Parámetros Estadísticos básicos, Resumen y Presentación de datos

Jhon Jairo Padilla, PhD.

Motivación

- Los resúmenes y las representaciones de datos son esenciales porque:
  - Enfocan al ingeniero en características importantes de los datos
  - Proporcionan ideas acerca del modelo que debería emplearse para la solución del problema
- Características básicas de los datos:
  - **Localización:** “Parte de en medio” de los datos.
  - **Dispersión o Variabilidad:** Cuál es el comportamiento o patrón que describe la variación en los valores de las muestras.

Representación de los datos

- Para mostrar las características básicas (Ubicación, Variabilidad) de los datos se utilizan Diagramas.
- Algunos de los más utilizados son:
  - Diagrama de Puntos
  - Diagramas de Tallo y Hoja
  - Distribuciones de Frecuencia e Histogramas
  - Gráficas de Caja
  - Gráficas de series de tiempo

Resumen de datos básico

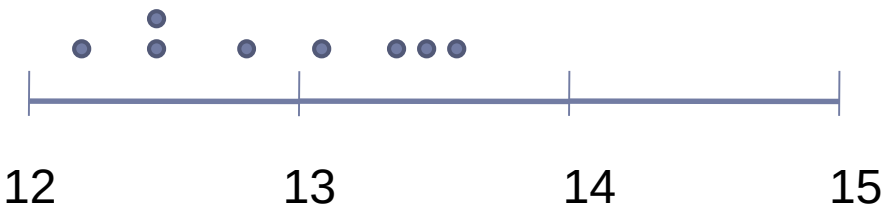
Diagrama de puntos, media, varianza

Diagrama de Puntos

- Se utilizan para ilustrar un número reducido de datos (hasta 20 observaciones)
- Si es muy pequeño el número de observaciones, es difícil identificar algún patrón de variabilidad específico

Muestras	12.6	12.9	13.4	12.3	13.6	13.5	12.6	13.1

Diagrama de Puntos:



Localización de los datos

- Es posible describir numéricamente las características básicas de los datos
- Localización:
  - Es la tendencia central de los datos
  - Se caracteriza con el promedio aritmético ordinario (o media)
  - Cuando los datos representan muestras (por lo general es así), la media aritmética se conoce como **media muestral.**

Localización: Media Muestral

- Definición:
  - Si las *n* observaciones de una muestra se denotan por x_1, $x_2, \ldots, xn$, entonces la **media muestral** es:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

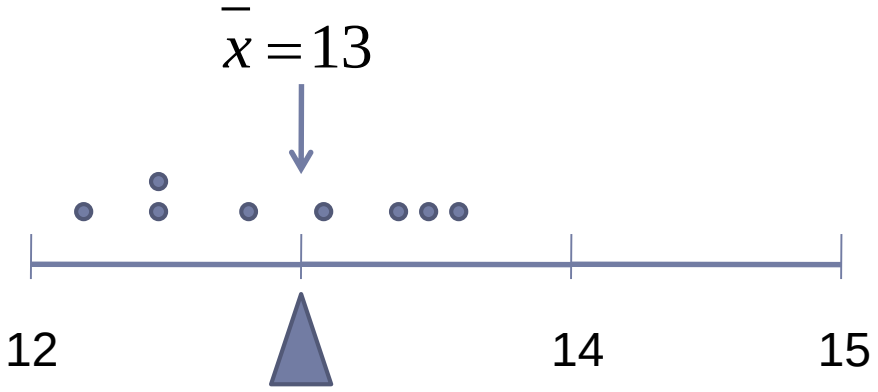
- Ejemplo:

Muestras	12.6	12.9	13.4	12.3	13.6	13.5	12.6	13.1

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{8} x_i}{8} = \frac{12.6 + 12.9 + \ldots + 13.1}{8} = 13.0$$

Localización: Media Muestral

- Interpretación física:
  - La **media muestral** puede verse como un **punto de equilibrio**.
  - Si cada observación representa *1* libra de masa colocada en el punto correspondiente sobre el eje x, entonces un punto de apoyo ubicado en $\bar{x}$ equilibraría exactamente este sistema de pesos.



Localización: Media Poblacional

- Si se toman todas las observaciones de una población para calcular el promedio, a este promedio se le llama **Media Poblacional** (μ).
- Si la población tiene un número finito (N) de elementos (observaciones), entonces la media poblacional es,

$$\overline{\mu} = \frac{\sum_{i=1}^{N} x_i}{N}$$

- **Nota:** La *media muestral* es una aproximación razonable de la *media poblacional* (siempre y cuando el número de muestras sea representativo)

Variabilidad de los datos: Varianza Muestral

- Definición:
  - Si $x_1, x_2, \ldots, xn$ es una muestra de observaciones, entonces la **varianza muestral** es

