UNIVERSIDAD DE CHILE
DEPARTAMENTO DE INGENIERIA INDUSTRIAL
MAGISTER EN GESTION Y DIRECCIÖN DE EMPRESAS

METODOS CUANTITATIVOS

PROFESORA: SARA ARANCIBIA C

2009

Introducción a la Estadística Descriptiva

¿Qué es la Estadística?

Estadística: Ciencia que trata de la recopilación, organización, presentación, análisis e interpretación de datos numéricos (estadísticas) con el fin de realizar una toma de decisiones más efectiva.

¿Quién utiliza la Estadística?

Las técnicas estadísticas se aplican de manera amplia en mercadotecnia, contabilidad, control de calidad y en otras actividades; estudios de consumidores; análisis de resultados; administración de instituciones; la educación; organismos políticos, médicos; y por otras personas que intervienen en la toma de decisiones.

DIVISIONES DE LA ESTADISTICA

Estadística descriptiva; Procedimientos estadísticos que sirven para organizar y resumir conjuntos de datos numéricos. Trata de la presentación de datos en gráficas o en distribuciones de frecuencias, y de aplicar diversos promedios y medidas de dispersión.

Estadística inferencial; Procedimientos estadísticos que sirven para deducir o inferir algo acerca de un conjunto de datos numéricos (población), seleccionando un grupo menor de ello (muestra).
Funciona tomando una muestra de una población y efectuando estimaciones acerca de una característica de esa población con base en los resultados de muestreo.

Población; Conjunto de todos los posibles individuos, personas,objetos o mediciones de interés científico

Para deducir algo acerca de una población, por lo general se toma una muestra de dicha población

Muestra; Una porción, o parte de una población de interés.

Tipos y ejemplos de variables

Datos

- **Cualitativos o de atributo (las no numéricas)**

  Ejemplos:
  Tipo de auto
  Color de cabello
  Género

- **Cuantitativos o numéricos**
  - **Discretos**

    Número de hijos
    Cantidad de empleados
  - **Continuos**

    Peso de un embarque
    Distancia entre dos ciudades

Problema

Supóngase que lo acaban de nombrar jefe de un proyecto para la X Región. El proyecto denominado “Parque para la Región” entre sus actividades debe recomendar la ubicación del parque y diseñarlo, debiendo considerar si éste habrá de orientarse hacia personas de todas las edades, sólo para niños o sólo para jubilados.
¿Cómo procedería para formular recomendaciones acerca de
a) la ubicación del parque
b) la orientación grupal (para todas las edades, jóvenes,personas mayores)?

NIVELES DE MEDICIÓN

Nivel de medición nominal

En este nivel se tienen dos o más categorías del ítem o variable.Las categorías no tienen orden o jerarquía. Lo que se mide es colocado en una u otra categoría, lo que indica solamente diferencias respecto a una o mas características.

Por ejemplo la variable sexo de la persona tiene sólo dos categorías:masculino y femenino.

Ninguna tiene mayor jerarquía que la otra, las categorías únicamente reflejan diferencias en la variable.
No hay orden de mayor a menor.

Si le asignamos una etiqueta o símbolo a cada categoría, esto identifica exclusivamente a la categoría.

Por ejemplo;
M=Masculino; F=Femenino

Si usamos números por ejemplo;
1=Masculino
2=Femenino
Los números utilizados en este nivel de medición tienen una función puramente de clasificación y no se puede manipular aritméticamente

Otros ejemplos: tipo de escuela a la que asiste (privada-pública), afiliación política (Partido A,Partido B, Partido C....), la carrera elegida, la raza, la ciudad de nacimiento, el canal de televisión preferido.

NOTA: Las categorías son los niveles donde serán caracterizados las unidades de análisis.

OBS: las categorías se consideran como mutuamente excluyentes y exhaustivas

Mutuamente excluyente: Una persona, objeto o medición se incluye solamente en una categoría.

Exhaustiva: Cada individuo, objeto, o medición debe aparecer en una categoría

Nivel de medición ordinal

En este nivel hay varias categorías, pero además éstas mantienen un orden de mayor a menor. Las etiquetas o símbolos de las categorías sí indican jerarquía. .

Calificación de estudiantes o personal diverso

Calificaciones	Número de personas
Excelente	6
Muy bien	28
Bien	25
Suficiente	17
Deficiente	0

Nivel de medición por Intervalos

Además del orden o jerarquía entre categorías, se establecen intervalos iguales en la medición. Las distancias entre categorías son las mismas a lo largo de toda la escala. Hay intervalo constante, una unidad de medida

Ejemplo Las puntuaciones en un cierto examen y las calificaciones en uno de historia o de matemáticas son ejemplos de la escala de medición de intervalo.

Obs: El cero en la medición, es un cero arbitrario, no es real (se asigna arbitrariamente a una categoría el valor cero y a partir de ésta se construye la escala)

Nivel de medición de razón

En este nivel, además de tenerse todas las características del nivel de intervalos (intervalos iguales entre las categorías y aplicación de operaciones aritméticas básicas y sus derivaciones), el cero es real, es absoluto (no es arbitrario). Cero absoluto implica que hay un punto en la escala donde no existe la propiedad

Ejemplos de estas mediciones sería la exposición a la televisión, el número de hijos, la productividad, las ventas de un producto y el ingreso.

Desde luego hay variables que pueden medirse en más de un nivel, según el propósito de medición. Por ejemplo la variable “antigüedad en la empresa”

Nivel de medición	Categorías
De razón	En días (o a K días)
Ordinal	Bastante antigüedad
	Regular antigüedad
	Poca antiguedad

Es muy importante indicar el nivel de medición de todas las variables e ítems de la investigación, porque dependiendo de dicho nivel se selecciona uno u otro tipo de análisis estadístico (por ejemplo, la prueba estadística para correlacionar dos variables de intervalo es muy distinta a la prueba para correlacionar dos variables ordinales).

1.-Resumen de datos:distribuciones de frecuencias y representaciones gráficas

La primera tarea es describir los datos, valores o puntuaciones obtenidas para cada variable

¿Cómo pueden describirse estos datos? Describiendo la distribución de las puntuaciones o frecuencias

¿Qué es una distribución de frecuencias?

Una distribución de frecuencias es un conjunto de puntuaciones ordenadas en sus respectivas categorías

Ejemplo:Distribución de frecuencia

Variable: Cooperación del personal para el proyecto de calidad de la empresa

Categorías	Códigos	Frecuencia
Si se ha obtenido la cooperación	1	91
No se ha obtenido la cooperación	2	5
No responden	3	26
Total		122

A veces , las categorías de las distribuciones de frecuencias son tantas que es necesario resumirlas en clases

Ingreso familiar ($)	Frecuencia
10.000-11.999	12
12.000-13.999	14
14.000-15.999	24
16.000-17.999	15
18.000-19.999	13
20.000-21.999	7
22.000-23.999	6
24.000-25.999	4
26.000-27.999	3
28.000-29.999	2
Total	100

¿Qué otros elementos contiene una distribución de frecuencias?

Las distribuciones de frecuencias pueden completarse agregando las frecuencias relativas y las frecuencias acumuladas. Las frecuencias relativas son los porcentajes de casos en cada categoría, y las frecuencias acumuladas son las que se van acumulando en cada categoría, desde la más baja hasta la más alta

Categorías	Códigos	Frecuencias absolutas	Frecuencias relativas%	Frecuencias acumuladas	Frecuencias Acum relativas
Si se ha obtenido la cooperación	1	91	74,6	91	74,6
No se ha obtenido la cooperación	2	5	4,1	96	78,7
No responden	3	26	21,3	122	100

¿Cómo presentar las distribuciones de frecuencia?
Pueden presentarse con los elementos más informativos para el lector y la verbalización de los resultados o un comentario o en forma de histograma o gráficas de otro tipo

Ejemplo: ¿ Se ha obtenido la cooperación del personal para el proyecto de calidad de la empresa?

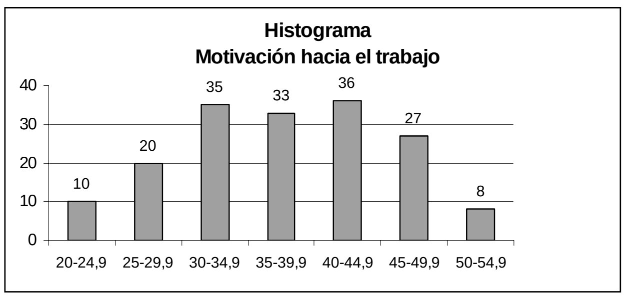
Categorías	N de organizaciones	Porcentajes
Si se ha obtenido la cooperación	91	74,6
No se ha obtenido la cooperación	5	4,1
No responden	26	21,3

Comentario: Prácticamente tres cuartas partes de las organizaciones si han obtenido la cooperación del personal. Llama la atención que poco más de una quinta parte no quiso comprometerse con su respuesta. Las organizaciones que no han logrado la cooperación del personal mencionaron como factores al ausentismo, rechazo al cambio.

