ISyE 3770
Chapter 6: Descriptive Statistics

Instructor: Dan Li

School of Industrial and Systems Engineering
Georgia Tech

Spring 2021

1. Descriptive Statistics
   - Numerical Summaries
     - Sample Mean
     - Sample Variance
     - Quantiles
     - Sample Range
   - Graphical Techniques
     - Stem-and-Leaf Diagram
     - Histogram
     - Frequency Plots
     - Box Plot
     - Probability Plot
2. Introduction to R

Descriptive Statistics

- An important aspect of statistics for organizing and summarizing the data in ways that facilitate is interpretation and subsequent analysis.
  - Sample mean & variance
  - Shape of the data: symmetric, skewed to the right or to the left.
  - Spread of the data: range, long or short tails.
  - Outliers: a few extreme values or points that appear separate from the rest of the data.
  - Different modes of the data: unimodal (one concentration), bimodal (two concentrations), etc.
  - Gaps in the data: they could be because of unrecorded data or two different subpopulations.

Data Types

- **Categorical or nominal data** - when we observe the frequency within several categories.
  - Example: For the attendance at a scientific conference, 20% are female researchers and 80% are male researchers.
- **Numerical data** - may be integer, real or complex numbers (observations).
  - Example: The lifetime of the computer chips from a production line (real observations).

Table of Contents

1. Descriptive Statistics
   - Numerical Summaries
     - Sample Mean
     - Sample Variance
     - Quantiles
     - Sample Range
   - Graphical Techniques
     - Stem-and-Leaf Diagram
     - Histogram
     - Frequency Plots
     - Box Plot
     - Probability Plot
2. Introduction to R

Sample Mean

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

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

- For the N observations in a population denoted by $x_1, x_2, \ldots, x_N$, the **population mean** is given by:

$$\mu = \sum_{i=1}^{N} x_i f(x) = \frac{\sum_{i=1}^{N} x_i}{N}$$

Sample Variance

- If n observations in a random sample are denoted by $x_1, x_2, \ldots, x_n$, the **sample variance** is:

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

- For the N observations in a population denoted by $x_1, x_2, \ldots, x_N$, the **population variance** is given by:

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

Sample Variance

- Sample Variance can be re expressed as follows

$$\begin{aligned}
\mathbb{E}(S^2) &= \mathbb{E}\left[\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}\right] \\
&= \frac{1}{n-1}\mathbb{E}\left(\sum_{i=1}^{n}(X_i-\bar{X})^2\right) \\
&= \frac{1}{n-1}\mathbb{E}\left(\sum_{i=1}^{n}\left(X_i^2+\bar{X}^2-2\bar{X}X_i\right)\right) \\
&= \frac{1}{n-1}\mathbb{E}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right) = \frac{1}{n-1}\left[\sum_{i=1}^{n}\mathbb{E}\left(X_i^2\right)-n\mathbb{E}\left(\bar{X}^2\right)\right]
\end{aligned}$$

Sample Variance

- The population variance is calculated with *N*, the population size. Why isn't the sample variance calculated with *n*, the sample size?
- The true variance is based on data deviations from the true mean, μ.
- The sample calculation is based on the data deviations from $\bar{x}$, not μ.
  - *X-bar* is an estimator of μ; close but not the same. So the *n*-1 divisor is used to compensate for the error in the mean estimation.

Sample Variance

- The sample variance is calculated with the quantity n-1.
- This quantity is called the “degrees of freedom”.
- Origin of the term:
  - There are n deviations from *x-bar* in the sample.
  - The sum of the deviations is zero.
  - n-1 of the observations can be freely determined, but the n^{th} observation is fixed to maintain the zero sum.

Quantiles

- **Percentiles** partition the data into 100 equal-size segments.
- The three **quartiles** partition the data into four equally sized counts or segments.
  - First or lower quartile (25 percentile): 25% of the data is less than q_1.
  - Second quartile (50 percentile): 50% of the data is less than q_2, the median.
  - Third or upper quartile (75 percentile): 75% of the data is less than q_3.

Sample Range

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

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

- It is the largest observation in the sample minus the smallest observation.

Table of Contents

1. Descriptive Statistics
   - Numerical Summaries
     - Sample Mean
     - Sample Variance
     - Quantiles
     - Sample Range
   - Graphical Techniques
     - Stem-and-Leaf Diagram
     - Histogram
     - Frequency Plots
     - Box Plot
     - Probability Plot
2. Introduction to R

Graphical Techniques

- Charts: Bar, Pareto and Pie
  - They are very common for displaying **categorical data**. Each bar corresponds to a different category. The height of the bar is proportional to the frequency of that category.
  - The Pareto chart displays the bars in decreasing order
- Stem-and-leaf plot
  - A stem-and-leaf plot is a display of the values/numbers in the data. Consider a numerical data set $x_1, x_2, \ldots, x_n$ for which each x_i consists of at least two digits.

