MANE 3332.04

Lecture 17, March 31

attendance
1 - E

Agenda

- Midterm not graded: still contacting students who missed exam
- Continue working on Technical Report One Assignment
- Chapter Six
- Attendance
- Questions?

Grades

25% mid-term

0 mid-term $\rightarrow$ $-100(.25) = \underline{-25}$

$\downarrow$

could make a $\underline{75}$

Schedule

Monday Lecture	Wednesday Lecture
3/31: Chapter 6	4/2: Chapter 5
4/7: Chapter 7 & 8	4/9: Chapter 8, Case 1
4/14: Chapter 8: Case 2	4/16: Chapter 8: Case 3
4/21: Chapter 9, case 1	4/23: Chapter 9, Case 2
4/28: Chpater 9, Case 3	4/30: Chapter 11
5/5: Chapter 11	5/7: Review

12 classroom sessions plus Final Exam

Handouts

- Chapter 6 Slides
- Chapter 6 Slides marked

Data Analysis

1) location of Data → mean, median, mode
2) Variability/spread → variance/standard dev.
3) Shape of data

Numerical Summaries

- Called Descriptive Statistics in Chapter 6
  - Descriptive statistics help us understand the location or central tendency of data and the scatter or variability in data
  - Included in all statistical software packages, R does a good job calculating descriptive statistics

Central Tendency

- Ostle, et. al. (1996) define central tendency as “the tendency of sample data to cluster about a particular numerical value”
- Population mean

Greek letter → μ

N - population size

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

hat notation

- Sample mean

n is now lower-case

$\bar{x}$ →

$$\bar{x} = \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$\hat{\mu}$ - est

estimator of μ

- Sample median - middle value → $\tilde{x}$
- Sample mode - most commonly occuring number(s)

$$\text{Range} = X_{max} - X_{min}$$

Measures of Variability

- There are several statistics that measure the variability or spread present in data
- Population variance

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

Calculator: σ_N or σ_n

- Sample variance

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

Calculator: σ_{n-1}

- Shortcut (Computational) Formula

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

- Standard deviation is often used because it is measured in the original units

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

all columns

midterm has 3 columns

R Function Su

- R code

  summary

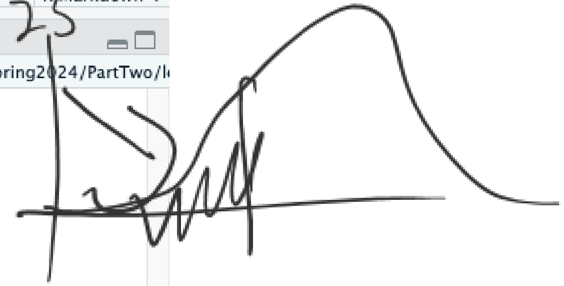
- Output is fi

```
26 ```{r}
27 summary(midterm)
28 ```
```

```
> knitr::opts_chunk$set(echo = TRUE)
> library(readxl)
> midterm <- read_excel("/Volumes/NO NAME/midterm.xlsx")
> View(midterm)
> summary(midterm)
 Participation      QuizAverage      MidtermExam
 Min.   :  2.941   Min.   :  0.00   Min.   :28.00
 1st Qu.: 67.500   1st Qu.: 60.00   1st Qu.:59.75
 Median : 87.941   Median : 80.00   Median :65.00
 Mean   : 77.096   Mean   : 74.42   Mean   :66.07
 3rd Qu.: 95.147   3rd Qu.: 92.00   3rd Qu.:74.75
 Max.   :100.000   Max.   :100.00   Max.   :92.00
                                    NA's   :5
>
```