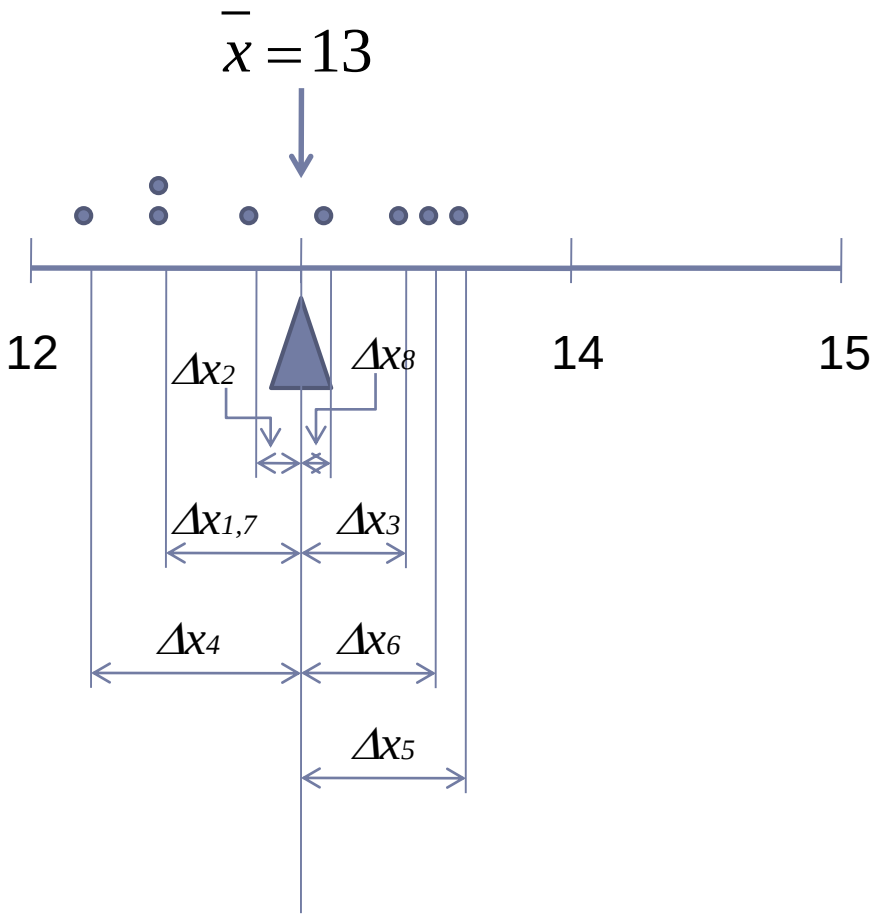
$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

  - Adicionalmente, la **desviación estándar muestral**, s, es la raíz cuadrada positiva de la **varianza muestral**.

- **Unidades:** Las unidades de la **varianza muestral** son las unidades originales de la variable al cuadrado, mientras que las unidades de la **desviación estándar** son las mismas de la variable original.
  - Ejemplo:
    - **Unidad original:** Voltios (V)
    - **Unidades varianza muestral:** V^2
    - **Unidades desviación estándar:** V

Cómo mide la variabilidad la varianza muestral?

- Entre mayor sea la cantidad de variabilidad de los datos, más grandes serán en magnitud absoluta algunas de las desviaciones $(x_i - \bar{x})$.
- Como la suma de las desviaciones siempre es cero, se requiere una medida de variabilidad cuya sumatoria no se anule. La **varianza muestral** permite esto.
- Por tanto:
  - Si s^2 es pequeña, hay relativamente poca variabilidad
  - Si s^2 es grande, la variabilidad es relativamente grande



Ejemplo: varianza muestral

i	x_i	$(x_i-\bar{x})$	$(x_i-\bar{x})^2$
1	12.6	-0.4	0.16
2	12.9	-0.1	0.01
3	13.4	0.4	0.16
4	12.3	-0.7	0.49
5	13.6	0.6	0.36
6	13.5	0.5	0.25
7	12.6	-0.4	0.16
8	13.1	0.1	0.01
suma	104.0	0.0	1.60

Suponiendo que las muestras sean medidas de voltaje tomadas de la salida de un circuito eléctrico, las unidades de x_i serán voltios (V).

Varianza muestral:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{1.60}{8-1} = \frac{1.60}{7} = 0.2286V^2$$

Desviación estándar muestral:

$$s = \sqrt{0.2286} = 0.48V$$

Variabilidad de los datos: Varianza Poblacional

- La **varianza poblacional** (σ^2) es análoga a la **varianza muestral**, pero la **varianza poblacional** tiene en cuenta las observaciones de toda la población (no solo las muestras).
- De igual forma la **desviación estándar poblacional** (σ) es la raíz cuadrada positiva de la **varianza poblacional**.
- Si la población es finita y se compone de N observaciones, la **varianza poblacional** puede definirse como

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \overline{\mu})^2}{N}$$

- Nota: La **varianza muestral** puede usarse como una estimación de la **varianza poblacional**

Comparación Varianza Muestral-Poblacional

- Varianza Muestral

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

- Varianza Poblacional

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{\mu})^2}{N}$$

Diferencia en el denominador

•La **varianza muestral** debería calcularse con respecto a la **media poblacional** (μ).
•En la práctica no se conoce la **media poblacional** (μ).
•Por tanto, se usa la **media muestral** para calcular la **varianza muestral**.
•Esto **ocasiona error** porque la media muestral no es igual a la media poblacional y además *la media muestral se encuentra más cerca de las muestras que la media poblacional.*

•**El error se compensa** disminuyendo en *1* el número de muestras que se coloca en el divisor de la fórmula de la varianza muestral (se toma *n-1* en lugar de *n*)

Grados de Libertad

- La sumatoria de las n desviaciones siempre da cero.
- Si se asignan los valores de cualesquiera $(n-1)$ desviaciones, la *n-ésima* desviación deberá tomar un valor obligatorio que mantenga la sumatoria de las desviaciones en cero.
- Por tanto, sólo $(n-1)$ desviaciones pueden ser determinadas libremente.
- Entonces, se dice que s^2 es una medida basada en $(n-1)$ **grados de libertad**.

