

UNIDAD 4

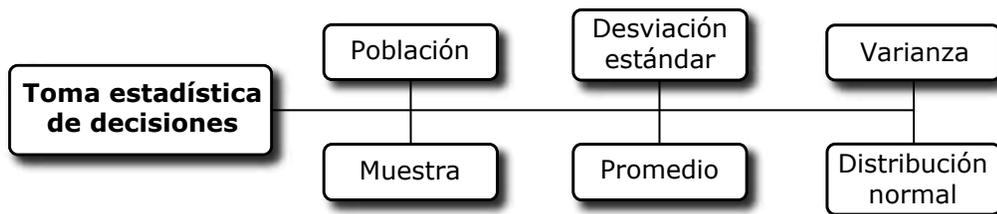
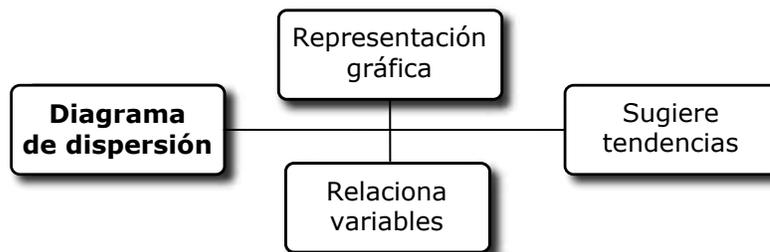
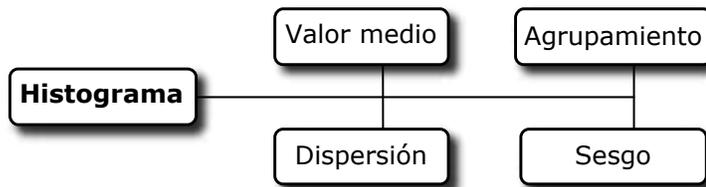
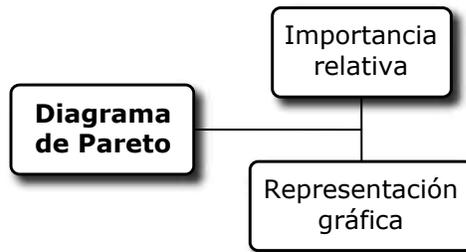
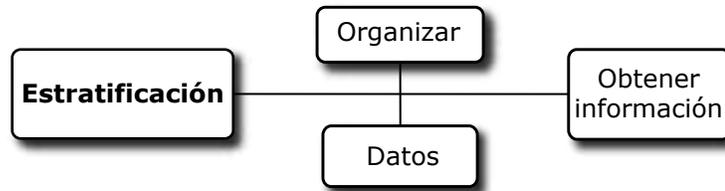
HERRAMIENTAS ESTADÍSTICAS DE CALIDAD

Objetivos

Al finalizar la unidad, el alumno:

- Distinguirá las diferentes herramientas estadísticas comúnmente utilizadas para identificar problemas de calidad dentro de un proceso.
- Utilizará diagramas de estratificación, de Pareto y de dispersión para identificar las causas de problemas de calidad.
- Interpretará histogramas, diagramas de dispersión y el coeficiente de correlación relacionados con el mejoramiento de la calidad.
- Ejemplificará en problemas de calidad la toma de decisiones basada en pruebas de hipótesis estadísticas.

CONCEPTOS CENTRALES



INTRODUCCIÓN

En la práctica, la calidad es una disciplina que se estudia por medio de herramientas estadísticas diseñadas para evaluar datos, obtener información, formar juicios y tomar decisiones en condiciones de incertidumbre, sobre productos y procesos. En esta unidad veremos los fundamentos de las más importantes.

La estratificación es una estrategia inicial que nos permite organizar y observar los datos de diferentes maneras, con el fin de ver los problemas desde varios puntos de vista y así distinguir las relaciones entre sí y entre las causas y sus efectos visibles. De esta manera, se pueden simplificar y comprender los problemas más complejos. No es una herramienta que deba usarse sola, siempre debe ir acompañada de las demás.

El Diagrama de Pareto es una técnica conocida en diversos medios, aparte de los relacionados con la calidad. Es una herramienta que frecuentemente se aplica después de la estratificación y que nos permite observar la importancia relativa de los factores de un problema; se puede utilizar para estudiar causas o efectos, costos, resultados o simplemente para ver cómo se acomodan los elementos que componen un fenómeno.

A continuación veremos cómo se construyen y qué información nos dan los histogramas. Esta herramienta se utiliza para presentar gráficamente el comportamiento de un proceso; con ella se busca determinar las características de su funcionamiento, saber si es aleatorio, si tiene variaciones y cómo se presentan.

Una vez establecidas las relaciones de causa y efecto es necesario determinar cómo influyen las causas en los efectos, para ello nos apoyamos en los diagramas de dispersión y en el cálculo del coeficiente de correlación. Estas herramientas nos permiten medir qué grado de influencia tiene una causa sobre diferentes efectos o cuál es la causa que mayor influencia ejerce sobre un efecto particular.

Al final de la unidad veremos los procedimientos que se siguen para la toma de decisiones basadas en la estadística, tal es el caso de la prueba de hipótesis sobre promedios basados en muestras. Esta herramienta es de gran importancia porque es de aplicación común en varias disciplinas; nos basta escuchar cualquier discusión sobre eventos de actualidad para oír la expresión: “en promedio...”, lo cual significa que alguien ha tomado una muestra, ha medido alguna variable de ella y pretende convencer a su audiencia de que la mayoría de la población se comporta o piensa igual que su promedio. Esta unidad está dedicada al estudio de los fundamentos que nos permiten aplicar correctamente estas herramientas.

4.1 Estratificación

Dos de los vicios más frecuentes en nuestro lenguaje son la *generalización* y la *simplificación*: “nunca llegas temprano” y “siempre logro mis objetivos”, son ejemplos de generalizaciones. Por otro lado, pareciera que los problemas más complejos (casi siempre en manos ajenas) pudieran resolverse fácilmente; si nos preguntaran qué opinamos acerca de un problema ajeno, diríamos: “¡fácil!, déjate de problemas, lo único que tienes que hacer es...” o “sigue un sencillo consejo para solucionar todos los problemas...”.

¿Qué es la estratificación?

La vida cotidiana, en general, y la vida profesional de forma constante se enfrentan a problemas complejos que no tienen soluciones simples, por el contrario son conflictivas o producen nuevos problemas rodeados por una nube de datos, opiniones antagónicas y alternativas. La **estratificación** es una herramienta que permite despejar esa nube para llegar al fondo del problema y destacar sus causas.

¿En qué consiste la estratificación?

Para la solución de un problema de calidad es necesario organizar la multitud de datos que usualmente aparecen como parte de los síntomas o de sus causas reales. La estratificación consiste en **agrupar las observaciones** de acuerdo con su naturaleza o con otros factores que influyen en el problema, como pueden ser: los tipos de fallas, la frecuencia con que se presentan, los lugares donde aparecen, las personas o situaciones involucradas, los proveedores, las máquinas, etcétera.

La estratificación es una estrategia inicial, no es por sí misma una herramienta que permita solucionar un problema, sino que es una representación para comprenderlo mejor, mediante la separación y clasificación de los datos disponibles. Un mismo conjunto de datos puede ser agrupado de diferentes maneras y su interpretación puede variar según el punto de vista de la clasificación.

Supongamos que en una fábrica se analiza un problema de puntualidad. Los datos iniciales podrían presentarse mediante una lista como la de la tabla 4.1.

Persona	L	M	M	J	V	L	M	M	J	V	L	M	M	J	V	L	M	M	J	V	L	Promedio
Smithe	2	4	1	3	2	0	5	1	4	3	0	3	2	1	1	4	5	2	3	0	3	2.3
Olafson	0	2	1	1	0	4	6	2	1	3	4	5	2	2	0	1	4	2	2	1	2	2.1
Krueger	5	7	3	4	6	4	8	4	3	7	3	9	5	6	3	6	7	5	3	5	4	5.1
Osha	2	5	1	3	2	4	6	1	0	3	2	4	2	1	3	4	5	2	0	4	3	2.7
Takima	5	6	4	2	3	3	4	2	5	1	0	6	5	4	3	5	6	2	4	3	2	3.6

Tabla 4.1 Minutos de retraso por persona durante los días laborales del mes.

Al analizarlo se observa que Krueger destaca por su promedio de 5.1 minutos de retraso diario, seguido por Takima con 3.6. La impuntualidad parece estar localizada en estas dos personas; sin embargo, la estratificación puede ser aplicada en forma sucesiva. Como primer paso se organizan los datos conforme a un criterio y como segundo, la misma organización podría ser refinada de

acuerdo con un segundo criterio. En nuestro caso, una vez detectado el fenómeno de impuntualidad destacada en dos personas, se podría refinar la tabla comparando el promedio de retrasos por día de la semana de todas las personas, con el fin de buscar algún patrón de conducta.

Persona	L	M	M	J	V	Promedio total
Smithe	1.8	4.3	1.5	2.3	1.5	2.3
Olafson	2.2	4.3	1.8	1.5	1.0	2.1
Krueger	4.4	7.8	4.3	4.0	4.3	5.1
Osha	3.0	5.0	1.5	1.0	3.0	2.7
Takima	3.0	5.5	3.3	3.8	2.5	3.6
Promedio por día	2.9	5.4	2.5	2.5	2.5	3.2

Tabla 4.2 Promedio de minutos de retraso por día y por persona del mes.

La estratificación del problema nos indica que los martes hay otro patrón en los retrasos repetitivos de todos los trabajadores; además de los retrasos de Krueger y Takima el promedio de impuntualidad de todas las personas se incrementa los martes casi al doble. Es probable que sean dos fenómenos separados, pero si no se hubiera analizado el conjunto de datos desde dos ángulos diferentes, no se hubiera detectado esta dualidad.

La estratificación es un instrumento que se aplica en problemas, especialmente en aquellos que son complejos, que tienen muchas causas que se relacionan unas con otras. Se utiliza en la solución de problemas de calidad para separar los factores técnicos como la maquinaria, la herramienta y la materia prima, de los factores humanos como el diseño o la ejecución del trabajo.

Algunas clasificaciones que pueden ayudar a organizar la estratificación son:

- Los tipos de materiales y proveedores.
- El momento de la producción: turno, día u hora.
- Las variables de proceso: temperatura, presión, caudal o nivel.
- El departamento: el área, la sección o la línea.
- Los operarios: experiencia, edad, género o turno.
- La maquinaria: tipo, antigüedad, herramienta, molde o cavidad.

Proceso de estratificación

1. **Recolectar datos con un objetivo.** No basta acumular mucha información, que con frecuencia es costosa, hay que investigar con propósitos definidos. Las variables y los puntos de vista deben seleccionarse antes de iniciar el acopio de datos.
2. **Representar gráficamente los datos y los puntos de vista.** Las gráficas y las tablas ayudan a descubrir patrones y relaciones que a veces resultan difíciles de ver cuando se representan sólo con números. Los diagramas más comunes son las gráficas de pastel, los histogramas

- y los diagramas de Pareto, pero no son los únicos, ya que hay otras formas ingeniosas para representar datos y destacar relaciones de causa y efecto.
3. **Determinar las causas posibles de la variación en los datos obtenidos y las desviaciones de lo normal.** Esto a veces lleva a nuevas investigaciones; por ejemplo, en el ejercicio anterior surge la siguiente pregunta: la impuntualidad no es un fenómeno generalizado sino localizado en dos personas, pero ¿por qué los martes todas las personas son impuntuales?
 4. **Profundizar en la estratificación.** Hay que examinar las diferencias entre los valores promedio y las variaciones entre clases o individuos.¹ Detectar nuevas diferencias o relaciones.
 5. **Sacar conclusiones.** Una vez que se establecen y se comprueban relaciones claras de causas y efectos se pueden emitir conclusiones y formular hipótesis para la solución de problemas.

La estratificación es un recurso preliminar que sirve para descubrir variaciones; cuando se profundiza en ella se adquiere un mejor conocimiento del problema y se acumulan datos para su solución.

EJERCICIO RESUELTO

Una empresa fabricante de accesorios para cortinas inició un proceso para mejorar la calidad de sus productos y recopiló datos (primer paso) acerca de las fallas y los departamentos que las originan. Los resultados se presentan en la tabla 4.3 (paso 2).

Falla	Departamento de troquelado	Departamento de fundición	Departamento de pintura	Total
Rotura en las esquinas	7	15		22
Poros		35	8	43
Chuecas	3	2		5
Rayadas	2		7	9
Incompletas	5	6	2	13
Total	17	58	17	92

Tabla 4.3 Clasificación de fallas por tipo y origen de la producción del día.

La tabla nos indica dos hallazgos: que las fallas más frecuentes son los poros y que el departamento de fundición produce el triple de fallas que los demás (paso 3). Dentro de este departamento se produce la mayoría de los poros, pero también la mayoría de las roturas de esquinas (paso 4). Lo primero sugiere un problema de proceso o de materiales de fundición, pero lo segundo podría ser un error de diseño (paso 5).

¹ Kume, Hitoshi, *Herramientas estadísticas básicas para el mejoramiento de la calidad*, Norma, México, 1992.

EJERCICIO 1

1. Se conoce el agrupamiento de observaciones de acuerdo con su naturaleza como:
 - a) Histograma.
 - b) Estratificación.
 - c) Correlación.
 - d) Diagrama de Pareto.

2. Una pequeña fábrica de resistencias eléctricas para calefacción industrial recibió cierto número de quejas de sus clientes debido a que algunas piezas no tienen el poder de calefacción especificado. El departamento de ingeniería hizo algunas investigaciones y encontró que, en efecto, 10% de la producción de los últimos tres meses no cumple con las especificaciones. Esta información sólo confirma el problema pero no ayuda a resolverlo, por lo que es necesario recopilar más datos. El gerente de ingeniería ordena una serie de pruebas y rastreos para determinar el origen (dónde y quién lo produce) y las causas (cómo y por qué). Los resultados de la investigación se encuentran en la tabla 4.4.

Operario	Mesa 1	Mesa 2	Mesa 3	Fallas totales
A	5	8	3	16
B	6	7	4	17
C	4	8	2	14
D	1	6	2	9
E	8	9	7	24
Total	24	38	18	80

Tabla 4.4 Informe sobre orígenes y causas de las fallas en la capacidad de calefacción, en resistencias promedio semanal por persona.

- La producción de la semana fue de 800 resistencias; se observó que los trabajadores no tienen lugar fijo de trabajo sino que el supervisor les asigna semanalmente un espacio a lo largo de una de las tres mesas existentes. ¿A qué conclusiones se puede llegar?
3. El mes pasado una fábrica de jugos de fruta compró su materia prima a diversos proveedores. Cada uno de ellos utiliza diferentes semillas, métodos de cultivo e irrigación. El rendimiento de cada lote comprado varía en función de la calidad de las frutas y de la máquina que se use para procesarlas. Hay cuatro proveedores: *Primero*, *Segundo*, *Tercero* y *Cuarto*; el jugo se procesa en cuatro máquinas diferentes: A, B, C y D, con resultados distintos.

