ISyE 6739 – Statistical Methods

Descriptive Statistics (Chapter 6)

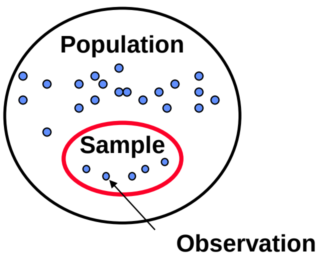
Instructor: Kamran Paynabar
H. Milton Stewart School of
Industrial and Systems Engineering
Georgia Tech

Kamran.paynabar@isye.gatech.edu
Office: Groseclose 436

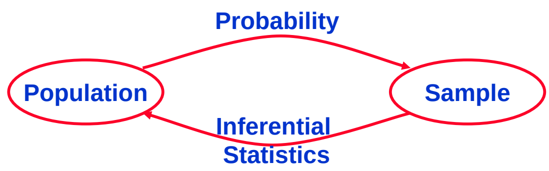
Population Vs. Sample

- **Population:** a finite well-defined group of ALL objects which, although possibly large, can be enumerated in theory
  (e.g. investigating ALL the bearings manufactured today).

- **Sample:** A sample is a SUBSET of a population
  (e.g. select 50 out of 1,000 bearings manufactured today).



Probability Vs. Statistics



Probability: given the information in the pail, what is in your hand?

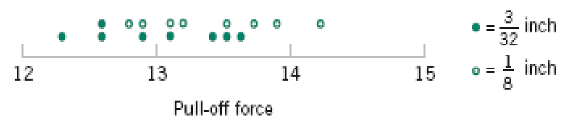
Statistics: given the information in your hand, what is in the pail?

List of Topics

- Descriptive Statistics (Ch.6)
  - Numerical Summaries:
    - Central Tendency
    - Variability
    - Position
  - Graphical Summaries:
    - Frequency Table and Histogram
    - Box plot
    - Time-series plot
    - stem-and-leaf diagram

Descriptive Vs. Inferential Statistics

- Descriptive Statistics:
  A set of statistical techniques used to organize, summarize, display, and describe important features of data



- Inferential (a.k.a. inductive) Statistics:
  A set of statistical methods that uses *sample* information to draw conclusion about the *population*

Descriptive Statistics

Numerical Summaries:
Central Tendency
Variability
Position

Numerical Summary of Data

Statistic: Any function of sampled observations is called a statistic

Central Tendency statistics

Mean $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

Median $\tilde{x}$

A value such that 50% of the data are at or above this value.

Mode $\hat{x}$

Observation with the highest frequency

Variability statistics

Range $R = x_{\max} - x_{\min}$

Variance $S^2 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$

Standard Deviation

$$S = \sqrt{\dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Sample Mean

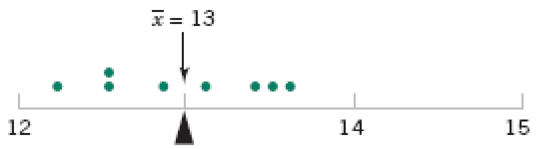
If the n observations in a sample are denoted by $x_1, x_2, \ldots, x_n$, the **sample mean** is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (6\text{-}1)$$

Let's consider the eight observations collected from the prototype engine connectors from Chapter 1. The eight observations are $x_1 = 12.6$, $x_2 = 12.9$, $x_3 = 13.4$, $x_4 = 12.3$, $x_5 = 13.6$, $x_6 = 13.5$, $x_7 = 12.6$, and $x_8 = 13.1$. The sample mean is

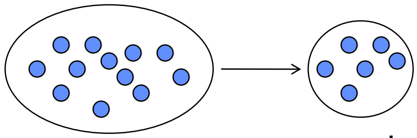
$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{8} x_i}{8} = \frac{12.6 + 12.9 + \cdots + 13.1}{8}$$

$$= \frac{104}{8} = 13.0 \text{ pounds}$$



Population Mean

N: population size



population

sample

n: sample size

Population mean $\mu = \dfrac{\sum_{i=1}^{N} x_i}{N}$

$\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$ **Sample mean**

The sample mean is a reasonable estimate of the population mean.

Sample Median

$\tilde{x}$ A value such that 50% of the data are at or above this value

How to calculate:

- Sort the data in ascending (or descending) order
- If n is an odd number, median is the $(n+1)/2^{th}$ number
- If n is an even number, median is the average of is the $n/2^{th}$ and $(n/2)+1^{th}$ numbers

Let's consider the eight observations collected from the prototype engine connectors from Chapter 1. The eight observations are $x_1 = 12.6$, $x_2 = 12.9$, $x_3 = 13.4$, $x_4 = 12.3$, $x_5 = 13.6$, $x_6 = 13.5$, $x_7 = 12.6$, and $x_8 = 13.1$.