Rango muestral

- Es una medida útil de la variabilidad.
- Es la diferencia entre la observación más grande y la más pequeña.
- Definición:
  - Si las n observaciones de una muestra se denotan por $x_1, x_2, \ldots, x_n$, entonces el **rango muestral** es

$$r = \max(x_i) - \min(x_i)$$

- Ejemplo:
  - En el ejemplo anterior, el rango muestral es

$$r = 13.6 - 12.3 = 1.3$$

- En general, cuando la variabilidad de los datos muestrales aumenta, el rango muestral se incrementa. Es un parámetro sencillo pero ignora la información del resto de datos muestrales.

Diagramas de Tallo y Hoja

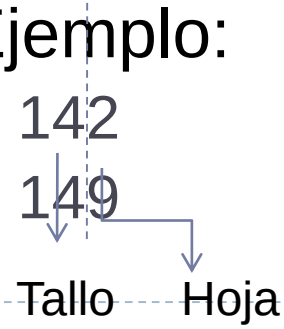
Motivación

- Los diagramas de puntos son útiles para hasta unas 20 observaciones.
- Ejemplo:
  - Suponga que se han tomado medidas de las longitudes de los paquetes (en bytes) que han sido recibidos en un computador durante una transferencia de un archivo
- Qué porcentaje de observaciones están por debajo de 120 bytes?
- Dibujar el diagrama de puntos….?

105	221	183	186	121	181	180	143
97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110
163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123
134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169
199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135
196	201	200	176	150	170	118	149

Construcción del diagrama de Tallo y Hoja

- Cada muestra debe estar representada con por lo menos dos dígitos
- Cada número se divide en dos partes:
  - Tallo (uno o más de los primeros dígitos, empezando por la izquierda)
  - Hoja (dígitos restantes)
- Ejemplo:
  - 142
  - 149



- Se deben elegir pocos tallos en comparación con el número de observaciones.
- Debe haber entre 5 y 20 Tallos

Diagrama de Tallo y Hoja: Ejemplo

Tallo	Hojas												Frecuencia
7	6												1
8	7												1
9	7												1
10	5	1											2
11	5	8	0										3
12	1	0	3										3
13	4	1	3	5	3	5							6
14	2	9	5	8	3	1	6	9					8
15	4	7	1	3	4	0	8	8	6	8	0	8	12
16	3	0	7	3	0	5	0	8	7	9			10
17	8	5	4	4	1	6	2	1	0	6			10
18	0	3	6	1	4	1	0						7
19	9	6	0	9	3	4							6
20	7	1	0	8									4
21	8												1
22	1	8	9										3
23	7												1
24	5												1

Rango en que se encuentran la mayoría de datos

Ubicación valor central

Efecto del número de Tallos

Tallo	Hojas								
6	1	3	4	5	5	6			
7	0	1	1	3	5	7	8	8	9
8	1	3	4	4	7	8	8		
9	2	3	5						

Tallo	Hojas				
6L	1	3	4		
6U	5	5	6		
7L	0	1	1	3	
7U	5	7	8	8	9
8L	1	3	4	4	
8U	7	8	8		
9L	2	3			
9U	5				

Tallo	Hojas		
6z	1		
6t	3		
6f	4	5	5
6s	6		
6e			
7z	0	1	1
7t	3		
7f	5		
7s	7		
7e	8	8	9
8z	1		
8t	3		
8f	4	4	
8s	7		
8e	8	8	
9z			
9t	2	3	
9f	5		
9s			
9e			

Otras características de los datos

- Los diagramas de tallo y hoja permiten determinar otras características de los datos:
  - Percentiles
  - Quartiles
  - Mediana
  - Moda

La mediana muestral

- Es una medida de tendencia central que divide los datos (ordenados) en dos partes iguales: La mitad queda debajo de la mediana y la mitad queda arriba de la mediana.

$$mediana = \frac{160+163}{2} = 161.5$$

Tallo	Hojas												Frecuencia
7	6												1
8	7												1
9	7												1
10	1	5											2
11	0	5	8										3
12	0	1	3										3
13	1	3	3	4	5	5							6
14	1	2	3	5	6	8	9	9					8
15	0	0	1	3	4	4	6	7	8	8	8	8	12
16	0	0	0	3	3	5	7	7	8	9			10
17	0	1	1	2	4	4	5	6	6	8			10
18	0	0	1	1	3	4	6						7
19	0	3	4	6	9	9							6
20	0	1	7	8									4
21	8												1
22	1	8	9										3
23	7												1
24	5												1

Total datos: 80
Mediana: entre datos 40 y 41

La moda muestral

- Es el valor de los datos que ocurre con mayor frecuencia.
- En el ejemplo la moda muestral es 158 ya que este valor ocurre 4 veces (más que ninguno otro).

Tallo	Hojas												Frecuencia
7	6												1
8	7												1
9	7												1
10	1	5											2
11	0	5	8										3
12	0	1	3										3
13	1	3	3	4	5	5							6
14	1	2	3	5	6	8	9	9					8
15	0	0	1	3	4	4	6	7	8	8	8	8	12
16	0	0	0	3	3	5	7	7	8	9			10
17	0	1	1	2	4	4	5	6	6	8			10
18	0	0	1	1	3	4	6						7
19	0	3	4	6	9	9							6
20	0	1	7	8									4
21	8												1
22	1	8	9										3
23	7												1
24	5												1

Cuartiles

- Cuando un conjunto ordenado de datos se divide en 4 partes iguales, los puntos de división se denominan **cuartiles**.
- Primer cuartil ó Cuartil inferior (q_1): el 25% de las observaciones están por debajo de él
- Segundo Cuartil (q_2): el 50% de las observaciones están por debajo de él. Es igual a la mediana.
- Tercer Cuartil ó Cuartil superior(q_3): el 75% de las observaciones están por debajo de él.