Si la fruta de *Primero* se procesa en la máquina A rinde 55% de jugo por kilo; 70% en la B; 50% en la C y 37% en la D. La fruta de *Segundo* rinde 60% en la A; 52% en la B; 35% en la C y 30% en la D. La fruta de *Tercero* rinde 75% en la A; 55% en la B; 80% en la C y 68% en la D. La fruta de *Cuarto* rinde 30% en la A; 43% en la B; 56% en la C y 85% en la D.

Construye una tabla de estratificación para iniciar el estudio del proceso e indica tus observaciones basadas en la tabla.

CALIDAD Y PRODUCTIVIDAD

4. ¿En qué consiste la estratificación?

5. ¿Cuáles son los pasos del proceso de estratificación?

4.2 Análisis o diagrama de Pareto

Wilfredo Pareto fue un economista italiano (1848-1923) el cual determinó que 85% de la riqueza de la ciudad de Milán estaba en manos de 15% de sus habitantes. Joseph Juran utilizó el mismo concepto, en la década de los cincuenta, para señalar que la mayoría de los problemas de calidad se debían a la minoría de causas, y a él se le debe el nombre de *principio de Pareto*. En la actualidad, se observa el principio de que en muchas actividades la mayoría de los resultados se deben a la minoría de las causas y a varios tipos de problemas: calidad, costos, eficiencia, ecología, seguridad, energía, etc. Se ha generalizado la expresión 80-20 para indicarlo. Esta frase significa, en general, que 80% de un fenómeno se debe a 20% de las causas, es decir, unas pocas causas producen la mayoría de los resultados.

¿A qué se refiere el diagrama de Pareto?

En un proceso pueden existir muchos problemas y cada problema puede tener varias causas. Puesto que los recursos son generalmente escasos, no es posible dedicarlos a la solución de todos ellos, sino que se deben aplicar en orden de importancia: los mejores recursos a los problemas más significativos. Para seleccionar ese conjunto de problemas o dar solución a las causas mayoritarias de un mismo problema se debe hacer un **análisis o Diagrama de Pareto (DP)**.

Los propósitos del DP son:

- Ordenar los datos más importantes de una situación (problemas y resultados).
- Identificar las causas que producen la mayor parte de los datos.

Una vez resuelto el problema principal, se vuelve a aplicar la técnica para localizar el siguiente, resolverlo y continuar así en un ciclo de mejora continua y ordenada. Un DP puede ser el primer paso de un proyecto de mejora y es muy útil para ilustrar las razones por las que se ataca, primero, cierto aspecto y después los demás; de esta manera se logra concentrar la atención de los involucrados sobre los objetivos más importantes.

¿Cuál es la importancia del Diagrama de Pareto?

El DP es un paso complementario de la estratificación. Muchas empresas se detienen en los beneficios de esta última, lo cual puede conducir al uso ineficaz de los recursos, pues es posible que se apliquen a problemas menores cuando podrían ser usados para solucionar problemas mayores. La cuantificación de la información encontrada en la estratificación, con frecuencia, sorprende a los responsables de la empresa: "sabíamos que había un problema, pero no conocíamos su magnitud", suele ser su expresión.

¿Cuál es la relación entre la estratificación y el DP?

Una ventaja adicional del DP es que permite evaluar los resultados de las acciones emprendidas. La simple inspección del diagrama, antes y después de haber actuado, indica la efectividad de los responsables y la trascendencia de los recursos empleados.

Pasos para la construcción de un diagrama de Pareto

1. Para la categoría o eje horizontal del diagrama se seleccionarán las variables de interés. Éstas pueden ser: tipos de defectos, equipos de trabajo, métodos de trabajo, personas, máquinas, clientes, proveedores, en fin, cualquier variable significativa para el proceso. Las categorías menores se pueden agrupar en una sola denominada "otras", pero se debe vigilar que su porcentaje relativo no sea mayor que ninguna categoría específica.
2. El eje vertical izquierdo se marca con unidades pertinentes a la categoría horizontal; por ejemplo, si hablamos de defectos provenientes de una materia prima comprada a diferentes proveedores, éstos ocupan el eje horizontal y las cantidades compradas en pesos o kilogramos se anotan en el eje vertical izquierdo. Se procura que la escala sea un múltiplo de diez, con el fin de que las líneas horizontales sirvan para ambas escalas.
3. El eje vertical derecho se marca con una escala de 0 a 100% en la misma longitud que la escala izquierda.
4. Se coloca una barra del tamaño de su magnitud sobre la categoría correspondiente. La magnitud en unidades se lee hacia la izquierda y su peso porcentual se lee hacia la derecha. Después se colocan las demás barras en orden descendente.
5. Del eje de la primera barra se traza una línea que represente el porcentaje de la segunda categoría hasta el eje de la misma. Éstas son las líneas de acumulados y cada una se inicia donde termina la anterior.

EJERCICIO RESUELTO

En la tabla 4.5 se presenta el resumen de los defectos observados durante la semana pasada en una línea de producción de vajillas.

Operario	Mesa 1	Mesa 2	Mesa 3	Fallas totales
A	5	8	3	16
B	6	7	4	17
C	4	8	2	14
D	1	6	2	9
E	8	9	7	24
Total	24	38	18	80

Tabla 4.5 Estratificación por tipo de defectos.

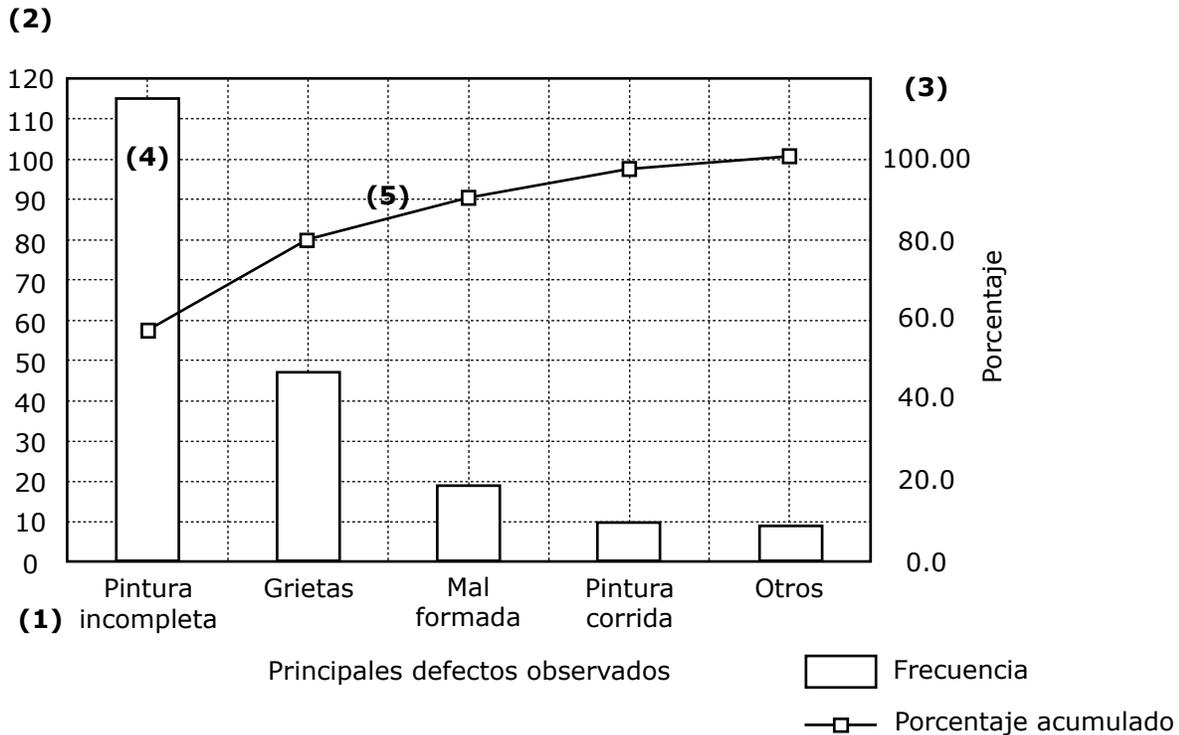


Figura 4.1 Diagrama de Pareto.

El eje horizontal contiene las cinco categorías (1) que agrupan los principales defectos observados.

La escala del eje vertical izquierdo es de cero a 120 defectos (2), un poco más grande que la de mayor frecuencia.

La escala vertical derecha no tiene la misma longitud que la anterior y es de cero a 100% (3), aunque la anotación puede ser mayor de 100% para ajustar las escalas de ambas líneas. La suma de porcentaje acumulado no puede ser mayor de 100.

Se procede a dibujar la gráfica (4) representando como barras verticales la suma de observaciones en cada categoría.

Se trazan las líneas del porcentaje acumulado a partir de la primera categoría (5).

La tabla y la gráfica nos indican que dos defectos: la pintura incompleta y las grietas, suman 81% de todos los defectos. Para estudiar el origen de las fallas, se podría hacer otro análisis llamado de *segundo nivel* porque proviene de los datos originales, pero se concentra en una sola categoría, o bien, analiza los datos pero con respecto a otras variables; por ejemplo, para indicar la cantidad de una falla en el eje vertical y los días de la semana o las máquinas en las categorías del eje horizontal.

EJERCICIO RESUELTO

La influencia de las máquinas en el defecto de pintura incompleta se determina con un DP de segundo nivel. Partiendo de la tabla de datos, se elabora la estratificación de defectos por máquina (tabla 4.6); se escoge una escala de frecuencia de 0 a 50, porque la máxima es 49 y esto permite que la escala de porcentaje coincida con la frecuencia.

Máquina	Cantidad	Porcentaje	Cantidad acumulada	Porcentaje acumulado
<i>B</i>	49	42.6	49	42.6
<i>C</i>	24	20.9	73	63.5
<i>A</i>	22	19.1	95	82.6
<i>D</i>	20	17.4	115	100.0
Total	115	100.0		

Tabla 4.6 Estratificación de pintura incompleta por máquina.

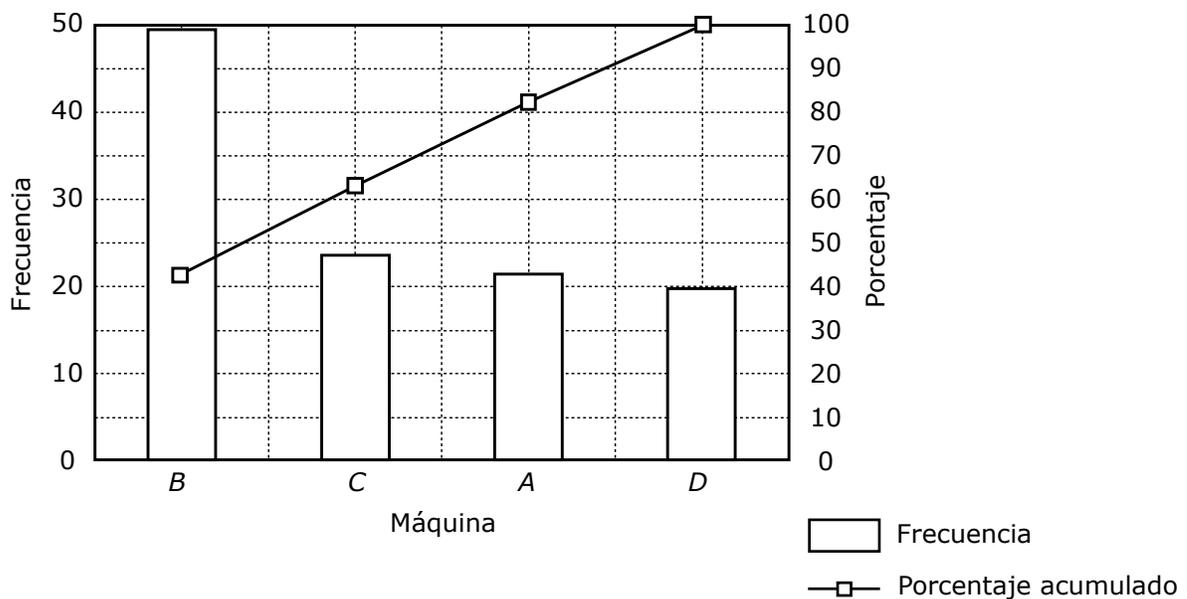


Figura 4.2 Diagrama de Pareto de segundo nivel.

Del primer análisis del DP se concluye que la mayor parte de los defectos son de pintura incompleta y, de éstos, la máquina *B* produce casi 43%.

EJERCICIO RESUELTO

Si se observa la tabla de datos iniciales, se puede ver que el día miércoles hay una concentración de fallas, sin que parezca que el turno sea un factor relevante; por ello se elabora un DP de segundo nivel para mostrar la influencia del día de la semana sobre los defectos de la máquina *B*.

Iniciamos con la tabla de estratificación:

Día	Cantidad	Porcentaje	Cantidad acumulada	Porcentaje acumulado
Miércoles	16	32.7	16	32.7
Martes	9	18.4	25	51.0
Jueves	9	18.4	34	69.4
Viernes	8	16.3	42	85.7
Lunes	7	14.2	49	100.0
Total	49	100.0		

Tabla 4.7 Estratificación de pintura incompleta en la máquina B, por día de la semana.

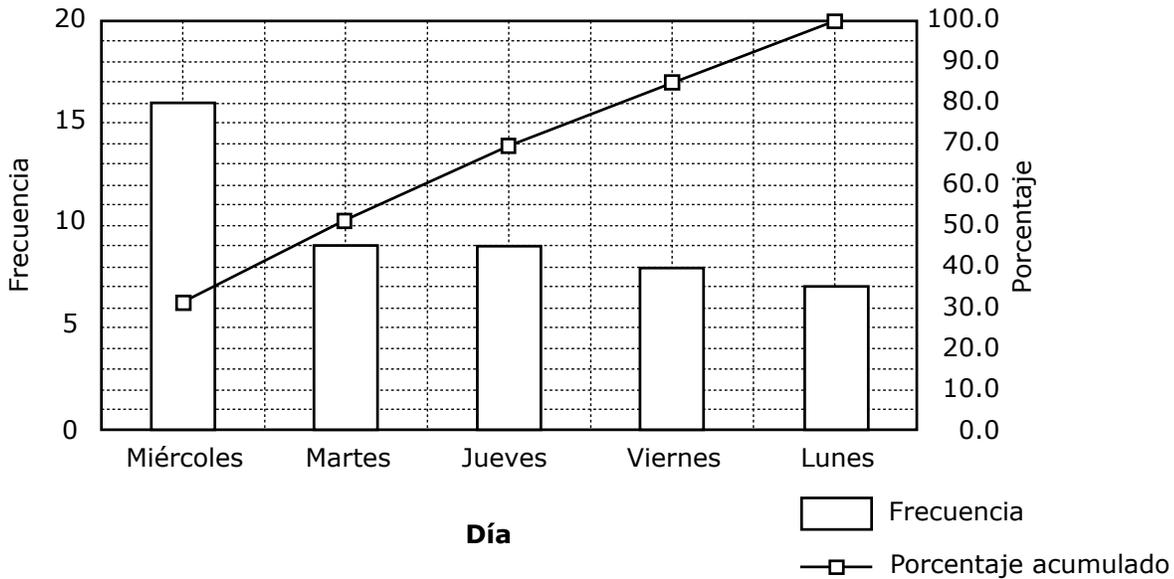


Figura 4.3 Diagrama de Pareto de segundo nivel.