Sample Mode

$\hat{x}$ Observation with the highest frequency

Let's consider the eight observations collected from the prototype engine connectors from Chapter 1. The eight observations are $x_1 = 12.6$, $x_2 = 12.9$, $x_3 = 13.4$, $x_4 = 12.3$, $x_5 = 13.6$, $x_6 = 13.5$, $x_7 = 12.6$, and $x_8 = 13.1$.

Sample Variance & Standard Deviation

If $x_1, x_2, \ldots, x_n$ is a sample of n observations, the **sample variance** is

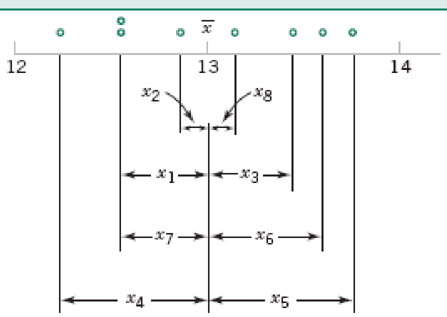
$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{6-3}$$

The **sample standard deviation,** s, is the positive square root of the sample variance.

How the sample variance measures variability through the deviations? $x_i - \bar{x}$

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

Easier to calculate



Example (pull-off force)

i	x_i		$x_i - \overline{x}$	$(x_i - \overline{x})^2$
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
	104.0	$\overline{x} = 13$	0.0	1.60

so the sample variance is

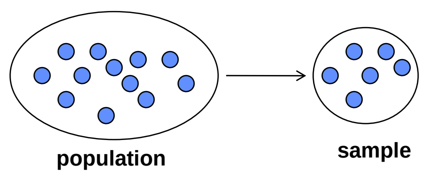
$$s^2 = \frac{1.60}{8-1} = \frac{1.60}{7} = 0.2286 \text{ (pounds)}^2$$

and the sample standard deviation is

$$s = \sqrt{0.2286} = 0.48 \text{ pounds}$$

Population Variance

N: population size



n: sample size

Population variance $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$ $s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$ **Sample variance**

The sample variance is a reasonable estimate of the population variance.

Sample Range

If the n observations in a sample are denoted by $x_1, x_2, \ldots, x_n$, the **sample range** is

$$r = \max(x_i) - \min(x_i) \qquad (6\text{-}6)$$

Let's consider the eight observations collected from the prototype engine connectors from Chapter 1. The eight observations are $x_1 = 12.6$, $x_2 = 12.9$, $x_3 = 13.4$, $x_4 = 12.3$, $x_5 = 13.6$, $x_6 = 13.5$, $x_7 = 12.6$, and $x_8 = 13.1$.

$$r = x_{\max} - x_{\min} = 13.6 - 12.3 = 1.3$$

Percentiles

To calculate i th $(1 < i < 99)$ percentile (P_i) :

- Sort the data in ascending order
- Calculate the rank as $r = (n+1)\times i/100$
- If r is integer, the i th percentile is the r^{th} sorted number
- If r is *non-integer*, the i th percentile is the average of is the floor$(r)^{th}$ and floor$(r)^{th}$+1 numbers

- P_{25} , P_{50} , and P_{75} are also known as first, second and third quartiles, respectively and denoted as Q_1 , Q_2 , and Q_3 .

Descriptive Statistics

Graphical Summaries:

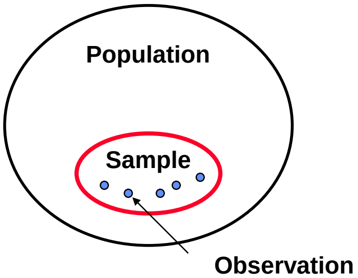
Frequency Table and Histogram
Box plot
Time-series plot
stem-and-leaf diagram

Frequency Table and Histogram

- **To construct a frequency table**
  1. **Find the range of the data**
     - **start the lower limit for the first bin just slightly below the smallest data value**
     - **b_0=<min(x), b_m=max(x),**
     - **$R = b_m - b_0$**
  2. **Divide this range into a suitable number of equal intervals**
     - **m=4 ~ 20, or $\sqrt{N}$ (N is the total number of observations)**
  3. **Count the frequency of each interval**
     - **if $b_{i-1} \leq x < b_i$**

