0.45.
- Calcule los valores de la recta de regresión y graficar sobre el diagrama de dispersión.
- Calcular el coeficiente de correlación e indicar lo que significa en este problema.

Variables aleatorias

Distribuciones de probabilidad para una variable aleatoria discreta

Distribución de probabilidad teórica de los resultados de lanzar un dado:

Cara del resultado	Probabilidad
1	$1/6$
2	$1/6$
3	$1/6$
4	$1/6$
5	$1/6$
6	$1/6$
Total	1

Dos distribuciones discretas muy importantes en los negocios son: La distribución binomial y la distribución de Poisson

Distribuciones de probabilidad para una variable aleatoria discreta

Distribución de probabilidad Binomial

Tiene cuatro propiedades esenciales:

1. En cada prueba del experimento sólo son posibles dos resultados: el suceso **A** (éxito) y su contrario **A'** (fracaso).
2. El resultado obtenido en cada prueba es independiente de los resultados obtenidos anteriormente.
3. La probabilidad del suceso A es constante, la representamos por **p**, y no varía de una prueba a otra. La probabilidad de **A'** es **1 - p** y la representamos por **q**.
4. El experimento consta de un número **n** de pruebas.

Aplicaciones:

En juegos de azar:

- ¿Qué probabilidad hay de que aparezca el color rojo 15 o más veces en 19 giros de la rueda de la ruleta?

TEMA 5. INTRODUCCIÓN A LA INFERENCIA



Variables aleatorias

Distribuciones de probabilidad para una variable aleatoria discreta

Distribución de probabilidad Binomial

Aplicaciones:

En control de calidad de un producto:

- ¿Qué probabilidad hay de que en una muestra de 20 llantas del mismo tipo ninguna esté defectuosa, si el 8% de todas las llantas producidas en cierta planta son defectuosas?

En la educación:

- ¿Qué probabilidad tiene un estudiante de aprobar un examen de 10 preguntas de opción múltiple (cada una de ellas de cuatro opciones) si adivina en cada pregunta? (Aprobar se define como lograr correcto el 60% de las preguntas, es decir acertar por lo menos 6 de cada 10)

En finanzas:

- ¿Cuál es la probabilidad de que cierta acción mostrará un aumento en su precio al cierre, en una base diaria durante 10 sesiones (consecutivas) de operaciones, si en realidad los cambios en el mercado accionario son aleatorios?

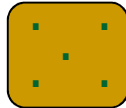
Variables aleatorias

Distribuciones de probabilidad para una variable aleatoria discreta

Distribución de probabilidad Binomial

Desarrollo del modelo matemático

- ¿Qué probabilidad hay de que en tres lanzamientos sucesivos de un dado salga la cara 5?



n = Número lanzamientos sucesivos 3

p = Probabilidad que salga 5, $p = 1/6$

$1 - p = q$ = Probabilidad de que no salga 5 $1 - p$ ó $q = 5/6$

¿Qué ocurre? Ninguno de los lanzamientos sea 5, en un lanzamiento salga un cinco, en 2 salga 5, en tres lanzamientos salga un 5. Y no puede haber otro resultado.

Si se lanza el dado solo tres veces e interesa conocer con que frecuencia ocurre el 5, ese valor no puede exceder el número de lanzamientos. Por lo tanto la escala de la variable binomial es este caso va de 0 a n .

Supongamos por ejemplo, que se lance el dado tres veces y se observa el siguiente resultado:

TEMA 5. INTRODUCCIÓN A LA INFERENCIA



Variables aleatorias

Distribuciones de probabilidad para una variable aleatoria discreta

Distribución de probabilidad Binomial

Desarrollo del modelo matemático

Primer lanzamiento	Segundo lanzamiento	Tercer lanzamiento
5	No cinco	5

La probabilidad de esta ocurrencia es decir, que salgan dos cincos en tres lanzamientos según esta secesión particular sería:

Primer lanzamiento	Segundo lanzamiento	Tercer lanzamiento
$p = 1/6$	$(1-p) = 5/6$	$p = 1/6$

Puesto que cada resultado es independiente de los demás la probabilidad de obtener la secuencia dada es:

$$p (1-p) p = p^2 (1-p)^1 = p^2 (1-p) = (1/6)^2 (5/6) = 5/216$$

Variables aleatorias

Distribuciones de probabilidad para una variable aleatoria discreta

Distribución de probabilidad Binomial

Desarrollo del modelo matemático

Por lo tanto, de los 216 resultados posibles e igualmente probables de lanzar un dado tres veces consecutivas, 5 tendrán la cara cinco en el primero y último lanzamientos, con una cara distinta a cinco (es decir, 1, 2, 3, 4 o 6) en el lanzamiento del medio y se obtendrá el orden particular anterior.

Sin embargo, ahora es posible preguntar ¿Cuántas series consecutivas diferentes puede haber para obtener dos caras cinco en $n = 3$ lanzamientos del dado?

$$\text{Esto sería } {}^nC_x = \frac{n!}{x! (n-x)!} = \frac{3!}{2! (3-2)!} = 3$$

Estas tres posibles series son

Secuencia 1 = 5, no cinco, 5 con probabilidad $p(1-p)p = p^2(1-p) = 5/216$

Secuencia 2 = 5, 5 no cinco con probabilidad $p p(1-p) = p^2(1-p) = 5/216$

Secuencia 3 = No cinco, 5, 5, con probabilidad $(1-p)p p = p^2(1-p) = 5/216$

TEMA 5. INTRODUCCIÓN A LA INFERENCIA

Variables aleatorias



Distribuciones de probabilidad para una variable aleatoria discreta

Distribución de probabilidad Binomial

Desarrollo del modelo matemático

Por lo tanto, la probabilidad de obtener exactamente 2 caras de cinco en tres lanzamientos de un dado es igual a:

(Número de secuencias posibles) X (probabilidad de la secuencia en particular)

$$(3) \times (5/216) = 15/216 = 0.0794 \text{ o } 7\%$$

Como a medida que aumenta n este proceso se haría muy laborioso de calcular, se ha desarrollado un modelo matemático que representa la distribución de probabilidad binomial para obtener el número de éxitos (X) conociendo los parámetros n y p:

$$P(X = x/n, p) = \frac{n!}{x! (n - x)!} p^x (1 - p)^{n - x}$$

Donde: $P(X = x/n, p)$ = Probabilidad de que $X = x$ conociendo n y p

n = tamaño de la muestra

p = probabilidad de éxito (1 - p) = probabilidad de fracaso

x = número de éxitos en la muestra $X = 0, 1, 2, 3, \dots, n$

TEMA 5. INTRODUCCIÓN A LA INFERENCIA



Variables aleatorias

Distribuciones de probabilidad para una variable aleatoria discreta

Distribución de probabilidad Binomial

Desarrollo del modelo matemático

Aplicando la fórmula al ejemplo de calcular la probabilidad de obtener exactamente 2 caras cinco en tres lanzamientos sucesivos de un dado sería:

$$\begin{aligned} P(X = 2 \text{ caras cinco}/n=3, p=1/6) &= \frac{3!}{2!(3-2)!} (1/6)^2 (1-1/6)^{3-2} \\ &= \frac{3!}{2!1!} (1/6)^2 (5/6)^1 \\ &= 3 (1/6) (1/6) (5/6) = 15/216 = 7\% \end{aligned}$$

Para evitar hacer estos cálculos que pueden ser tediosos se han desarrollado tablas donde podemos leer directamente las probabilidades asociadas para una n determinada y una p .

TEMA 5. INTRODUCCIÓN A LA INFERENCIA

Variables aleatorias



Distribuciones de probabilidad para una variable aleatoria discreta

Distribución de probabilidad Binomial

Ejercicio 5.1

Usando las tablas de distribución binomial, determine lo siguiente:

- a. Si $n = 4$ y $p = 0.12$, entonces $P(X = 0/n = 4, p = 0.12) = ?$
- b. Si $n = 10$ y $p = 0.40$ entonces $P(X = 9/n = 10, p = 0.40) = ?$
- c. Si $n = 7$ y $p = 0.83$ entonces $P(X = 6/n = 7, p = 0.83) = ?$

Ejercicio 5.2

La probabilidad que un paciente no se recupere de una operación particular es de 0.10.

- a. ¿Cuál es la probabilidad de que exactamente 2 de los siguientes 8 pacientes que sufran esta operación no se recuperen?
- b. ¿Cuál es la probabilidad de que sólo un paciente de los 8 no se recupere?

Ejercicio 5.3

En una cadena de comida rápida en base a la experiencia anterior se sabe que el 7% de todos los comprobantes de comida están equivocados.

- a. Si se selecciona una muestra de 5 comprobantes cual es la probabilidad de que:
 - a.1. ¿Exactamente uno este equivocado?
 - a.2. ¿Por lo menos dos estén equivocados?
 - a.2. ¿No más de 2 estén equivocados?
- b. Que suposiciones sobre la distribución de probabilidad se necesitan para solucionar a?

TEMA 5. INTRODUCCIÓN A LA INFERENCIA

Variables aleatorias



Distribuciones de probabilidad para una variable aleatoria discreta

Distribución de probabilidad Binomial

Ejercicio 5.4

Con base en la experiencia anterior, la impresora principal de una universidad, funciona adecuadamente el 90% del tiempo. Si se hace una muestra aleatoria de 10 inspecciones:

- ¿Cuál es la probabilidad de que de un lote de 6 piezas defectuosas se puedan reprocesar satisfactoriamente por lo menos 3 de ellas?
- ¿Cuál es la probabilidad de que ninguna de ellas se pueda reprocesar?
- ¿Cuál es la probabilidad de que todas ellas se puedan reprocesar?

Media de la Distribución binomial

$$\mu_x = np$$

Desviación típica

$$\sigma_x = \sqrt{np(1-p)}$$

Variables aleatorias

Distribuciones de probabilidad para una variable aleatoria discreta Distribución de probabilidad Binomial

Ejemplo:

En 10 lanzamientos consecutivos de un dado cuantos lados 5 esperaríamos que salieran?

$$\overline{u_x} = 10 (1/6) = 10/6 = 1,7$$

$$\sigma_x = \sqrt{10 (1/6) (5/6)} = \sqrt{8,333}$$

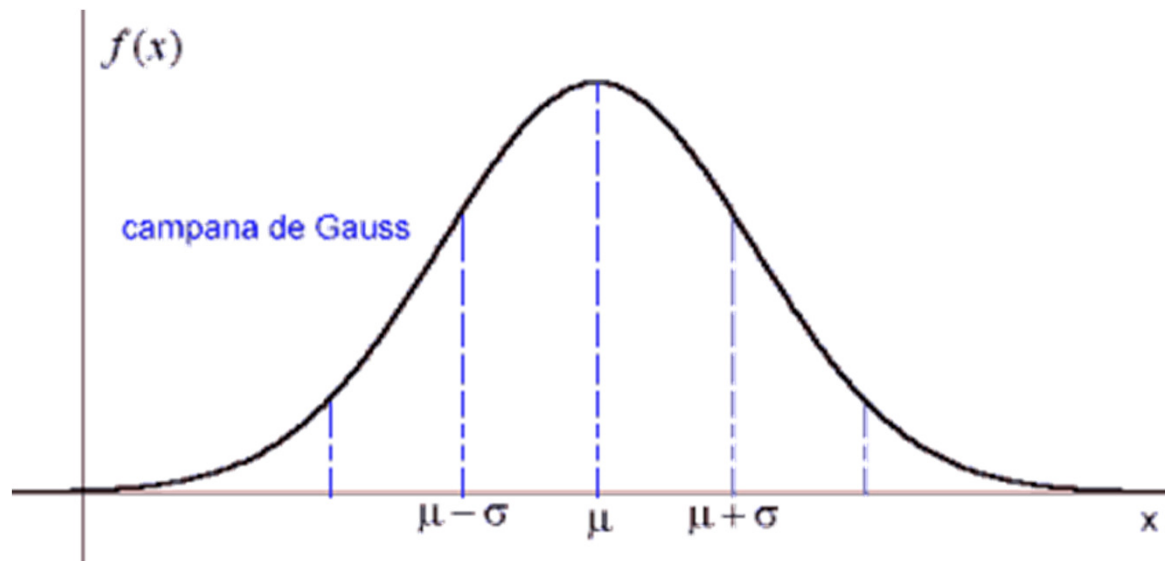
Variables aleatorias

Distribuciones de probabilidad para una variable aleatoria continua

La distribución normal

Muchas variables aleatorias continuas presentan una función de densidad cuya gráfica tiene forma de campana.

