

MANE 3332.01

Lecture 17

Agenda

- Midterm exams are not graded; still contacting students who missed
- Linear Combination Practice Problems (assigned 10/28, due 10/30)
- Chapter Six
- Attendance
- Questions?

Question 1 (1 point)



Consider the linear combination $Y = -9.312 + (1.341) \cdot X_1 + (-7.679) \cdot X_2$. The random variables X_1 and X_2 are independent with $E(X_1) = 29.81$, $E(X_2) = 31.347$, $V(X_1) = 10.172$, and $V(X_2) = 14.69$. What is the value of $E(Y)$?

☐ -200.7384

☐ -210.0504

☒ The correct answer is not provided.

☐ -847.9337

☐ -196.1867

☐ -108.4759

☐ 884.5179

$$E(Y) = ?$$

$$= E(-9.312 + 1.341X_1 - 7.679X_2)$$

$$= E(-9.312) + 1.341E(X_1) - 7.679E(X_2)$$

$$= -9.312 + 1.341(29.81) - 7.679(31.347)$$

$$= -210.0502 \text{ (double check)}$$

$$= -210.0504$$

Question 3 (1 point)



Consider the linear combination $Y = -1.175 + (8.234)X_1 + (-7.984)X_2$. The random variables X_1 and X_2 are not independent but have $\text{COV}(X_1, X_2) = 0.968$ with $E(X_1) = 46.609$, $E(X_2) = 28.538$, $V(X_1) = 16.96$, and $V(X_2) = 14.786$. What is the value of $E(Y)$?

☒ 154.7561

☐ 2092.3895

☐ 207.3443

☐ -138.3194

☐ 1442.3197

☐ The correct answer is not provided.

☐ 20.4222

$$E(Y) = E(-1.175 + 8.234X_1 - 7.984X_2)$$

$$= E(-1.175) + 8.234E(X_1) - 7.984E(X_2)$$

$$= -1.175 + 8.234(46.609) - 7.984(28.538)$$

$$= \cancel{154.4209}$$

$$= 154.7561$$

Question 5 (1 point)



$$Y = cX \rightarrow V(Y) = c^2 V(X)$$

Consider the linear combination $Y = -9.958 + (4.99) \cdot X_1 + (-2.877) \cdot X_2$. The random variables X_1 and X_2 are independent with $E(X_1) = 25.678$, $E(X_2) = 27.144$, $V(X_1) = 12.378$, and $V(X_2) = 10.547$. What is the value of $V(Y)$?

☐ 40.0819

☐ 8.2126

☐ 220.9146

☒ 395.5123

☐ The correct answer is not provided.

☐ -174.2093

$$V(Y) = V(-9.958 + 4.99X_1 - 2.877X_2)$$

$$= V(-9.958) + V(4.99X_1) + V(-2.877X_2)$$

$$= 0 + (4.99)^2 V(X_1) + (-2.877)^2 V(X_2)$$

$$= (4.99)^2 (12.378) + (-2.877)^2 (10.547)$$

$$= 395.51232$$

Question 7 (1 point)



Consider the linear combination $Y = -6.185 + (4.743)X_1 + (-6.834)X_2$. The random variables X_1 and X_2 are not independent but have $\text{COV}(X_1, X_2) = -3.084$ with $E(X_1) = 46.12$, $E(X_2) = 29.62$, $V(X_1) = 12.638$, and $V(X_2) = 13.8$. What is the value of $V(Y)$?

☐ 10.1391

☐ 928.8141

☒ -360.204

☐ 728.8867

☒ The correct answer is not provided.

☒ -103.2758

$$V(c_0 + c_1 X_1 + c_2 X_2) = c_1^2 V(X_1) + c_2^2 V(X_2) + 2c_1 c_2 \text{COV}(X_1, X_2)$$

$$\begin{aligned} V(Y) &= V(-6.185) + 4.743^2 V(X_1) + (-6.834)^2 V(X_2) \\ &\quad + 2(4.743)(-6.834) \text{COV}(X_1, X_2) \\ &= 4.743^2 (12.638) + (-6.834)^2 (13.8) \\ &\quad + 2(4.743)(-6.834)(-3.084) \\ &= 1128.7416 \end{aligned}$$

Handouts

- Chapter 6 Slides
- Chapter 6 Slides marked

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 *variance/std.dev*
 - Included in all statistical software packages, R does a good job calculating descriptive statistics