Ahora la escala izquierda puede ser de 0 a 20 y podrá coincidir con las marcas de la escala derecha.

La conclusión es que nuestro problema está en la máquina B y que algo está sucediendo durante ambos turnos del miércoles. El DP nos ha servido para identificar con claridad dónde y cuándo se origina el problema; después se puede pasar a la etapa de búsqueda de soluciones. Un error típico de la aplicación del DP es identificar el problema principal en la primera etapa y no profundizar en las etapas subsecuentes para encontrar sus causas principales, sino saltar de inmediato a dar soluciones aparentemente obvias.

EJERCICIO 2

1. El principio de Pareto se refiere a que:
 - a) La acumulación de riqueza tiende a aumentar por su propia inercia.
 - b) La mayoría de los resultados se deben a la minoría de las causas.
 - c) El empobrecimiento del pueblo se debe siempre a las mismas causas.
 - d) Generalmente un problema se debe a la mayoría de las causas.

2. Una fábrica de camisas desea mejorar la calidad de sus productos y desarrolla un estudio para encontrar cuáles son los defectos más importantes. En la siguiente tabla aparecen los resultados de la investigación. Desarrolla los DP necesarios para identificar el problema más importante y sus causas probables.

Defectos	Turno			Total
	1	2	3	
Hilo	13	12	27	52
Corte	6	9	9	24
Armado	5	10	5	20
Tela	6	5	3	14
Costura	6	4	3	13

3. ¿Cuál es uno de los propósitos de los DP?
 - a) Encontrar la relación entre las posibles causas y efectos de un problema.
 - b) Determinar cuáles son los problemas más importantes de una situación.
 - c) Cuantificar todos los síntomas de un problema y sus relaciones internas.
 - d) Presentar en forma gráfica las causas y soluciones de un problema.

4. ¿Qué es un DP de segundo nivel?
 - a) Es aquel que comprueba los datos de un DP de primer nivel partiendo de datos diferentes.
 - b) Es un análisis posterior al de primer nivel que se hace con datos nuevos para comprobar.
 - c) Es un DP que utiliza los mismos datos, pero se concentra en una sola categoría.
 - d) Es un DP analizado por el supervisor de la persona que elaboró el diagrama original.

5. ¿Cuál es el error típico al utilizar los DP?
 - a) No profundizar en las etapas subsecuentes para encontrar las causas principales de un problema.
 - b) Fallar al elaborar el orden descendente de la tabla de frecuencias.
 - c) Que no coincidan correctamente los porcentajes acumulados con las frecuencias.
 - d) Que no coincida la escala de frecuencia con la de porcentaje, dificultando la lectura de la gráfica.

4.3 Histogramas

En la estratificación y en el DP se organizan las variables de un proceso de mayor a menor; el histograma es otra representación gráfica del comportamiento de un proceso, pero en este caso las diferentes observaciones de una misma variable se grafican alrededor de un valor medio o central. Lo importante es **ver qué tanto se dispersa una misma variable** alejándose de dicho valor central.

¿Para qué se utiliza el histograma?

Para aumentar la calidad de un proceso es necesario reducir al mínimo su variación; el histograma es una **gráfica de barras** que permite determinar cuántas veces se repite un mismo valor, es decir, con qué frecuencia se repite una medición dentro de un rango y qué tan ceñidos o dispersos están los valores de su promedio. Si se comparan histogramas de un mismo proceso, a lo largo del tiempo, se determina si hay algún patrón de comportamiento.

El histograma se inicia contando los resultados de un proceso que se consideran iguales; cada observación se marca en una forma para registro y luego las marcas se resumen en tablas de números; finalmente se agrupan en forma gráfica. De esta manera resulta más fácil su interpretación que leyendo listados de números.

Construcción de un histograma

1. Determinar el **rango** de las observaciones. ¿Entre qué valores extremos pueden aparecer las observaciones que se hacen? Por ejemplo, si se quisiera determinar las tallas de una prenda juvenil, se podría establecer el rango entre la talla 5 y la 11, es decir, todas o casi todas las tallas usadas por el segmento de mercado denominado "juvenil" caen entre estos dos valores. Una máquina que corta tramos de tubo para frenos que deben medir 30 cm, puede cortarlos entre 29.5 y 30.5 cm por su variación de proceso. El **rango** es el **valor mínimo y máximo** que adopta una variable.
2. Determinar el número de **clases**². Las clases son el número de intervalos en los que se divide el rango, es decir, cuántas columnas tendrá el histograma. En el ejemplo de las tallas, el rango es de 5 a 11 y está dividido en 13 clases: 5, 5½, 6, 6½, y así sucesivamente hasta 11. Las longitudes del tubo de frenos se pueden dividir (colocar en clases) separadas 0.2 cm una de otra: la uno contiene las piezas que miden entre 29.5 y 29.7; la dos contiene las de 29.71 a 29.9; la tres las de 29.91 a 30.1; la cuatro las de 30.11 a 30.3 y la cinco las de 30.31 a 30.5.

En la práctica, el número mínimo recomendable de clases es cinco y el máximo 15, aunque se procurará que sea una cantidad impar para que el valor central caiga en la barra de en medio de la gráfica y las demás columnas queden simétricas. Una cantidad inicial podría ser la raíz cuadrada del número de datos; por ejemplo, si hay 50 datos, se pueden usar siete clases.

² Un método más formal para determinar el número de clases necesarias para clasificar un conjunto de datos está dado por la regla de Sturges, la cual se expresa de la siguiente manera: $C = 3.3 (\log n) + 1$, donde C = número de clases y n = tamaño del conjunto de datos que se examinan.

Los límites de cada clase se llaman **intervalos** y se fijarán de modo que no haya confusión al anotar las observaciones; por ejemplo, si se fijan dos intervalos como: “entre 12 y 13, y entre 13 y 14”, cuando aparezca una lectura = 13, no se sabrá si ponerla en la primera clase o en la segunda. Lo correcto sería: 12.01 y 13; 13.01 y 14.

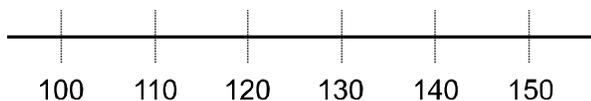
3. Contar las observaciones en cada clase. Se puede usar una tabla como la siguiente:

Clase	Intervalos	Marcas de cuenta (una marca por cada observación dentro del intervalo)	Frecuencia (número de observaciones)	Frecuencia relativa (porcentaje del total)
1	100 a 109	XXX	3	15
2	110 a 119	XXXX	4	20
3	120 a 129	XXXXXX	6	30
4	130 a 139	XXXXX	5	25
5	140 a 149	XX	2	10

Tabla 4.8 Marcador de observaciones.

Se pueden ampliar los intervalos de las clases en los extremos del rango porque generalmente contienen pocos elementos; así, la clase 5 del ejemplo anterior termina en 149, pero podría ampliarse hasta incluir “150” o bien podría ser 140 y mayor. En el otro extremo, la clase 1 de 100 a 109, podría ser “menor de 109”.

4. Dibujar el histograma. El eje horizontal se divide en el número de intervalos; en el ejemplo, hay cinco intervalos y los límites se acotan con el valor inferior de cada uno.



5. Ahora hay que trazar el eje vertical. Su altura está determinada por la cuenta máxima de observaciones en un intervalo; en el ejemplo, la clase 3 contiene seis observaciones, entonces el eje vertical puede tener seis o siete divisiones:

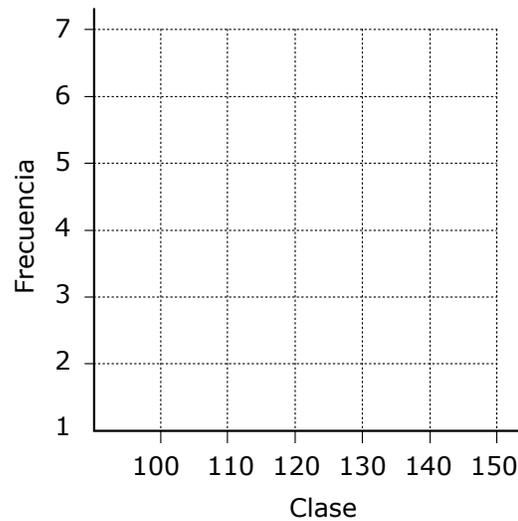


Figura 4.4 Quinto paso para la elaboración del histograma: trazar la cuadrícula.

6. A continuación, se dibuja una columna en cada clase; su altura es la cantidad de observaciones dentro del intervalo (figura 4.5).

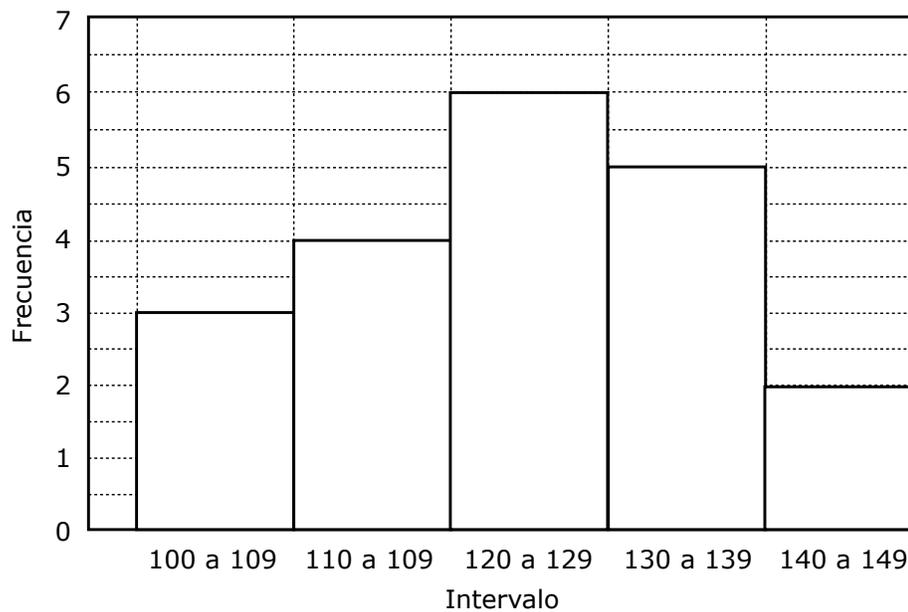


Figura 4.5 Histograma.

Si el histograma está bien construido y contiene un número suficiente de datos es útil para observar el comportamiento de la **población**, también sirve para explicar el comportamiento de un fenómeno o de las variables de un proceso. Con un histograma se pueden observar variables como la dimensión media de una pieza y la dispersión de la misma medida en todas las piezas; estos datos son útiles para mejorar la calidad de los procesos.

Recordemos que cuando hablamos de *población* nos referimos a todos los elementos de un conjunto homogéneo; por ejemplo, toda la producción de clavos de una máquina durante un turno o la cantidad de lluvia diaria durante un mes en determinada ciudad, etcétera.

EJERCICIO RESUELTO

Una empresa empaedora de arroz y frijol compra esos productos por tonelada y los empa en bolsas de un kilo; tiene tres máquinas dosificadoras iguales, pero compradas con tres años de diferencia cada una. Es importante que cada bolsa contenga exactamente el peso indicado, porque si contiene menos es seguro que los supermercados le rechazarán lotes completos de entrega. Por otro lado, si las bolsas contienen más peso del indicado la empresa pierde, por entregar más producto. Con objeto de conocer cuál es la situación, se toma una muestra de 50 bolsas de la máquina más nueva y se verifica el peso de cada una con una báscula de precisión. Los datos para construir un histograma se muestran en la tabla 4.9.

Envasadora Montserrat

Pesos observados en una muestra de 50 bolsas de arroz (gramos por bolsa)

999	1000	1001	1001	1001	<table border="1"> <thead> <tr> <th>Clase</th> <th>Frecuencia</th> </tr> </thead> <tbody> <tr> <td>Hasta 998</td> <td>1</td> </tr> <tr> <td>999</td> <td>5</td> </tr> <tr> <td>1000</td> <td>10</td> </tr> <tr> <td>1001</td> <td>17</td> </tr> <tr> <td>1002</td> <td>10</td> </tr> <tr> <td>1003</td> <td>4</td> </tr> <tr> <td>1004</td> <td>2</td> </tr> <tr> <td>1005</td> <td>1</td> </tr> </tbody> </table>	Clase	Frecuencia	Hasta 998	1	999	5	1000	10	1001	17	1002	10	1003	4	1004	2	1005	1
Clase	Frecuencia																						
Hasta 998	1																						
999	5																						
1000	10																						
1001	17																						
1002	10																						
1003	4																						
1004	2																						
1005	1																						
1001	1003	1000	1002	1000																			
1000	999	1003	1002	1001																			
1002	1002	1001	1001	1004																			
1001	1000	1002	1003	1002																			
999	1004	1000	1000	1001																			
1002	1001	1002	1001	1002																			
1000	999	1001	1001	1003																			
1001	998	1002	1000	1005																			
999	1000	1001	1001	1001																			

Tabla 4.9

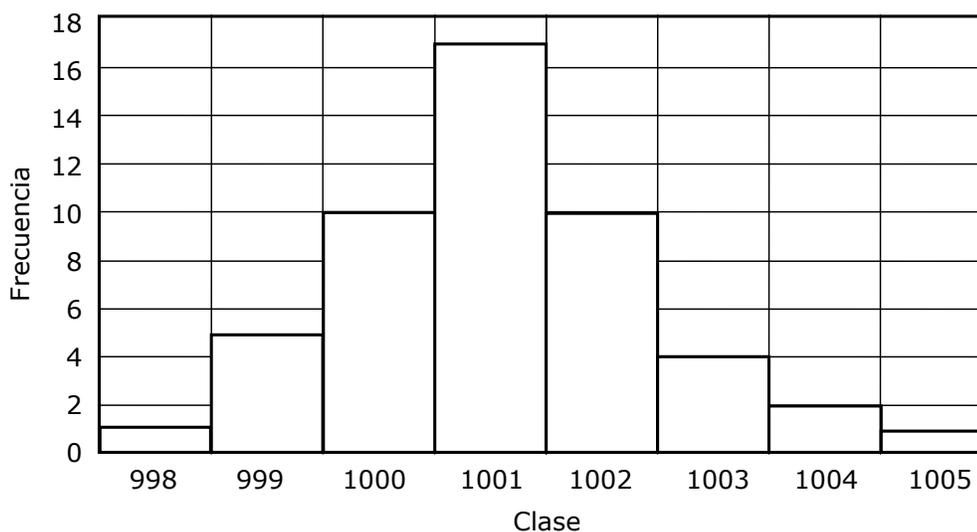


Figura 4.6 Histograma de la máquina empaedora nueva.