Stem-and-Leaf Diagram

- Stem & leaf diagrams are better for large sets.
- Steps to construct a stem-and-leaf diagram:
  1) Divide each number (x_i) into two parts: a **stem**, consisting 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 the stems and leaves on the display.

Example for Stem-and-Leaf Diagram

To illustrate the construction of a stem-and-leaf diagram, consider the alloy compressive strength data in Table 6-2.

Table 6-2 Compressive Strength (psi) of Aluminum-Lithium 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	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

Figure 6-4 Stem-and-leaf diagram for Table 6-2 data. Center is about 155 and most data is between 110 and 200. Leaves are unordered.

Histogram

- The histogram is a bar chart for numerical data, where each bar represents the frequency of the numerical data in the interval representing the bar. The length of all intervals is equal and it is called bandwidth.
  - Note: For a good presentation of the data using a histogram we have to choose the bandwidth carefully.

Histogram

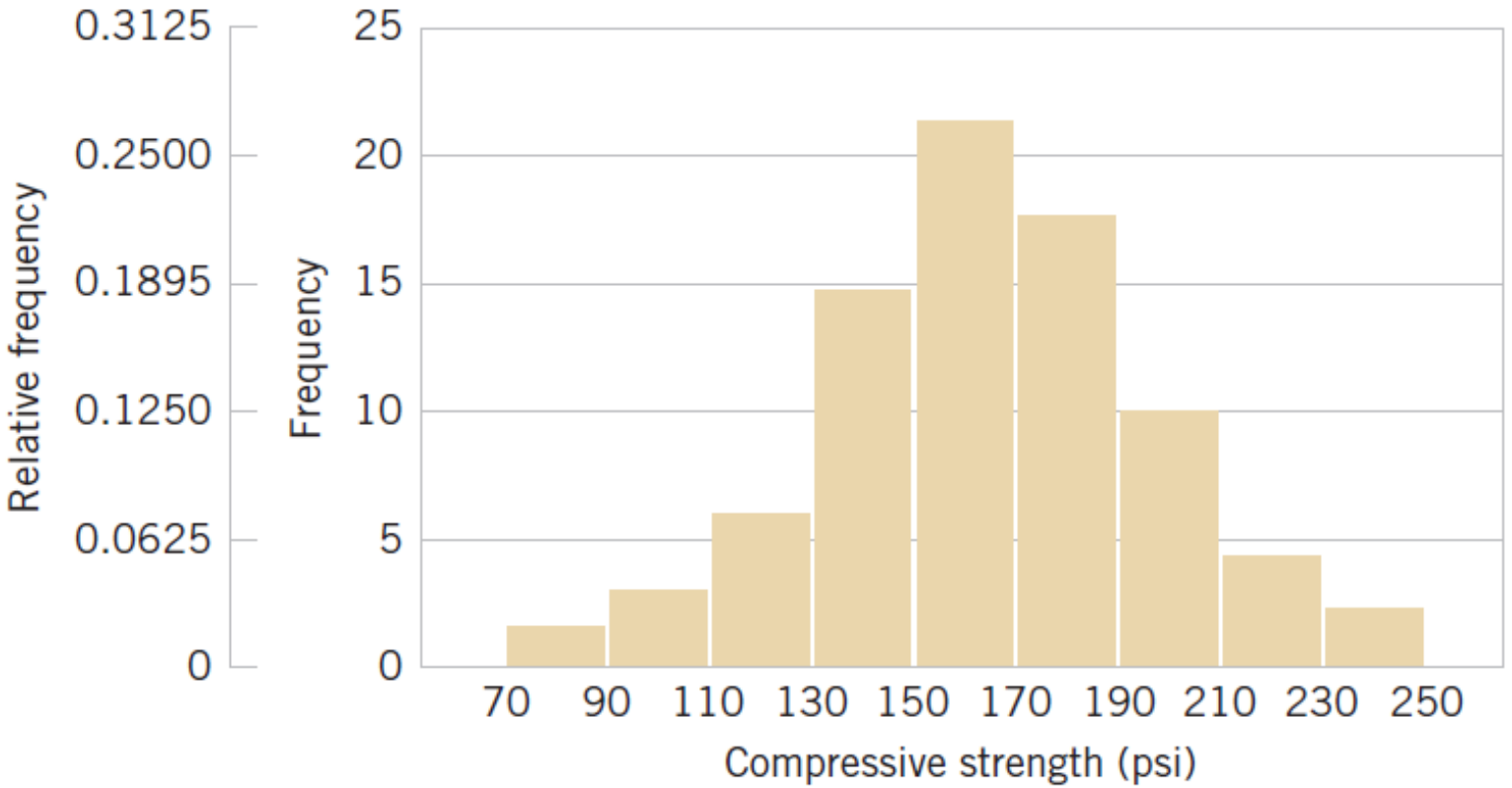


Figure 6-7 Histogram of compressive strength of 80 aluminum-lithium alloy specimens. Note these features – (1) horizontal scale bin boundaries & labels with units, (2) vertical scale measurements and labels, (3) histogram title at top or in legend.