Q1 →

25



$x = Q_1$

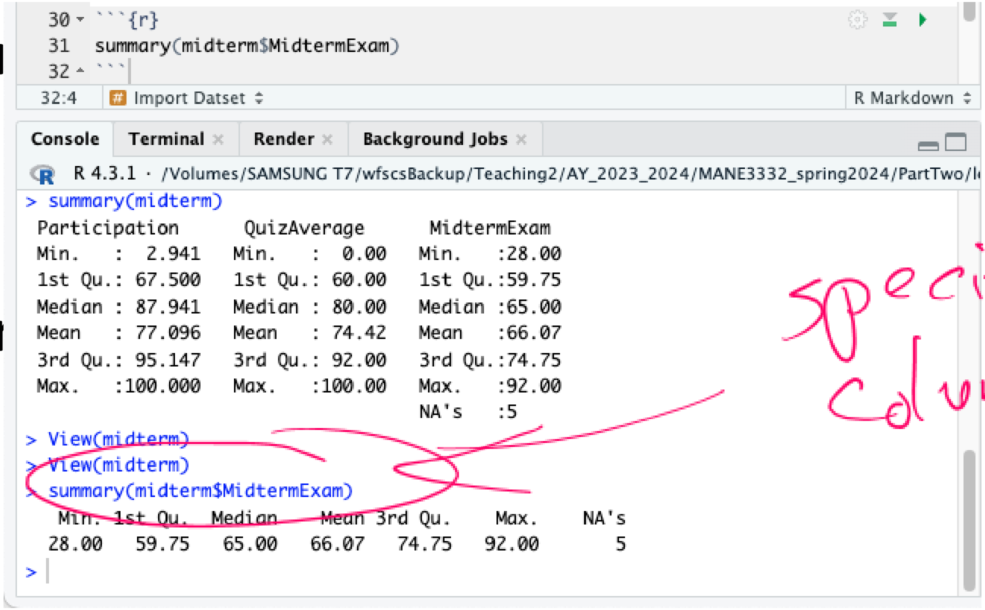
Descriptive Statistics

R Function Su

- R code

  `summary`

- Output is fr



specific
column

Descriptive Statistics

R Funct

- Sum
- Desc
- Desc
- R Co

  lib

  des

imported library

```
34 ```{r}
35 library(psych)
36 describe(midterm)
37
```

Description: df [3 × 13]

mean absolute deviation

forecasting

	vars	n	mean ($\bar{x}$)	sd (s)	median ($\tilde{x}$)	trimmed	mad
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Participation	1	33	77.10	25.65	87.94	81.35	16.13
QuizAverage	2	33	74.42	23.49	80.00	77.48	23.72
MidtermExam	3	28	66.07	13.73	65.00	66.62	13.34

3 rows | 1-8 of 13 columns

37:4 Import Datset R Markdown

- Psych package output from Spring 2024

Describe() Output

$\sigma^2 \sim (x-\mu)^2 \to 2^{nd}$ order

Description: df [3 × 13]

$\tilde{x}$

median	trimmed	mad	min	m... (Max)	range	skew (3rd order)	kurtosis (4th order)	se
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
87.94	81.35	16.13	2.94	100	97.06	-1.31	0.79	4.47
80.00	77.48	23.72	0.00	100	100.00	-1.24	1.28	4.09
65.00	66.62	13.34	28.00	92	64.00	-0.46	0.39	2.59

3 rows | 6-14 of 13 columns

Describe Output

Standard error
Chapters 8 & 9
$\frac{s}{\sqrt{n}}$

Calculating Quantiles

$$x_p = x_{(i)} + d\left[x_{(i+1)} - x_{(i)}\right]$$

$$p(n+1) = i + d$$

2.3.2 Sample Quantiles

In Example 2.8, we consider an ogive for the plated bracket data. The point (1.55, 0.567) is on that ogive, so we estimate that 56.7% of the sampled population of brackets weighed at most 1.55 ounces. Weights associated with other percentages can also be estimated by locating the appropriate point on the ogive. In general, if the point (x, p) is on the ogive, we can use x as an estimate of the weight with $100p\%$ of the population values at or below it. This estimate, called the 100pth *sample quantile*, is denoted x_p.

If two persons (or computer programs) use different groupings to obtain an ogive, the resulting quantiles will differ. To remedy this deficiency, an algebraic procedure is required.

THE 100pth SAMPLE QUANTILE

Several definitions of sample quantiles are used. We use the one that agrees with the default values output by the UNIVARIATE procedure in SAS®. Also, the definition used here is consistent with our definition of the sample median.

Suppose a sample of size n is obtained from some population associated with a continuous variable. For $0 < p < 1$, let $p(n+1) = i + d$, with i the integer part of $p(n+1)$ and $0 \le d < 1$ the decimal part. If $1 \le i < n$, and $d = 0$, the 100pth sample quantile is $x_{(i)}$. If $1 \le i < n$ and $0 < d < 1$, interpolate linearly between $x_{(i)}$ and $x_{(i+1)}$. In either case, the 100pth *sample quantile* is

$$x_p = x_{(i)} + d\left[x_{(i+1)} - x_{(i)}\right] \tag{2.4}$$

when $1 \le i < n$. If $i = 0$ or n, the 100pth sample quantile does not exist. If $100p$ is an integer, the corresponding quantile is called a *percentile*.