Representación gráfica de la función de densidad de la distribución normal



Distribuciones de probabilidad para una variable aleatoria continua

Importancia de la distribución normal

Se debe principalmente a que hay muchas variables asociadas a fenómenos naturales que siguen el modelo de la normal:

1. *Caracteres morfológicos* de individuos (personas, animales, plantas,...) de una especie, p.ejm. tallas, pesos, envergaduras, diámetros, perímetros,...
2. *Caracteres fisiológicos*, por ejemplo: efecto de una misma dosis de un fármaco, o de una misma cantidad de abono.
3. *Caracteres sociológicos*, por ejemplo: consumo de cierto producto por un mismo grupo de individuos, puntuaciones de examen.
4. *Caracteres psicológicos*, por ejemplo: cociente intelectual, grado de adaptación a un medio,...
5. *Errores* cometidos al medir ciertas magnitudes.
6. *Valores estadísticos* muestrales, por ejemplo : la media.
7. *Otras distribuciones* como la binomial o la de Poisson son aproximaciones normales, ...
8. Y en general cualquier característica que se obtenga como suma de muchos factores.

Distribuciones de probabilidad para una variable aleatoria continua

Propiedades de la distribución normal

1. Tiene forma de campana y de apariencia simétrica
2. Sus medidas de tendencia central (media, mediana, moda, rango medio, son idénticas.
3. Puede tomar cualquier valor ($-\infty, +\infty$)
4. Son más probables los valores cercanos a uno central que llamamos media **U**
5. Conforme nos separamos de ese valor **U** , la probabilidad va decreciendo de igual forma a derecha e izquierda (es simétrica).
6. Conforme nos separamos de ese valor **U** , la probabilidad va decreciendo de forma más o menos rápida dependiendo de un parámetro **σ** , que es la desviación típica.

TEMA 5. INTRODUCCIÓN A LA INFERENCIA

Variables aleatorias



Distribuciones de probabilidad para una variable aleatoria continua

El modelo matemático de la distribución normal

$$f(X) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{1}{2}\left(\frac{x - u_x}{\sigma_x}\right)^2}$$

Donde: e = Constante matemática 2.7182

π = Constante matemática 3.14159

u_x = Media de la población

σ_x = Desviación típica de la población

X = Cualquier valor de la variable aleatoria continua donde
 $-x < X < +x$

Variables aleatorias

Distribuciones de probabilidad para una variable aleatoria continua

Estandarización de la distribución normal

Para calcular cualquier área bajo una distribución normal, Gauss, la estandarizó de tal forma que solo se necesita una tabla. Utilizando la fórmula de transformación:

$$Z = \frac{X - u_x}{\sigma_x}$$

Una distribución normal estándar es aquella cuya variable aleatoria Z siempre tiene una media $u = 0$ y una desviación típica $\sigma_z = 1$.

Variables aleatorias

Distribuciones de probabilidad para una variable aleatoria continua Estandarización de la distribución normal

Ejemplo:

Mediante una investigación de mercados realizada en España se encontró que el gasto mensual promedio en familias de tres personas es de 450 euros al mes con una desviación típica de 80 euros. Los gastos mensuales siguen una distribución normal.

- ¿Qué porcentaje de familias gasta menos de 350 euros?
- ¿Qué porcentaje de familias gasta entre 250 y 350 euros?
- ¿Qué porcentaje de familias gasta mas de 350 euros?
- ¿Qué porcentaje familias gasta mas de 460 euros?
- ¿Qué porcentaje de familias se gasta menos de 460 euros?
- Que porcentaje de familias gasta menos de 250 euros o mas de 470 euros?

TEMA 5. INTRODUCCIÓN A LA INFERENCIA

Variables aleatorias

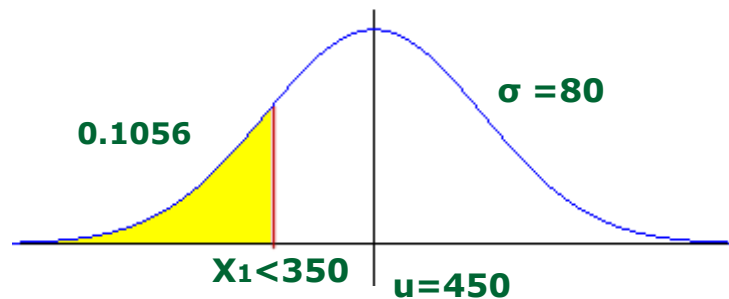


Distribuciones de probabilidad para una variable aleatoria continua

Estandarización de la distribución normal

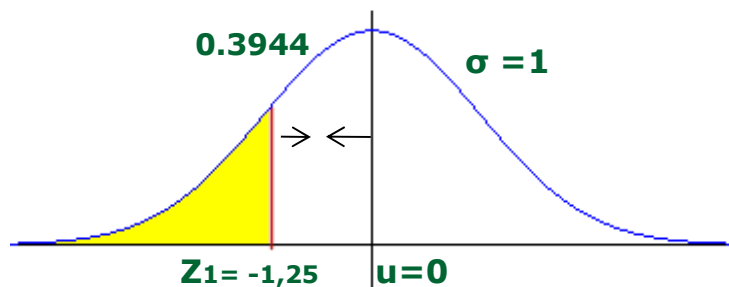
a. ¿Qué porcentaje de familias gasta menos de 350 euros al mes?