Cuartiles

- Si no coinciden en el valor de una observación, los cuartiles se calculan como un valor intermedio entre las dos observaciones vecinas a ellos.

$$q_1 = muestra(m_1) = muestra(\frac{n+1}{4})$$

$$q_3 = muestra(m_2) = muestra(\frac{3(n+1)}{4})$$

$$q_1 = muestra(\frac{80+1}{4}) = muestra(20.25)$$

$$q_1 = 143.50$$

$$q_3 = muestra(\frac{3(80+1)}{4}) = muestra(60.75)$$

$$q_3 = 181.0$$

Tallo	Hojas												Frecuencia
7	6												1
8	7												1
9	7												1
10	1	5											2
11	0	5	8										3
12	0	1	3										3
13	1	3	3	4	5	5							6
14	1	2	3	5	6	8	9	9					8
15	0	0	1	3	4	4	6	7	8	8	8	8	12
16	0	0	0	3	3	5	7	7	8	9			10
17	0	1	1	2	4	4	5	6	6	8			10
18	0	0	1	1	3	4	6						7
19	0	3	4	6	9	9							6
20	0	1	7	8									4
21	8												1
22	1	8	9										3
23	7												1
24	5												1

Rango Inter-Cuartílico (IQR)

- Es una medida de la variabilidad.
- Es menos sensible a los valores extremos que el rango muestral ordinario.

$$IQR = q_3 - q_1$$

Percentil

- El percentil 100k-ésimo es un valor de los datos tal que aproximadamente el 100k% de las observaciones está en este valor o por debajo del mismo.

Distribuciones de Frecuencia e Histogramas

Características

- Una distribución de frecuencia es un resumen de datos más compacto que un diagrama de tallo y hoja.
- Contrucción:
  - Dividir el rango de datos en intervalos (intervalos de clase ó celdas). Las celdas también pueden ser categorías (joven, anciano, viejo, etc)
  - Características de los intervalos:
    - En lo posible deben ser del mismo ancho (mejor interpretación visual).
    - Debe existir cierto criterio para elegir el número de intervalos (NI): Número de observaciones, grado de dispersión de los datos.
    - Valor razonable: $5 \leq NI \leq 20$
    - Criterio práctico: $NI = \sqrt{n}$; n: número de observaciones

Ejemplo: Selección de Intervalos

- Conjunto de datos inicial

105	221	183	186	121	181	180	143
97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110
163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123
134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169
199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135
196	201	200	176	150	170	118	149

- n=80

  $NI = \sqrt{n} = \sqrt{80} \cong 9$
- Entonces un buen número de celdas sería entre 8 y 9
- Valores inicial y final:
  - Menor observación: 76
  - Mayor observación: 245
- Luego, se escoge:
  - Valor Inicial: 70 (<76)
  - Valor final: 250 (>245)
- Rango: 250 – 70 = 180 unidades
- Ancho celda: 180 / 9 = 20

Ejemplo: Selección de Intervalos

Intervalo	Frecuencia	Frecuencia relativa	Frecuencia Acumulada
$70 \leq x < 90$	2	0.0250	0.0250
$90 \leq x < 110$	3	0.0375	0.0625
$110 \leq x < 130$	6	0.0750	0.1375
$130 \leq x < 150$	14	0.1750	0.3125
$150 \leq x < 170$	22	0.2750	0.5875
$170 \leq x < 190$	17	0.2125	0.8000
$190 \leq x < 210$	10	0.1250	0.9250
$210 \leq x < 230$	4	0.0500	0.9750
$230 \leq x < 250$	2	0.0250	1.0000