EXAMPLE 2.18

Suppose we want to find the 43rd percentile of the sample of plated weights in Table 2.1. Since there are $n = 75$ observations in the sample and $p = 0.43$, we find $p(n+1) = (0.43)(75+1) = 32.68$. Letting $i = 32$ and $d = 0.68$, we use Equation (2.4) to obtain $x_{0.43} = x_{(32)} + (0.68)(x_{(33)} - x_{(32)})$. The 32nd ordered value in Figure 2.1(b) is $x_{(32)} = 1.50$ and the 33rd ordered value is $x_{(33)} = 1.51$. Thus, the 43rd percentile for these data is $x_{0.43} = 1.50 + (0.68)(1.51 - 1.50) = 1.5068 \approx 1.507$. Using this as a point estimate of the population percentile, we can say that approximately 43% of the plated brackets produced on the day the data were collected had weights of 1.507 ounces or less. ■

The Sample Median Is a Percentile

Suppose we want to find the 50th percentile and the data set contains n values. When n is even, $(0.50)(n+1) = (n/2) + (0.50)$, with $n/2$ a positive integer. Using Equation (2.4) with $i = n/2$ and $d = 0.50$, $x_{0.50} = x_{(i)} + (0.50)[x_{(i+1)} - x_{(i)}] = [x_{(i)} + x_{(i+1)}]/2$. When n is odd, $(0.50)(n+1) = (n+1)/2$, with $(n+1)/2$ a positive integer. Using Equation (2.4) with $i = (n+1)/2$ and $d = 0$, we find $x_{0.50} = x_{(i)}$. But, this is precisely how the sample median was defined. Thus, $\tilde{x} = x_{0.50}$.

SAMPLE QUARTILES

The percentiles $x_{0.25}$, $x_{0.50}$, and $x_{0.75}$ are known as the *first*, *second*, and *third sample quartiles*, respectively. These quantities are often denoted q_1, q_2, and q_3.

EXAMPLE 2.19

Consider the plated bracket weights in Table 2.1. Using the ordered stem-and-leaf display presented in Figure 2.1(b), we find the following.

(a) *First Quartile:* Since $(0.25)(75+1) = 19$, $q_1 = x_{0.25} = x_{(19)} = 1.46$.

(b) *Second Quartile (Median):* Since $(0.50)(75+1) = 38$, $q_2 = \tilde{x} = x_{0.50} = x_{(38)} = 1.53$.

reference for calculating quantiles

1) Sort data

2) find $p = .5$ (median)

$x_1 = 6, x_2 = 4, \dots, x_8 = 6$ → order in data

$x_{(1)} = 3, x_{(2)} = 4, x_{(3)} = 4, \dots, x_{(8)} = 7$

Qu

8 observations from binomial distribution with

[redacted]

sorted

6, 4, 5, 7, 3, 5, 4, 6 → 3, 4, 4, 5, 5, 6, 6, 7

$x_{(4)}$ $x_{(5)}$

Quantile Example

$p = .5,\ n = 8$

$p(n+1) = i + d$

$i \rightarrow$ integer

$d \rightarrow$ decimal remainder

$.5(8+1) = \underline{4.5}$

$\rightarrow i = 4,\ d = .5$

$X_{.5} = X_{(4)} + .5\left[X_{(5)} - X_{(4)}\right]$

$= 5 + .5\left[5 - 5\right]$

$= \underline{5}$

What am I doing? linear interpolation

(2) add to note card for final

Exploratory Data (Graphical) Analysis

- Exploratory data analysis (EDA) is the use of graphical procedures to analyze data.
- John Tukey was a pioneer in this field and invented several of the procedures
- Tools include stem-and-leaf diagrams, box plots, time series plots and digidot plots , dot plots

look at graph and get data values

Stem and Leaf Diagram

- Excellent tool that maintains data integrity
- The stem is the leading digit or digits
- The leaf is the remaining digit
- Make sure to include units
- R Code

```
stem(midterm$MidtermExam)
```