$$P(X_1 < 350) = ?$$



Transformamos esta distribución normal en una distribución normal estandarizada mediante:

$$Z_1 = \frac{X_1 - u}{\sigma} = \frac{350 - 450}{80} = -1,25$$



Se observa que a la medición X_1 le corresponde una medición Z_1 . Esto significa que desde la media $u=0$ hasta Z_1 hay -1,25 desviaciones típicas. Este valor lo puedo mirar en la tabla y es: 0.3944. Como la mitad de la curva vale 0.50, entonces para calcular $P(x_1 < 350)$ es decir el área sombreada tendríamos:

$$P(X_1 < 350) = 0.50 - 0.3944 = 0.1056 \text{ o } 10,56\%$$

TEMA 5. INTRODUCCIÓN A LA INFERENCIA

Variables aleatorias



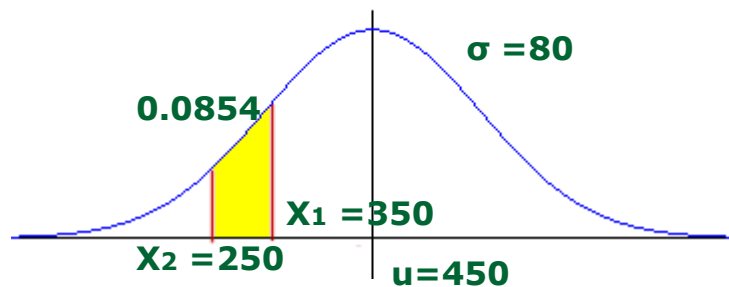
Distribuciones de probabilidad para una variable aleatoria continua

Estandarización de la distribución normal

b. ¿Qué porcentaje de familias gasta entre 250 y 350 euros al mes?

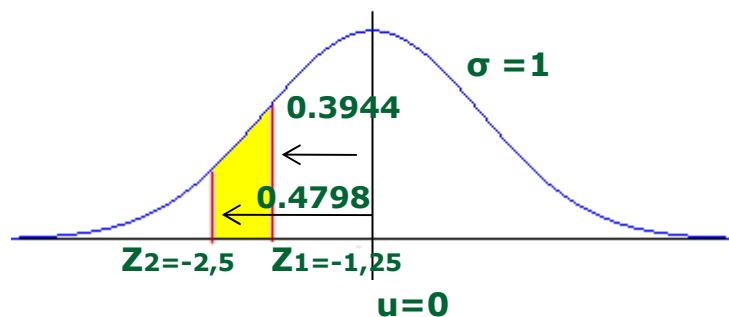
$$P(250 < X_1 < 350) = ?$$

Transformamos esta distribución normal en una distribución normal estandarizada y calculamos los valores Z1 y Z2



$$Z_1 = \frac{X_1 - u}{\sigma} = \frac{350 - 450}{80} = -1,25$$

$$Z_2 = \frac{X_2 - u}{\sigma} = \frac{250 - 450}{80} = -2,5$$



Z1 en tablas = 0.3944; Z2 en tablas = 0.4798

$$P(X_1 < 350) = 0.50 - 0.3944 = 0.1056 \text{ o } 10,56\%$$

$$\begin{aligned} \text{Por tanto } P(250 < X < 350) &= \text{Valor de Z2} - \text{Z1} \\ &= 0.4798 - 0.3944 = 0.0854 \\ &= 8,54\% \end{aligned}$$

TEMA 5. INTRODUCCIÓN A LA INFERENCIA

Variables aleatorias

Distribuciones de probabilidad para una variable aleatoria continua

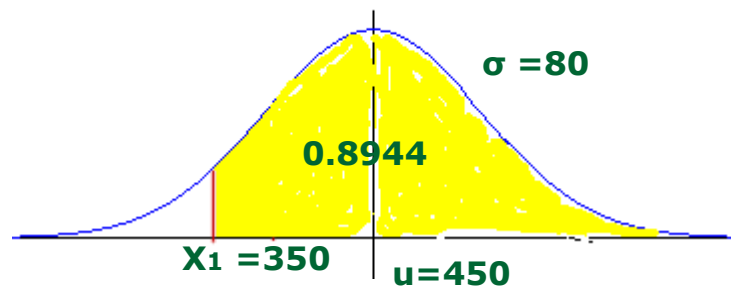
Estandarización de la distribución normal

c. ¿Qué porcentaje de familias gasta mas de 350 euros al mes?

$$P(X_1 > 350) = ?$$

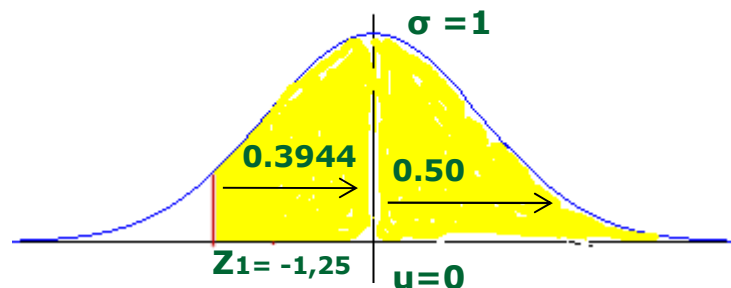
Transformamos esta distribución normal en una distribución normal estandarizada y calculamos el valor Z_1 .

$$Z_1 = \frac{X_1 - u}{\sigma} = \frac{350 - 450}{80} = -1,25$$



Z_1 en tablas = 0.3944;

$$P(X_1 > 350) = 0.3944 + 0.50 = 0.8944 \text{ o } 89,44\%$$



El porcentaje de familias que se gastan más de 350 euros al mes es del 89,44%.

TEMA 5. INTRODUCCIÓN A LA INFERENCIA

Variables aleatorias

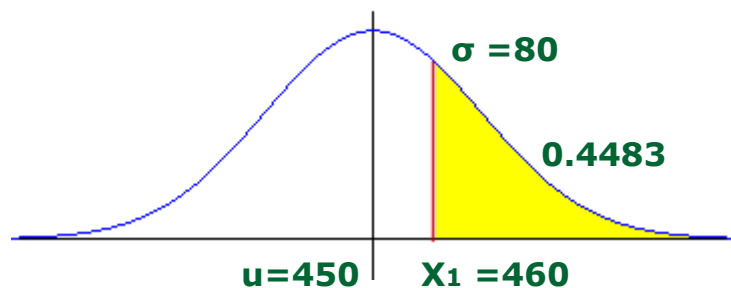


Distribuciones de probabilidad para una variable aleatoria continua

Estandarización de la distribución normal

d. ¿Qué porcentaje de familias gasta mas de 460 euros al mes?

$$P(X_1 > 460) = ?$$

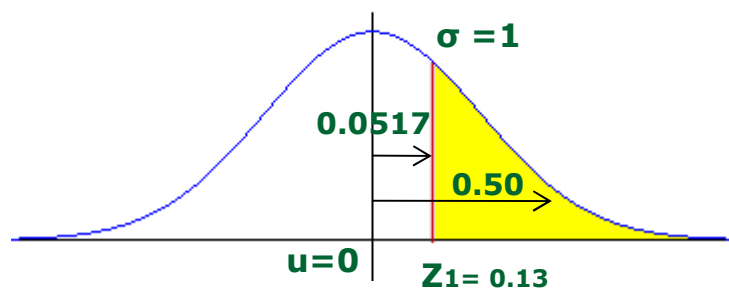


Transformamos esta distribución normal en una distribución normal estandarizada y calculamos el valor Z_1 .

$$Z_1 = \frac{X_1 - u}{\sigma} = \frac{460 - 450}{80} = 0.13$$

Z_1 en tablas = 0.0517;

$$P(X_1 > 460) = 0.50 - 0.0517 = 0.4483 \text{ o } 44,83\%$$



El porcentaje de familias que se gastan más de 460 euros al mes es del 44,83%.