Poor Choices in Drawing Histograms

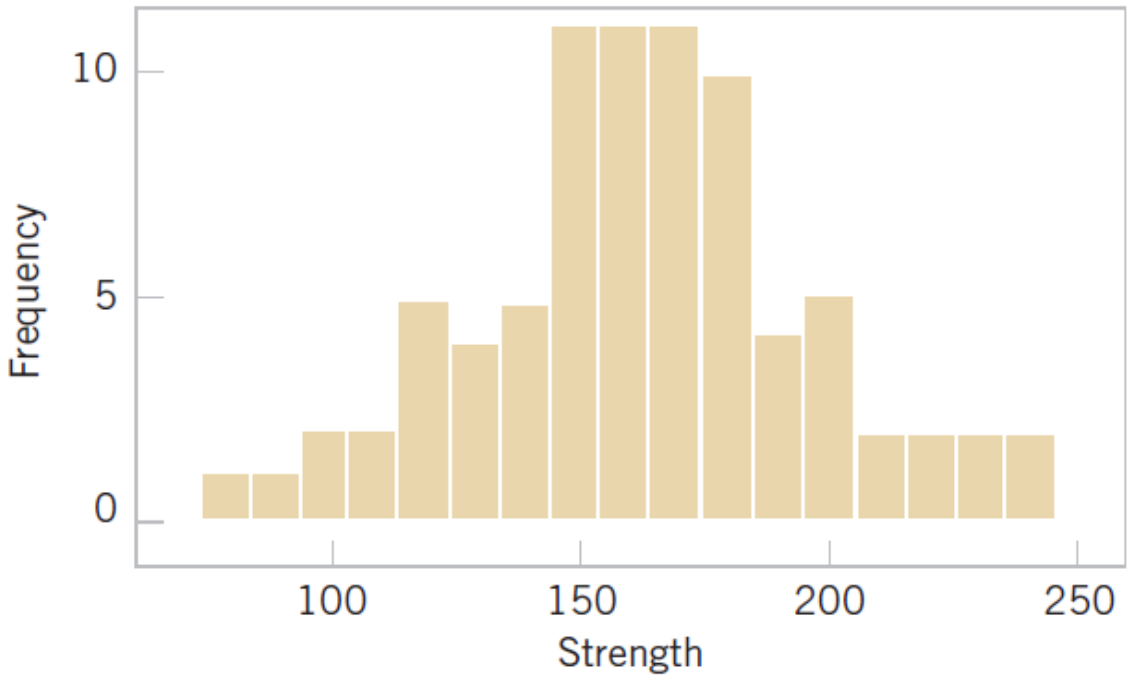


Figure 6-8 Histogram of compressive strength of 80 aluminum-lithium alloy specimens. Errors: too many bins (17) create jagged shape, horizontal scale not at class boundaries, horizontal axis label does not include units.

Frequency Distributions

- A frequency distribution is a compact summary of data, expressed as a table, graph, or function.
- The data is gathered into **bins or cells**, defined by **class intervals**.
- The **number of classes**, multiplied by the class interval, should exceed the range of the data. The square root of the sample size is a guide.
- The boundaries of the class intervals should be convenient values, as should the **class width**.

Frequency Distribution Table

Frequency Distribution for the data in Table 6-2

Considerations:

Range = 245 – 76 = 169

$\sqrt{80} = 8.9$

Decisions:

Number of classes = 9

Class width = 20

Range of classes = 20 * 9 = 180

Starting point = 70

Table 6-4 Frequency Distribution of Table 6-2 Data

Class	Frequency	Relative Frequency	Cumulative Relative Frequency
70 ≤ x < 90	2	0.0250	0.0250
90 ≤ x < 110	3	0.0375	0.0625
110 ≤ x < 130	6	0.0750	0.1375
130 ≤ x < 150	14	0.1750	0.3125
150 ≤ x < 170	22	0.2750	0.5875
170 ≤ x < 190	17	0.2125	0.8000
190 ≤ x < 210	10	0.1250	0.9250
210 ≤ x < 230	4	0.0500	0.9750
230 ≤ x < 250	2	0.0250	1.0000
	80	1.0000	