Example: Forged Piston Rings for Engines

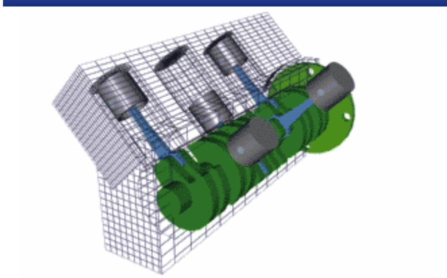
- **Population:**
  - **The inside diameter of forged piston rings(mm)**
  - **One sample that includes 125 observations were collected**



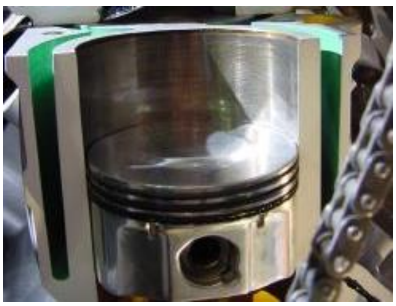
Forged Piston-Ring Inside Diameter (mm)

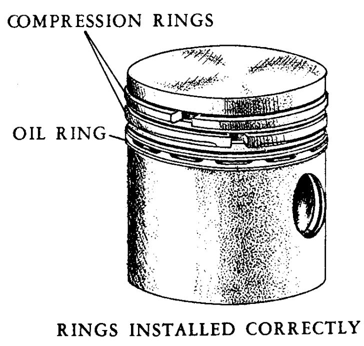
Observations				
74.030	74.002	74.019	73.992	74.008
73.995	73.992	74.001	74.011	74.004
73.988	74.024	74.021	74.005	74.002
74.002	73.996	73.993	74.015	74.009
73.992	74.007	74.015	73.989	74.014
74.009	73.994	73.997	73.985	73.993
73.995	74.006	73.994	74.000	74.005
73.985	74.003	73.993	74.015	73.988
74.008	73.995	74.009	74.005	74.004
73.998	74.000	73.990	74.007	73.995
73.994	73.998	73.994	73.995	73.990
74.004	74.000	74.007	74.000	73.996
73.983	74.002	73.998	73.997	74.012
74.006	73.967	73.994	74.000	73.984
74.012	74.014	73.998	73.999	74.007
74.000	73.984	74.005	73.998	73.996
73.994	74.012	73.986	74.005	74.007
74.006	74.010	74.018	74.003	74.000
73.984	74.002	74.003	74.005	73.997
74.000	74.010	74.013	74.020	74.003
73.988	74.001	74.009	74.005	73.996
74.004	73.999	73.990	74.006	74.009
74.010	73.989	73.990	74.009	74.014
74.015	74.008	73.993	74.000	74.010
73.982	73.984	73.995	74.017	74.013

Piston Rings



V - The cylinders are arranged in two banks set at an angle to one another





Frequency Distribution for Piston-Ring Diameter

- Data range: b_0= 73.965 <min(x); b_N =max(x)=74.030
- N=125; # of Bin m=13, Interval=(74.030-73.965)/13=0.005
- Count for each bin: $b_{i-1} \leq x < b_{i,}$

Table 2-3 Frequency Distribution for Piston-Ring Diameter

Ring Diameter, x (mm)	Tally	Frequency	Cumulative Frequency	Relative Frequency	Cumulative Relative Frequency
$73.965 \leq x < 73.970$	1	1	1	0.008	0.008
$73.970 \leq x < 73.975$		0	1	0.000	0.008
$73.975 \leq x < 73.980$		0	1	0.000	0.008
$73.980 \leq x < 73.985$	1111 111	8	9	0.064	0.072
$73.985 \leq x < 73.990$	1111 1111	10	19	0.080	0.152
$73.990 \leq x < 73.995$	1111 1111 1111 1111	19	38	0.152	0.304
$73.995 \leq x < 74.000$	1111 1111 1111 1111 111	23	61	0.184	0.488
$74.000 \leq x < 74.005$	1111 1111 1111 1111 11	22	83	0.176	0.664
$74.005 \leq x < 74.010$	1111 1111 1111 1111 11	22	105	0.176	0.840
$74.010 \leq x < 74.015$	1111 1111 111	13	118	0.104	0.944
$74.015 \leq x < 74.020$	1111	4	122	0.032	0.976
$74.020 \leq x < 74.025$	11	2	124	0.016	0.992
$74.025 \leq x \leq 74.030$	1	1	125	0.008	1.000
	Total	125		1.000	