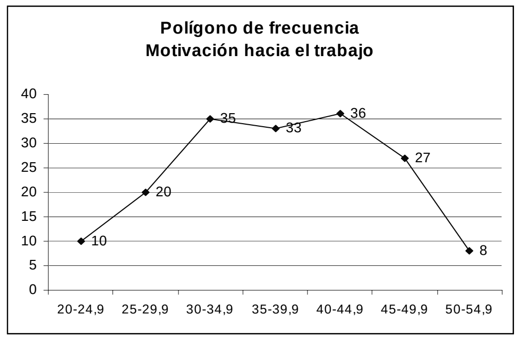
Representación gráfica de una distribución de frecuencia

Tres diagramas que representan de manera adecuada, una distribución de frecuencias son el histograma, el polígono de frecuencias y el polígono de frecuencias acumuladas

Un histograma describe una distribución de frecuencias utilizando una gráfica de barras (rectángulos verticales adyacentes), en la que la altura de cada barra es proporcional a la frecuencia de la clase que representa



El polígono de frecuencias consiste en una línea poligonal formada por segmentos de recta que unen los puntos determinados por la intersección de la vertical del punto medio de clase, y la horizontal de la frecuencia de clase.



Un polígono de frecuencias acumuladas (ojiva) se utiliza cuando se desea determinar cuántas observaciones se encuentran por encima o debajo de ciertos valores.

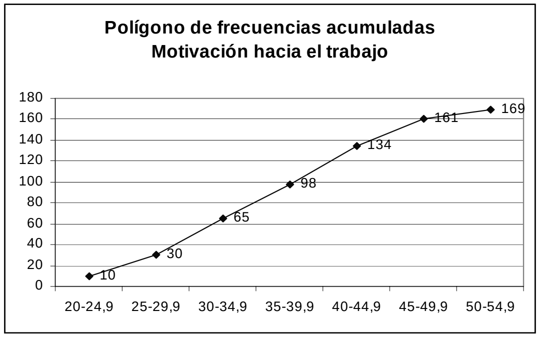


Diagrama de Pareto

Un diagrama de Pareto se asemeja a un histograma, excepto que es una gráfica de barras de frecuencias de una variable cualitativa. Los diagramas de Pareto se usan en el control de procesos para tabular las causas asociadas con variaciones de causas atribuibles en la calidad del producto. Los diagramas de Pareto permiten que tanto equipos de trabajadores como gerentes se concentren en las áreas más importantes en las que se necesitan acciones correctivas

Ejemplo: Se encontró que los refrigeradores que no fueron aprobados en la inspección final en una planta ensambladora de aparatos eléctricos durante el último mes tenían defectos debidos a las siguientes causas: ensamble, acabado de laca, fallas eléctricas, abolladuras u otras causas. Realizar un diagrama de Pareto

Diagrama de Pareto de : Defectos

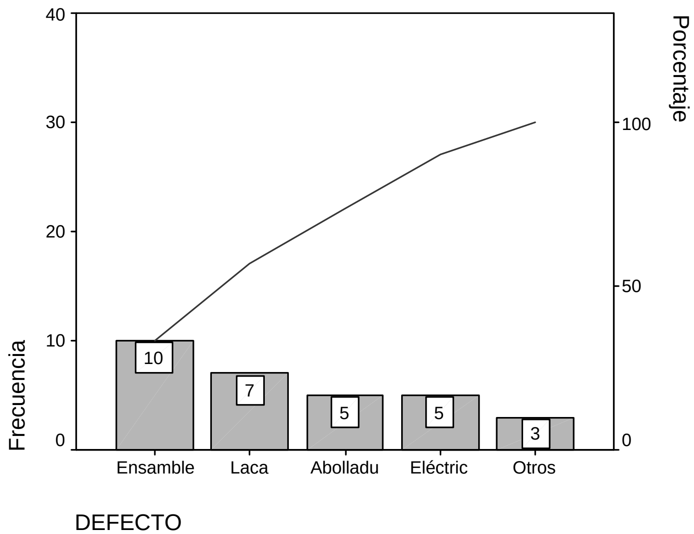


Diagrama de tallo y hojas

Técnica de análisis exploratorio de datos que clasifica y ordena simultáneamente los datos cuantitativos y proporciona una perspectiva de la forma de la distribución.

Para ilustrar el uso del diagrama de tallo y hojas, consideremos los datos de la tabla

Cantidad de preguntas contestadas en forma correcta en una prueba de aptitud

112	72	69	97	107
73	92	76	86	73
126	128	118	127	124
82	104	132	134	83
92	108	96	100	92
115	76	91	102	81
95	141	81	80	106
84	119	113	98	75
68	98	115	106	95
100	85	94	106	119

Primero ordenamos los datos, de acuerdo con los dígitos iniciales de cada uno, en el lado izquierdo de una línea vertical. A la derecha de esa recta se anota el último dígito de cada dato, conforme se recorren las calificaciones en el orden en que fueron anotadas

6	9 8
7	2 3 6 3 6 5
8	6 2 3 1 1 0 4 5
9	7 2 2 6 2 1 5 8 8 5 4
10	7 4 8 0 2 6 6 0 6
11	2 8 5 9 3 5 9
12	6 8 7 4
13	2 4
14	1

Con esta organización de los datos, es fácil clasificar los dígitos de cada renglón en su rango correspondiente. Al hacerlo se llega al diagrama de tallo y hojas que vemos a continuación

6	8 9
7	2 3 3 5 6 6
8	0 1 1 2 3 4 5 6
9	1 2 2 2 4 5 5 6 7 8 8
10	0 0 2 4 6 6 6 7 8
11	2 3 5 5 8 9 9
12	4 6 7 8
13	2 4
14	1

Tablas de contingencia

Las tablas de frecuencia pueden organizar datos de sólo una variable a la vez. Si se desea examinar o comparar dos variables, una tabla de contingencia resulta de mucha utilidad.

Supongamos que se recolecta información sobre el número de pasajeros de una línea aérea, también se obtuvieron datos sobre las edades de los pasajeros y el número de vuelos en los que se registraron cada año. Ambas variables pueden verse en detalle mediante una tabla de contingencia.

	Número de vuelos por año			
Edad	1-2	3-5	Mayor de 5	Total
Menor de 25	1 (2%)	1 (2%)	2 (4%)	4 (8%)
25-40	2 (4%)	8 (16%)	10 (20%)	20 (40%)
41-65	1 (2%)	6 (12%)	15 (30%)	22 (44%)
66 y más	1 (2%)	2 (4%)	1 (2%)	4 (8%)
Total	5 (10%)	17 (34%)	28 (56%)	50 (100%)

También se podría mostrar la información en cada celda según el género. Todas las otras características relevantes de los pasajeros también podrían incorporarse en la tabla. Dicho perfil descriptivo podría ser muy útil en la identificación del pasajero típico y en el establecimiento de políticas de marketing efectivas.

2.- Descripción de los datos: medidas de tendencia central

¿Qué es un promedio?
Valor que representa un conjunto de datos.
Señala un centro de los valores.
Una denominación más precisa que promedio es medida de tendencia central

Las medidas de tendencia central, que por lo común se emplean en administración y economía son :

- la media aritmética
- la mediana
- la moda
- la media geométrica

Media: (Mean) La media (promedio) es la medida de tendencia central más utilizada y puede definirse como el promedio aritmético de una distribución. Es una medida solamente aplicable a mediciones por intervalos o de razón

Media de una muestra

$$\overline{\mathrm{X}} = \frac{\sum X}{n}$$

$$donde \sum X \ \ indica \ la \ suma \ de \ todos \ las \ X$$

n es el número total de valores en la muestra

Media de una población

$$\mu = \frac{\sum X}{N}$$

μ indica la media poblacional.

N es el número total de observaciones en la población

Ejemplo: El ingreso anual (en dólares) de una muestra de varios empleados en una industria son: 42900, 49100,38300 y 56800
La media muestral es 187100/4=46775
¿La media obtenida es un dato estadístico o un parámetro?
Estadístico porque es un valor muestral

Nota:Cualquier característica medible de una muestra se llama dato estadístico. Cualquier característica medible de una población, como la media, se denomina parámetro.