Cumulative Frequency Plot

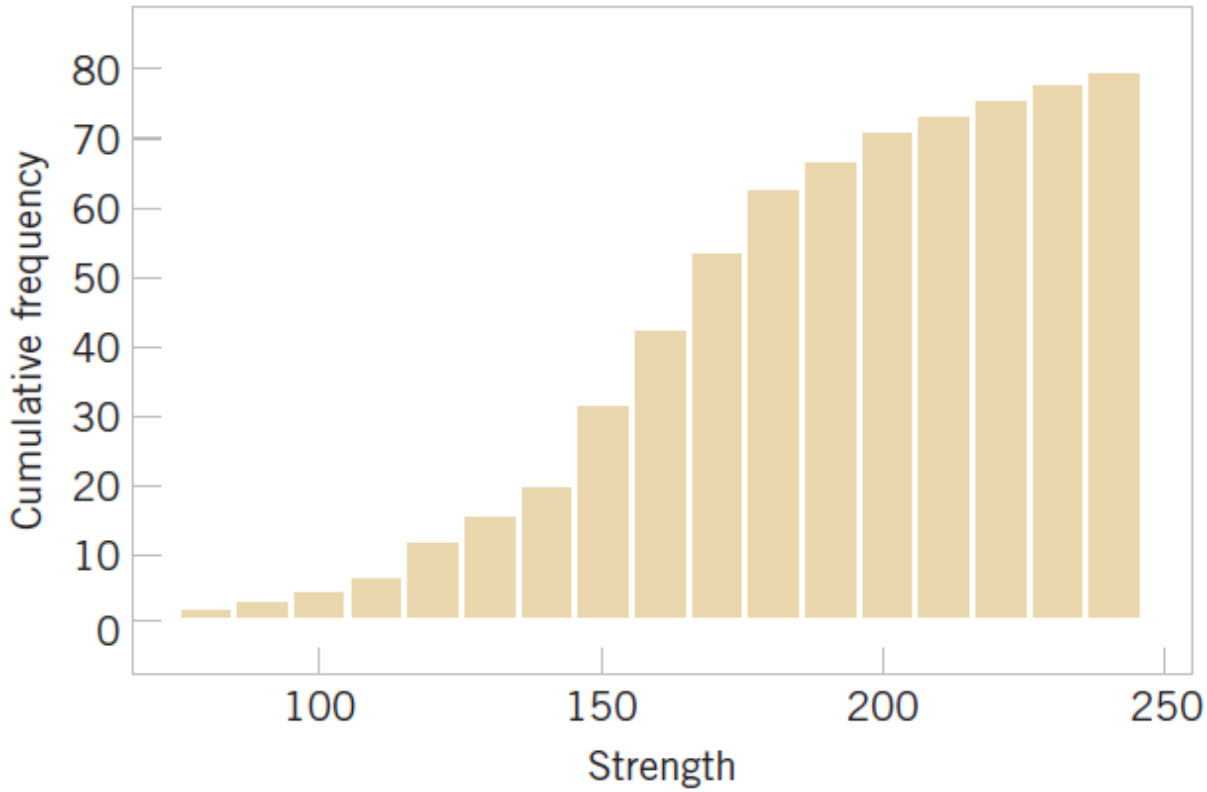
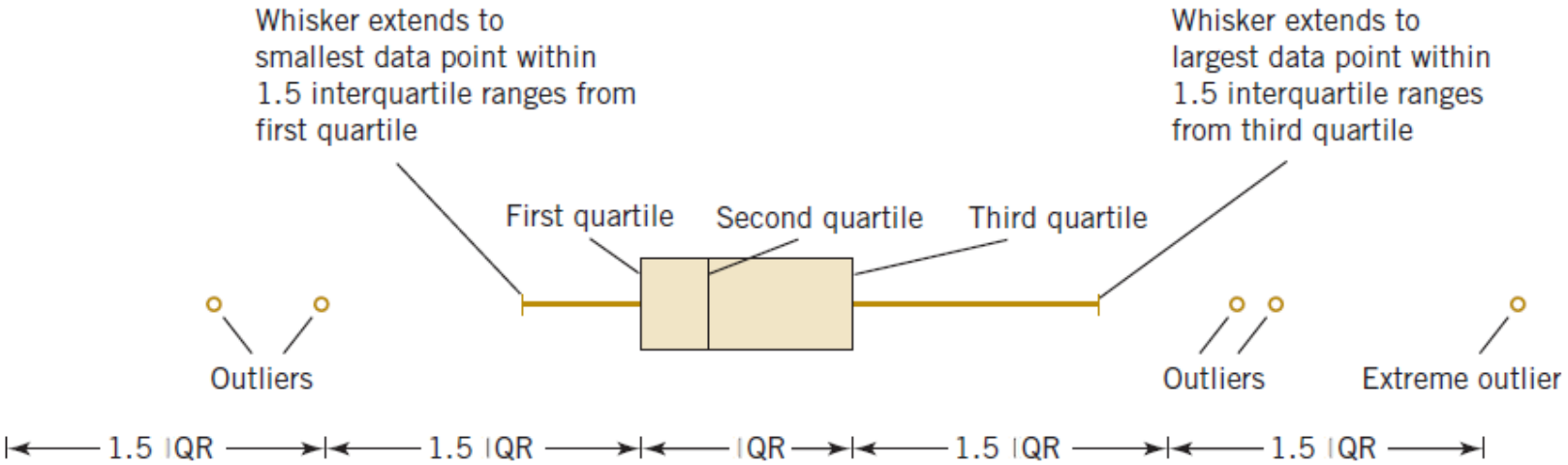