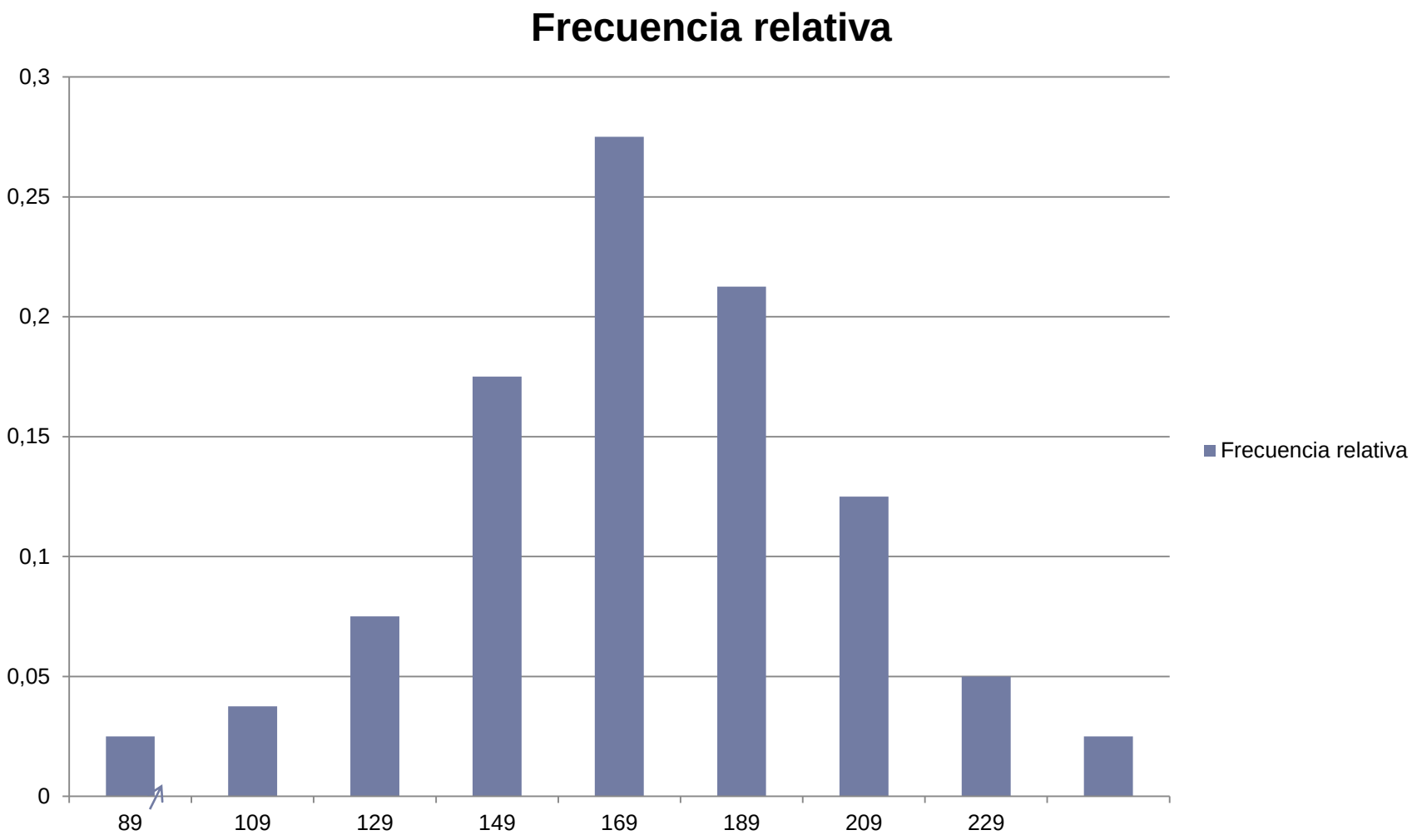
- Distribución de frecuencia relativa:
  - Se determina dividiendo la frecuencia observada en cada intervalo de clase por el número de observaciones (proporción sobre el total)
- Distribución de frecuencia Acumulada:
  - Es la sumatoria acumulada de las frecuencias relativas (proporción acumulada).

Ejemplo: Selección de Intervalos

Intervalo	Frecuencia	Frecuencia relativa	Frecuencia Acumulada
$70 \leq x < 90$	2	0.0250	0.0250
$90 \leq x < 110$	3	0.0375	0.0625
$110 \leq x < 130$	6	0.0750	0.1375
$130 \leq x < 150$	14	0.1750	0.3125
$150 \leq x < 170$	22	0.2750	0.5875
$170 \leq x < 190$	17	0.2125	0.8000
$190 \leq x < 210$	10	0.1250	0.9250
$210 \leq x < 230$	4	0.0500	0.9750
$230 \leq x < 250$	2	0.0250	1.0000

- Ejemplos de Análisis:
  - La mayoría de muestras se encuentran entre 130 y 190 bytes
  - El 97.5% de las muestras están por debajo de 230 bytes.

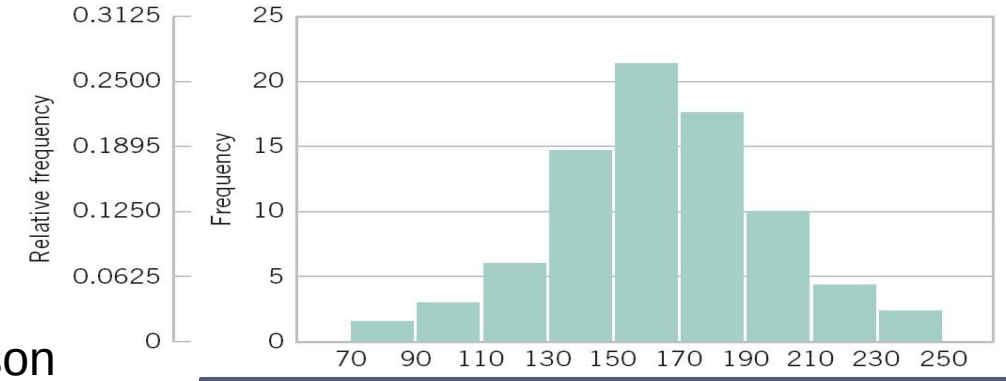
Histograma


Frecuencia relativa
0,3
0,25
0,2
0,15
0,1
0,05
0
89
109
129
149
169
189
209
229
Frecuencia relativa

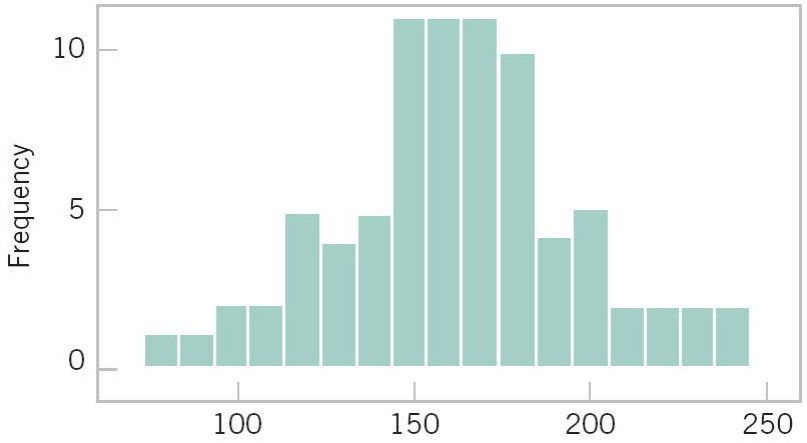
Efecto de diferentes anchos de intervalos en el Histograma