La Mediana:(Median) es el valor céntrico en un conjunto de valores ordenados de menor a mayor o de mayor a menor.Una forma fácil de localizar la posición del elemento medio para datos no agrupados es por medio de

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

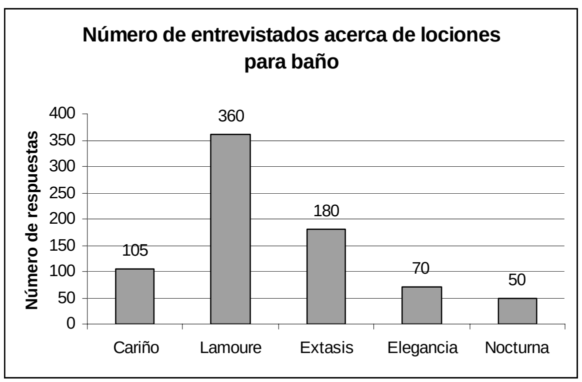
- . Es una medida de tendencia central propia de los niveles de medición ordinal, por intervalos y de razón
- . No es influenciada por valores extremos
- . 50%de las observaciones son mayores que la mediana
- No necesita ser uno de los valores del conjunto de datos
- Es única para un conjunto de observaciones

Ejemplo: Supóngase que intenta adquirir una casa en un condominio de la comuna de La Reina. El agente de ventas le indicó que el precio promedio de las casas disponibles en este momento es de 4225 UF.
Si tuviera un presupuesto máximo de 2800UF, podría pensar que está fuera de sus posibilidades. Sin embargo al verificar los precios individuales de las casas podría cambiar de idea. Los precios son 2300UF, 2600UF, 3500UF y 8500UF.
El precio 8500UF está haciendo que la media se incline hacia arriba, por lo que es un promedio no representativo. Un precio entre 2600 y 3500 es un promedio más representativo. En casos como éste la mediana proporciona una medida más exacta de la tendencia central

<u>La moda</u>: (mode): Valor de la observación que aparece con más frecuencia

•La moda es útil en especial al describir los niveles nominal y ordinal de medición (aunque puede determinarse para todos los niveles de datos)
•Un conjunto de datos puede tener más de una moda

Ejemplo: Una empresa ha desarrollado cinco lociones para baño. En el diagrama se muestran los resultados de una investigación de mercado diseñada para determinar qué loción para baño prefieren los consumidores



La mayor cantidad de respuestas favoreció a la llamada Lamoure, según lo indica la barra más alta. Por tanto tal producto es la moda

Uso de media, mediana y moda

Datos de la población
<u>La moda</u> puede ser útil como medida descriptiva de un grupo de la población, aunque sólo si existe una moda claramente perceptible.

<u>La mediana</u> es siempre una medida excelente para representar el nivel típico de los valores observados, como los índices salariales, de una población. Esto es así independientemente de la existencia de más de una moda o de que la distribución de la población sea asimétrica o simétrica.

<u>La media aritmética</u> es excelente como valor representativo de una población, aunque sólo si la población es claramente simétrica. En datos no simétricos, los valores extremos distorsionarán el valor de la media como valor representativo.

Uso de media, mediana y moda

Datos muestrales
La moda: La moda no es una medida aceptable de posición respecto de datos muestrales, porque su valor puede variar ampliamente de una muestra a otra.

La mediana: es mejor que la moda, porque su valor es más estable entre muestra y muestra.

La media: es la más estable de las tres medidas

Ejemplo: Se han recopilado los índices salariales de los 650 empleados por hora de una empresa manufacturera. La medida más representativa del índice salarial típico es la mediana, porque en este caso está implicada una población y la mediana no se ve relativamente afectada por la posible falta de simetría de los índices salariales.

Ejemplo:Una muestra aleatoria de n=100 índices salariales se obtiene en una compañía con varios miles de empleados por hora El índice salarial más representativo de estos varios miles de empleados es la media muestral. Aunque es improbable que la media muestral sea exactamente igual al índice salarial medio de toda la población, por lo general se hallará mucho más cerca de la media poblacional que la mediana muestral como estimador del índice salarial mediano de la población

Nota: Es mejor usar la mediana que la media como medida de tendencia central, cuando un conjunto de datos contiene valores extremos. Otra medida que se usa a veces cuando hay valores extremos es la media recortada al 5%, que se obtiene eliminando el 5% de los valores mayores y el 5% de los menores, en un conjunto de datos, procediendo entonces a calcular el promedio de los valores restantes.

Media Geométrica

La media geométrica (MG) tiene una amplia aplicación en los negocios y en la economía.
Hay dos usos principales de la MG:
1)Para promediar porcentajes, índices y cifras relativas
2) Para determinar el incremento porcentual promedio en ventas, producción u otras actividades o series económicas de un periodo a otro.

La media geométrica (MG) de un conjunto de n números positivos se define como la raíz n-ésima del producto de los n números. Proporciona una medida precisa de un cambio porcentual promedio en una serie de números

$$MG = \sqrt[n]{(X_1)(X_2)(X_3)\ldots\ldots\ldots\ldots(X_n)}$$

Ejemplo: Para ilustrar el empleo de la media geométrica en promedios de porcentajes , supóngase que las utilidades obtenidas por una compañía constructora en cuatro proyectos fueron de 3, 2 , 4 y 6%, respectivamente. ¿Cuál es la media geométrica de las ganancias?

$$MG = \sqrt[n]{(X_1)(X_2)(X_3)(X_4)}$$
$$= \sqrt[4]{(3)(2)(4)(6)}$$
$$= \sqrt[4]{144}$$
$$= 3{,}46$$

La media geométrica 3,46 da una cifra más conservadora para las utilidades porque no se ve tan afectada por los valores extremos. En realidad, será igual o menor que la media aritmética

Ejemplo:

Suponga que usted recibe un 5 por ciento de aumento en su salario este año y un 15 por ciento de aumento el año próximo. El incremento porcentual promedio es 9.886, no 10.0. Empezamos por calcular la media geométrica. Recuerde, por ejemplo, que un aumento de 5 por ciento en el salario es 105 o 1.05. Lo escribiremos como 1.05.

Esto se puede verificar suponiendo que su salario mensual sea de $3.000 al empezar y que usted reciba dos aumentos de 5 por ciento y de 15 por ciento.

Aumento 1 = $3.000 (.05) = $150.00
Aumento 2 = $3.150 (.15) = 472.50
Total $622.50

Su aumento total de salario es $622.50. Esto es equivalente a:
$3.000.00 (.09886) = $296.58
$3.296.58 (0.9886) = $325.90
$622.48 que es aproximadamente $622.50

Para determinar el incremento porcentual promedio en ventas, exportaciones, población u otras series se utiliza la siguiente fórmula

INCREMENTO PORCENTUAL PROMEDIO EN EL TIEMPO

$$GM = \sqrt[n]{\frac{Valor\ al\ final\ del\ periodo}{Valor\ al\ principio\ del\ periodo}} - 1$$

n: tiempo transcurrido

Ejemplo

La población de Haarlan, Alaska, era de dos personas en 1990, en 2000 fue de 22. ¿Cuál es la tasa de incremento anual promedio de este periodo?
Solución
Hay 10 años entre 1990 y 2000 por lo que n = 10. La fórmula (3-6) para la media geométrica aplicada a este tipo de problema queda:

$$MG = \sqrt[10]{\frac{22}{2}} - 1 = 1.271 - 1 = 0.271$$

El valor final es .271. La tasa anual de incremento es 27.1 por ciento. Esto significa que la tasa de crecimiento en Haarlan es 27.1 por ciento por año.

Ejercicio:
El Director ejecutivo de la empresa Airlines desea determinar la tasa de crecimiento promedio en los ingresos con base en las cifras entregadas en la tabla: si la tasa de crecimiento promedio es menor que el promedio industrial del 10% se asumirá una nueva campaña publicitaria . Determine la tasa de crecimiento promedio.

Año	Ingreso Miles de US$	Porcentaje del año anterior
1992	50	
1993	55	
1994	66	
1995	60	
1996	78	

Media Ponderada

La media ponderada de un conjunto de números denotados por $X_1, X_2, X_3, \ldots, X_n$, con ponderaciones $W_1, W_2, W_3, \ldots, W_n$ se calcula como sigue:

$$\overline{X}_w = \frac{w_1 X_1 + w_2 X_2 + \ldots.. + w_n X_n}{w_1 + w_2 + \ldots.. + w_n}$$

Esto puede abreviarse

$$\overline{X}_w = \frac{\sum (w_i X_i)}{\sum wj}$$

Ejemplo:
Una empresa comercial paga a sus vendedores $6.50, $7.50, u $8.50 (dólares) por hora.
Podría llegarse a la conclusión de que la media de los sueldos (por hora) es $7.50, obtenida al calcular (6.50+7.50+8.50)/3. Esto es cierto sólo si hay el mismo número de vendedores que perciben $6.50, $7.50, y $8.50
Sin embargo, supóngase que 14 empleados de ventas ganan $6.50, a 10 se les paga $7.50,y 2 obtienen $8.50.
Para encontrar la media se debe calcular:

$$\overline{X}_w = \frac{14(6.50) + 10(7.50) + 2(8.50)}{14 + 10 + 2}$$

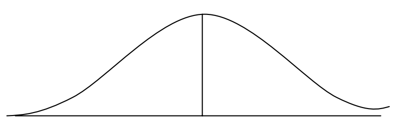
La media ponderada de los sueldos por hora es $7.04

3. Estadísticos de la distribución: Asimetría y curtosis

La asimetría es una estadística necesaria para conocer qué tanto nuestra distribución se parece a una distribución teórica llamada curva normal

El sesgo o asimetría (Skewness) es la carencia de forma simétrica en la gráfica de un conjunto de datos

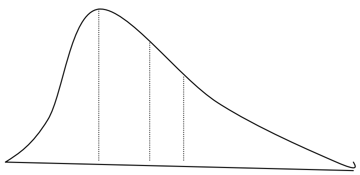
•Si no existe asimetría o sesgo en los datos, son iguales la media, la mediana y la moda. La mitad de los valores están por encima de estos promedios y la mitad por debajo de ellos (asimetría =0).