Figure 6-10 Cumulative histogram of compressive strength of 80 aluminum-lithium alloy specimens. Comment: Easy to see cumulative probabilities, hard to see distribution shape.

Box Plot or Box-and-Whisker Chart

- A box plot is a graphical display showing shape, outliers, center, and spread (SOCS).
- It displays the 5-number summary: *min*, q_1, *median*, q_3, and *max*.



Box Plots and Quartiles

- Sample Median is a measure of the central tendency and divides the data into two equal parts, half above and half below–what is the other?
    - If the number of observations is eve, the median is halfway between the two central values
- When an ordered set of data is divided into four equal parts, the division points are called **quartiles**.

Box Plots and Quartiles

- The sample median is also known as the second quartile, q_2.
- The first of lower quartile, q_1 is a value that has approximately 25% of observations below it and approximately 75% of the observations above.
- The third of upper quartile, q_3 is a value that has approximately 75% of observations below it and approximately 25% of the observations above.
- When the number of observations is even, q_1 and q_3 are calculated as the $(n+1)/4$ for q_1 and $3(n+1)/4$ for q_3 ordered observations and interpolate as needed.

Box Plot Example

- The "cold start ignition time" of an automobile engine is being investigated by a gasoline manufacturer. The following times (in seconds) were obtained for a test vehicle:
  1.75, 1.92, 2.62, 2.35, 3.09, 3.15, 2.53, 1.91
  - Calculate the sample mean, sample variance, and sample standard deviation.
  - Construct a box plot of the data.
- **Descriptive Statistics**

Variable	N	Mean	Median	TrMean	StDev	SE Mean
time	8	2.415	2.440	2.415	0.534	0.189

Box Plot Example

- Construct a box plot of the data.
  $1.75, 1.92, 2.62, 2.35, 3.09, 3.15, 2.53, 1.91$

Variable	Minimum	Maximum	Q1	Q3
time	1.750	3.150	1.912	2.973

Constructing a Probability Plot

To construct a probability plot:

- Sort the data observations in ascending order: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$.
- The observed value $x_{(j)}$ is plotted against the observed cumulative frequency $(j - 0.5)/n$.
- The paired numbers are plotted on the probability paper of the proposed distribution.

If the paired numbers form a straight line, then the hypothesized distribution adequately describes the data.

Example 6-7: Battery Life

The effective service life (X_j in minutes) of batteries used in a laptop are given in the table. We hypothesize that battery life is adequately modeled by a normal distribution. To this hypothesis, first arrange the observations in ascending order and calculate their cumulative frequencies and plot them.

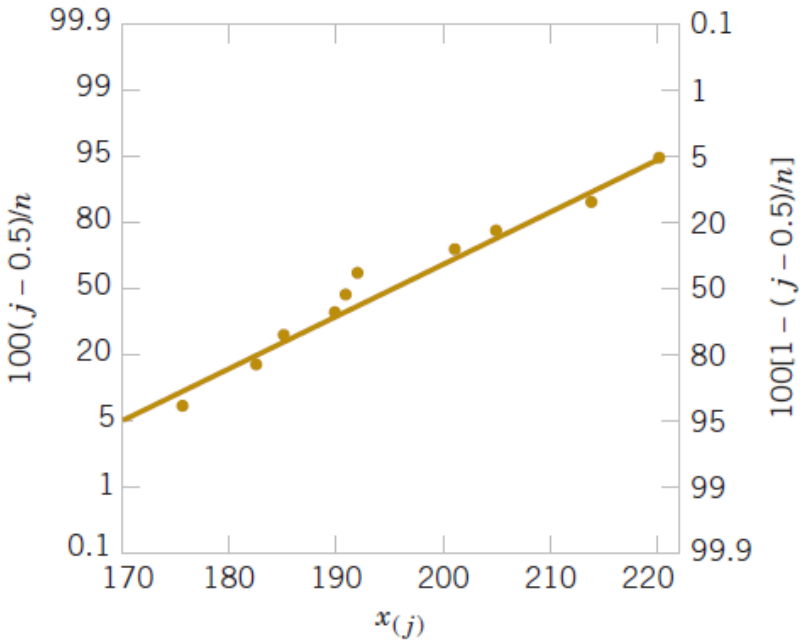


Figure 6-22 Normal probability plot for battery life.