Los Histogramas son más estables para conjuntos grandes de datos (≥75)

Para el ejemplo, los dos histogramas brindan aproximadamente la misma información (n=80)

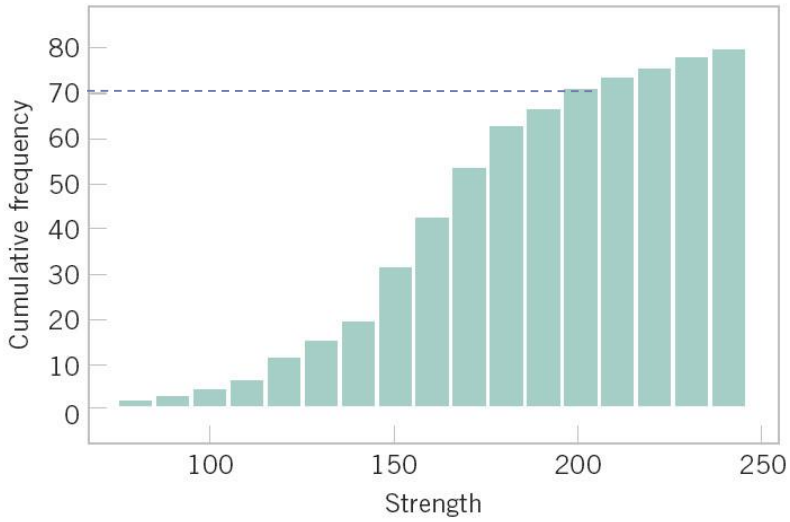


Histograma con 9 celdas



Histograma con 17 celdas

Gráfico de la Distribución acumulada



- Se puede observar por ejemplo que hay 70 muestras menores o iguales que 200 bytes.

Utilidad de los Histogramas

- Muestran una información más resumida de los datos que los Diagramas de Tallo-Hoja (Se pierde cierta información).
- Se gana en brevedad y facilidad de interpretación.
- Son más efectivos para tamaños de muestras relativamente grandes ($\geq$75).
- Con tamaños de muestras grandes, el histograma es un indicador confiable de la forma general de la población.

Tipos de Histogramas según la forma

- Histogramas simétricos:
  - La media y la mediana coinciden
  - Si además, los datos tienen una sola moda, se dice que son *unimodales.* Por tanto, la media, la moda y la mediana

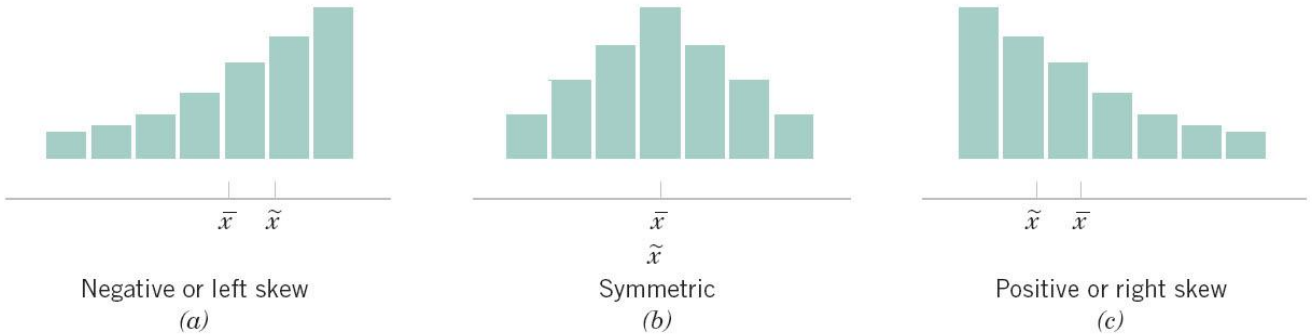


Figure 6-11
Histograms for symmetric and skewed distributions.

- Histogramas Sesgados:
  - Son asimétricos con una cola larga a uno de los lados
  - Sesgo a la izquierda: moda > mediana > media
  - Sesgo a la derecha: moda < mediana < media

Histograma con categorías- Diagrama de Pareto

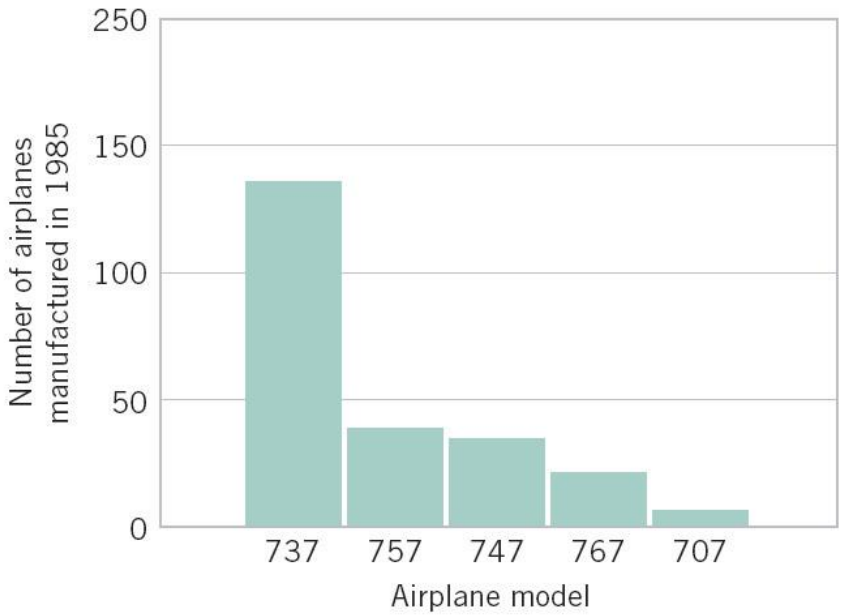


Figure 6-12
Airplane production in 1985. (*Source:* Boeing Company.)