TEMA 5. INTRODUCCIÓN A LA INFERENCIA

Variables aleatorias

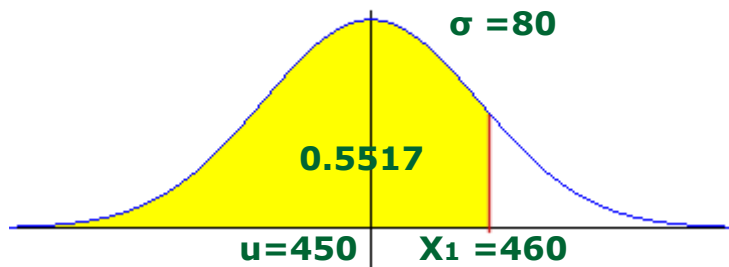


Distribuciones de probabilidad para una variable aleatoria continua

Estandarización de la distribución normal

e. ¿Qué porcentaje de familias se gasta menos de 460 euros al mes?

$$P(X_1 < 460) = ?$$

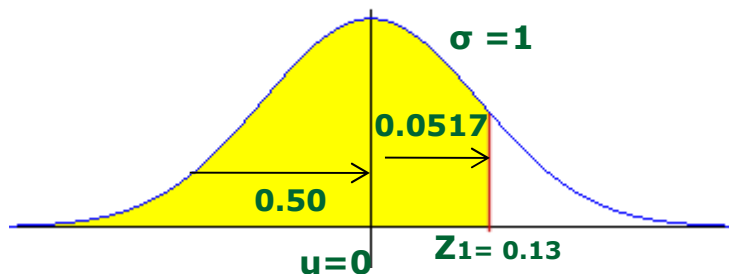


Transformamos esta distribución normal en una distribución normal estandarizada y calculamos el valor Z_1 .

$$Z_1 = \frac{X_1 - u}{\sigma} = \frac{460 - 450}{80} = 0.13$$

Z_1 en tablas = 0.0517;

$$P(X_1 < 460) = 0.50 + 0.0517 = 0.5517 \text{ o } 55,17\%$$



El porcentaje de familias que se gastan menos de 460 euros al mes es del 55,17%%.

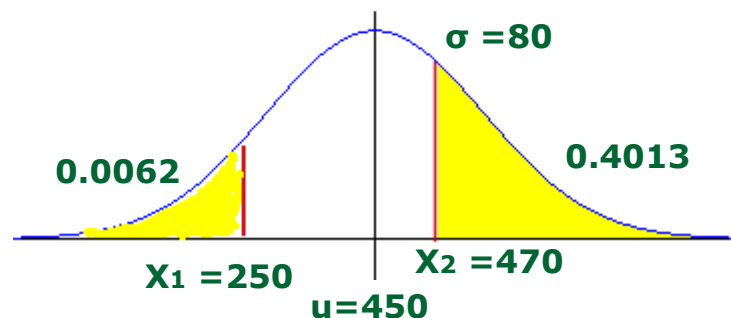
TEMA 5. INTRODUCCIÓN A LA INFERENCIA

Variables aleatorias

Distribuciones de probabilidad para una variable aleatoria continua Estandarización de la distribución normal

f. ¿Qué porcentaje de familias se gasta menos de 250 euros o mas de 470 euros al mes?

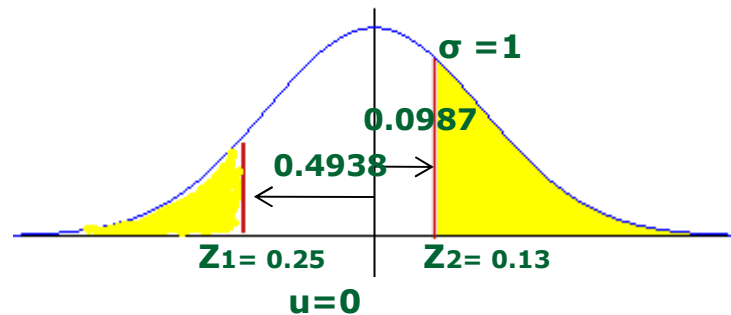
$$P(X_1 < 250) \text{ o } P(X_2 > 470) = ?$$



Transformamos esta distribución normal en una distribución normal estandarizada y calculamos los valores Z_1 y Z_2