3 Characteristics of data

- 1) location
- 2) Spread/variability
- 3) Shape of data

Central Tendency / Location

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

✓ Greek letters,
Capital N

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

hat notation

$\hat{N}$ is an estimator of N

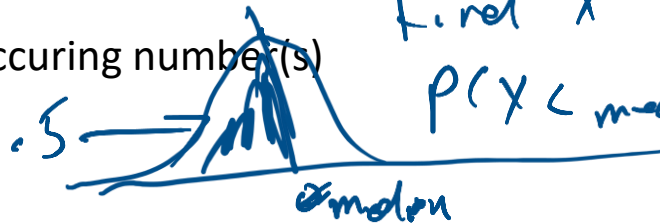
- Sample mean
don't use Greek letters
lower case n

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

- Sample median - middle value
- Sample mode - most commonly occurring number(s)

find x s.t. $P(X < x) = .5$

$P(X < \text{median}) = .5$



Measures of Variability

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

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

- Sample variance

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

- Shortcut (Computational) Formula

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

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

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

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

Some short
calculations

σ_N - Population

σ_n - Sample

R Function Summary - Data Frame

R code

```
summary(midterm)
```

Output is from Spring 2024 results

```
26 ~~~{r}
27 summary(midterm)
28 ~~~|
28:4 Import Dataset ↕ R Markdown ↕
```

Console Terminal × Render × Background Jobs ×

R 4.3.1 · /Volumes/SAMSUNG T7/wfscsBackup/Teaching2/AY_2023_2024/MANE3332_spring2024/PartTwo/

```
> 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

```
> |
```

Descriptive Statistics

R Function Summary - Variable

R code

```
summary(midterm$MidtermExam)
```

Output is from Spring 2024 results

```
30 ~~~{r}
31 summary(midterm$MidtermExam)
32 ~~~
```

32:4 Import Dataset R Markdown

Console Terminal Render Background Jobs

R 4.3.1 · /Volumes/SAMSUNG T7/wfscsBackup/Teaching2/AY_2023_2024/MANE3332_spring2024/PartTwo/1

```
> 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

> View(midterm)
> View(midterm)
> summary(midterm$MidtermExam)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 28.00  59.75   65.00   66.07  74.75   92.00     5

> |
```

Descriptive Statistics

R Function Describe

Summary() does not report variability

Describe() has to be imported

Describe() is part of the package psych

R Code for descriptive statistics using psych package

```
library(psych)
```

```
describe(midterm)
```

Psych package output from Spring 2024

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

Description: df [3 × 13]

	vars <dbl>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>	trimmed <dbl>	mad <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 Dataset R Markdown

Describe() Output

$$\sigma^2 = \frac{1}{N} \sum (x - \mu)^2 \rightarrow 2^{\text{nd}} \text{ moment about } \mu$$

Describe Output, part 2

3rd order moment skewness

4th order moment kurtosis

Description: df [3 x 13]

	median <dbl>	trimmed <dbl>	mad <dbl>	min <dbl>	m... <dbl>	range <dbl>	skew <dbl>	kurtosis <dbl>	se <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

mean absolute deviation

$$\text{mad} = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$$

$$se = \frac{s}{\sqrt{n}}$$

s standard error

Describe Output

Attendance
1-A

Calculating Quantiles

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 100

of the population values at or below it. This estimate, called the 100

th *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 100 TH 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 $0 \leq d < 1$ the decimal part. If $1 \leq i < n$, and $d = 0$, the 100

th sample quantile is $x_{(i)}$. If $1 \leq i < n$ and $0 < d < 1$, interpolate linearly between $x_{(i)}$ and $x_{(i+1)}$. In either case, the 100

th *sample quantile* is

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

when $1 \leq i < n$. If $i = 0$ or n , the 100

th sample quantile does not exist. If 100

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, $\bar{x} = x_{0.50}$.

SAMPLE QUANTILES

The percentiles $x_{0.25}$, $x_{0.50}$, and $x_{0.75}$ are known as the *first*, *second*, and *third sample quantiles*, respectively. These quantiles 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.

- First Quartile*: Since $(0.25)(75+1) = 19$, $q_1 = x_{0.25} = x_{(19)} = 1.46$.
- Second Quartile (Median)*: Since $(0.50)(75+1) = 38$, $q_2 = \bar{x} = x_{0.50} = x_{(38)} = 1.53$.

reference for calculating quantiles

Quantile Example

8 Observations from
binomial distribution with
 $n = 10$ and $p = .5$.
6, 4, 5, 7, 3, 5, 4, 6

Quantile Example

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

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 and Leaf Example

R output of a Stem and Leaf diagram

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