Stem a

- R ou

units

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

2 | 8
3 |
4 | 4
5 | 11566
6 | 1333444667
7 | 2247
8 | 00147
9 | 2
```

stem leaf

stem - 7

leaf - 4

units → 7(10)

value $7(10)+4 = 74$

Stem and Leaf Plot of Midterm Exam Scores

Histogram

- A histogram is a barchart displaying the frequency distribution information
- There are three types of histograms: frequency, relative frequency and cumulative relative frequency

counts

% of observations

total %

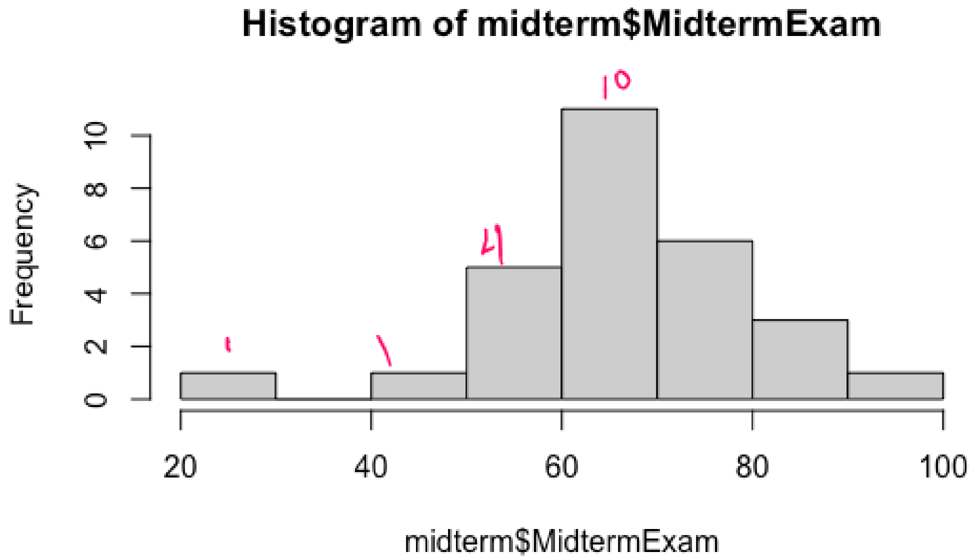
- R code