Table 6-6 Calculations for Constructing a Normal Probability Plot

j	$x_{(j)}$	$(j-0.5)/10$	$100(j-0.5)/10$
1	176	0.05	5
2	183	0.15	15
3	185	0.25	25
4	190	0.35	35
5	191	0.45	45
6	192	0.55	55
7	201	0.65	65
8	205	0.75	75
9	214	0.85	85
10	220	0.95	95

Probability Plot on Standardized Normal Scores

A normal probability plot can be plotted on ordinary axes using z-values. The normal probability scale is not used.

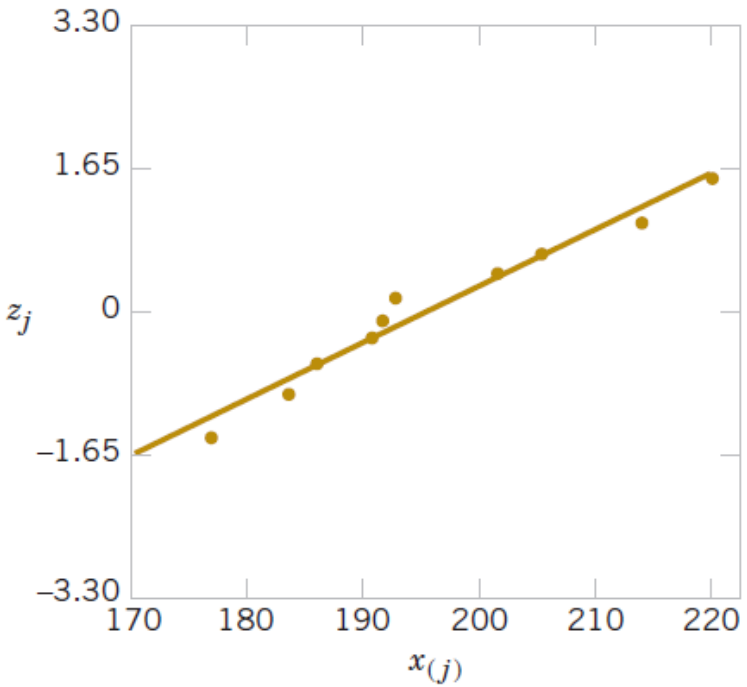


Figure 6-23 Normal Probability plot obtained from standardized normal scores. This is equivalent to Figure 6-19.

Table 6-6 Calculations for Constructing a Normal Probability Plot

j	$x_{(j)}$	$(j-0.5)/10$	z_j
1	176	0.05	-1.64
2	183	0.15	-1.04
3	185	0.25	-0.67
4	190	0.35	-0.39
5	191	0.45	-0.13
6	192	0.55	0.13
7	201	0.65	0.39
8	205	0.75	0.67
9	214	0.85	1.04
10	220	0.95	1.64