- Un histograma por categorías, donde las categorías se ordenan de mayor frecuencia a menor frecuencia (izq-der) se denomina Diagrama de Pareto.

Gráficas de Caja

Características

- Describe varias características de un conjunto de datos:
  - Centro
  - Dispersión
  - Desviación de la simetría
  - Observaciones inusuales alejadas del centro de datos (Puntos Atípicos)

Significado de los gráficos de caja

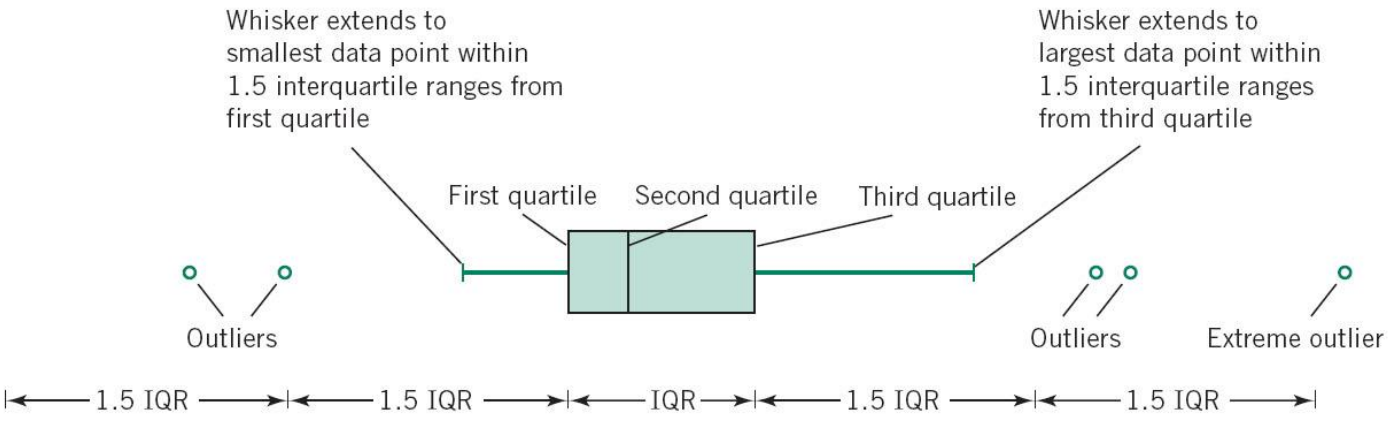
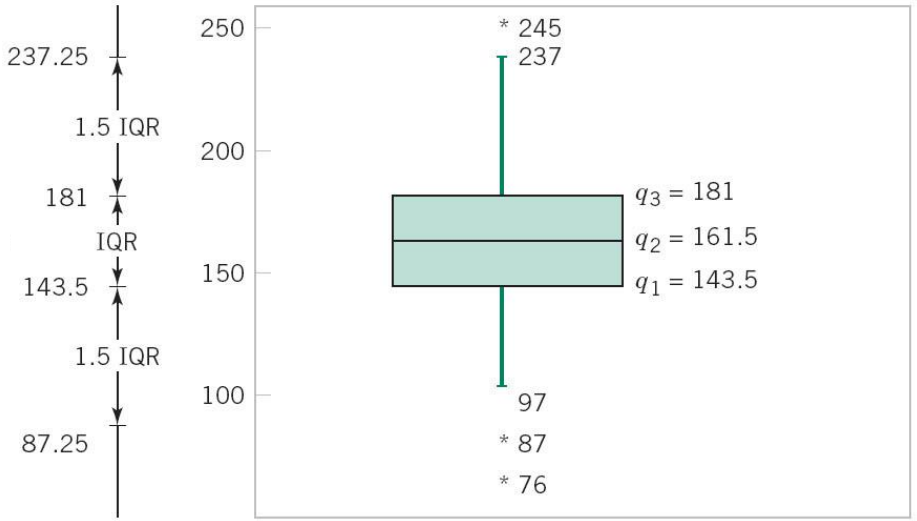


Figure 6-13 Description of a box plot.

- Whisker: Bigote
- Interquartile range (IQR): Rango Intercuartílico
- Outliers: Puntos atípicos
- Extreme outlier: Punto atípico extremo

Ejemplo



- La distribución de los datos es bastante simétrica alrededor del valor central.