Se observa que la mayoría de las bolsas pesan 1001 g y los demás pesos están uniformemente repartidos alrededor de esa medida. Veamos ahora el resto de la planta:

Envasadora Montserrat

Pesos observados en una muestra de 96 bolsas de arroz (gramos por bolsa).

999	1003	1003	995	1000	998	1001	1002	Clase	Frecuencia
1000	999	1000	997	1000	1003	1004	1003	995	4
998	996	1002	997	998	1003	996	1000	996	14
995	997	996	1000	1004	996	998	1000	997	6
1001	1000	998	1005	996	996	1002	999	998	9
996	1001	1002	1002	999	997	1004	999	999	8
1004	996	996	998	1003	1001	997	1001	1000	14
996	998	999	1004	1003	1004	1001	1000	1001	9
1000	1005	1003	997	995	998	1002	1002	1002	7
1004	1005	995	996	999	999	1001	1000	1003	10
1004	1003	996	996	1000	1001	1004	996	1004	11
1004	1005	1000	1004	1001	1003	1000	998	1005	4

Tabla 4.10 Datos para la construcción del histograma de empaclado en todas las máquinas.

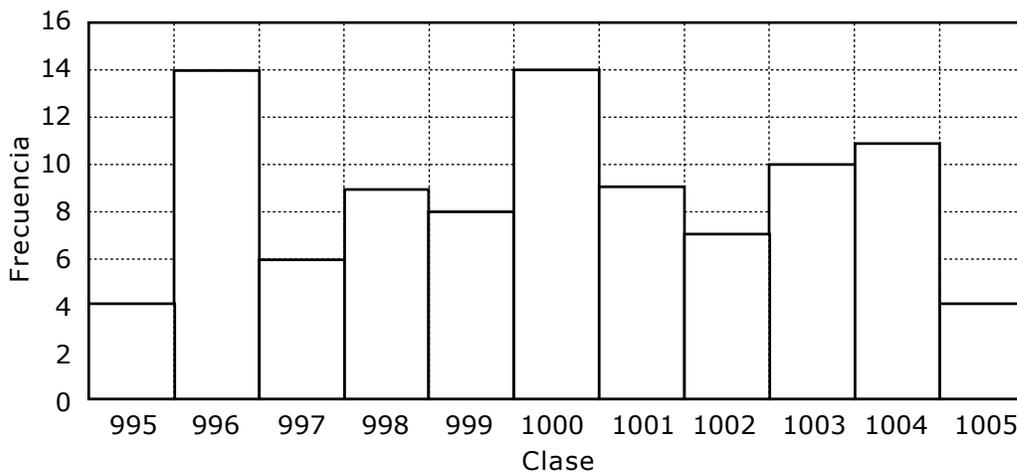


Figura 4.7 Histograma multimodal.

En este ejemplo se nota que hay tres “picos” de frecuencia; parece haber más bolsas que contienen 996, 1 000 y 1 004 g que los demás pesos, ¿qué significaría esto?

Como se indicó en el planteamiento del ejercicio, hay tres máquinas y es probable que cada una tenga, centrado en las cantidades que aparecen como “picos”, un promedio de exactitud diferente. El histograma sirvió para descubrir que las máquinas se están comportando en forma distinta, aunque el promedio sea 1 000 g por bolsa. A este tipo de histogramas o, para ser más precisos, a este

tipo de fenómenos se les denomina **variación multimodal** y sirven para indicar que las muestras provienen de dos o más diferentes orígenes.

Los histogramas proporcionan diversos tipos de información:

- a) Los picos indican la cantidad de elementos que hay dentro de la medida especificada (la clase o casilla). Si la muestra proviene del mismo origen, el pico indica la **tendencia central** de un grupo de mediciones, es decir, anuncia el promedio o valor central. Los datos más frecuentes coinciden con el promedio, los demás datos aparecen como columnas simétricas a ambos lados y caen en declive hacia los extremos. En el ejercicio anterior tenemos tres máquinas con tres valores centrales y por eso los picos son diferentes que el promedio general.

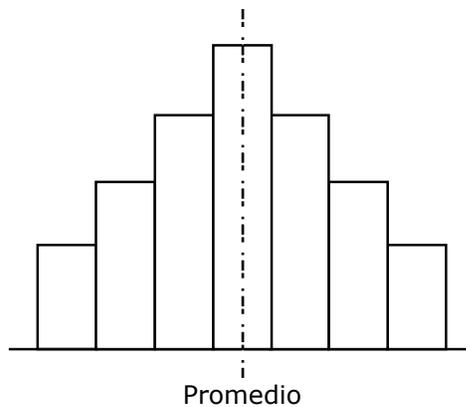


Figura 4.8 Histograma normal.

- b) Indican si el comportamiento de las mediciones es **simétrico**; si hay la misma cantidad de elementos a ambos lados de la media o si hay asimetrías que puedan indicar otro tipo de problemas. Si la gráfica no es simétrica, se dice que el proceso tiene **sesgo**. La figura presenta un sesgo hacia la izquierda, es decir, las observaciones más frecuentes no coinciden con el promedio, sino que caen a su izquierda.

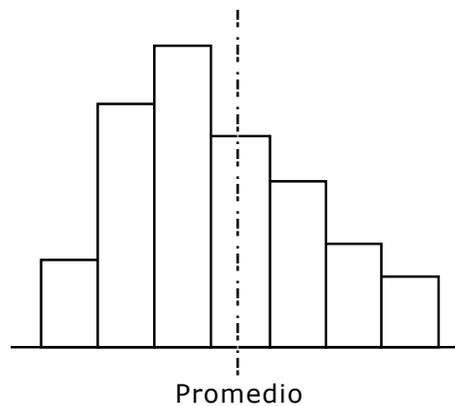


Figura 4.9 Histograma sesgado.

- c) Ilustran si el proceso está **centrado**. En el ejercicio anterior, la gráfica nos indica que sólo 14 bolsas de las 96, pesaron 1 000 g, lo cual indica que el proceso no está centrado porque la mayoría de las mediciones no coincidieron con el promedio. Un ejemplo de esto podría ilustrarse mediante la edad de dos personas: si una tiene 20 y la otra tiene diez, en promedio tienen 15, pero ninguna de ellas tiene esa edad, la muestra no está centrada. Lo estaría si tomamos una muestra de alumnos de tercer año de secundaria; en ella, la mayoría tendría 15 años, en coincidencia con el promedio; unos cuantos tendrían 14 o menos y otros pocos tendrían 16 o más; por tanto, la muestra está centrada.
- d) Varios picos indican **variaciones multimodales**, lo cual significa que las mediciones pueden variar por diversas causas o provienen de varios procesos independientes. La variación también se puede deber a diferencias en los operadores de una misma máquina.

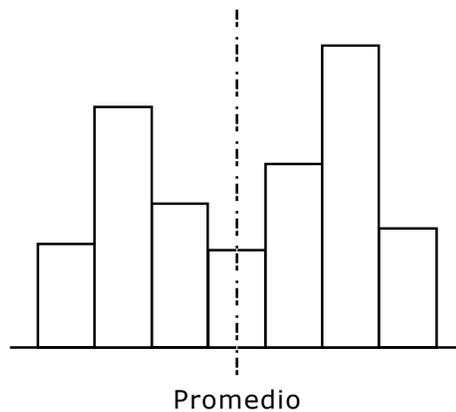


Figura 4.10 Histograma bimodal.

- e) Los **acantilados** son caídas bruscas de las mediciones. En la práctica, los acantilados se presentan cuando un proceso no puede rebasar cierto límite; por ejemplo, si se hace un histograma de calificaciones escolares, no se encontrarán observaciones mayores de diez ni menores de cinco porque el proceso está limitado dentro de esos valores.

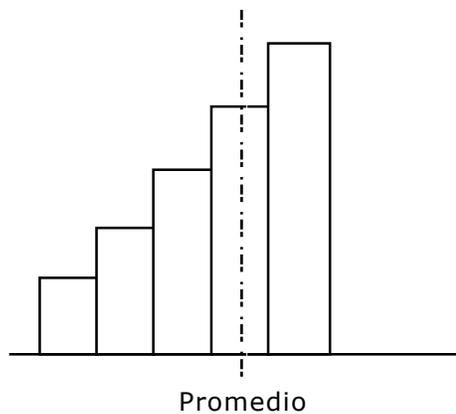


Figura 4.11 Histograma acantilado.

- f) La **variabilidad** de los datos. En los casos anteriores hemos visto que hay una cantidad variable de observaciones dentro de cada clase, la mayor frecuencia coincide con la media, pero hay ocasiones en las que todas las clases contienen casi el mismo número de observaciones. Se dice que esa muestra contiene mucha variabilidad porque no hay una frecuencia que destaque alrededor del valor central.

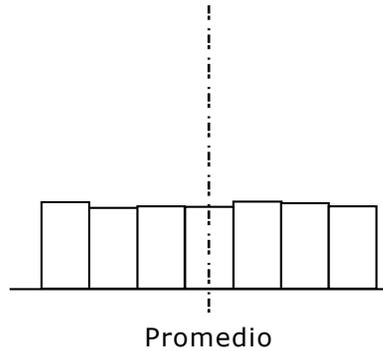


Figura 4.12 Histograma con mucha variabilidad.

- g) También es posible encontrar procesos con poca variabilidad, como el caso de una muestra en la cual casi todas las observaciones coinciden con el promedio y unas cuantas aparecen arriba o debajo de él.

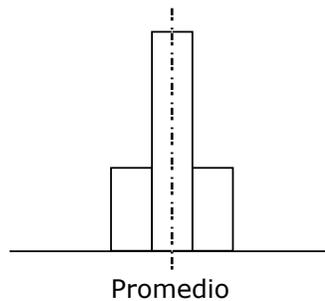


Figura 4.13 Histograma con poca variabilidad.

Ésta es una condición deseable en cualquier proceso; la poca variabilidad indica que está bajo control, que la muestra y la población de donde salió son homogéneas.

El histograma y la curva de distribución normal

Si la variable de una población puede cambiar al azar conforme aumenta el número de elementos que se miden para formar un histograma, la frecuencia, alrededor del valor central, también aumenta; es decir, cada vez aparecen más elementos cuya medida coincide con la media y se agudiza el perfil de las barras del histograma.

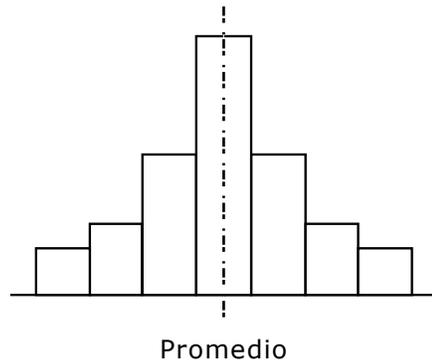


Figura 4.14 Histograma con un gran número de elementos.

Por otro lado, si los límites entre una clase y otra se hacen más cerrados, el perfil del histograma va adquiriendo la forma característica de la curva de distribución normal (figura 4.15).

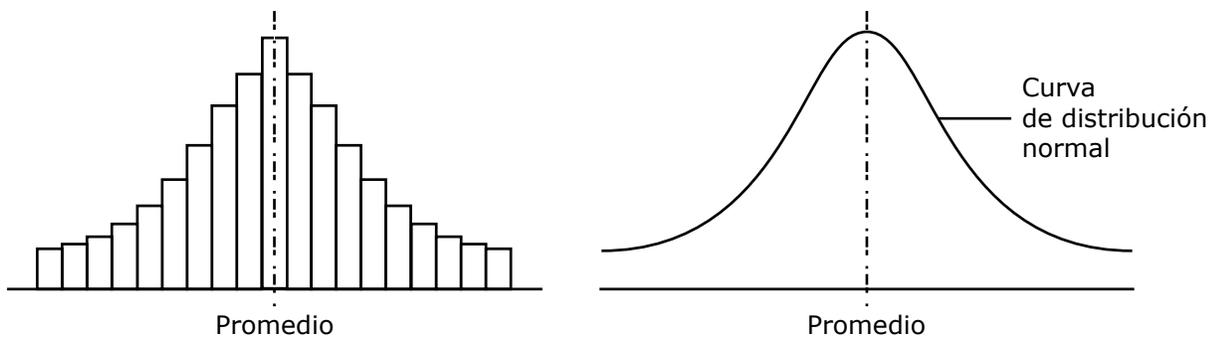


Figura 4.15 El histograma con clases muy angostas se aproxima poco a poco hasta llegar a la curva de distribución normal.

La curva de distribución normal indica el comportamiento de cualquier medida que pueda variar exclusivamente por efecto del azar; por ejemplo, en una máquina siempre habrá variaciones en la medida de las piezas que produce, por mucho cuidado que se tenga al ajustarla. Esas variaciones se deben y se distribuyen al azar alrededor del valor medio de la medida que se toma.

EJERCICIO 3

1. En un histograma, ¿qué son las clases?
 - a) Las categorías en las que se divide el problema.
 - b) El número de intervalos en los que se divide el rango.
 - c) La cantidad de divisiones en las que se reparte la variable.
 - d) La distancia entre cada subdivisión del rango de la variable.

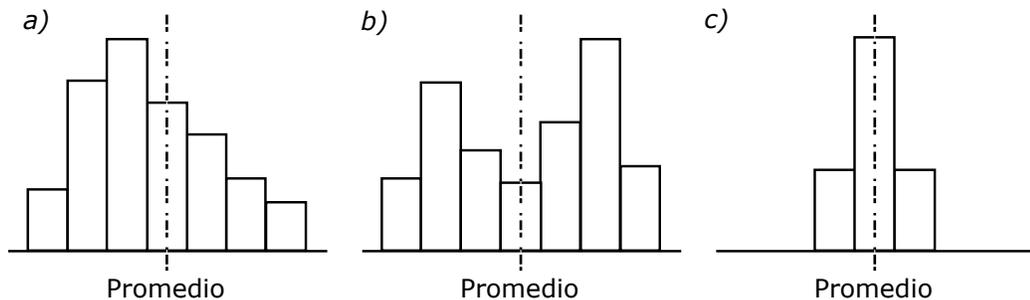
2. Una fábrica de paneles de madera produce tableros de 10 mm de espesor trabajando una máquina en tres turnos. El departamento de ingeniería toma unas muestras y mide su espesor; los resultados se anotan a continuación. Elabora un histograma con esos datos.

Novopanel

Fábrica de aglomerados de madera

9.8 10.0 10.0 10.1 10.2 9.9 10.1 9.8
 10.3 10.2 9.9 10.5 10.3 10.2
 9.9 9.6 9.7 10.4 10.1 10.1 10.0 10.0
 10.4 10.0 10.2 9.9
 9.7 9.9 10.3 10.1
 9.8 10.0 10.2 9.7 10.1 9.6 10.0 9.8

3. Observa los siguientes histogramas y señala qué indican.



4. ¿Cuáles son las características de un histograma con mucha variabilidad?
 - a) Hay muchos elementos en los extremos.
 - b) Tiene pocos elementos cerca del valor medio.
 - c) Las mediciones cambian de una muestra a otra.
 - d) Hay casi el mismo número de elementos en todas las clases.