Probability Plot Variations

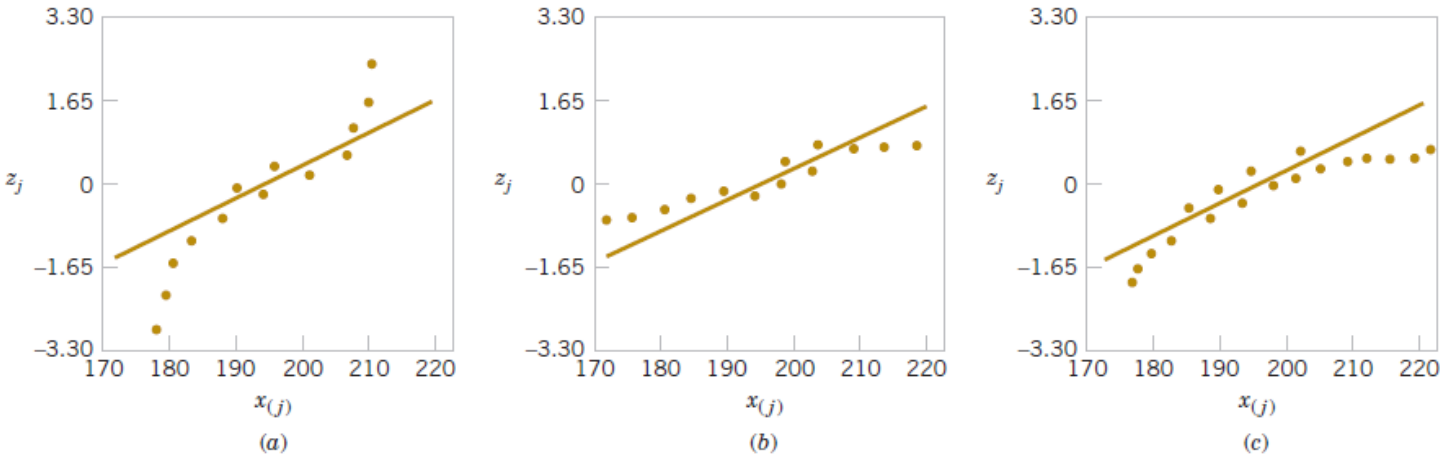


Figure 6-24 Normal probability plots indicating a non-normal distribution.
(a) Light tailed distribution
(b) Heavy tailed distribution
(c) Right skewed distribution

Table of Contents

1. Descriptive Statistics
   - Numerical Summaries
     - Sample Mean
     - Sample Variance
     - Quantiles
     - Sample Range
   - Graphical Techniques
     - Stem-and-Leaf Diagram
     - Histogram
     - Frequency Plots
     - Box Plot
     - Probability Plot
2. Introduction to R

What is R?

- R is a software for statistical computation and graphics
- It consists of a language plus a run-time environment with graphics, a debugger, access to certain system functions, and the ability to run programs stored in script files
- Free software
- OS: Windows, Unix, Linux, Mac
- Homepage: http://www.r-project.org

Installing Rstudio

- Go to https://rstudio.com/products/rstudio/download/
- Download and install the FREE version for your system (Mac/Windows/Ubuntu)
- Ask TAs for help if you have questions

R Example

- A Quick Overview of R
- The tragedy that befell the space shuttle *Challenger* and its astronauts in 1986 led to a number of studies to investigate the reasons for mission failure. Attention quickly focused on the behavior of the rocket engine's O-rings.
- Here is the data consisting of observations on x = O-ring temperature (F) for each test) O-ring or actual launch of the shuttle rocket engine:
- 84 49 61 40 83 67 45 66 70 69 80 58 68 60 67 72 73 70 57 63 70 78 52 67 53 67 75 61 70 81 76 79 75 76 58 31

R Example

- Construct a stem-and-leaf display of the data. What appears to be a representative temperature value?
- Construct a histogram.
- Would you describe any observation as being far from the rest of the data (an outlier)?
- Construct a boxplot based on the numerical summaries of these data.

R Example

- Type the numbers in a txt file and name is shuttle.txt
- Below is the some example code
  - Note that # implies a comment

```
# Read the data in a vector from the file `shuttle.txt'':

temp = scan(``shuttle.txt'')

# Construct an histogram of the data:

hist(temp,nclass=10,main =``Histogram of the O-ring temperature'')

# Construct a stem-and-leaf plot

stem(temp)
```

R Example

```
# Construct a boxplot
boxplot(temp)
# Obtain numerical summaries:
summary(temp)
# Obtain the sample variance of the data
var(temp)
```

Data With R

- Objects: vector, factor, array, matrix, data.frame, ts, list
- Mode (numerical, character, complex, and logical); Length
- Read data stored in text (ASCII) files
  read.table(), scan(), and *read.fwf()*
- Saving data
  write(x, file="data.txt"), write.table() write in a file a *data.frame*
- Generating data

Data With R

- Generating data

```
> x <- 1:30
> x <-  seq(1,4,0.5)
> c(1,1.5,2,2.5,3,3.5,4)
> rep(1,5)
> gl(2,3,10)        (generate factor levels)
> y <- matrix(c(1,2,3,1,2,3),3,2)

# 3 by 2 Matrix  # Filled in column by column
> y[1,]     # is the first row
> y[,2]     # is the second column
> y[3,1]   # is the element on row 3 and column 1
> apply(y,2,mean) #gives us the column means
> apply(y,1,mean) #gives us the row means
```

expand.grid() creates a *data.frame*

Another Example with R

- **The following data are the temperatures of effluent at discharge from a sewage treatment facility on consecutive days:**

  43 47 51 48 52 50 46 49
  45 52 46 51 44 49 46 51
  49 45 44 50 48 50 49 50

  (a) Calculate the sample mean, sample median, sample variance, and sample standard deviation.

  (b) Construct the histogram, box plot and normal quantile-quantile (QQ) plot (or normal probability plot) of the data set.

  (c) Do you think the data may follow a normal dist?