Media
mediana
Moda

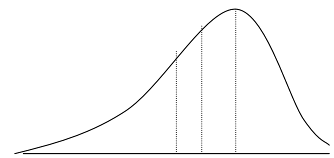
Cuando la asimetría es positiva quiere decir que hay más valores agrupados hacia la izquierda de la curva (por debajo de la media)

- La moda es el valor que corresponde al punto más alto
- La media es el mayor de los tres promedios



Cuando la asimetría es negativa significa que los valores tienden a agruparse hacia la derecha de la curva (por encima de la media)

- La moda es el valor que corresponde al punto más alto de la distribución
- La media es el más pequeño de los tres promedios



Nota: Si la distribución es muy asimétrica, la media no sería un promedio útil. La mediana y la moda son más representativas

Coeficiente de asimetría:

Karl Pearson desarrolló una medida para evaluar el sesgo de unas distribución, denominada coeficiente de asimetría (C.A)

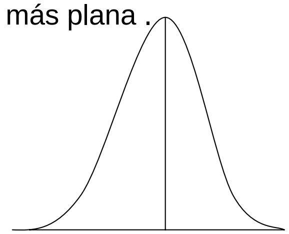
$$C.A = \frac{3(media - mediana)}{desviación \quad estándar}$$

Ejemplo: Las duraciones de estadía en el piso de cancerología de un hospital, se organizaron en una distribución de frecuencias. La duración media fue 28 días, la mediana 25 días y la duración modal 23 días Se calculó la desviación estándar de 4,2 días ¿Cuál es el coeficiente de asimetría?

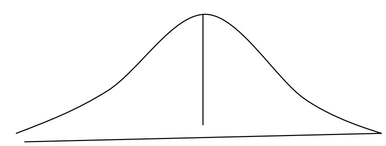
El coeficiente de asimetría por lo general se encuentra entre -3 y 3. En este caso es 2,4 e indica un grado importante de asimetría con sesgo positivo.

Curtosis;(Kurtosis) mide el grado de agudeza de una distribución

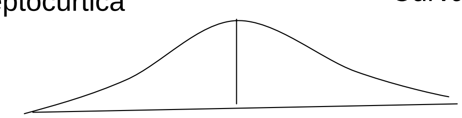
Cuando la curtosis es cero, significa que se trata de una curva normal, . Si es positiva, quiere decir que la curva o distribución es más levantada. Si es negativa, quiere decir que la curva es más plana .



Curva leptocúrtica



Curva mesocúrtica



Curva plalticúrtica

Nota: La asimetría y la curtosis requieren mínimo un nivel de medición por intervalos

4. Medidas de Posición: Cuantiles

Los cuantiles son valores de la distribución que la dividen en partes iguales, es decir, en intervalos, que comprenden el mismo número de valores. Los más usados son los cuartiles, los deciles y los percentiles.

PERCENTILES: Son 99 valores que dividen en cien partes iguales el conjunto de datos ordenados. Ejemplo, el percentil de orden 15 deja por debajo al 15% de las observaciones, y por encima queda el 85%.

CUARTILES: Son los tres valores que dividen al conjunto de datos ordenados en cuatro partes iguales, son un caso particular de los percentiles.

DECILES: son los nueve valores que dividen al conjunto de datos ordenados en diez partes iguales, son también un caso particular de los percentiles.

La determinación de cuartiles con frecuencia es de utilidad. Por ejemplo muchas escuelas de posgrados admitirán sólo aquellos estudiantes que estén en el 25% superior (sobre el tercer cuartil) de los candidatos. Las empresas, con frecuencia, desean señalar las plantas cuyos deficientes registros de producción los colocan por debajo del cuartil inferior

El lugar del P-ésimo percentil se halla

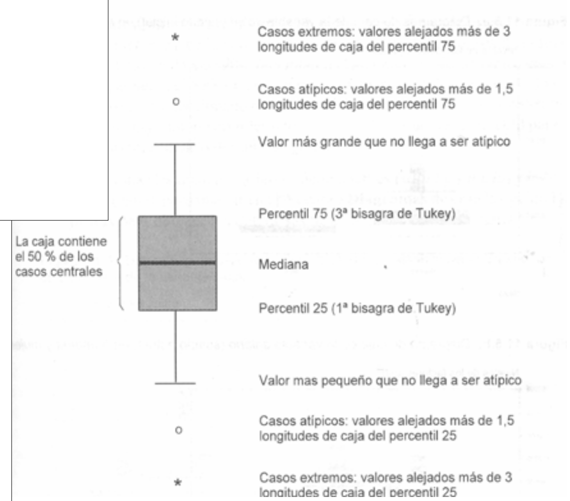
$$L_p = (n+1)\frac{P}{100}$$

Donde L_p es el sitio del percentil deseado en una serie ordenada
n es el número de observaciones
P es el percentil deseado

La amplitud intercuartil es la diferencia entre el tercer y primer cuartil

Diagrama de caja

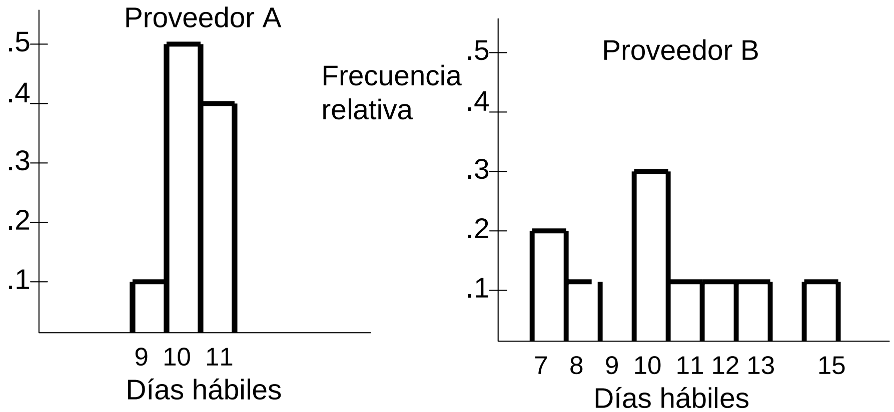
El diagrama de caja permite resumir los datos en una gráfica



5.- Medidas de dispersión

Con frecuencia es conveniente contar con medidas de dispersión o de la variabilidad de los valores de los datos. Por ejemplo, suponga que usted es un agente de compras de una importante empresa manufacturera, y que con regularidad coloca pedidos con dos proveedores distintos. Ambos le indican que necesitan alrededor de 10 días hábiles para surtir sus pedidos. Después de varios meses de trabajar así encuentra usted que el promedio de días necesarios para surtir los pedidos es realmente, unos 10 para cada proveedor. Los histogramas que resumen la cantidad de días hábiles requeridos para surtir los pedidos se ven en la figura.

Aunque la cantidad promedio es, más o menos, de 10 en ambos casos ¿ tienen éstos el mismo grado de confiabilidad para entregar a tiempo? ¿Qué proveedor prefiere usted?



Examinaremos varias medidas que describirán la dispersión o variabilidad de los datos; la amplitud total, la varianza, la desviación estándar, dispersión relativa

¿Por qué estudiar la dispersión?

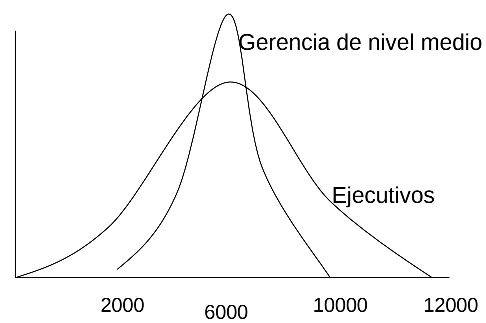
1. Al aplicar una medida de dispersión es posible evaluar la confiabilidad del promedio que se está utilizando
Una dispersión pequeña indica que los datos se encuentran acumulados cercanamente, por ejemplo, alrededor de la media aritmética. Por tanto la media se considera bastante representativa de los datos. Esto es la media es un promedio confiable.
2. Una medida de dispersión permite apreciar cuán dispersas están dos o más distribuciones.