```
hist(midterm$MidtermExam)
```

→ needs lots of data to get shape

Histogram E

- R output



Histogram of Midterm Exam Scores

estimators to be robust → not overly influenced by extreme values

Boxplot

Inter Quartile Range

$IQR = Q_3 - Q_1$

median

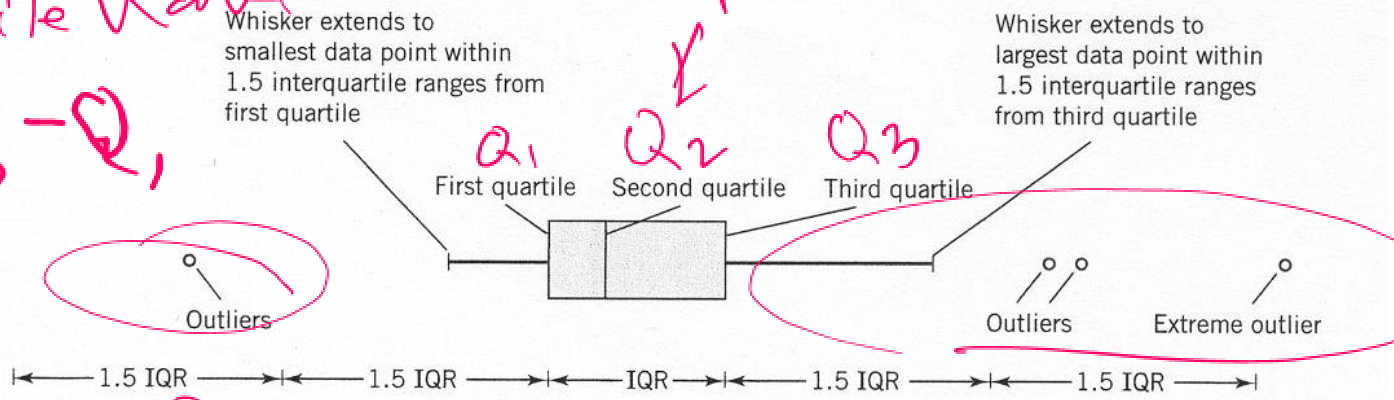


Figure 2-11
Description of a box plot.

Box Plot 1 → Famous ~~Station~~ Statistician

Box plot with explanation

Box & whiskers

Q1: which plant has highest Quality index? plant 1 has highest QI ✓

Box Plot 2

Q2: what is difference in QI between plants 2 & 3? by location

same median, so no difference

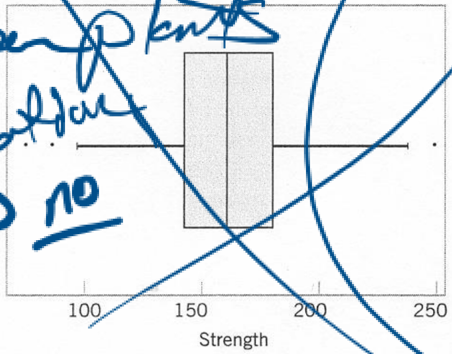


Figure 2-12 Box plot for compressive strength data in Table 2-2.

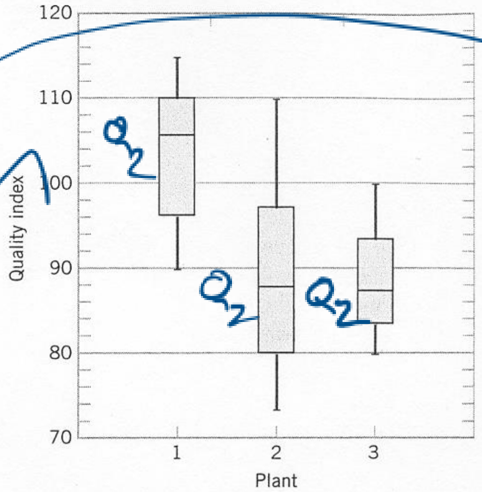


Figure 2-13 Comparative box plots of a quality index at three plants.

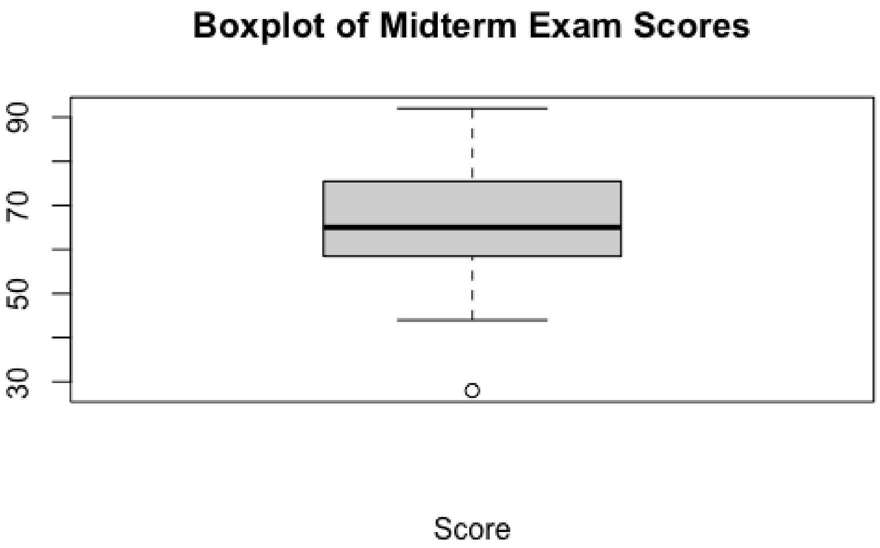
examples of boxplots

Q3: which plant has most variability?