Histogram for Piston-ring Diameter Data - A graphical display of the frequency table

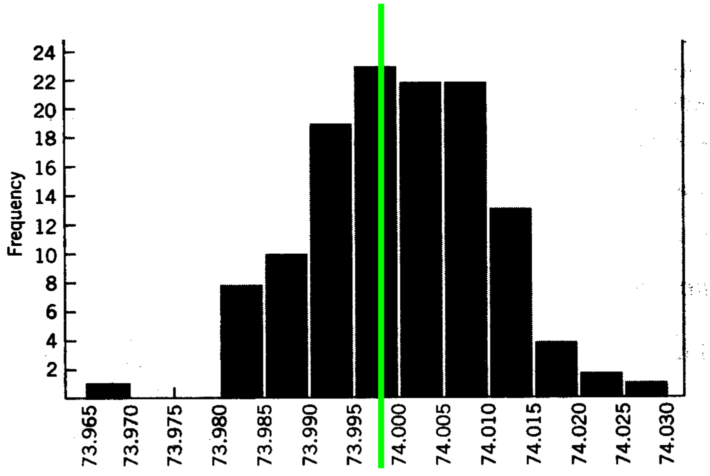
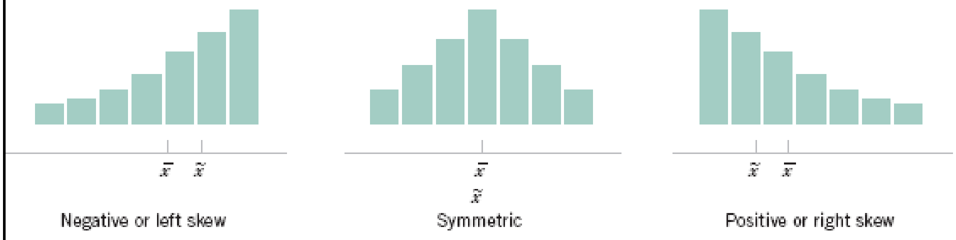


Figure 2-4 Histogram for piston-ring diameter data.

Interpretation based on Histogram

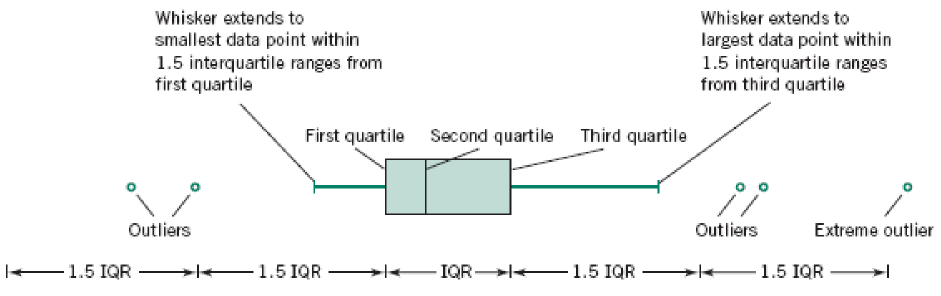
Three Properties of Sample Data

- **Shape:**
  - **roughly symmetric and unimodal**
- **The center tendency or location**
  - **the points tend to cluster near 74mm.**
- **Scatter or spread range**
  - **variability is relatively high** (min=73.967; max=74.030)

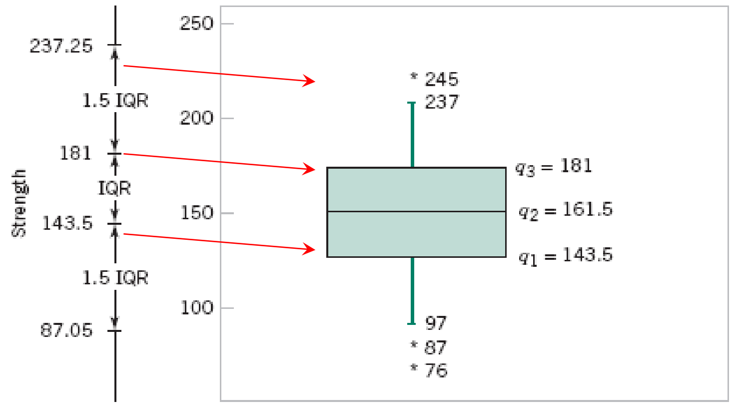


Box Plots

- The box plot is a graphical display that simultaneously describes several important features of a data set, such as center, spread, departure from symmetry, and identification of observations that lie unusually far from the bulk of the data (outliers).