Medidas de dispersión

Amplitud total (rango)
Se trata de la diferencia entre los valores mayor y menor de un conjunto de datos

Amplitud total= Valor más alto- Valor más bajo

Ejercicio: Los costos anuales de viaje para ejecutivos y gerentes medios en una empresa se organizaron en distribuciones de frecuencias y se representaron por medio de polígonos de frecuencias
1. ¿Cuál es la media aritmética de los costos de viaje para ejecutivos? ¿ Par los gerentes de nivel medio?
2. ¿Cuál es la amplitud total para los ejecutivos?¿ Y para los gerentes de nivel medio?
3. Compare la dispersión de las dos distribuciones y explique lo que indica



Nota: Un defecto importante de la amplitud total es que se basa sólo en dos valores. No toma en consideración todos los datos

Varianza y desviación estándar

La varianza y la desviación estándar se basan en las desviaciones con respecto a la media

Varianza:(Variance) Media aritmética de las desviaciones cuadráticas con respecto a la media

Varianza poblacional

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N} \qquad \text{o bien} \quad \sigma^2 = \frac{\sum \mathrm{X}^2}{\mathrm{N}} - \left(\frac{\sum X}{N}\right)^2$$

donde

σ^2 = es el símbolo para la varianza de una población

X es el valor de la observación en la población

μ es la media de la población

N es el número total de observaciones en la población

En general es difícil interpretar el significado del valor de una varianza, porque las unidades en las que se le expresa son valores elevados al cuadrado. Es más frecuente el uso de la raíz cuadrada

Desviación estándar:(Standard deviation) Raíz cuadrada de la varianza.

$$\sigma = \sqrt{\frac{\sum (X-\mu)^2}{N}} \qquad \text{o bien} \quad \sigma = \sqrt{\frac{\sum \mathrm{X}^2}{\mathrm{N}} - \left(\frac{\sum X}{N}\right)^2}$$

Una desviación estándar pequeña para un conjunto de valores indica que éstos se encuentran localizados cerca de la media. Por el contrario, una desviación estándar grande revela que las observaciones están muy dispersas con respecto a la media

Varianza muestral

La fórmula para la varianza muestral utilizada como estimador de la varianza poblacional es

$$s^2 = \frac{\sum (X-\overline{X})^2}{n-1} \qquad \text{o bien} \quad \mathrm{s}^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1}$$

¿Por qué se hizo al denominador esta modificación? Puede demostrarse que si se hubiera calculado la varianza muestral utilizando sólo n en el denominador, el resultado subestimaría la varianza poblacional. Esto es, la varianza muestral sería un estimador sesgado de la varianza poblacional

Desviación estándar muestral

La desviación estándar de una muestra se utiliza como un estimador de la desviación estándar de la población

$$s = \sqrt{\frac{\sum (X-\overline{X})^2}{n-1}} \qquad \text{o bien} \quad \mathrm{s} = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1}}$$

Interpretación y usos de la desviación estándar

Por lo común la desviación estándar se emplea como una medida para comparar la dispersión en dos o más conjuntos de observaciones. Por ejemplo, la desviación estándar de las cantidades quincenales invertidas en el plan de participación de las utilidades de una empresa se ha calculado como 7,51 dólares. Supóngase que tal empresa tiene una rama en el sur. Si la desviación estándar para otro grupo de empleados en el oeste es 10,47 dólares y las medias son aproximadamente iguales, esto indica que las cantidades invertidas por los empleados del sur no se dispersan tanto como las de los empleados del oeste (porque 7,51<10,47). Ya que las cantidades invertidas por los empleados del sur se acumulan a la media, el valor medio para estos trabajadores es una medida más confiable que la media para el grupo del oeste.

La desviación estándar se interpreta como cuánto se desvía- en promedio- de la media un conjunto de puntuaciones

Supóngase que un obtuvo para su muestra una media de ingreso familiar de $800000 y una desviación estándar de $100000. La interpretación es que los ingresos familiares de la muestra se desvían en promedio respecto a la media en cien mil pesos..

La desviación estándar sólo se utiliza en variables medidas por intervalos o de razón

Teorema de Chebyshev
El matemático ruso Chebyshev desarrolló un teorema que permite determinar la proporción mínima de los valores que se encuentra dentro de un número específico de desviaciones estándares con respecto a la media.
Por ejemplo, con base en el teorema de Chebyshev, al menos tres de cada cuatro valores, o 75%, deben encontrarse entre la media más dos desviaciones estándares y la media menos dos desviaciones estándares.
Esta relación se aplica sin importar la forma de la distribución.
Además, al menos ocho de cada nueve valores, o 89,9%, se encontrarán entre la media más tres desviaciones estándares y la media menos tres desviaciones. Al menos 24 de 25 valores, o 96%, se encontrarán entre la media y menos cinco desviaciones.

En términos generales, el teorema de Chebyshev establece que;

Teorema de Chebyshev; Para un conjunto cualquiera de observaciones (muestra o población), la proporción mínima de los valores que se encuentran dentro de k desviaciones estándares desde la media es al menos $1-1/k^2$

donde k es una constante mayor que 1.

Ejemplo: suponga que la cantidad media quincenal depositada por los empleados de una empresa en el plan de participación de utilidades de la empresa fue $51,04 y se obtuvo una desviación estándar de $7,51. Al menos ¿qué porcentaje de las contribuciones se encuentran a una distancia de más dos desviaciones estándares y menos dos desviaciones estándares de la media?
Solución; Aproximadamente 75%, que se obtiene al calcular

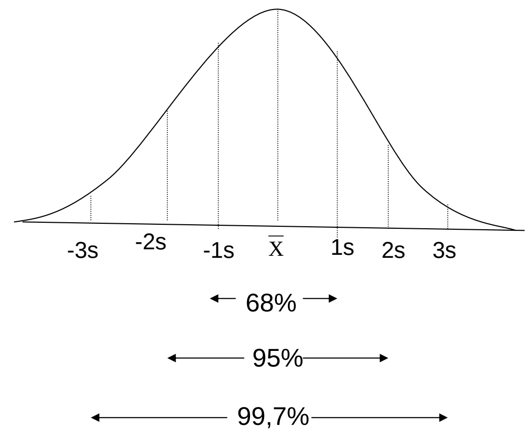
$$1-1/k^2 = 1-1/2^2 = 1-\frac{1}{4} = 0{,}75$$

El teorema de Chebyshev se refiere a cualquier conjunto de valores; esto es, la distribución de los valores puede tomar cualquier forma. Sin embargo, para una curva de distribución simétrica en forma de campana podemos ser más precisos al explicar la dispersión.

Para una curva de distribución simétrica en forma de campana, se puede ser preciso al explicar la dispersión con respecto a la media

Regla empírica (regla normal): Para una distribución de frecuencias simétrica de campana aproximadamente 68% de las observaciones se encontrará a más y menos una desviación estándar de la media; aproximadamente 95% de las observaciones se encontrarán a más y menos dos desviaciones estándares desde la media; y prácticamente todas las observaciones 99,7% se encontrarán a más y menos tres desviaciones desde la media

Curva simétrica de campana, que muestra las relaciones entre desviación estándar y media



Ejemplo:
Se observa que las cuentas de energía eléctrica de una zona residencial correspondiente al mes de junio tienen una distribución normal. Si se calcula que la media de estas cuentas es de \$84, con una desviación estándar de \$24, entonces se desprende que aproximadamente 68% de las cantidades facturadas se encuentran entre \$60 y \$108. Asimismo de desprende que aproximadamente 95% de las cantidades facturadas se hallan entre \$36 y \$132

Nota:
El concepto de desviación estándar es muy importante en los negocios y en la economía. Por ejemplo, en finanzas la desviación estándar se utiliza como medida de riesgo relacionada con varias oportunidades de inversión.
Mediante el uso de la desviación estándar para medir la variabilidad en las tasas de rendimiento ofrecidas por diferentes inversiones, el analista financiero puede medir el nivel de riesgo que tiene cada activo financiero.

Ejercicio: Markus Boggs es gerente de Inversiones S:A. Markus estaba interesado en las tasas de rendimiento de los últimos cinco años de dos diferentes fondos mutuos. Fondo mutuo1 mostró, durante un período de cinco años, tasas de rendimiento del 12, 10, 13, 9 y 11 % mientras que Fondo mutuo 2 arrojó 13, 12, 14, 10 y 6%. Un cliente se acercó a Boggs y expresó su interés en uno de estos fondos mutuos ¿ Cual debería escoger Boggs para su cliente? Ayuda: Una inversión más segura es la que tiene un grado menor de riesgo (el riesgo se mide por la desv estándar)

Dispersión Relativa

El coeficiente de variación (C.V) es una medida muy útil cuando:
1.-Los datos están en unidades diferentes (como dólares y días de inasistencia).
2.- Los datos están en las mismas unidades, pero las medias muy distantes (como sucede con los ingresos de los ejecutivos y los ingresos de los empleados no calificados)
3.- Cuando se desea comparar la variabilidad de dos conjuntos de datos

Coeficiente de variación: Indica la magnitud relativa de la desviación estándar en comparación con la media de la distribución, expresada como porcentaje.