Q2 → rectangle is largest & whiskers are longest

Box Plot 3

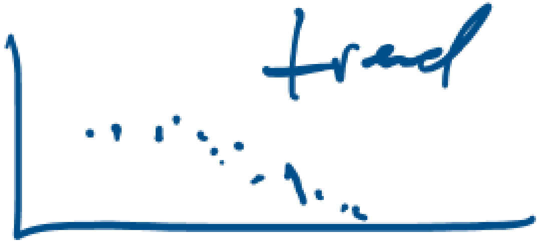
- R code for [illegible]
  `boxplot` [illegible] `='Score',m`
  `ain='Box` [illegible] `es')`
- R Box Plot [illegible]



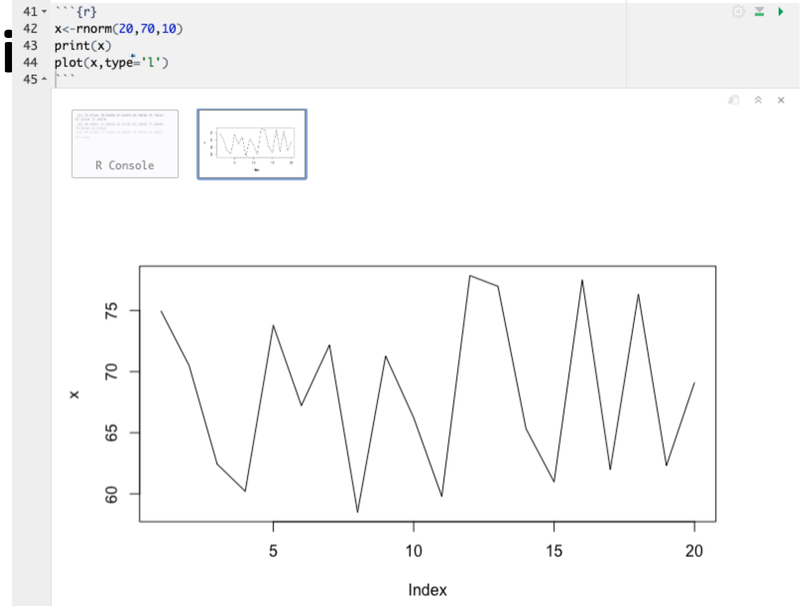
Boxplot of Midterm Exam Scores

Time Series Plot

- 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 time
- Excellent tool for detecting:
  - trends,
  - cycles,
  - other non-random patterns



Time Series Plot i



Time Series Plot

Probability Plotting

- **Probability plotting** is a graphical method of determining whether sample data conform to a hypothesized distribution
- Used for validating assumptions
- Alternative to hypothesis testing

ignore this ancient history

Construction

1. Sort the data from smallest to largest, .

2. $x_{(1)}, x_{(2)}, \dots, x_{(n)}$

3. Calculate the observed cumulative frequency $(j - 0.5)/n$