```
2 | 8
3 |
4 | 4
5 | 11566
6 | 13334446679
7 | 2247
8 | 00147
9 | 2
```

Stem and Leaf Plot of Midterm Exam Scores

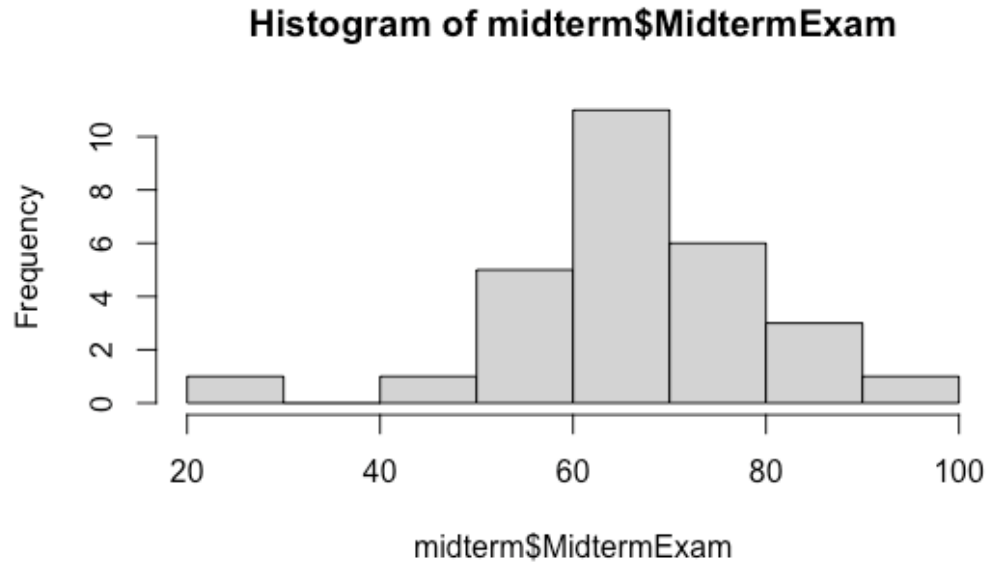
Histogram

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

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

Histogram Example

R output of histogram



Histogram of Midterm Exam Scores

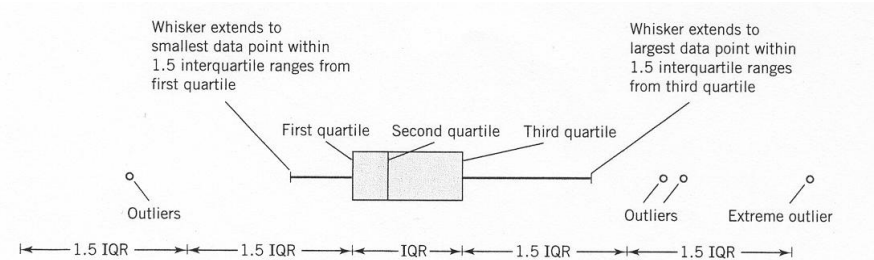
Boxplot

Graphical display that simultaneously describes several important features of a data set such as center, spread, departure from symmetry and outliers

Requires the calculation of quantiles (quartiles)

Box Plot 1

Figure 2-11
Description of a box plot.



Box plot with explanation

Box Plot 2

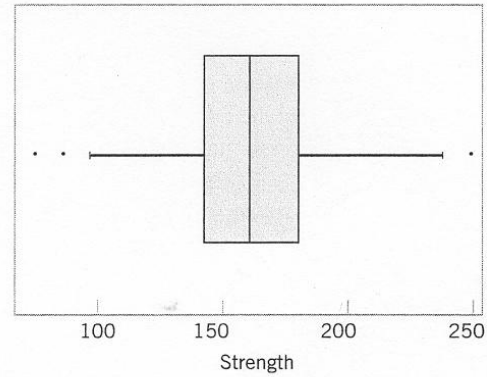


Figure 2-12 Box plot for compressive strength data in Table 2-2.

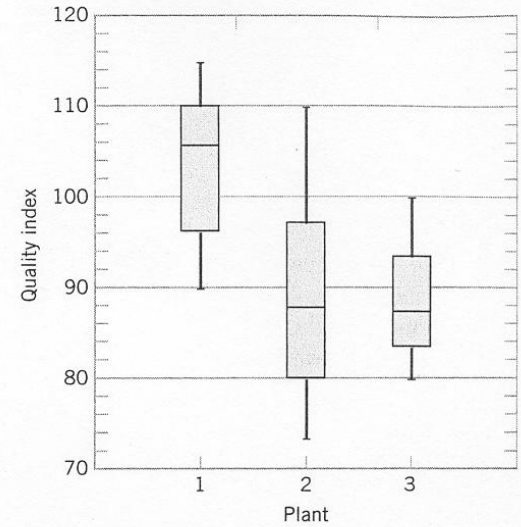


Figure 2-13 Comparative box plots of a quality index at three plants.

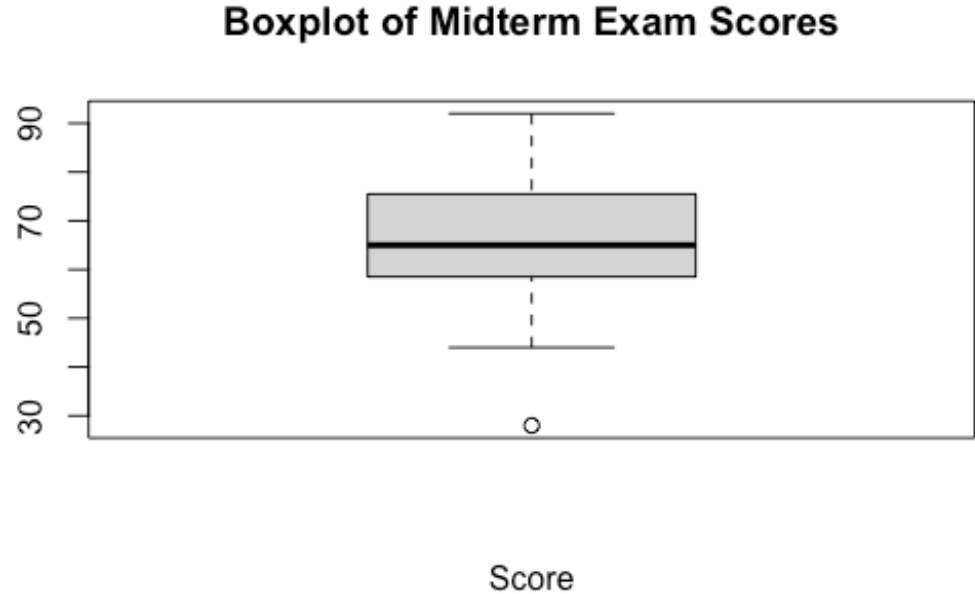
examples of boxplots

Box Plot 3

R code for Box Plot