Población $$C.V = \frac{\sigma}{\mu}.100$$

Muestra $$C.V = \frac{s}{\overline{X}}.100$$

Ejemplo:
Se va a comparar la variación en los ingresos anuales de ejecutivos con variación en los ingresos de trabajadores no calificados.
Para una muestra de ejecutivos, $\overline{X} = \$500000 \ y \ s = \$50000$
Para una muestra de trabajadores no calificados, $\overline{X} = \$12000 y \ s = \$1200$
Uno se ve tentado a afirmar que hay mayor dispersión en los ingresos anuales de los ejecutivos porque $50000>$1200. Sin embargo, las medias están tan distantes que se necesitan convertir las estadísticas a coeficientes para efectuar una comparación significativa de la variación en los ingresos anuales.

Para los ejecutivos; CV= 10%
Para los trabajadores no calificados; CV= 10%

No existe diferencia en la dispersión relativa de los dos grupos

Ejemplo

Un estudio de las calificaciones obtenidas en un curso interno sobre principios de administración y los años de servicio de los empleados inscritos en el curso, dio como resultado estas estadísticas:
la calificación media fue 200; la desviación estándar 20. La media del número de años de servicio fue 18 años, la desviación estándar de 2,16 años. Compare la dispersión relativa de las dos distribuciones.

Para las calificaciones: C.V= 10%.
Para los años de servicio: C.V = 12%

6. PUNTUACIONES Z (Medida de localización relativa)

Las puntuaciones "z" son transformaciones que se pueden hacer a los valores o puntuaciones obtenidas, con el propósito de analizar su distancia respecto a la media, en unidades de desviación estándar.

$$Z = \frac{X - \overline{X}}{s}$$

donde X es la puntuación o valor a transformar

$\overline{X}$ es la media de la distribución

s la desviación estándar de ésta.

Z es la puntuación transformada en unidades de desviación estándar

Una puntuación "Z" nos indica la dirección y grado en que un valor individual obtenido se aleja de la media, en una escala de unidades de desviación estándar.

Las puntuaciones Z son el método más comúnmente utilizado para estandarizar la escala de una variable medida en un nivel por intervalos

Ejemplo:
Supongamos que en una distribución de frecuencias obtuvimos una media de 60 y una desviación estándar de 10, y deseamos comparar a una puntuación de 50 con el resto de la distribución. Entonces, transformamos esta puntuación o valor en una puntuación Z. Tenemos que:
$X=50$, $\bar{X}=60$, $s=10$
la puntuación Z correspondiente a un valor de 50 es:
$Z= (50-60)/10= -1$

Podemos decir que el valor 50 está localizado a una desviación estándar por debajo de la media de la distribución
(el valor “30” está a tres desviaciones estándar por debajo de la media)

Estandarizar los valores permite comparar puntuaciones de dos distribuciones diferentes.
Por ejemplo, podemos comparar una distribución obtenida en una preprueba con otra obtenida en una postprueba (en un contexto experimental).

Supongamos que se trata de un estímulo que incrementa la productividad. Un trabajador obtuvo en la preprueba una productividad de 130 (la media del grupo fue de 122,5 y la desviación estándar de 10). Y en la postprueba obtuvo 135 (la media del grupo fue de 140 y la desviación estándar de 9,8).
¿mejoró la productividad del trabajador?
Aparentemente la mejoría no es considerable

Sin transformar las dos calificaciones en puntuaciones Z no podemos asegurarlo porque los valores no pertenecen a la misma distribución. Entonces transformamos ambos valores a puntuaciones Z, los transformamos a una escala común, donde la comparación es válida.

El valor de 130 en productividad es en términos de unidades de desviación estándar igual a $z= (130-122,5)/10=0,75$
y el valor de 135 corresponde a una puntuación
$Z =(135-140)/9,8= -0,51$

Como podemos observar, en términos absolutos 135 es una mejor puntuación que 130, pero no en términos relativos (en relación a sus respectivas distribuciones)

La variable estandarizada expresa la posición de un elemento en una distribución dada, tanto en relación a la media como a la desviación estándar quedando libre de toda ambigüedad.

Ejemplo:
Un alumno obtuvo las siguientes notas:
Historia=5; Francés=4; Matemáticas=4,5; Castellano=6
¿Qué opinión puede dar Ud de estas notas?
Si las compara con la nota 4, que es el promedio de las notas de 1 a 7 y la que se exige como mínimo para aprobar, puede concluir que ese alumno tiene una buena nota en Historia y el mínimo en Francés.
Pero ¿puede decir si el alumno dentro de su curso es de los mejores?.

Para discernir esto es necesario acudir al promedio y a la desviación estándar de las notas de esos ramos de todos los alumnos.

	Historia	Francés	Matemáticas	Castellano
Promedio de curso	6,5	4	4,5	5,6
Desviación estándar	1	2	3	2
Var estandarizada Z	-1,5	0	0	0,2

Aquí vemos que ha cambiado la faz de las notas, ya que un 5 en Historia significa que está muy por debajo del resto de sus compañeros, en cambio el 4 en Francés y el 4,5 en Matemáticas indican que está dentro del promedio del curso.

La distribución de puntuaciones Z no cambia la forma de la distribución original, pero si modifica las unidades originales a unidades de desviación estándar. La distribución de puntuaciones Z tiene una media 0 y una desviación estándar 1.

Las puntuaciones Z son un elemento descriptivo adicional que podemos agregar para analizar nuestros datos

Detección de valores atípicos

A veces un conjunto de datos tiene uno o más elementos con valores demasiado grandes o demasiados pequeños. A los valores extremos como éstos se les llama **valores atípicos**.
Un valor atípico puede ser un elemento para el cual se haya anotado su valor en forma errónea. Si es así, puede corregirse antes de proseguir el análisis. También, un valor atípico puede ser uno que por error se incluyó en el conjunto de datos y, en estos casos, debe eliminarse. Puede ser tan sólo un elemento poco común que se haya anotado en forma correcta y que sí pertenece al conjunto de datos.

Los valores estandarizados (valores z) pueden emplearse para identificar los valores atípicos.
Para identificar valores atípicos se recomienda considerar que cualquier elemento con un valor z inferior a -3 o superior a 3 sea tratado como un valor atípico. La exactitud de esos elementos se podrá revisar después para determinar si pertenecen al conjunto de datos o no.

OBS: Con los diagramas de caja también se pueden identificar los valores atípicos, pero no necesariamente los mismos valores que aquellos menores que 3 o mayores que 3 en los valores z. SE puede usar cualquiera de esos métodos o ambos. Se trata de identificar valores que podrían no pertenecer al conjunto de datos.

NUMEROS INDICES

Un número índice es una cifra relativa (expresada en forma de porcentaje), que representa las variaciones medias en precio, cantidad o valor, de uno o más ítem en una época dada, respecto del período base.
Los índices tratan de cuantificar variaciones, y no expresan si los precios son altos, o si se ha producido mucho. Sólo pretenden comparar cifras con otras que se consideran como referencia.

Los índices se utilizan entonces, para comparaciones en el tiempo y en el espacio.
Los índices no miden sino que indican una cierta tendencia.

Indices simples

En muchos casos es preciso expresar las diversas cifras de una serie anual o mensual en función de una que considera como base. En este caso, se trata de índices simples, que representan el porcentaje de cada cifra de la serie, respecto del valor observado en el período base.

Si se designa por Xo el valor de la observación para el período base y por Xt el de los restantes períodos, el índice It para el período t queda expresado por el cuociente de ambas observaciones.

Generalmente se acostumbra expresar los índices en forma porcentual, de modo que la relación anterior se multiplica por 100.

$$I_t = 100 \frac{X_t}{X_0}$$

Energía Eléctrica:
Producción total e índice de consumo eléctrico industrial

		Indice de consumo eléctrico			
Periodo	Producción total (Mill de kWh)	Base Prom 1994=100	Base Prom 1996=100	Base Prom 1998=100	% aumento anual
1994	21.929,10	100,0	97,8	85,5	
1995	23.629,80	107,8	105,4	92,1	7,8
1996	22.411,70	102,2	100,0	87,4	-5,2
1997	23.953,50	109,2	106,9	93,4	6,9
1998	25.649,80	117,0	114,4	100,0	7,1

Fuente: Boletín Mensual del Banco Central de Chile (Nov 1999

Para calcular el porcentaje de aumento A entre dos cantidades Xo y Xt conviene recordar la siguiente relación

$$A = 100\left(\frac{X_t}{X_0} - 1\right)$$

Utilización de los índices

Las actividades económicas precisan de algunos indicadores que permitan cuantificar en forma sintética el desarrollo de los acontecimientos. Los organismos estatales o privados han confeccionado números índices para diferentes fenómenos económicos. El uso de esos indicadores es variado y depende en cada oportunidad de su contenido específico.