Comparación de varios espacios muestrales

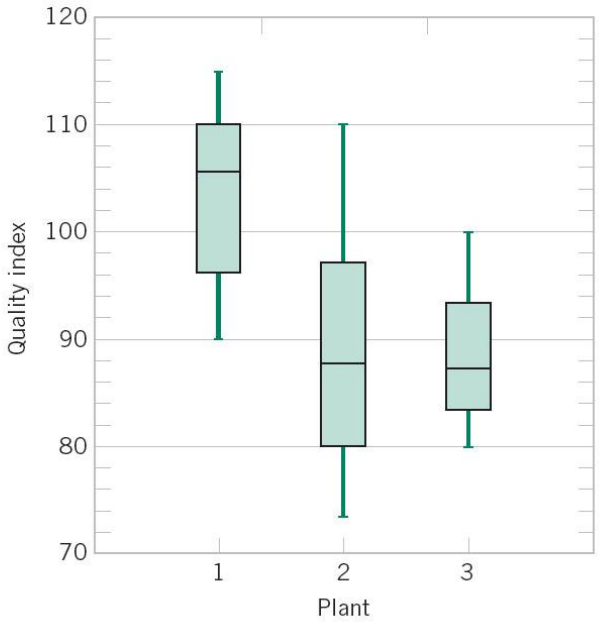


Figure 6-15
Comparative box plots of a quality index at three plants.

- En este ejemplo se comparan los índices de calidad de 3 plantas diferentes. Cada caja representa el espacio muestral de cada planta.
- Conclusiones:
  - Planta1 tiene mejor índice de calidad
  - Plantas 2 y 3 tienen índice de calidad similar
  - La planta 2 tiene una variabilidad muy grande

Gráficas de Series de Tiempo

Características

- Una serie de datos en que las observaciones se registran en el orden en que ocurren.
- Eje vertical: variable observada
- Eje horizontal: tiempo
- Permite observar tendencias, ciclos y otras características generales que no se observan en otros diagramas

Tendencia ascendente

Picos en el 2do trimestre

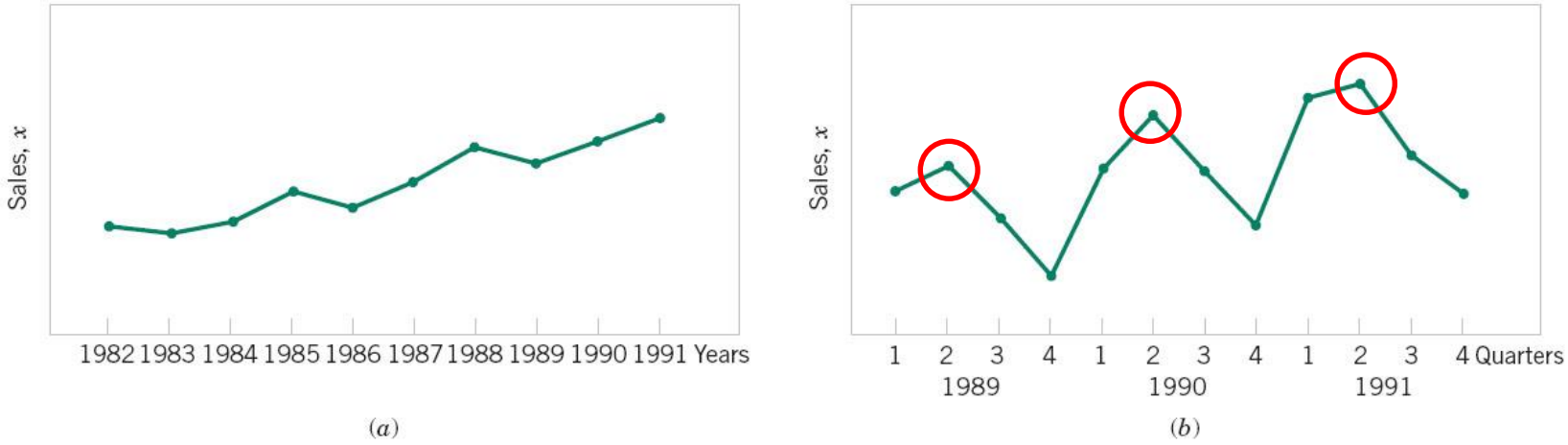


Figure 6-16 Company sales by year (*a*) and by quarter (*b*).

Diagrama punto-dígito (digidot)

Es una combinación del diagrama de tallo-hoja con una gráfica de series de tiempo.

Ejemplo:

La tabla del ejemplo de los bytes de una transmisión de un archivo.

Se agrega la información del orden en que ocurrieron.

Diagrama punto-dígito

Ejemplo:

Muestras de la concentración en un proceso químico.

Se observa cada hora.

Después de 20 horas, la concentración tiene valores por debajo de 85 (g/l)

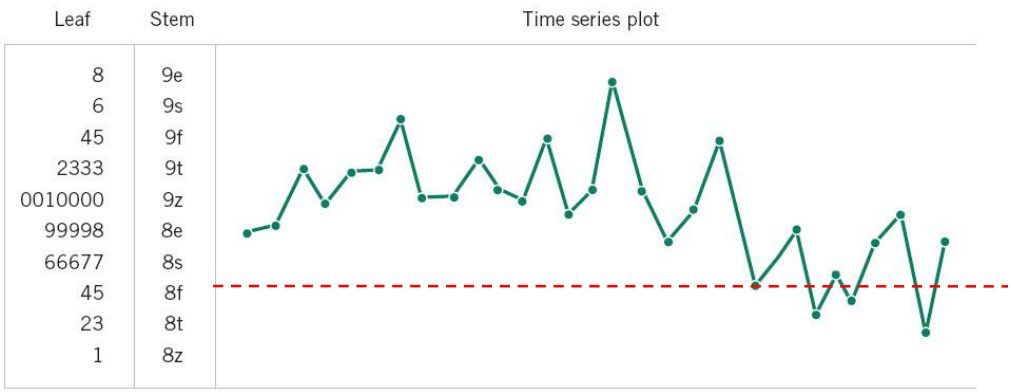


Figure 6-18 A digi-dot plot of chemical process concentration readings, observed hourly.