5. ¿Qué es el rango?
- a) Los límites dentro de los cuales aparece una variable.
 - b) El orden en el cual se acomodan las mediciones.
 - c) El número de casillas para clasificar los datos.
 - d) La cantidad de elementos dentro de una casilla.

4.4 Diagramas de dispersión

Hemos visto que para solucionar un problema es necesario detectar y modificar sus causas; sin embargo, esta manipulación de causas para afectar un resultado no es una serie de ensayos de prueba y error “a ver cuál sirve”, sino un procedimiento que determina, primero, cuáles causas están relacionadas con un determinado efecto y luego cuál tiene mayor influencia sobre un resultado específico. Para encontrar ese grado de influencia se utilizan los **diagramas de dispersión**.

¿En qué consiste un diagrama de dispersión?

En un diagrama de dispersión se grafican pares de valores relacionados entre sí. Las variables que se manipulan se llaman **variables independientes** y los resultados obtenidos se denominan **variables dependientes**, porque dependen de las primeras. Por ejemplo, la resistencia mecánica (variable dependiente) de una aleación depende del contenido de un elemento (variable independiente); la capacidad de carga de una grúa depende de la potencia de su motor; el peso de un pollo de engorda depende de la cantidad de carbohidratos ingeridos.

Se pueden elaborar diagramas de dispersión que indiquen una relación aparente entre dos variables, pero esto no significa que una dependa de la otra, sino que ambas dependen, simultáneamente, de una tercera. Por ejemplo, si se grafica la estatura de un grupo de jóvenes contra su peso, se encontrará que a mayor estatura, mayor peso, pero ¿depende la estatura del peso? No, lo que pasa es que ambos dependen de la edad de la persona. La variable independiente es la edad y el peso o la estatura son las variables que dependen de aquella.

¿Por qué es importante encontrar la relación de dependencia entre dos variables y no sólo la correlación?

Es importante establecer la relación de **dependencia** entre dos variables y no sólo la **relación** entre sí, porque si se busca un determinado efecto, se debe manipular la variable independiente y no las otras variables que no modifican el efecto deseado. De esta manera, un criador de pollos podrá cambiar el contenido de los alimentos que da a sus aves y al mismo tiempo éstos reaccionarán con un diferente peso y estatura.

Procedimiento para construir un diagrama de dispersión

1. Seleccionar las variables que se pondrán en la gráfica y obtener datos en pares, es decir, para cada valor de una variable independiente debe corresponder otro de la variable dependiente.
2. Trazar los ejes de la gráfica. La variable independiente va sobre el eje horizontal (o eje de las X); la variable dependiente va sobre el eje vertical (o eje de las Y). Anotar el nombre de las variables junto al eje que las representa.
3. Marcar la escala y las unidades. La escala comprende todos los valores que se utilizaron, desde el mínimo hasta el máximo de ambas variables. No forzosamente las gráficas se inician desde cero. Algunas variables pueden tomar valores negativos; en este caso, se necesitan ejes con valores negativos a la izquierda del cero en el eje de las X y debajo de cero en el eje de las Y. El tamaño de las unidades de una escala puede ser diferente al tamaño de las unidades en la otra.

Si una variable sólo nos interesa en un rango que no incluye el cero, por ejemplo, el espesor de una lámina de asbesto-cemento que varía entre 5.8 y 6.3 mm, entonces la gráfica sólo contendrá estos valores o un rango ligeramente mayor, digamos, desde 5.5 hasta 6.5 mm, para que el eje tenga diez divisiones.

4. Colocar los pares de datos. Un punto en la gráfica está definido por sus coordenadas: un valor de X y otro de Y . El cruce de las coordenadas localiza el punto. Si varios pares de datos se repiten, se trazan círculos concéntricos sobre el primer punto.
5. Identificar la gráfica, las escalas, los ejes y, en caso necesario, los puntos anotados.

EJERCICIO RESUELTO

Un agricultor cultiva tomate en un terreno con suficiente irrigación. Él desea incrementar el rendimiento por hectárea y para ello, en diferentes parcelas, aplica un fertilizante que previamente fue seleccionado por el método de estratificación, pero que ahora se dosificará en diferentes cantidades con el fin de escoger la mejor opción, es decir, la que mayor peso de tomates produce. La gráfica de los pares de datos se muestra en la tabla 4.11.

Rancho productor de tomates

Kilogramo de fertilizante por hectárea	Toneladas de tomate producido
9.5	4.32
9.8	4.37
9.9	4.37
10.0	4.40
10.1	4.41
10.2	4.41
10.3	4.42
10.4	4.44
10.5	4.45

Tabla 4.11 Producción de tomates contra fertilizante utilizado.

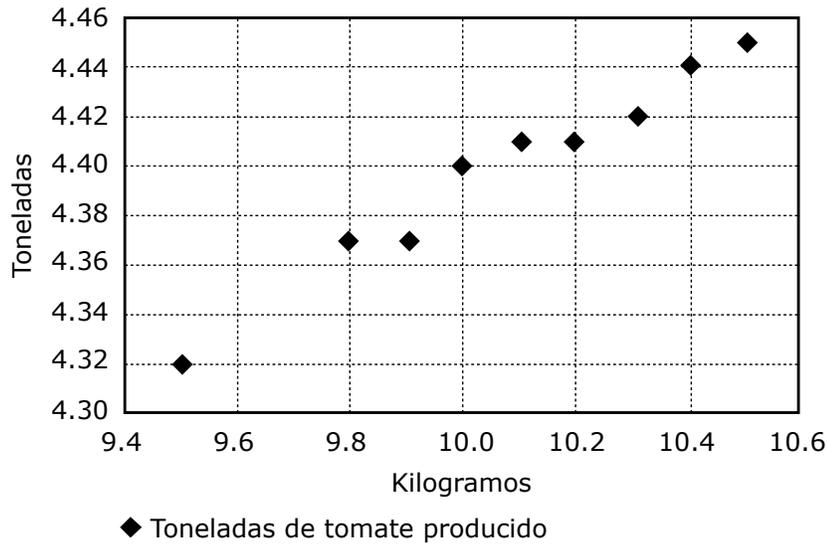


Figura 4.16 Diagrama de dispersión entre fertilizante usado y producción de tomate por hectárea.

Se puede concluir que existe una relación directa entre la cantidad de fertilizante aplicada a la tierra y el rendimiento de las plantas de tomate.

Interpretación de los diagramas de dispersión

La relación entre dos variables puede ser:

- **Positiva**, si para un aumento en la variable independiente se obtiene un *aumento* en la variable dependiente.
- **Negativa**, si para un aumento en la variable independiente se obtiene una *disminución* en la variable dependiente.
- **No lineal**, si la relación no se asemeja a una línea recta.
- **Sin relación**, si los pares de puntos se dispersan por toda la gráfica y para un mismo valor de variable independiente, aparecen varios valores de la dependiente.

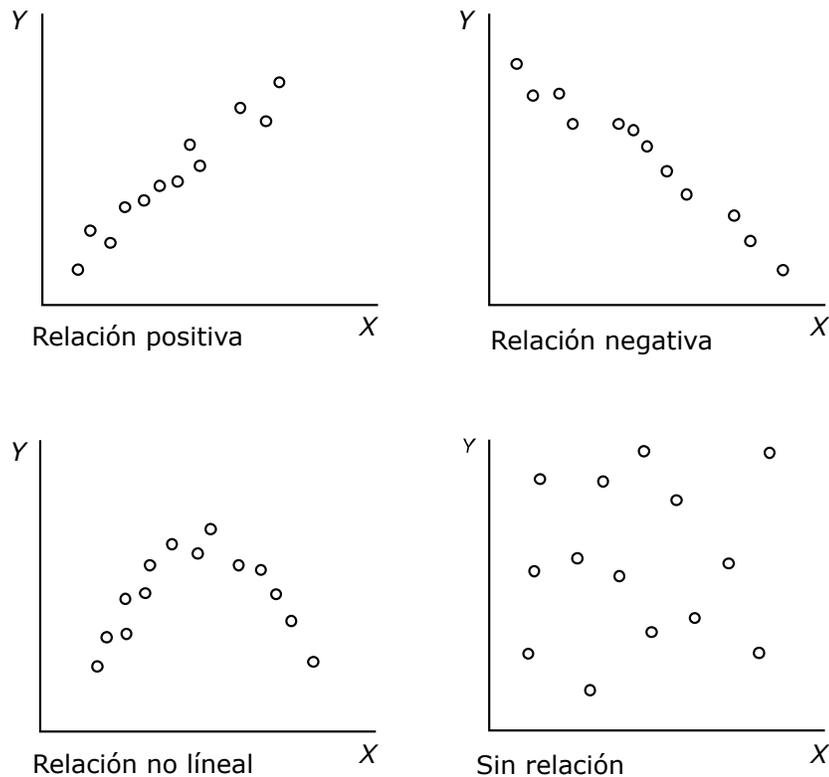


Figura 4.17 Tipos más comunes de correlación entre variables.

Para interpretar correctamente un diagrama de dispersión es necesario observar la tendencia general de los puntos y no un punto aislado que pudiera provenir de un error de captura; otro aspecto que se observará es la posición de las variables sobre la gráfica: la variable independiente se colocará sobre el eje horizontal y la dependiente sobre el vertical. La lectura de una gráfica es: “el efecto Y , depende de la causa X ” o bien, “ X influye en Y ”.

¿Cómo se interpreta un diagrama de dispersión?

Errores típicos de los diagramas de dispersión

El error más frecuente en el uso o interpretación de los diagramas de dispersión es considerar que hay relación entre dos variables cuando, en realidad, no la hay o ambas dependen de una tercera. Por ejemplo, no sería correcto establecer una relación directa entre el gasto total del rancho productor de tomates y la producción por hectárea, porque gastar más no significa que se gaste en fertilizante, se podría gastar en otra cosa. El aumento en la producción y en gastos podría ser una simple coincidencia de eventos independientes.

Cuando se trata de problemas de calidad, con presión para solucionarlos, es frecuente suponer que si dos fenómenos coinciden, el primero depende del segundo, cuando en realidad el segundo depende del primero. Por ejemplo, en una planta donde se ha observado que el número de piezas defectuosas aumenta conforme aumenta el número de horas extra: ¿qué tienen que ver las horas extra con la cantidad de defectos? Se podría decir que tantas horas de trabajo

continuo causan descuidos y errores (como es el caso de los accidentes de trabajo), cuando lo que pasa es lo contrario, es decir, cuando hay más errores, se requieren más horas extra para corregirlos.

Para no cometer esta equivocación, el efecto de coincidencia debe rastrearse hacia las causas de ambos. En el ejemplo anterior, si la cantidad de pedidos de nuevos clientes aumenta, es posible que la fábrica decida trabajar tiempo extra en proporción directa con las nuevas cantidades de pedidos; en este caso, ambos fenómenos, los defectos y el tiempo extra, dependen de un tercero: los nuevos pedidos.

Coeficiente de correlación

Para decir que dos variables están relacionadas entre sí, es necesario calcular el **coeficiente de correlación** r entre los elementos que componen todos los pares. Si r se calcula entre 0.9 y 1.0, se dice que hay una fuerte correlación; entre 0.8 y 0.89 es buena; entre 0.7 y 0.79 es regular, y por debajo de 0.7 es mala. Se calcula así:

$$r = \frac{S_{xy}}{\sqrt{(S_{xx})(S_{yy})}}$$

La S no es varianza, sólo es una letra para abreviar las operaciones que siguen. De donde:

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n}$$

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}$$

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n}$$

Esta presentación simbólica fue desarrollada para facilitar el cálculo manual. En la actualidad se dispone de calculadoras de mano y paquetes de cómputo que calculan el coeficiente de correlación directamente de las series de datos; en el siguiente ejemplo se ejecutan los cálculos en forma manual.

EJEMPLO RESUELTO

Pinturas Tlapalli

Efecto del aditivo QP contra el tiempo de secado de la pintura

Cantidad de aditivo	Tiempo de secado
0	14.0
1	11.0
2	10.0
3	8.0
4	7.5
5	7.0
6	6.5
7	6.0
8	5.0
9	5.5
10	4.0

Tabla 4.12 Tiempo de secado de una pintura para diferentes cantidades de aditivo.

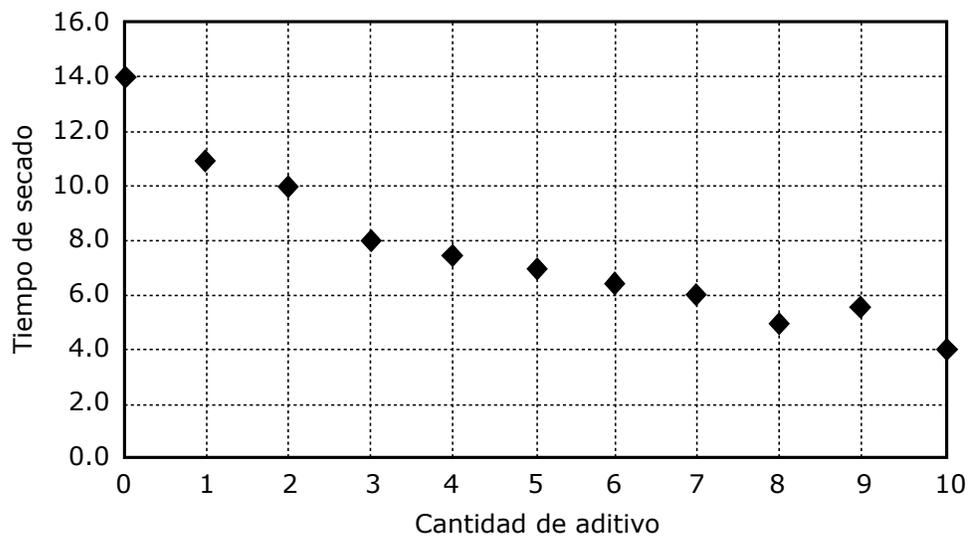


Figura 4.18 Tiempo de secado contra cantidad de aditivo.

- Mediante un diagrama de dispersión se observa la relación entre el tiempo de secado y la cantidad de aditivo utilizado.
- Se calcula el coeficiente de correlación entre las variables y se interpreta.