replaced by computer graphics

For the normal distribution find z_j that satisfies

$$\frac{j - 0.5}{n} = P(Z \le z_j) = \Phi(z_j)$$

3. Plot z_j versus $x_{(j)}$ on special graph paper

weakness: Subjectivity

Usage

- If the … tion is corre…

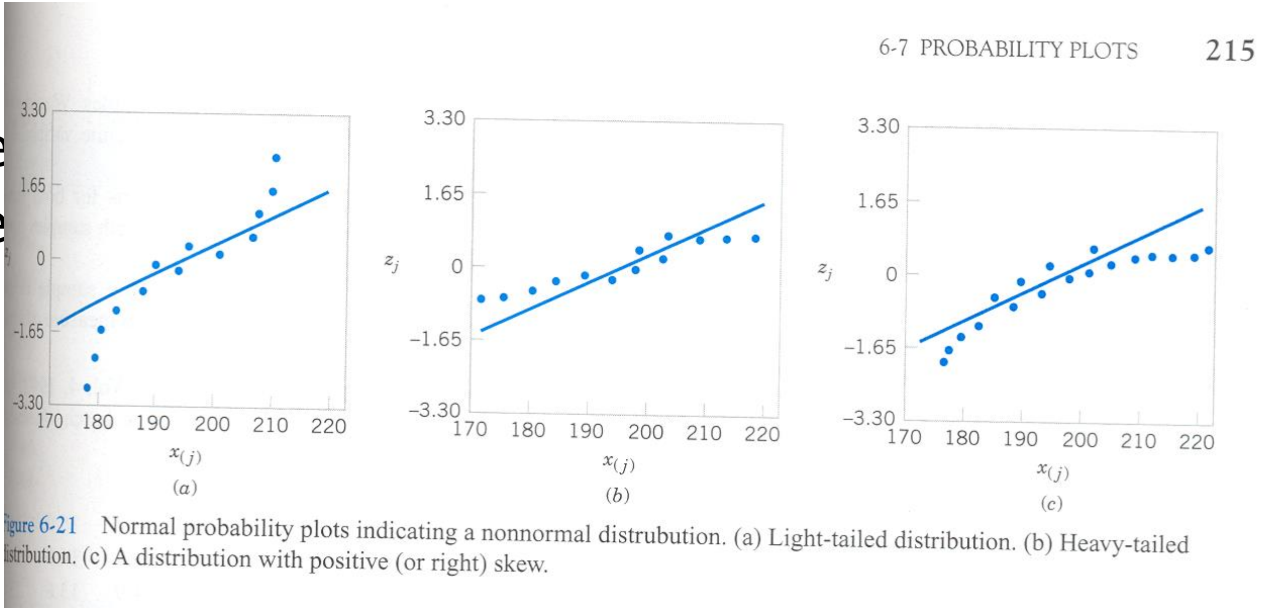
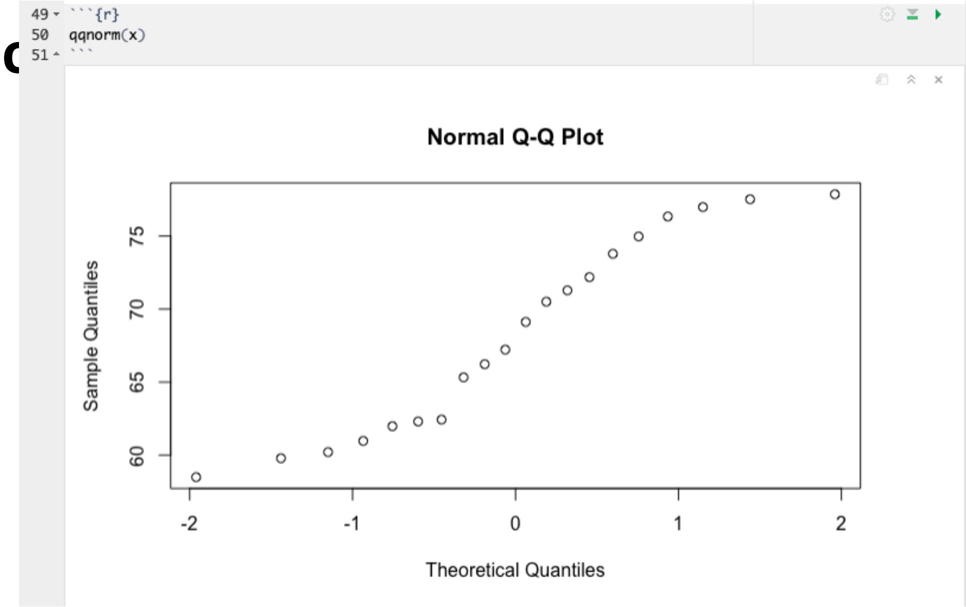


Figure 6-21 Normal probability plots indicating a nonnormal distrubution. (a) Light-tailed distribution. (b) Heavy-tailed distribution. (c) A distribution with positive (or right) skew.

normal probability plots from textbook, figure 6.21 on page 215

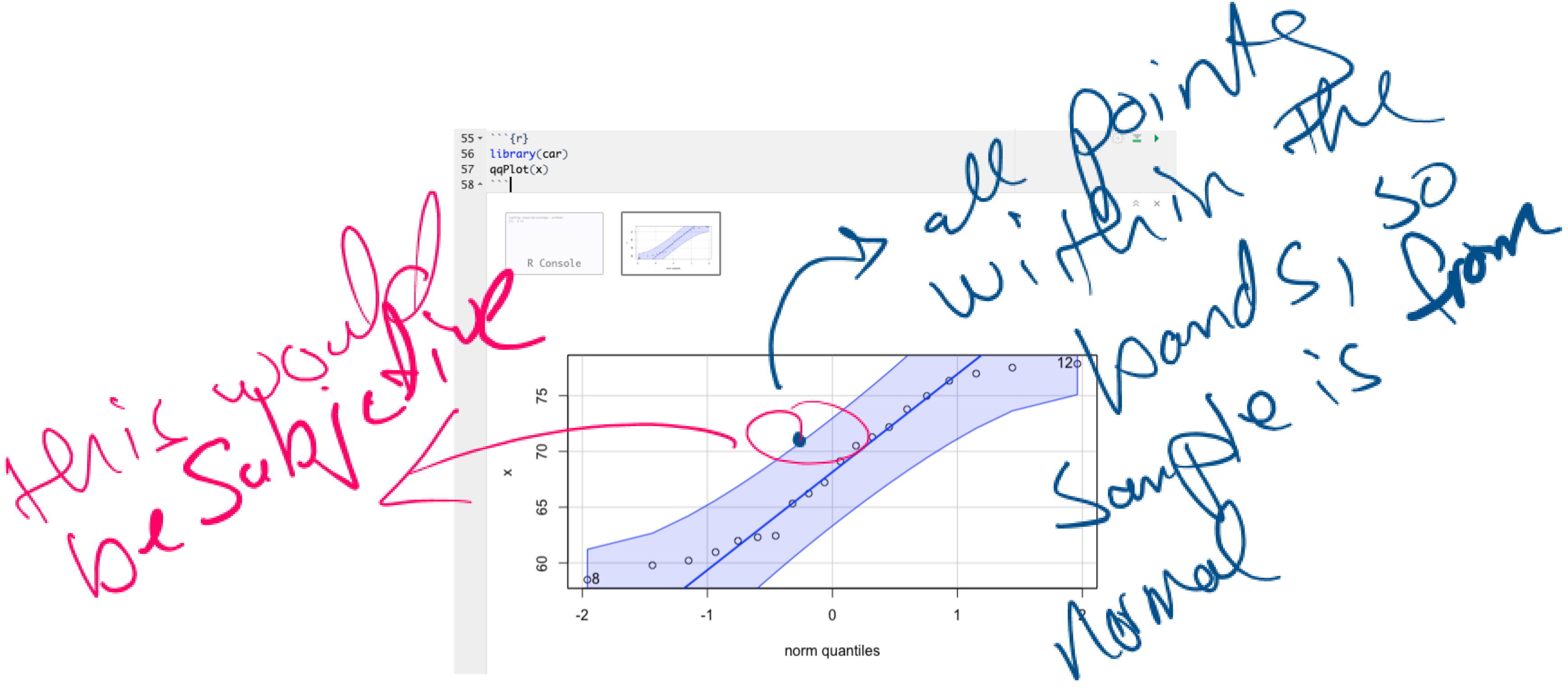
Probability Plo



Normal Probability Plot

Probability Plot Example 2

- Difficulty from example one is how close to straight is “good enough”
- Add confidence bands to normal probability plot
  - Requires package car to be added to R