Las ventas comerciales, los salarios pagados, el ingreso nacional, los presupuestos fiscales, etc, han sufrido variaciones por diferentes causas, una de las cuales es el constante aumento de los precios. Por lo tanto, es necesario eliminar dicho efecto para obtener series depuradas, que se refieran a valores reales.

Se precisa entonces, deflactar las series expresadas en valores monetarios de cada periodo, con el objeto de que representen valores monetarios de igual poder adquisitivo.

Para la deflactación debería utilizarse el índice más apropiado, el que no siempre existe. Por lo general, se emplean índices de precios al por mayor o al consumidor, que pese a no ser siempre estrictamente aplicables, permiten eliminar en forma relativamente aceptable, la deformación introducida en la serie por las variaciones de los precios.

CALCULO DEL IPC

Los economistas definen inflación como un aumento generalizado y sostenido en el tiempo del nivel general de precios en la economía. Es decir, en palabras más simples, hay inflación cuando existe una tendencia general de los precios a subir.
El Indice de Precios al Consumidor o IPC es el indicador mensual que mide la inflación en Chile. El IPC del país está compuesto de una canasta de bienes y servicios y cada uno de ellos tiene una ponderación distinta de acuerdo con la importancia de éstos en el presupuesto familiar de los chilenos.

El IPC es el Índice de Precios al Consumidor e intenta reflejar la variación de la inflación en un determinado período. El cuadro señala que el índice de IPC para enero de 1999 era de 99,67 puntos, partiendo de un índice de 100 en Dic de 1998, lo que comúnmente se denomina "base 100".

¿Cuánto ha variado la inflación entre Dic de 1998 y Octubre del 2000. Para calcularlo, ocupamos la siguiente fórmula:

Variación IPC Dic 1998 a Oct 2000 = ((Índice IPC Oct 2000/Índice IPC Dic 1998) –1) X 100 = 6,46%,
La variación es igual a 6,46%, que equivale a la inflación acumulada en ese período.

¿Cuánto ha variado la inflación entre Abril del 2000 y Julio del 2000? Para calcularlo, ocupamos la siguiente fórmula:

Variación IPC Abr 2000 a Julio 2000 = ((Índice IPC Julio 2000/Índice IPC Abr 2000) –1) X 100

=(104,91/104,31 -1)x100=0,5752%

Observe que la inflación acumulada de Mayo, Junio y Julio no es la suma de las variaciones mensuales (es decir 0,2109+0,2295+0,1336=0,574), sino que algo mayor 0,5752 %

Por otra parte, también se puede determinar el índice conociendo la inflación del período. Considere la inflación en Enero de 1999, equivalente a –0,33% y en Feb de 1999 equivalente a 0,07%. Si la base 100 se define en Dic 1998. ¿ Cuál es el índice de IPC acumulado en Enero y Febrero?

Debemos multiplicar
Base 100x (1+IPC Enero)=100x(1+-0,0033)=99,67=Indice IPC Enero 1999

Base 100x (1+IPC Enero)x(1+IPC Feb)=
100x(1+-0,0033)x(1+0,0007)=99,74= Indice Feb 1999

Ejercicio:

a) Considere la siguiente serie de valores que representan ingresos en pesos corrientes (pesos de dic de cada año). Dado el índice IPC (base dic 1998=100) calcule el deflactor IPC base dic1999=100 y luego estime la serie de ingresos en base dic 1999=100.
b) Resuelva el ejercicio anterior usando alguna forma equivalente

Año	Ingresos corrientes	IPC
	($ dic de cada año)	(dic 1998=100)
1989	116664	37,5093
1990	131678	47,7601
1991	258790	56,6713
1992	185600	63,8659
1993	245679	71,6789
1994	456876	78,0909
1995	345721	84,4933
1996	245696	90,0969
1997	234567	95,5428
1998	236458	100
1999	345672	102,31

Año	Ingresos corrientes	IPC
	($ dic de cada año)	(dic 1998=100)
1989	116664	37,5093
1990	131678	47,7601
1991	258790	56,6713
1992	185600	63,8659
1993	245679	71,6789
1994	456876	78,0909
1995	345721	84,4933
1996	245696	90,0969
1997	234567	95,5428
1998	236458	100
1999	345672	102,31

Notar que el deflactor del IPC base 1999 es igual

$$\text{Deflactor IPC base dic 99} = 1 + \frac{\text{IPCdic añot - IPCdic año1999}}{\text{IPC dic año1999}}$$

Lo que equivale a

$$\text{Deflactor IPC base dic 99} = \frac{\text{IPCdic añot}}{\text{IPC dic año1999}}$$

	A	B	C	A/C
Año	Ingresos corrientes	IPC	Deflactor IPC	Ingreso en $ constantes
	($ dic de cada año)	(dic 1998=100)	(base dic 99)	(base dic1999=100)
1989	116664	37,5093	0,367	318212
1990	131678	47,7601	0,467	282076
1991	258790	56,6713	0,554	467200
1992	185600	63,8659	0,624	297322
1993	245679	71,6789	0,701	350667
1994	456876	78,0909	0,763	598571
1995	345721	84,4933	0,826	418622
1996	245696	90,0969	0,881	279001
1997	234567	95,5428	0,934	251181
1998	236458	100	0,977	241920
1999	345672	102,31	1,000	345672

$$VALORES \quad REALES = \frac{Valores \quad nominales}{Deflactor}$$

b) Formas equivalentes de estimar los ingresos reales en base dic 1999 son las siguientes

	A	B	C	(A/B)*102,31
Año	Ingresos corrientes	IPC	A/B	Ingreso en $ constantes
	($ dic de cada año)	(dic 1998=100)		(base dic1999=100)
1989	116664	37,5093	3110,269	318212
1990	131678	47,7601	2757,071	282076
1991	258790	56,6713	4566,509	467200
1992	185600	63,8659	2906,089	297322
1993	245679	71,6789	3427,494	350667
1994	456876	78,0909	5850,566	598571
1995	345721	84,4933	4091,697	418622
1996	245696	90,0969	2727,019	279001
1997	234567	95,5428	2455,099	251181
1998	236458	100	2364,580	241920
1999	345672	102,31	3378,673	345672

	A	B	C	(A/C)*100
Año	Ingresos corrientes	IPC	IPC	Ingreso en $ constantes
	($ dic de cada año)	(dic 1998=100)	(dic 1999=100)	(base dic1999=100)
1989	116664	37,5093	36,662	318212
1990	131678	47,7601	46,682	282076
1991	258790	56,6713	55,392	467200
1992	185600	63,8659	62,424	297322
1993	245679	71,6789	70,061	350667
1994	456876	78,0909	76,328	598571
1995	345721	84,4933	82,586	418622
1996	245696	90,0969	88,063	279001
1997	234567	95,5428	93,386	251181
1998	236458	100	97,742	241920
1999	345672	102,31	100,000	345672

Aplicación. La Unidad de Fomento (UF)

La Unidad de Fomento (UF) es una medida reajustable basada en la variación del Indice de Precios al Consumidor (IPC).

- **El Ministerio de Hacienda en su Decreto Supremo Nº 40 del 2 de enero de 1967, creó la Unidad de Fomento. El valor de la UF se reajustaba el primer día de cada trimestre según la variación del IPC del trimestre anterior.**

- **El Decreto Supremo Nº 280 del 12 de Mayo de 1975 estableció que la UF pasaría a reajustarse en forma mensual.**

Aplicación. La Unidad de Fomento (UF)

- **El Decreto Supremo Nº 613 del 14 de julio de 1977 estableció que su valor se reajustaría en forma diaria a partir del 1º de Agosto de dicho año.**

- **Mediante el acuerdo del 08-Ene-1990 y según las normas del Banco Central la UF continúa reajustándose en forma diaria, siendo calculada a principios de cada mes para el período comprendido entre el día 10 de dicho mes y el día 9 del mes siguiente, de acuerdo a la tasa promedio geométrica de la variación del IPC del mes anterior.**

Cálculo de la UF

El valor de la UF para cada día se determinará sobre la base del valor del día anterior, según la siguiente fórmula:

$$UF_{día} = UF_{día-1} \times Rd$$

Donde:
Rd = Factor de reajuste diario del valor de la Unidad de Fomento.
d = Nº de días comprendidos en el período para el cual se calcula el valor diario de la UF.
vIPC-1 = Variación porcentual del IPC registrada en el mes inmediatamente anterior.