<i>n</i>	<i>x</i> cantidad de aditivo	<i>y</i> tiempo de secado	$x_i y_i$	x_i^2	y_i^2
1	0	14.0	0.0	0	196.00
2	1	11.0	11.0	1	121.00
3	2	10.0	20.0	4	100.00
4	3	8.0	24.0	9	64.00
5	4	7.5	30.0	16	56.25
6	5	7.0	35.0	25	49.00
7	6	6.5	39.0	36	42.25
8	7	6.0	42.0	49	36.00
9	8	5.0	40.0	64	25.00
10	9	5.5	49.5	81	30.25
11	10	4.0	40.0	100	16.00
Total	55	84.5	330.5	385	735.75
Promedio	5	7.7			

Tabla 4.13 Cálculo del coeficiente de correlación.

$$S_{xy} = 330.5 - \frac{(55)(84.5)}{11} = -92$$

$$S_{xx} = 385 - \frac{(55)^2}{11} = 110$$

$$S_{yy} = 735.75 - \frac{(84.5)^2}{11} = 86.6$$

$$r = \frac{-92}{\sqrt{(110)(86.6)}} = -0.94$$

El signo negativo indica una correlación negativa, es decir, la variable dependiente disminuye conforme aumenta la variable independiente. En este caso, el tiempo de secado disminuye al aumentar el aditivo. Puesto que r se acerca a 1, la correlación es fuerte.

EJERCICIO 4

1. El rendimiento de un automóvil con motor de gasolina en km/l depende de diversos factores: la cilindrada del motor, el peso de la unidad y el octanaje del combustible. Utilizando los siguientes datos elabora la gráfica de dispersión de cada variable contra rendimiento, calcula el coeficiente de correlación e interpreta el resultado para cada par de variables.

Laboratorios automotrices de referencia

Reporte de rendimiento de gasolina

a) contra cilindrada		b) contra peso de la unidad		c) contra octanaje	
cm ³	km/l	peso (kg)	km/l	octanos	km/l
800	16.0	1100	16.0	80	13.5
850	16.5	1320	15.8	85	13.9
900	15.8	1450	15.0	90	14.2
1100	17.0	1600	14.0	93	14.5
1200	14.0	1670	13.9	95	14.8
1500	14.5	1700	13.0		
1700	15.0	1800	12.5		

2. Si el cálculo del coeficiente de correlación entre dos variables resulta 0.85, se puede decir que la dependencia entre las variables es:
- Fuerte.
 - Buena.
 - Regular.
 - Mala.
3. ¿Qué se grafica en un diagrama de dispersión?
- Cualquier tipo de variables.
 - Variables que existan al mismo tiempo.
 - Pares de valores relacionados entre sí.
 - Medidas independientes de procesos.
4. ¿Qué relación debe existir entre las variables graficadas en un diagrama de dispersión?
- La variable dependiente cambia proporcionalmente con los cambios en la variable independiente.
 - Ambas variables cambian al mismo tiempo, pero una no depende de la otra, son independientes.
 - Ambas variables dependen mutuamente entre sí; al cambiar cualquiera de ellas, cambia la otra.
 - La variable independiente afecta, pero no es directamente proporcional a la variable dependiente.

4.5 Toma estadística de decisiones

¿Cómo se determinan las dimensiones de las tallas de ropa o zapatos? Son las medidas promedio tomadas de muchas personas de diversas edades, pero ¿acaso alguien midió las proporciones de todos los habitantes de un país para sacar las tallas?: no, solamente se midieron algunas personas seleccionadas cuidadosamente por ser las más representativas de toda la población.

La **toma estadística de decisiones** se basa en tomar medidas de una variable o un atributo sobre una muestra que pertenece a un grupo más grande. La información que se saca de una muestra es útil para tomar decisiones acerca del total de la población y para aceptar o rechazar hipótesis formuladas sobre ella; por ejemplo, el tamaño más frecuente de calzado en niños de 12 años es 25 cm.

Conceptos básicos

La **población** es el total de individuos que conforman un grupo homogéneo de objetos que se estudian; por ejemplo, un lote completo de producción de una fábrica de ruedas para bicicletas, todos los jóvenes de 14 años que estudian secundaria en una determinada ciudad; las mujeres entre 40 y 50 años de edad de un sector social específico; los pistones de motor producidos ayer por la máquina número 2 de una planta o el consumo por casa de energía eléctrica en una colonia.

Una parte representativa de una población se llama **muestra**; por ejemplo, algunas piezas extraídas de un lote de producción, algunos jóvenes de 14 años o ciertas mujeres entre 40 y 50 años.

La **variable** es una cualidad observable de un objeto, la cual puede ser medida con una escala continua; por ejemplo, dimensiones, peso, color, temperatura, resistencia eléctrica, resistencia mecánica, dureza superficial, contenido de azufre, etc. La variable toma diferentes valores según el individuo que se observa; es decir, cada pieza de un lote o cada persona de un grupo pueden tener diferente valor que la pieza o la persona de junto.

Un **atributo** es una cualidad observable que sólo puede adoptar dos estados discretos (que no tiene valores intermedios): sí o no, aceptado o rechazado, en favor o en contra, presente o ausente, dentro o fuera de ciertos límites, etcétera.

La **media**, **media aritmética** o el **promedio** es la medida de la tendencia central de una variable en una población o en una muestra; por ejemplo, la estatura media de los jóvenes de 14 años es de 1.65 m; el diámetro promedio de los pistones fabricados por la máquina número 3, ayer en el primer turno, fue de 41.538 mm; la media aritmética de las resistencias eléctricas producidas la semana pasada fue de 1 050 ohms.

La **media poblacional** se denomina μ (letra griega, se pronuncia "mu" o "miu") y es igual al promedio de una variable de la población total; es la suma de los valores de una variable medida en todos los elementos de una población, dividida entre el número total (N) de elementos en la población:³

$$\mu = \frac{X_1 + X_2 + X_3 + \dots + X_N}{N} = \frac{\sum_{i=1}^N X_i}{N}$$

³ Se usa X (mayúscula) para indicar el valor de una variable. No confundirla con X (qui), la letra griega.

La media muestral se denomina \bar{X} (equis barra, equis tilde o equis media) y es igual al promedio de una variable en la muestra y se calcula así:

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

La diferencia es que en la media muestral se miden los valores de la variable sólo en los elementos de la muestra, y se divide entre la cantidad (n) de individuos tomados de la muestra.

Se utilizan letras griegas para referirnos a la población y letras latinas para indicar la muestra (N para la población y n para la muestra).

Las **medidas de dispersión** son indicadores que sirven para saber cómo se distribuyen las observaciones alrededor de la media; por ejemplo, si bien el promedio de estatura de los jóvenes de 14 años es 1.65 m, no todos miden eso, algunos miden más y otros miden menos. Dentro de las medidas de dispersión las más útiles son:

- La **varianza o variancia** es el promedio de las desviaciones elevadas al cuadrado.

La varianza de la población es:

$$\sigma^2 (\text{sigma al cuadrado}) = \frac{(X_1 - \mu)^2 + (X_2 - \mu)^2 + (X_3 - \mu)^2 + \dots + (X_N - \mu)^2}{N} = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N}$$

- Para que el cálculo quede en las mismas unidades de las observaciones, se extrae la raíz cuadrada de la varianza y se obtiene así la desviación estándar:

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

Si se opera sobre una muestra, la varianza y la desviación estándar de la muestra son:

- varianza de la muestra

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

- desviación estándar de la muestra

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

La **distribución normal** es una distribución de datos continuos que produce una curva simétrica en forma de campana. Si un histograma de eventos al azar se divide en clases cada vez más pequeñas y contiene un número muy grande de dichos eventos, la línea que une el borde superior de cada barra vertical es la curva de distribución normal; por ejemplo, en la manufactura de rines de 14 pulgadas para llanta de automóvil, la inmensa mayoría medirá 14 pulgadas de diámetro, algunos medirán 14.05 o 13.95 y sólo unos cuantos medirán 14.1 o 13.9; en una fábrica de ollas de peltre, el espesor del vitrificado promedio será de 0.03 mm, algunas piezas tendrán 0.02 o 0.04 mm y muy pocas habrán sido fabricadas con 0.01 o 0.05 mm de espesor.

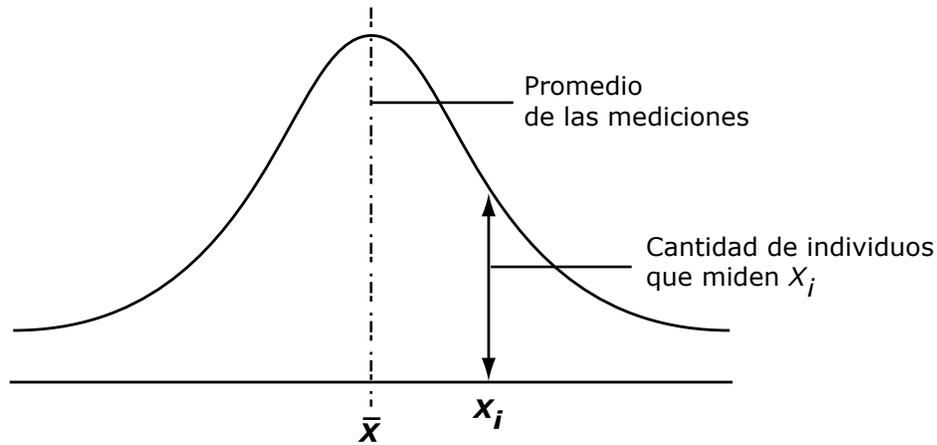


Figura 4.19 Elementos de la curva de distribución normal.

EJERCICIO RESUELTO

Se utilizan dos máquinas diferentes para producir minidiscomos para computadora; como surgieron ciertos problemas con el diámetro de los discos, que debería ser de 3.0 pulgadas, se tomó una muestra de la producción de cada una. Las mediciones de los diámetros se anotan en la tabla 4.14. Calcula el promedio, la varianza y la desviación estándar. Si el objetivo disminuyera la variación alrededor de la media, ¿cuál sería el mejor proceso?

	Máquina 1	Máquina 2	Diferencia Máquina 1 - media	Diferencia²	Diferencia Máquina 2 - media	Diferencia²
	3.41	3.28	-0.1500	0.0225	-0.0300	0.0009
	3.74	3.07	0.1800	0.0324	-0.2400	0.0576
	3.89	3.3	0.3300	0.1089	-0.0100	0.0001
	3.65	3.25	0.0900	0.0081	-0.0600	0.0036
	3.22	3.75	-0.3400	0.1156	-0.4400	0.1936
	3.65	3.2	0.0900	0.0081	-0.1100	0.0121
	3.33	3.3	-0.2300	0.0529	-0.0100	0.0001
Total	24.89	23.15	Total	0.3485	Total	0.2680
Media	3.56	3.31	entre n - 1 > (= s ²)	0.0581	entre n - 1 > (= s ²)	0.0447

Tabla 4.14

La varianza de las muestras de la máquina 1 es 0.0581 y la de la máquina 2 es 0.0447; por lo tanto, el mejor proceso es el de la máquina 2, ya que produce con menor variación. Las

desviaciones estándar de la muestra son:

$$s_1 = \sqrt{0.0581} = 0.241 \text{ y } s_2 = \sqrt{0.0447} = 0.211$$

respectivamente.

Unidades de la distribución normal

Para comparar cualquier unidad de medida, ya sean metros, litros o kilos, con una distribución normal, es necesario codificar los datos. A esta codificación se le llama estandarización.

¿Cómo se compara la forma de medir una variable con la distribución normal?

Puesto que la curva de distribución normal es una línea que representa cualquier tipo de variable, no tiene dimensiones, pero para utilizarla es necesario convertir las variables que se manejan en la práctica a este valor estandarizado. Z o K es tal valor estandarizado y es la coordenada sobre el eje horizontal entre el promedio y el valor observado en un elemento de la población. Se puede decir que es la distancia desde el centro de la curva, hasta un punto cualquiera de ésta.

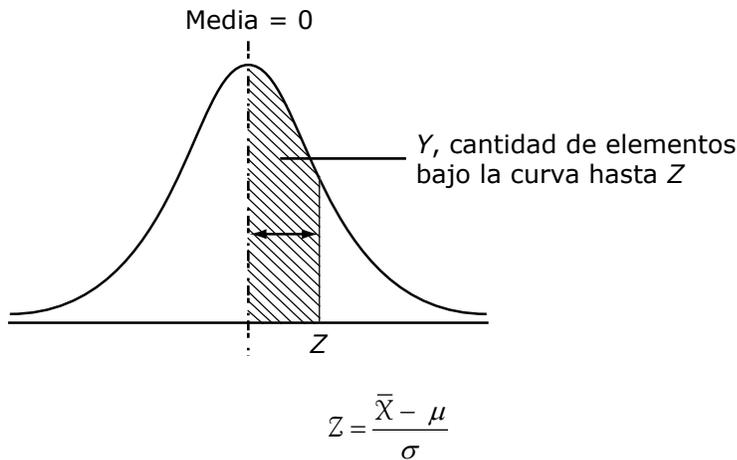


Figura 4.20 Z sobre la curva de distribución normal.

La distribución normal estandarizada tiene una media de cero. El área total bajo la curva es 1 y el área desde la media hasta una coordenada Z se puede consultar en las tablas del apéndice y en algunas calculadoras de mano. Hay dos formas de tabular el área: algunas tablas la presentan desde el centro de la curva, es decir, desde la media hacia la derecha, con números positivos que van desde 0.0000 hasta 0.5000; considerando que la curva es simétrica, si el área que se busca está del lado izquierdo, simplemente se busca su equivalente del lado derecho.

Otras tablas presentan el área desde el extremo izquierdo de la curva (0.0000) y van avanzando hasta llegar a la media (0.5000) y continúan tabulando valores hasta el extremo derecho de la curva (1.0000). Para saber de cuál se trata, hay que ver el esquema que aparece en la parte superior de las tablas; por ejemplo:

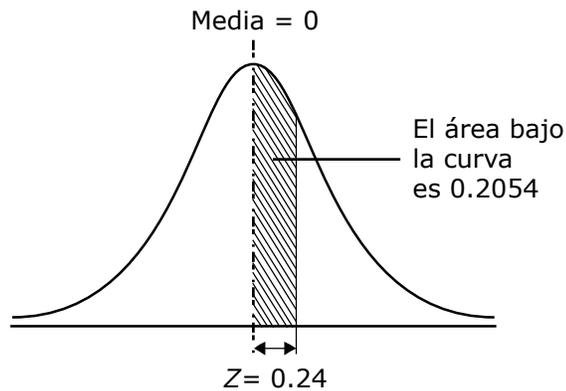


Figura 4.21 El área bajo la curva se consulta en las tablas correspondientes de distribución normal.

Si consideramos los datos del ejercicio anterior, y suponiendo que las dimensiones de la población de todos los discos producidos varían de acuerdo con una distribución normal, se podrían estandarizar los datos de la siguiente manera:

Puesto que el diámetro de los discos debería ser 3.0 pulg, $\mu = 3.0$; la máquina 1 los produce con un promedio de $\bar{X}_{M1} = 3.56$ pulg y la máquina 2 con $\bar{X}_{M2} = 3.31$ pulg; se anota sobre una gráfica de distribución normal cuáles serían las posiciones de μ , \bar{X}_{M1} , \bar{X}_{M2} . Considerando que s^2 (la varianza de la muestra) es igual a 0.0581, la desviación estándar s será la raíz cuadrada, es decir, 0.241 y 0.2114, respectivamente.