$$Z_1 = \frac{X_1 - u}{\sigma} = \frac{250 - 450}{80} = -2,5$$

$$Z_2 = \frac{X_2 - u}{\sigma} = \frac{470 - 450}{80} = 0.25$$



Z_1 en tablas = 0.4938; Z_2 en tablas = 0.0987

$$P(X_1 < 250) = 0.50 - 0.4938 = 0.0062 \text{ o } 0.62\%$$

$$P(X > 470) = 0.50 - 0.0987 = 0.4013$$

$$P(X_1 < 250) \text{ o } P(X_2 < 470) = 0.0062 + 0.4013 = 0.4075 = 40,75\%$$

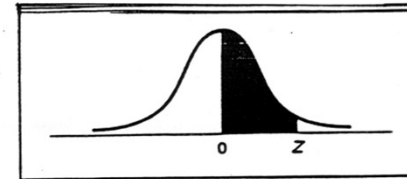
TEMA 5. INTRODUCCIÓN A LA INFERENCIA

Distribuciones de probabilidad continua

LA CURVA NORMAL

TABLA 2

Área debajo de la curva normal



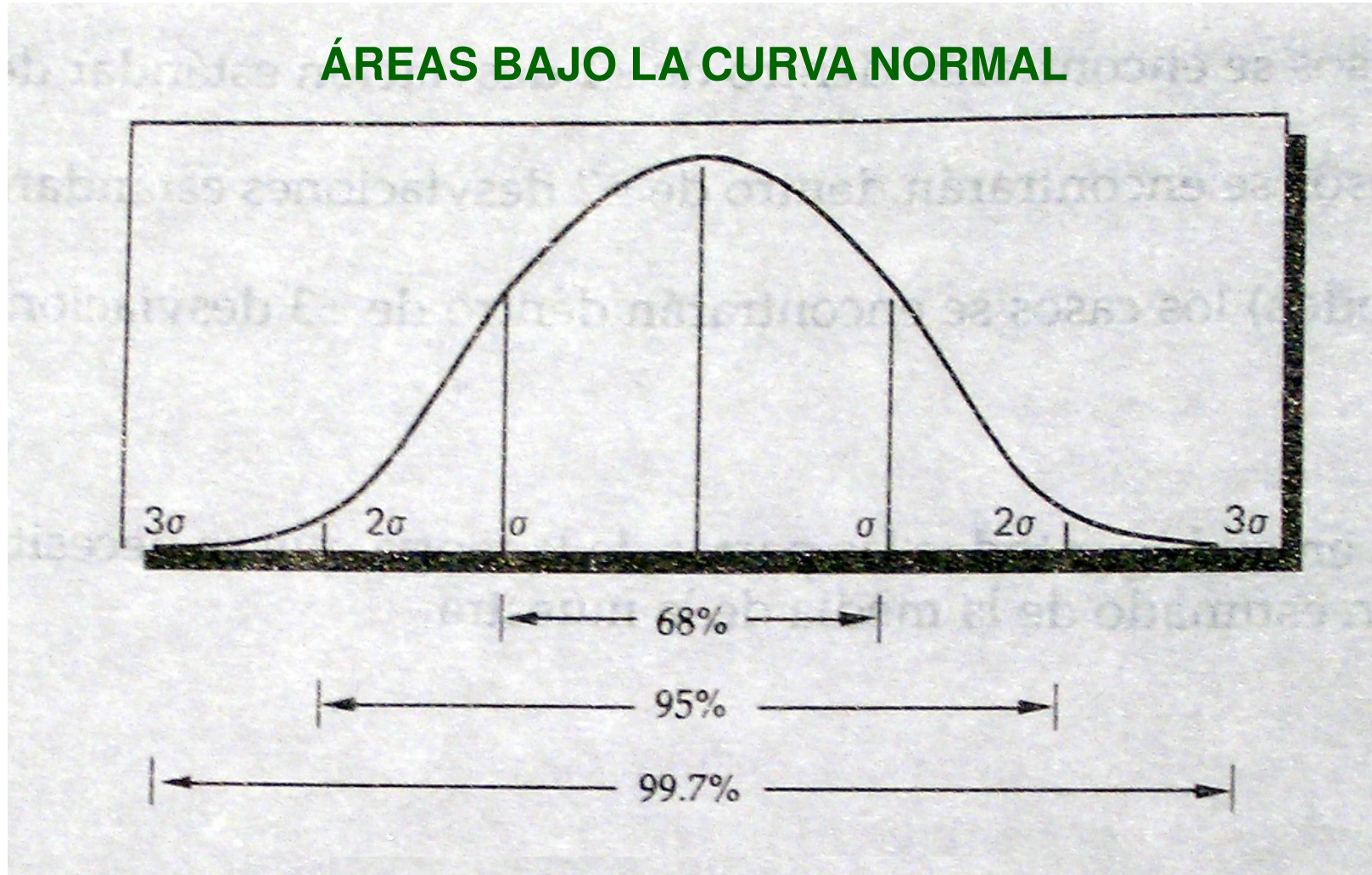
La entrada representa el área debajo de la distribución normal estándar desde la media a Z

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.49865	.49869	.49874	.49878	.49882	.49886	.49889	.49893	.49897	.49900
3.1	.49903	.49906	.49910	.49913	.49916	.49918	.49921	.49924	.49926	.49929
3.2	.49931	.49934	.49936	.49938	.49940	.49942	.49944	.49946	.49948	.49950
3.3	.49952	.49953	.49955	.49957	.49958	.49960	.49961	.49962	.49964	.49965
3.4	.49966	.49968	.49969	.49970	.49971	.49972	.49973	.49974	.49975	.49976
3.5	.49977	.49978	.49978	.49979	.49980	.49981	.49981	.49982	.49983	.49983
3.6	.49984	.49985	.49985	.49986	.49986	.49987	.49987	.49988	.49988	.49989
3.7	.49989	.49990	.49990	.49990	.49991	.49991	.49992	.49992	.49992	.49992
3.8	.49993	.49993	.49993	.49994	.49994	.49994	.49994	.49995	.49995	.49995
3.9	.49995	.49995	.49996	.49996	.49996	.49996	.49996	.49996	.49997	.49997

TEMA 5. INTRODUCCIÓN A LA INFERENCIA

Distribuciones de probabilidad continua

ÁREAS BAJO LA CURVA NORMAL



Distribuciones de probabilidad continua

AREAS BAJO LA CURVA NORMAL

1. El 68.3% de los casos se encontrarán dentro de ± 1 desviación estándar
2. El 95.4% de los casos se encontrarán dentro de ± 2 desviaciones estándar
3. El 99.7% de los casos se encontrarán dentro de ± 3 desviaciones estándar

Variables aleatorias

Distribuciones de probabilidad para una variable aleatoria continua Ejercicios de cálculo de áreas de la distribución normal

Ejercicio 5.5:

Mediante una investigación de mercados realizada en Madrid, se encontró que el consumo promedio de refrescos en familias de tres personas es de 3.5 litros a la semana con una desviación típica de 1.5 litros. Se sabe que el consumo de refrescos sigue una distribución normal.

- ¿Qué porcentaje de familias consume menos de 3.5 litros semana?
- ¿Qué porcentaje de familias consume entre 2 y 3.50 litros semana?
- ¿Qué porcentaje de familias consume más de 4 litros semana?
- ¿Qué porcentaje familias consume mas de 3 litros semana?
- ¿Qué porcentaje de familias consume menos de 2 litros semana?
- Que porcentaje de familias consume menos de 2.50 litros o más de 4,5 litros semana?