R Code for this Data Set

```
dat1 <- c(43, 47, 51, 48, 52, 50, 46, 49,
    45, 52, 46, 51, 44, 49, 46, 51,
    49, 45, 44, 50, 48, 50, 49, 50);

mean(dat1);          median(dat1);
var(dat1);           sqrt(var(dat1));

(48.125   49   7.244565   2.691573)

stem(dat1);    stem(log(dat1));
hist(dat1);
boxplot(dat1);
qqnorm(dat1);
```

Stem-and-Leaf Plot

> stem(dat1)

The decimal point is at the |

42 | 0

44 | 0000

46 | 0000

48 | 000000

50 | 0000000

52 | 00

> stem(log(dat1))

The decimal point is 1 digit(s) to the left of the |

37 | 688

38 | 11333

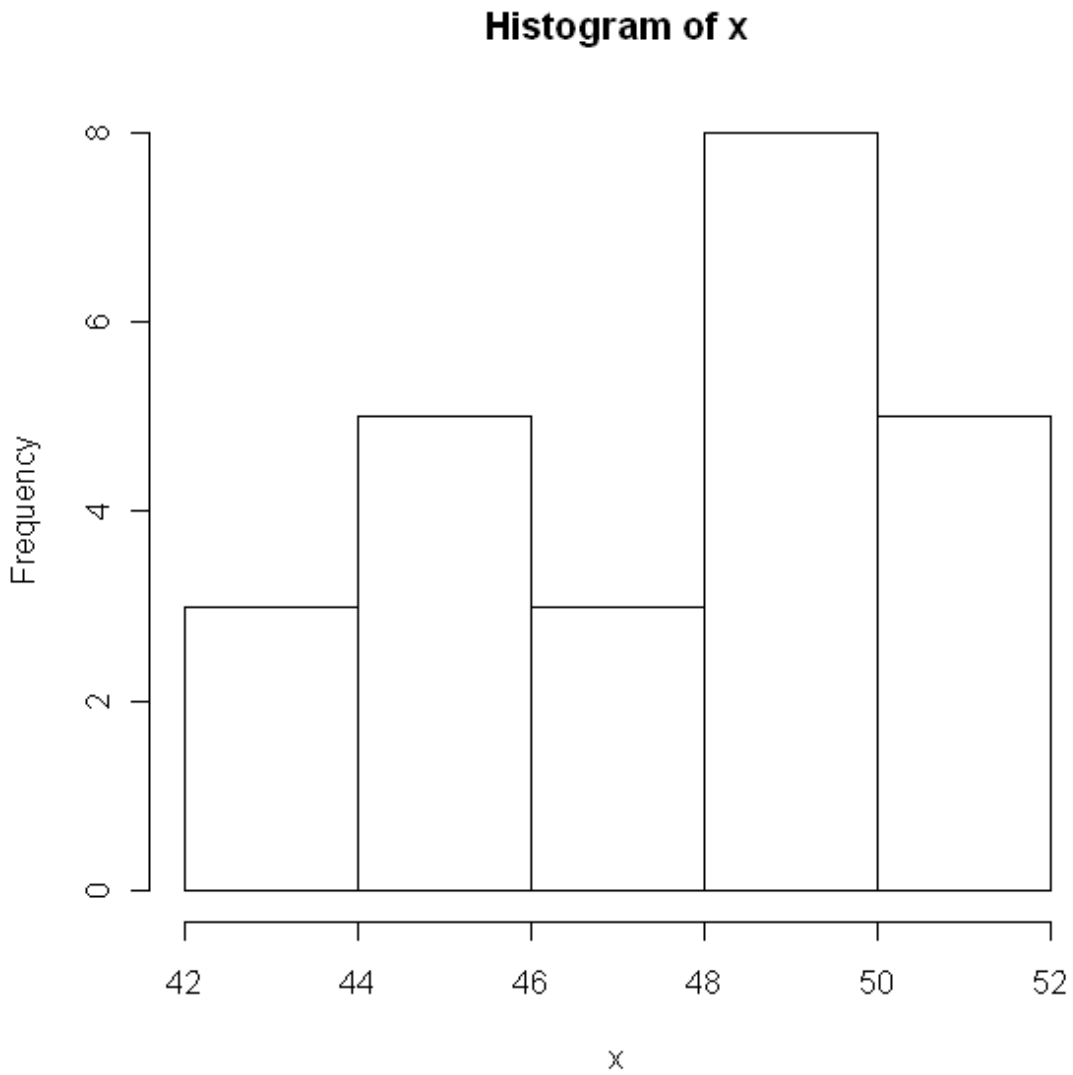
38 | 5779999

39 | 1111333

39 | 55

```
> log(dat1)
 [ 1] 3.761200 3.850148 3.931826 3.871201 3.951244 3.912023 3.828641 3.891820
 [ 9] 3.806662 3.951244 3.828641 3.931826 3.784190 3.891820 3.828641 3.931826
[17] 3.891820 3.806662 3.784190 3.912023 3.871201 3.912023 3.891820 3.912023
```

Histogram



Box Plot Histogram