Primero hay que codificar los datos:

$$Z_{M1} = \frac{\bar{X} - \mu}{\sigma} = \frac{3.56 - 3.00}{\sqrt{0.0581}} \quad 2.32$$

$$Z_{M2} = \frac{\bar{X} - \mu}{\sigma} = \frac{3.31 - 3.00}{0.2114} \quad 1.4664$$

Luego se consulta la tabla y se encuentra el área bajo la curva hasta la cota $Z = 2.32$ y $Z = 1.47$ (redondeando).

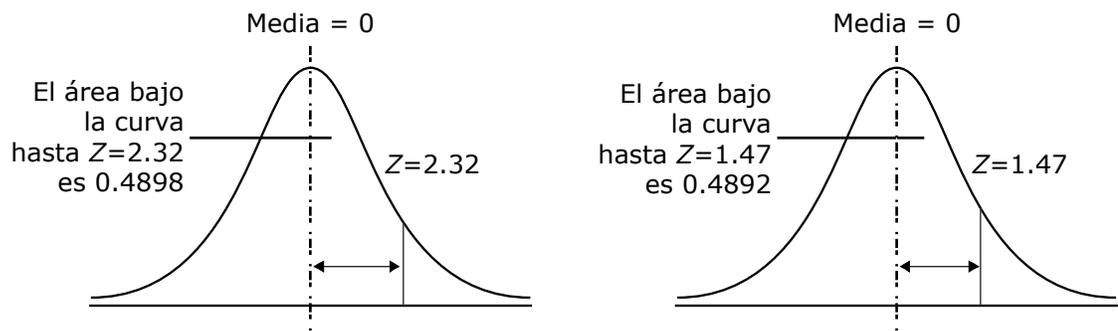


Figura 4.22 El área bajo la curva se consulta en tablas para cada valor de Z .

La tabla al final de la unidad indica el área desde el centro (la media) hasta el valor $+Z$; si Z es negativa, como la curva es simétrica, se consulta en la tabla como si fuera un valor positivo, pero se considera del lado izquierdo de la media.

Puesto que 100% de las mediciones están bajo la curva, se puede interpretar el área debajo de ésta como la probabilidad de que una medida se presente; por ejemplo, si se observa la figura 4.22, primera parte, se puede decir que 48.98% de todas las mediciones que se hagan de los discos de la máquina 1 caerán entre la media y la cota 2.32; en la gráfica de la derecha, se puede decir que 42.92% de todas las mediciones de los discos de la máquina 2 caerán entre la media y la cota 1.47.

EJERCICIOS RESUELTOS

1. Un lote muy grande de tornillos producidos ayer se ha hecho sospechoso de tener una longitud mayor a la especificada de 20 mm; el inspector de calidad toma al azar una muestra de diez piezas y la mide; encuentra que el promedio en la muestra es 19.98 mm con una desviación estándar de 0.1; ¿qué porcentaje de toda la producción se encontrará entre el promedio medido en la muestra y el promedio esperado de la población total?

$$Z = \frac{\bar{X} - \mu}{\sigma} = \frac{19.98 - 20.00}{0.1} = -0.20$$

Con este dato se consulta la tabla y se encuentra el valor 0.0793, es decir, 7.93% de la producción de ayer se encontrará entre 19.98 y 20 mm.

2. Con los mismos datos, ¿qué porcentaje de la población se encontrará entre 19.90 y 20.10 mm?

$$Z_1 = \frac{\bar{X} - \mu}{\sigma} = \frac{19.90 - 20.00}{0.1} = -1 \quad \text{y} \quad Z_2 = \frac{\bar{X} - \mu}{\sigma} = \frac{20.10 - 20.00}{0.1} = 1$$

0.3413 (de la zona a la izquierda de la media) + 0.3413 (de la zona a la derecha de la media) = 0.6826, es decir, 68.26% de la población estará entre esos valores.

3. ¿Qué porcentaje de la población estará entre 19.95 y 20.15 mm si la varianza es 0.02 y el promedio es 20.00?

Convertir primero varianza en desviación estándar:

$$s = \sqrt{s^2} = \sqrt{0.02} = 0.141$$

Luego calcular cada una de las Z :

$$-Z_1 = \frac{\bar{X} - \mu}{\sigma} = \frac{19.95 - 20.00}{0.141} = -0.35 \quad \text{y} \quad Z_2 = \frac{20.15 - 20.00}{0.141} = 1.06$$

Consultar la tabla para cada ordenada Z : $\text{área}_1 = 0.1368$ y $\text{área}_2 = 0.3554$; se suman ambas y el total es 0.4922, por tanto, 49.22% de la población medirá entre 19.95 y 20.15 mm.

La **hipótesis** ($H_{\text{subíndice}}$) es la suposición que se hace del valor medio de una variable en una población, pero como es muy difícil medir todos sus elementos se toma una muestra y el valor medio encontrado se usa para aceptar o rechazar la hipótesis; por ejemplo, se formula la hipótesis de que la medida más común de calzado para alumnos de segundo año de secundaria es 26.5 cm y una fábrica toma una muestra de 10% de todos los alumnos de ese grado en la capital y mide su calzado. Si el promedio de la medida tomada sobre la muestra está entre 26.3 y 26.7 cm ($\pm 1\%$ de 26.5 cm), se aceptará la hipótesis.

La decisión fundamental en asuntos de calidad es la **aceptación** o **rechazo** de un lote de producción (la población), basándonos en la medición de variables o atributos de una muestra y comparándolos con valores esperados, deseados o aceptables. Si la producción de que se habla es pequeña o puede ser detenida para inspección de todos sus elementos, la decisión se puede tomar con un alto grado de confianza; es decir, si se pudieran medir todos los objetos de un conjunto, sería difícil cometer el error de aceptar lo que debería haberse rechazado o rechazar lo que debería ser aceptado.

Sin embargo, cuando los lotes de producción son muy grandes o éste no puede detenerse, es necesario tomar muestras para medirlas y tomar decisiones basándonos en una parte del total. En la práctica, es imposible hacer inspecciones basadas en 100% de los artículos y 100% de sus variables; para tal efecto, se cuenta con instrumentos estadísticos que permiten tomar decisiones rápidas y confiables, aunque sólo se mida una parte del total.

Se puede tomar una buena decisión si se cuenta con toda la información, pero no siempre es así; el propósito de la estadística es reducir el grado de incertidumbre en las decisiones, a pesar de no contar con toda la información. La principal herramienta para la toma de decisiones en condiciones de incertidumbre es la **prueba de hipótesis**.

La prueba de hipótesis comienza formulando una suposición "nula" H_0 y una suposición alterna H_A . Tomemos como ejemplo el de una envasadora de granos que vende arroz en bolsas de dos kilos:

$$H_0: \mu = 2 \quad H_A: \mu \neq 2$$

La **hipótesis nula** dice que el promedio de contenido en todas las bolsas (toda la población) es de 2 kg; se prueba contra la hipótesis alterna de que el promedio de contenido de todas las bolsas *no* es de 2 kg (nótese el signo). La hipótesis nula puede ser rechazada o no rechazada, pero nunca se puede "aceptar" una hipótesis nula como "verdadera".

Cuando *no se rechaza* una hipótesis, lo que se está diciendo es que con la muestra tomada no hay suficiente evidencia para rechazarla. Podría suceder que el promedio real de todas las bolsas (toda la población) fuera de 1.95 kilos, pero por errores, de muestreo o de medición, el promedio de la muestra se reportará de 2 kg; esto no prueba que el total sea de 2 kg.

En la prueba de hipótesis, se supone que H_0 es verdadera hasta que se demuestra que es falsa y esto se logra sólo cuando la diferencia entre el valor observado en la muestra y el esperado por la hipótesis es suficientemente significativo; es decir, cuando esa diferencia es mayor que una diferencia debida al azar o a errores de muestreo.

Se dice que **una diferencia es estadísticamente insignificante** si la diferencia entre el valor medio de la población y el valor medio de la muestra es suficientemente pequeño como para atribuirlo a un error de muestreo o al azar.

En la práctica, la media de la población es lo que se busca; ese valor se plantea como la hipótesis y luego se toma una muestra de dicha población. Se mide el valor medio en ésta y se compara con la hipótesis suponiendo que ambas se comportan en una distribución normal, si la diferencia no es significativa, es decir, es menor que un porcentaje preestablecido de tolerancia, se acepta la hipótesis; si la diferencia es significativa, se rechaza la hipótesis. El **error de muestreo** es el nivel de confianza con el que deseamos aceptar o rechazar la hipótesis.

De una población distribuida en forma estándar, 95% de los elementos están entre $Z_1 = -1.96$ y $Z_2 = 1.96$, ya que el área de cada una hasta la media es 0.475 y el área entre ambas es $2 \times 0.475 = 0.95 = 95\%$. Esto se puede interpretar también diciendo: "existe 95% de probabilidad de que los resultados de la muestra caigan dentro de $Z = \pm 1.96$ si la hipótesis nula es verdadera".

Los valores señalados para $Z = \pm 1.96$ son **valores críticos** porque determinan la zona de rechazo; lo que falta para 100% es 5 y se denomina **nivel de significancia o valor alfa** porque señala los límites dentro de los cuales una diferencia entre los valores de la muestra y los valores esperados de la hipótesis son suficientemente significativos como para rechazar la hipótesis nula. Ese 5% está repartido en los extremos de la curva, 2.5% a cada lado, porque 95% corresponde a la zona central alrededor de la media.

Se puede señalar cualquier valor crítico, pero los más frecuentes son 95, 98 y 99% y sus respectivas cotas Z son: ± 1.96 , ± 2.34 y ± 2.58 .

EJERCICIO RESUELTO

Es común ver que las botellas de refrescos no están llenas hasta el mismo nivel; si se anuncia que un refresco contiene 16 onzas (oz) y una muestra de 25 unidades resulta con un promedio de 15.85 oz y una desviación estándar de 1 oz, ¿qué probabilidad hay de que esa media sea sólo un error de muestreo y de que realmente todas las botellas tengan una media de llenado de 16 oz?

Codificando:

$$Z_1 = \frac{\bar{X} - \mu}{\sigma} = \frac{15.85 - 16.00}{1} = -0.15$$

Puesto que, como consumidores, sólo nos interesa si las botellas están menos llenas que lo que indica su etiqueta, la prueba se aplica sobre un solo extremo de la curva de distribución normal. Para $Z = -0.15$, el área es 0.0596; si el llenado tiene únicamente variaciones al azar, 5.96% de todas las botellas estarán llenas entre 15.85 y 16 oz; si se considera que también habrá botellas con más de 16 oz y que esta parte tiene 50% de probabilidad de suceder, $5.96 + 50 = 55.96\%$ de las botellas contendrán 15.85 oz o más y sólo $100 - 55.96 = 44.04\%$ (aproximadamente una de cada dos) contendrán menos. En una prueba rápida, si el gerente de la tienda encuentra más de una botella con menos llenado en cada dos que observe, deberá rechazar el lote completo porque es probable que tenga errores de llenado, más allá de lo debido, por variaciones al azar.

EJERCICIO 5

1. ¿En qué se basa la toma estadística de decisiones?
 - a) En la teoría de que el promedio es una medida representativa del total.
 - b) En que la mayoría de los elementos de una población es igual a la media.
 - c) En que no se puede medir el total de una población, pero sí una parte de ella.
 - d) En tomar medidas de una muestra que pertenece a un grupo más grande.

2. ¿Qué es una variable?
 - a) Una cualidad observable que puede ser medida sobre una escala continua.
 - b) Cualquier valor de un objeto que puede cambiar con respecto al tiempo.
 - c) Una medida representativa de una característica de un proceso en marcha.
 - d) La capacidad de un parámetro de adoptar valores a lo largo del tiempo.

3. ¿Qué es la varianza o variancia?
 - a) Es la variabilidad de un rango de datos; se calcula sacando la diferencia entre el valor más pequeño y el más grande.
 - b) Es una medida de dispersión que representa el promedio del cuadrado de las diferencias con respecto a la media.
 - c) Es el número de elementos de un conjunto que caen por arriba o por debajo de la media de la población de que se trate.
 - d) Es la cantidad de elementos que caen cercanos a la media aritmética de una muestra o de una población.

4. Considere la siguiente tabla que representa la recaudación diaria en la caja 12 de un supermercado.

Recaudación \$/día	12 500	11 380	4 960	15 789	8 321	10 483	9 148	12 357	11 987
-----------------------	--------	--------	-------	--------	-------	--------	-------	--------	--------

- ¿Cuál es la media de recaudación en esta caja?
- a) 10 679
 - b) 10 976
 - c) 10 796
 - d) 10 769
-
5. Con los mismos datos, ¿cuál es la desviación estándar de la recaudación diaria? (considérese la muestra como si fuera población).
 - a) 3 062
 - b) 3 620
 - c) 3 206
 - d) 3 026

AUTOEVALUACIÓN

1. Al gerente de una planta productora de piezas de plástico para la industria automotriz se le informó que un determinado producto aparece con excesos en el espesor de la pintura, según la siguiente tabla:

Máquina	L	M	M	J	V	L	M	M	J	V	L	M	M	J	V	L	M	M	J	V	L	Promedio
A	0.2	0.4	0.1	0.3	0.2	0.0	0.5	0.1	0.4	0.3	0.0	0.3	0.2	0.1	0.1	0.4	0.5	0.2	0.3	0.0	0.3	
B	0.0	0.2	0.1	0.1	0.0	0.4	0.6	0.2	0.1	0.3	0.4	0.5	0.2	0.2	0.0	0.1	0.4	0.2	0.2	0.1	0.2	
C	0.5	0.7	0.3	0.4	0.6	0.4	0.8	0.4	0.3	0.7	0.3	0.9	0.5	0.6	0.3	0.6	0.7	0.5	0.3	0.5	0.4	
D	0.2	0.5	0.1	0.3	0.2	0.4	0.6	0.1	0.0	0.3	0.2	0.4	0.2	0.1	0.3	0.4	0.5	0.2	0.0	0.4	0.3	
E	0.5	0.6	0.4	0.2	0.3	0.3	0.4	0.2	0.5	0.1	0.0	0.6	0.5	0.4	0.3	0.5	0.6	0.2	0.4	0.3	0.2	

La mayor desviación aparece:

- En la máquina E.
 - Los lunes.
 - En la máquina C.
 - Los martes.
2. Una oficina de atención a clientes atiende las ventas por teléfono y recibe sus quejas; ¿qué criterios de clasificación de quejas son necesarios para informar a producción de las fallas en los productos?
- Antigüedad del producto y tipo de falla.
 - Fecha de queja y nombre de la operadora.
 - Cantidad de fallas reportadas al día.
 - Localización del cliente y tipo de falla.
3. En la construcción del diagrama de Pareto, ¿qué contienen los ejes verticales izquierdo y derecho?
- El izquierdo tiene las categorías que se estudian y el derecho las unidades defectuosas acumuladas.
 - El izquierdo es el porcentaje de defectos y el derecho la cantidad de observaciones.
 - El izquierdo es el número de eventos observados y el derecho indica la frecuencia de cada uno.
 - El izquierdo es el número de unidades del eje horizontal y el derecho es el porcentaje acumulado.
4. Una constructora ha inspeccionado su última obra multifamiliar y reporta los defectos encontrados en la siguiente tabla de hallazgos; puesto que se observa que los datos están fuera de orden, una vez que estén ordenados, ¿cuál es el porcentaje acumulado hasta “ventanas atoradas”?