Estimación mediante intervalos de confianza

RECORDAMOS CONCEPTOS BÁSICOS.....

Parámetro.

Es una descripción resumida de una medida de la población objetivo. Un parámetro indica el valor real que se obtendría si se realizara un censo.

Ejemplos:

La edad promedio de los consumidores de cerveza en España.

La cuota de mercado (en porcentaje) de leche de Central Lechera Asturiana.

Estimación mediante intervalos de confianza

RECORDAMOS CONCEPTOS BÁSICOS.....

Estadístico.

Es una descripción resumida de una característica o medida de la muestra. Los estadísticos muestrales se utilizan para estimar los parámetros poblacionales.

Ejemplo:

En una muestra de 1000 amas de casa se encontró que el consumo promedio de leche entera es de 3,5 litros a la semana en el hogar.

Estimación mediante intervalos de confianza

CONCEPTOS BÁSICOS

Nivel de Confianza.

Es la probabilidad de que un intervalo de confianza incluya al parámetro de la población que se está estimando.

Nivel de Significancia.

Al estimar un parámetro de la población, mediante el uso de un estadístico muestral, el nivel de significancia es el error que estamos dispuestos a permitir en la estimación.

Estimación mediante intervalos de confianza

CONCEPTOS BÁSICOS

Intervalo de confianza.

Es el rango en que se encontrará el parámetro poblacional. suponiendo un nivel de confianza determinado.

Factor de corrección de la población finita (fcpf).

Es un dato para la sobreestimación de la varianza en un parámetro poblacional (por ejemplo, una media o una proporción) cuando el tamaño de la muestra representa 5% o más del tamaño de la población.

Estimación mediante intervalos de confianza

Símbolos que se emplean para parámetros y estadísticos

Variable	Población	Muestra
Media o promedio de una variable continua	μ	$\bar{x}$
Varianza de una variable continua	σ^2	s^2
Proporción (Medición dicotómica o mediciones binomiales) Proporción que responde "si"	π	ρ
Proporción que responde "no"	$(1 - \pi)$	$(1 - \rho)$ ó q
Varianza de la proporción	σ^2	s^2

Estimación mediante intervalos de confianza

Veamos el siguiente ejemplo:

Una población compuesta por los alumnos que están tomando un curso de Investigación comercial, dictado por un profesor. El profesor recolecto tres tipos de información de sus alumnos:

1. Número de estudiante de cada alumno (columna 1).
2. La edad de cada uno de ellos (columna 2).
3. La declaración a una pregunta sobre la intención de tomar o no el curso de Estrategias de marketing en el siguiente año.

La codificación que se utilizó fue 1= sí 0 = no (columna 3). (ver cuadro página siguiente).

TEMA 5. INTRODUCCIÓN A LA INFERENCIA

Estimación mediante intervalos de confianza

CENSO DE EDAD Y ELECCIÓN DEL CURSO DE ESTRATEGIAS DE MARKETING COMO OPCION PARA LOS ESTUDIANTES

X1 = Edad. Variable continua

X2 = Elección de curso. Variable discreta.

La población total N = 50

No. estudiante	Edad (X1)	Elección del curso E. M. 1= sí 2= no	$X = X1 - \mu$	$X^2 = (X1 - \mu)^2$
1	25	1	1.3	1.69
2	27	0	3.3	10.89
3	29	1	5.3	28.09
4	31	1	7.3	53.29
5	25	0	1.3	1.69
6	29	0	5.3	28.09
7	27	0	3.3	10.69
8	24	0	0.3	0.09
9	27	1	3.3	10.89
10	28	1	4.3	18.49
11	33	0	9.3	86.49
12	29	1	5.3	28.09
13	26	0	2.3	5.29
14	28	0	4.3	18.49
15	28	1	4.3	19.49
16	26	0	2.3	5.29
17	26	1	2.3	5.29
18	36	1	12.3	151.29
19	29	0	4.3	18.49
20	26	0	2.3	5.29
21	21	0	-2.7	7.29
22	19	0	-4.7	22.09
23	24	0	0.3	0.09
24	22	0	-1.7	2.89
24	20	1	-3.7	13.69
26	22	0	-1.7	2.89
27	19	1	-4.7	22.09
28	20	0	-3.7	13.69
29	19	0	-4.7	22.09
30	24	0	0.3	0.09
31	25	0	1.3	1.69
32	22	1	-1.7	2.89
33	20	0	-3.7	13.69

Estimación mediante intervalos de confianza

CENSO DE EDAD Y ELECCIÓN DEL CURSO DE ESTRATEGIAS DE MARKETING COMO OPCION PARA LOS ESTUDIANTE

Calculamos los parámetros poblacionales para la variable aleatoria continua: Edad

No. estudiante	Edad (X ₁)	Elección del curso E. M. 1= sí 2= no	X = X ₁ - μ	X ² = (X ₁ - μ) ²
34	21	1	-2.7	7.29
35	21	0	-2.7	7.29
36	23	1	-0.7	0.49
37	21	0	-2.7	7.29
38	23	0	-0.7	0.49
39	18	0	-5.7	32.49
40	21	1	-2.7	7.29
41	19	0	-4.7	22.09
42	23	0	-0.7	0.49
43	22	1	-1.7	2.89
44	19	0	-4.7	22.09
45	20	0	-3.7	13.69
46	20	0	-3.7	13.69
47	21	0	-2.7	7.29
48	20	1	-3.7	13.69
49	19	0	-4.7	22.09
50	18	0	-5.7	32.49

$$\sum X_1 = 1,184$$

$$\sum X_2 = 17$$

$$\sum (X_1 - \mu) = 0$$

$$\sum (X_1 - \mu)^2 = 844.09$$

Media: $\mu = \sum_{i=1}^N X_i / N = 23.7$ Varianza: $\sigma^2 = \sum_{i=1}^N (X_i - \mu)^2 / N = 16.9$

Desviación estándar: $\sigma = \sqrt{\sigma^2} = 4.1$

Estimación mediante intervalos de confianza

Supongamos que de la población (N) de 50 estudiantes, seleccionamos al azar una muestra (n) de 5 estudiantes