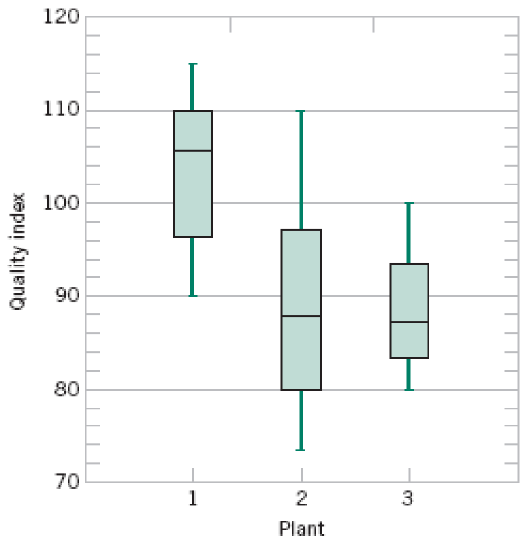
Example (Table 6-2)



Box plot for compressive strength data in Table 6-2.

Example

Comparative box plots of a quality index at three plants.



Time Series Plot

- A time series or time sequence is a data set in which the observations are recorded in the order in which they occur.
- A time series plot is a graph in which the vertical axis denotes the observed value of the variable (say x) and the horizontal axis denotes the time (which could be minutes, days, years, etc.).
- When measurements are plotted as a time series, we often see patterns like trends, cycles, or other broad features of the data

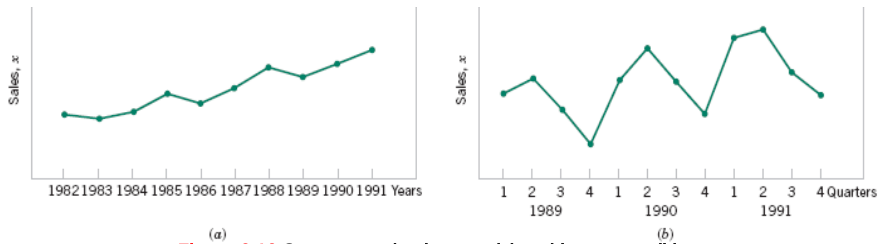


Figure 6-16 Company sales by year (a) and by quarter (b).

Example

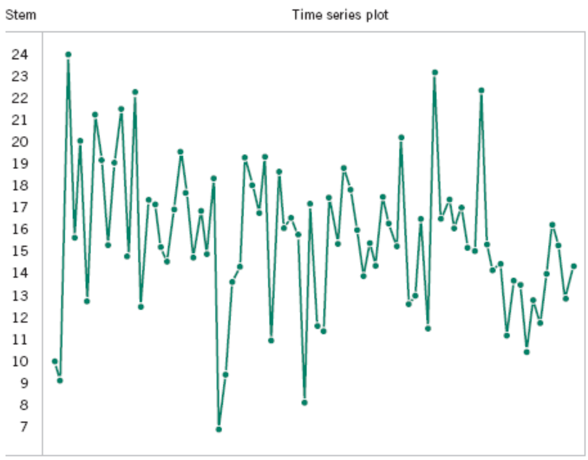


Figure 6-17 A digidot plot of the compressive strength data.

Stem-and-Leaf Diagrams

A **stem-and-leaf diagram** is a good way to obtain an informative visual display of a data set $x_1, x_2, \ldots, x_n$, where each number x_i consists of at least two digits. To construct a stem-and-leaf diagram, use the following steps.

Steps for Constructing a Stem-and-Leaf Diagram

(1) Divide each number x_i into two parts: a **stem**, consisting of one or more of the leading digits and a **leaf**, consisting of the remaining digit.

(2) List the stem values in a vertical column.

(3) Record the leaf for each observation beside its stem.

(4) Write the units for stems and leaves on the display.

Stem-and-Leaf Diagrams

Example 6-4

Table 6-2 Compressive Strength (in psi) of 80 Aluminum-Lithium Alloy Specimens

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

Stem-and-Leaf Diagrams

Figure 6-4 **Stem-and-leaf diagram for the compressive strength data in Table 6-2.**

Stem	Leaf	Frequency
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

Stem : Tens and hundreds digits (psi); Leaf: Ones digits (psi)

Stem-and-Leaf Diagrams

Example 6-5

Stem	Leaf
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

(a)

Stem	Leaf
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

(b)

Stem	Leaf
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	

(c)