Estratificación por tipo de defectos

Tipo de defecto encontrado	Cantidad
Desnivel de pisos	19
Plomería con fugas	115
Fallas eléctricas	47
No prende calentador	9
Ventanas atoradas	10

- a) 95.5
b) 9.5
c) 5
d) 10
5. Utilizando la misma tabla del ejercicio anterior, ¿cuál es el tercer defecto más importante?
- a) Fallas eléctricas.
b) No prende calentador.
c) Plomería con fugas.
d) Desevel de pisos.
6. ¿Qué es un histograma?
- a) Es un diagrama de Pareto, pero con los valores centrados sobre la escala horizontal.
b) Una gráfica de barras que indica la frecuencia con la que se repiten los valores dentro de un rango.
c) Es una gráfica de dispersión, pero en vez de puntos se dibujan barras verticales.
d) Es un diagrama en el cual se indican los porcentajes acumulados de las observaciones.
7. ¿Qué es el rango de las observaciones?
- a) Es el promedio alrededor del cual se sitúan las observaciones.
b) Es la cantidad de categorías que se indican en un histograma.
c) Son los límites absolutos que se pueden tolerar en una variable.
d) Son los valores máximo y mínimo que puede adoptar una variable.
8. Estadísticamente hablando, ¿qué es la población?
- a) La cantidad de personas que vive en una localidad.
b) El recinto físico en el que vive la gente de un pueblo.
c) Una parte de los elementos del universo estadístico.
d) Todos los elementos de un conjunto homogéneo.
9. ¿Cómo se le llama a un histograma en el cual la barra de mayor frecuencia no coincide con el promedio?
- a) Sesgado.
b) Simétrico.
c) Bimodal.
d) Multimodal.

10. ¿Cómo se presenta un histograma cuyos datos tienen mucha variabilidad?
- No hay una frecuencia que destaque alrededor del valor central.
 - Las barras tienen tamaños muy diferentes y no hay dos iguales.
 - Hay categorías vacías dentro del rango de las observaciones.
 - Hay muchas observaciones en los extremos del rango.
11. En un proceso que se desempeña al azar, ¿qué sucede cuando los límites entre una clase y otra son muy cercanos?
- Aumenta el número de clases y se dificulta la elaboración y lectura.
 - Disminuye el número de observaciones en cada clase y disminuye la precisión.
 - El perfil del histograma se aproxima a la curva de distribución normal.
 - Aparecen con mayor facilidad los sesgos y acantilados de la gráfica.
12. ¿Qué indica la curva de distribución normal?
- Cómo se acomodan los datos de un proceso con variación natural.
 - El reparto de las observaciones alrededor del promedio de lo que se mide.
 - La variación de los datos por causa de la variación en las mediciones.
 - El comportamiento de cualquier medida que varíe por efecto del azar.
13. ¿En qué consiste que la relación entre dos variables sea negativa?
- No hay una relación de dependencia entre la variable independiente y la otra.
 - Al aumentar la variable independiente, disminuye la variable dependiente.
 - Al aumentar la variable dependiente, disminuye la variable independiente.
 - Al disminuir la variable independiente, disminuye la variable dependiente.
14. El rendimiento de una parcela de cultivo de papa guarda cierta relación con la cantidad de agua que se aplica durante las primeras ocho semanas después de la siembra. Los datos se muestran en la siguiente tabla:

Agua aplicada m ³	200	190	185	205	170	215	191	211	230	180
Rendimiento ton	65	54	48	68	30	80	60	73	90	45

El coeficiente de correlación entre estas variables es:

- 0.899
 - 0.989
 - 0.998
 - 0.897
15. Un examen hecho sobre el contenido de cereal en diez cajas etiquetadas con 454 g ha revelado que el promedio de contenido es de 435 g con una desviación estándar de 20 g ¿cuánto vale Z ?
- 19.0
 - 20
 - 0.95
 - 3.0

CALIDAD Y PRODUCTIVIDAD

16. Un fabricante de cables eléctricos industriales ofrece su producto en rollos de 100 m con una tolerancia de más o menos 1 m. En el último lote recibido, el promedio de longitud en una muestra de nueve rollos fue de 92.5 m, con una desviación estándar de 8 m. Formula la hipótesis nula.
- a) $H_0 = 100$ m
 - b) $H_0\mu = 92.5$ m
 - c) $H_0 = 92.5$ m
 - d) $H_0\mu = 100$ m
17. Considerando los datos anteriores, ¿dentro de qué límites de Z no se podría rechazar la hipótesis nula anterior con 95% de confianza?
- a) 1.96
 - b) ± 1.96
 - c) ± 10 g
 - d) 10 g

RESPUESTAS DE LOS EJERCICIOS

EJERCICIO 1

1. b)
2. Casi la mitad de las fallas se produjeron en la mesa 2; todos los trabajadores incurren en más fallas que su promedio cuando trabajan en esa mesa. Es probable que alguna herramienta esté descalibrada. El trabajador *D* es mejor que el promedio; el trabajador *E* es peor que el promedio.
3. Porcentaje de rendimiento de jugo por kilo de fruta, por proveedor y máquina

Proveedores	Máquina A	Máquina B	Máquina C	Máquina D	Promedio
Primero	55	70	50	37	53.0
Segundo	60	52	35	30	44.3
Tercero	75	55	80	68	69.5
Cuarto	30	43	56	85	53.5
Promedio	55.0	55.0	55.3	55.0	

Conclusiones: todas las máquinas tienen el mismo rendimiento promedio, las variaciones provienen de los proveedores; el mejor es *Tercero*, el peor es *Segundo* y los demás están iguales a media tabla.

4. La estratificación consiste en agrupar las observaciones de acuerdo con su naturaleza o con otros factores que influyen en el producto.
5. Recolectar datos, representar gráficamente los datos, determinar las posibles causas, profundizar en la estratificación y elaborar conclusiones.

EJERCICIO 2

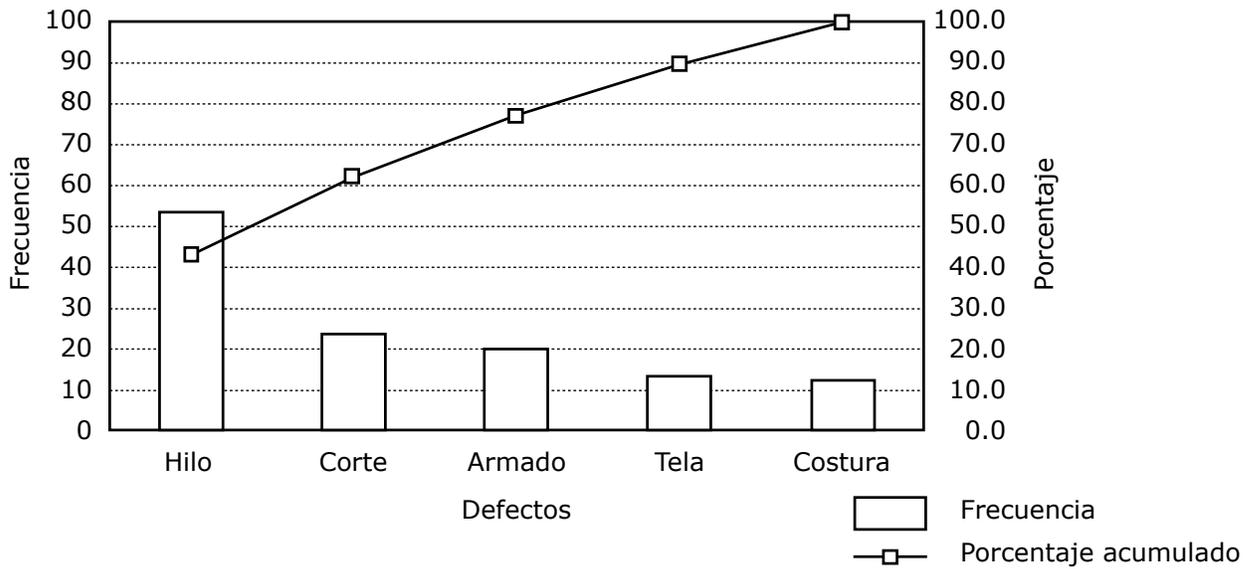
1. b)
2. El primer paso debe ser organizar los datos para poder manejarlos. Para ello se elabora una tabla de estratificación de los defectos por turno:

Defectos	Turno			Total
	1	2	3	
Hilo	13	12	27	52
Corte	6	9	9	24
Armado	5	10	5	20
Tela	6	5	3	14
Costura	6	4	3	13

Defectos	Cantidad	Porcentaje	Cantidad acumulada	Porcentaje acumulado
Hilo	52	42.3	52	42.3
Corte	24	19.5	76	61.8
Armado	20	16.3	96	78.0
Tela	14	11.4	110	89.4
Costura	13	10.5	123	100.0
Total	123	100.0		

CALIDAD Y PRODUCTIVIDAD

A continuación se elabora el DP inicial, con el total de defectos:



La escala izquierda debe llegar hasta 100 y coincide con la escala derecha.

A primera vista el hilo es el mayor problema y la tentación para cambiar de proveedor y de investigar en los tres turnos parece obvia; sin embargo, un DP de segundo nivel para verificar el desempeño de los turnos revela lo siguiente:

Turno	Cantidad	Porcentaje	Cantidad acumulada	Porcentaje acumulado
Tercero	27	51.9	27	51.9
Primero	13	25.0	40	76.9
Segundo	12	23.1	52	100.0
Total	52	100.0		

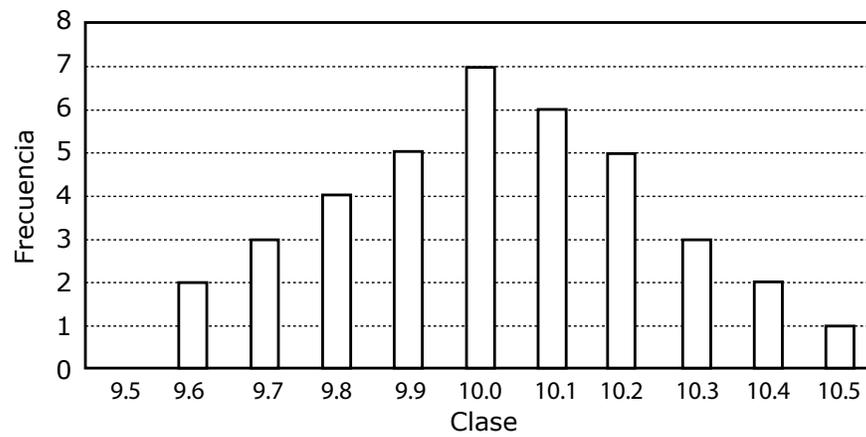
Parece que el problema del hilo no es por el hilo en sí mismo, ya que si así fuera aparecería en igual proporción en los tres turnos; más bien resulta que el *uso del hilo* en el tercer turno está causando el problema.

3. b)
4. c)
5. a)

EJERCICIO 3

1. b)
- 2.

Clase	Frecuencia
9.5	0
9.6	2
9.7	3
9.8	4
9.9	5
10.0	7
10.1	6
10.2	5
10.3	3
10.4	2
10.5	1



3.
 - a) Histograma con sesgo.
 - b) Histograma bimodal.
 - c) Histograma con poca variabilidad.
4. d)
5. a)

CALIDAD Y PRODUCTIVIDAD

EJERCICIO 4

1.

a) Rendimiento contra cilindrada.

<i>n</i>	<i>X</i> cilindrada	<i>y</i> rendimiento	$x_i y_i$	x_i^2	y_i^2
1	800.0	16.0	12 800.0	640 000.0	256.00
2	850.0	16.5	14 025.0	722 500.0	272.25
3	900.0	15.8	14 220.0	810 000.0	249.64
4	1 100.0	17.0	18 700.0	1 210 000.0	289.00
5	1 200.0	14.0	16 800.0	1 440 000.0	196.00
6	1 500.0	14.5	21 750.0	2 250 000.0	210.25
7	1 700.0	15.0	25 500.0	2 890 000.0	225.00
Total	8 050.0	108.8	123 795.0	9 962 500.0	1 698.14

$$\begin{aligned} S_{xy} &= -1325 \\ S_{xx} &= 705\,000 \\ S_{yy} &= 7.08 \\ r &= -0.59 \end{aligned}$$

El coeficiente de correlación es negativo: a mayor cilindrada, menor rendimiento. No hay una clara correlación entre las variables, hay un motor de 1 100 que rinde más que el de 800.

b) Rendimiento contra peso.

<i>n</i>	<i>x</i> peso	<i>y</i> rendimiento	$x_i y_i$	x_i^2	y_i^2
1	1 100.0	16.0	17 600	1 210 000	256.0
2	1 320.0	15.8	20 856	1 742 400	249.6
3	1 540.0	15.0	23 100	2 371 600	225.0
4	1 600.0	14.0	22 400	2 560 000	196.0
5	1 670.0	13.9	23 213	2 788 900	193.2
6	1 700.0	13.0	22 100	2 890 000	169.0
7	1 800.0	12.5	22 500	3 240 000	156.3
Total	10 730.0	100.2	151 769	16 802 900	1 445.1

$$\begin{aligned} S_{xy} &= -1823 \\ S_{xx} &= 355\,342.9 \\ S_{yy} &= 10.8 \\ r &= -0.93 \end{aligned}$$

El coeficiente de correlación es negativo: a mayor peso, menor rendimiento. Hay una clara correlación entre las variables. El coeficiente de correlación es alto.

c) Rendimiento contra octanaje.

<i>n</i>	<i>x</i> octanaje	<i>y</i> rendimiento	$x_i y_i$	x_i^2	y_i^2
1	80.0	13.9	1 112.0	6 400.00	193.21
2	85.0	14.0	1 190.0	7 225.00	196.00
3	90.0	15.0	1 350.0	8 100.00	225.00
4	93.0	15.8	1 469.4	8 649.00	249.64
5	95.0	16.0	1 520.0	9 025.00	256.00
Total	443.0	74.7	6 641.4	39 399.00	1 119.85

$$\begin{aligned} S_{xy} &= 23.0 \\ S_{xx} &= 149.2 \\ S_{yy} &= 3.83 \\ r &= 0.961 \end{aligned}$$

El coeficiente de correlación es positivo: a mayor octanaje, mayor rendimiento. Hay una clara correlación entre las variables. El coeficiente de correlación es alto.

2. b)
3. c)
4. a)

EJERCICIO 5

1. d)
2. a)
3. b)
4. d)
5. a)

RESPUESTAS DE LA AUTOEVALUACIÓN

1. d)
2. a)
3. d)
4. a)
5. d)
6. b)
7. d)
8. d)
9. a)
10. a)
11. c)
12. d)
13. b)
14. b)
15. c)
16. d)
17. b)