No. estudiantes	Edad X ₁	Elección del curso E.M. 1= si 0= No (X ₂)
32	22	1
17	26	1
05	25	0
37	21	0
41	19	0

Los estadísticos calculados para esta muestra son:

$$\text{Media: } \bar{X} = \sum_{i=1}^n x_i/n = 22.6$$

$$\text{Varianza } S^2 = \sum_{i=1}^n \frac{(X_1 - \bar{X})^2}{n-1} = 8.3$$

$$\text{Desviación estándar: } S = \sqrt{S^2} = 2.88$$

Estimación mediante intervalos de confianza Variable continua

Supongamos, que la población que vamos a analizar es muy grande y no conocemos el valor de la media de la población. Entonces, conociendo los estadísticos muestrales la podemos estimar a partir de estos, de la siguiente manera:

$$[\bar{x} - (Z) (s/\sqrt{n}) \leq u \leq \bar{x} + (z) (s/\sqrt{n})] = (1 - \alpha)$$

Donde:

μ = Media de la población (parámetro a estimar)

$\bar{X}$ = Media de la muestra S = desviación típica de la muestra

n = Tamaño de la muestra

Z = Número de desviaciones z asociados a un nivel de confianza dado.

$(1 - \alpha)$ = Nivel de confianza con el que se realiza la estimación

Estimación mediante intervalos de confianza Variable continua

En el ejemplo que venimos manejando se tomo una muestra de 5 alumnos y su edad promedio fue de 22,6 años y su desviación típica de 2,88 años. Con un nivel de confianza del 95,53% ¿Dentro de que valores se espera que estaría la media de la población de estudiantes?

$$[22,6 - (1,96) (2,88/\sqrt{5}) \leq u \leq 22,6 + (1,96) (2,88/\sqrt{5})] = (1- 0.0447)$$

$$[20.0 \leq \mu \leq 25,2] = 95,53\%$$

Como podemos observar la media μ de la variable edad de la población de 50 estudiantes es de 23,7 años, que estaría contemplada dentro de este intervalo cuando realizamos la estimación con los datos muestrales.

Estimación mediante intervalos de confianza Variable continua

Ejercicio 5.6

Con el propósito de conocer el consumo anual per cápita de refrescos en España la compañía Coca Cola realizó una investigación de mercados con 1500 entrevistas y si encontró que el consumo anual per cápita es de 130 litros con una desviación típica de 35 litros.

- Con un nivel de confianza del 95,53% estimar el intervalo donde se encuentra el consumo promedio real per cápita de refrescos en España.
- Realice la misma estimación con un nivel de confianza del 99%
- ¿Qué sucede con los dos intervalos calculados? Explícalo.

Estimación mediante intervalos de confianza Variable discreta

Media:

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

Para nuestro ejemplo sería:

$$p = \frac{1 + 1 + 0 + 0 + 0}{5} = 0.4$$

Por tanto, $q = (1 - P) = (1 - 0.4) = 0.6$

Estimación mediante intervalos de confianza Variable discreta

La varianza para la proporción será:

$$s^2 = \frac{\sum_{i=1}^n (X_i - p)^2}{gl}$$
$$= \frac{(1 - 0.4)^2 + (1 - 0.4)^2 + (0 - 0.4)^2 + (0 - 0.4)^2 + (0 - 0.4)^2}{4} = 0.3$$

La desviación estándar:

$$s = \sqrt{s^2} = \sqrt{0.3} = 0.548$$

Estimación mediante intervalos de confianza Variable discreta

La varianza para la proporción será:

$$s^2 = \frac{\sum_{i=1}^n (X_i - p)^2}{gl}$$
$$= \frac{(1 - 0.4)^2 + (1 - 0.4)^2 + (0 - 0.4)^2 + (0 - 0.4)^2 + (0 - 0.4)^2}{4} = 0.3$$

La desviación estándar:

$$s = \sqrt{s^2} = \sqrt{0.3} = 0.548$$

Nótese que la fórmula de la varianza y desviación típica es:

$$s^2 = pq \left(\frac{n}{n-1} \right) \text{ ó } npq/n-1 = (0.4) (0.6) (5/4) = 0.3$$

$$s = \sqrt{npq/n-1}$$

Estimación mediante intervalos de confianza Variable discreta

Como no conocemos π , la verdadera proporción poblacional ni tampoco conocemos $\sqrt{\pi(1-\pi)/n}$ la desviación estándar de la distribución muestral, se estima entonces de la siguiente manera:

$[p - (Z) (\sqrt{p(1-p)/n-1}) \leq \pi \leq p + (Z) (\sqrt{p(1-p)/n-1})] = (1 - \alpha)$ En el ejemplo:

$[0.4 - (1,96)(\sqrt{0.4(1-0.4)/4}) \leq \pi \leq 0.4 + (1,96)(\sqrt{0.4(1-0.4)/4})] = (95\%)$

$[0.2824 \leq \pi \leq 0.5176] = 95\%$

Es decir que el verdadero parámetro poblacional π estaría entre el 0.2824 y 0.5176. Este intervalo es muy amplio puesto que el tamaño de la muestra es pequeño y el valor de la variable $p(1-p)$ o pq de la variable es grande. Lo que podemos decir es que en 95.53 muestras de 100 esperaríamos que la media verdadera de la proporción estuviese contenida en este intervalo. la verdadera media es 0.27, que está contenida en este intervalo.

Estimación mediante intervalos de confianza Variable Discreta

Ejercicio 5.7

Con el propósito de conocer su cuota de mercado un fabricante de coches nacional realizó un estudio con 1000 consumidores y les preguntó la marca de coche que poseían. De todos los entrevistados 100 respondieron que utilizaban su marca.

- Con un nivel de confianza del 95,53% estimar el intervalo donde se encuentra la cuota de mercado real de este fabricante
- Realice la misma estimación con un nivel de confianza del 90%
- ¿Qué sucede con los dos intervalos calculados? Explícalo.

Estimación mediante intervalos de confianza

VALORES DE Z PARA LOS NIVELES DE CONFIANZA MÁS USADOS

Nivel de Confianza	Valor de Z
$(1 - \alpha) = 90\%$	1,645
$(1 - \alpha) = 95\%$	1,960
$(1 - \alpha) = 95,53\%$	2,000
$(1 - \alpha) = 99\%$	2,575



Fin del curso