  - If all points are within the band, we are 95% confident that the sample is from a normal distribution. However if one or more points are not within band, the data is not from a normal distribution

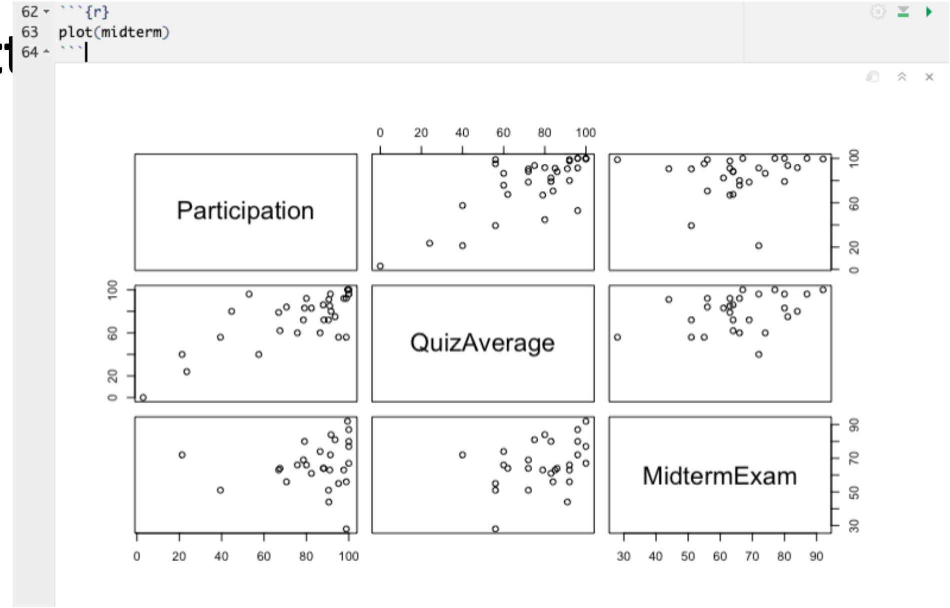


QQ Plot with band

not cover ~~need~~ Chapter 5

Multivariate Data

Matrix of Scatt



Scatter Plots

Covariance in R

```{r}
```
midterm_NA <- na.omit(midterm)
print(cov(midterm_NA))
```

```
             Participation QuizAverage MidtermExam
Participation     340.16778    193.7847    28.75699
QuizAverage       193.78474    269.0899    81.17460
MidtermExam        28.75699     81.1746   188.43915
```

Covariance Matrix
```

Correlation

```{r}
print(cor(midterm_NA))
```

```
              Participation QuizAverage MidtermExam
Participation     1.0000000   0.6405076   0.1135825
QuizAverage       0.6405076   1.0000000   0.3604839
MidtermExam       0.1135825   0.3604839   1.0000000
```

Correlation Matrix