Factor de reajuste diario del valor de la Unidad de Fomento.

$$Rd = \left(1 + \frac{vIPC_{-1}}{100}\right)^{\frac{1}{d}}$$

Ejercicio: Estime el valor de la UF diaria del mes de Marzo del 2009

Regresión lineal simple

Definición: Es un modelo matemático para estimar el efecto de una variable sobre otra. Está asociado con el coeficiente r de Pearson.

Hipótesis a probar: Correlacionales y causales

Variables: Dos. Una se considera como independiente y otra como dependiente.

Nivel de medición de las variables: Intervalos o razón

Procedimiento e interpretación: La regresión lineal se determina en base al diagrama de dispersión. Los diagramas de dispersión son una manera de visualizar gráficamente una correlación. Este diagrama puede ser resumido a una línea

Conociendo la línea y la tendencia, podemos predecir los valores de una variable conociendo los de la otra variable.
Esta línea se expresa mediante la ecuación de regresión lineal

$$\hat{Y} = a + bX$$

donde Y "sombrero " es un valor de la variable dependiente que se desea predecir, "a" es la ordenada en el origen y "b" la pendiente o inclinación.
Los programas y paquetes de análisis estadístico por computadora que incluyen la regresión lineal proporcionan los datos de a y b.
Para predecir un valor de Y se sustituyen los valores correspondientes en la ecuación

Nota; La regresión lineal es útil con relaciones lineales, no con relaciones curvilíneas

Método de mínimos cuadrados para el ajuste de una línea de regresión

La ecuación lineal que representa al modelo de regresión lineal simple es

$$\mathrm{Y_i} = \alpha + \beta X_i + \varepsilon_i$$

Donde
Yi =Valor de la variable dependiente en la i-ésima observación
α =primer parámetro de la ecuación de regresión, el cual indica la intersección con el eje Y.
β = segundo parámetro de la ecuación de regresión, el cual indica la pendiente de la línea de regresión
Xi= el valor especificado de la variable independiente en la i-ésima observación.
εi=error de muestreo aleatorio en la i-ésima observación

Los parámetros α y β del modelo de regresión lineal se estiman con los valores de a y b, que se basan en los datos muestrales. Así, la ecuación de regresión lineal basada en datos muestrales que se usa para estimar un solo valor (condicional) de la variable dependiente, donde el "sombrero" sobre la Y indica que se trata de un valor estimado es

$$\hat{Y} = a + bX$$

De acuerdo con el criterio de mínimos cuadrados, la línea de regresión del menor ajuste (y la mejor ecuación) es aquella para la cual se reduce al mínimo la suma de las desviaciones cuadradas entre los valores estimado y real de la variable dependiente para los datos muestrales.

$$Mín \sum (Y_i - \hat{Y}_i)^2$$

Las fórmulas de cálculos por las cuales pueden determinarse los valores de a y b en la ecuación de regresión para la ecuación que satisface el criterio de mínimos cuadrados son

$$b = \frac{\sum XY - n\overline{X}\,\overline{Y}}{\sum X^2 - n\overline{X}^2} \qquad a = \overline{Y} - b\overline{X}$$

Una vez formulada la ecuación de regresión, puede servir para estimar el valor de la variable dependiente dado el valor de la variable independiente. Sin embargo, esta estimación sólo debe realizarse dentro del rango de los valores de la variable independiente originalmente muestreada, ya que no existe base estadística para suponer que la línea de regresión es adecuada fuera de estos límites.

Valor residual

Nota:Para un valor X de la variable independiente, el valor Y "sombrero" de la línea de regresión suele denominarse valor ajustado de la variable dependiente.La diferencia entre el valor Y observado y el valor Y sombrero ajustado se llama el residual de esa observación y se denota por e

$$e = Y - \hat{Y}$$

Una gráfica de residuales se obtiene trazando los residuales e respecto de la variable independiente X o alternativamente respecto delos valores Y sombrero de la línea de regresión ajustada.

Error estándar de estimación

Error estándar de estimación: Mide la dispersión de los valores observados, con respecto a la recta de regresión. Se determina por medio de la ecuación:

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

El error estándar de la estimación, mide cuán cerca de la recta de regresión se encuentran los valores reales. Cuando el error estándar es pequeño, indica que las dos variables están relacionadas muy de cerca.
Por otra parte el coeficiente de correlación mide la fuerza de la asociación entre dos variables. Cuando los puntos sobre el diagrama de dispersión parecen cercanos a la línea recta, se observa que el coeficiente de correlación tiende a ser grande

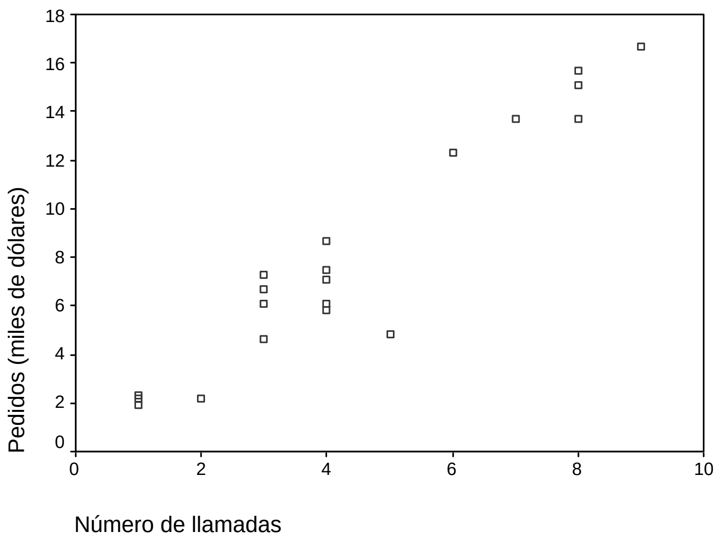
Ejemplo: El gerente de ventas de una compañía se está preparando para una reunión de ventas, y le gustaría mostrar al grupo de vendedores la forma como se relaciona el número de llamadas a clientes con el valor anual de pedidos que se reciben. De sus registros recolectó la siguiente información muestral para el último año.

Número de llamadas	Pedidos (miles de dólares)		Número de llamadas	Pedidos (miles de dólares)
5	4,8		2	2,2
4	6,1		4	7,1
6	12,3		4	8,7
7	13,7		8	13,7
8	15,7		1	2,3
1	2,2		3	4,6
3	7,3		9	16,7
4	5,8		3	6,1
1	1,9		4	7,5
3	6,7		8	15,1

A partir de estos datos muestrales
a) Realice un gráfico de dispersión
b) Determine el coeficiente de correlación
c) Desarrolle una ecuación de regresión para predecir el valor de pedidos a partir del número de llamadas
d) Cuál es el valor de pronóstico de pedidos si se realizan cinco llamadas?

Solución

Gráfico de dispersión



b) Coeficiente de correlación r = 0,956
Nivel de significancia s = 0,000
La correlación es significante al nivel 0,01

c) $\hat{Y} = a + bX$

$Número \quad de \quad pedidos = -0{,}105 + 1{,}85\,(llamadas)$

d) $\hat{Y} = a + bX$

$Número \quad de \quad pedidos = -0{,}105 + 1{,}85\,(5) = 9{,}145$

Notar que el coeficiente de determinación es 0,915, indicando que aproximadamente 91,5% de la variación en los pedidos se explica por los números de llamadas

Ejemplo:¿Cuál es la relación entre la cantidad gastada por semana en alimentos y el tamaño de una familia? ¿Las familias grandes gastan más en alimentos?. Una muestra de 10 familias en una ciudad reveló los siguientes tamaños de familia e importes en dinero gastados en alimentos, en cierto periodo.

Tamaño de la familia	Cantidad gastada en alimentos
3	99
6	104
5	151
6	129
6	142
3	111
4	74
4	91
5	119
3	91

a) Determine la ecuación de regresión
b) Obtenga el error estándar de la estimación
c) estime la cantidad de dinero que una familia de cuatro personas gastará en alimentos

Solución:

a) $\hat{Y} = a + bX$

$Cantidad \quad gastada = 60.4 + 11{,}3\,(tamaño)$

b) $S_{Y,X} = \sqrt{\dfrac{\sum (Y - \hat{Y})}{n-2}} = \sqrt{\dfrac{3467{,}2}{8}} = 20{,}818$

C) Cantidad gastada = 60,4+11,3 (4)=105,6

Obs: En este ejemplo el coeficiente de correlación es r=0,589 y s=0,037. La correlación es significativa al nivel 0,05

Consideraciones de base para la regresión lineal

1. Para cada valor de X existe un grupo de valores de Y, y estos valores se distribuyen en forma normal.
2. Las medias de estas distribuciones normales de valores Y se encuentran todas en la línea de regresión
3. Las desviaciones estándares de dichas distribuciones normales son iguales.
4. Los valores Y son estadísticamente independientes. Esto significa que al seleccionar una muestra, los valores Y seleccionados para un valor de X específico no dependen de los valores Y para cualquier otro valor X

Representación gráfica de las consideraciones básicas de la regresión