```
boxplot(midterm$MidtermExam,xlab='Score',main='Boxplot of Midterm Exam Scores')
```

R Box Plot output



Boxplot of Midterm Exam Scores

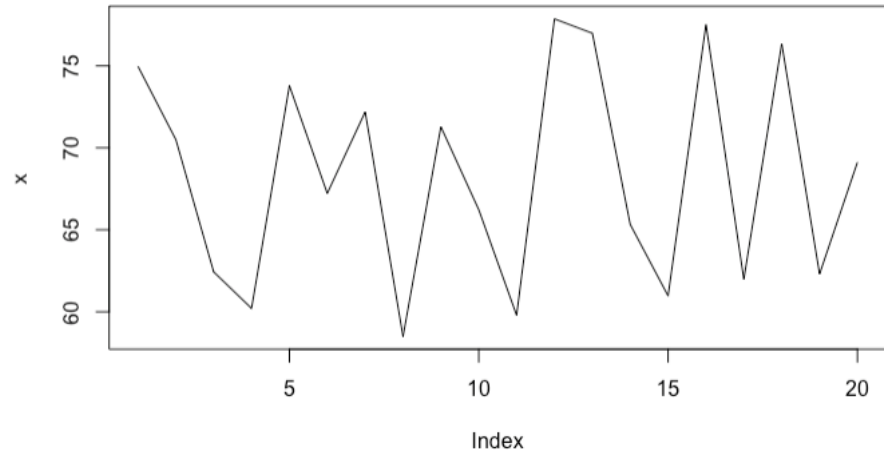
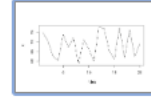
Time Series Plot

- 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 in R

```
41 {r}  
42 x<-rnorm(20,70,10)  
43 print(x)  
44 plot(x,type='l')  
45 }
```

R Console



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

Construction

1. Sort the data from smallest to largest, .
2. $x_{(1)}, x_{(2)}, \dots, x_{(n)}$
3. Calculate the observed cumulative frequency $(j - 0.5)/n$

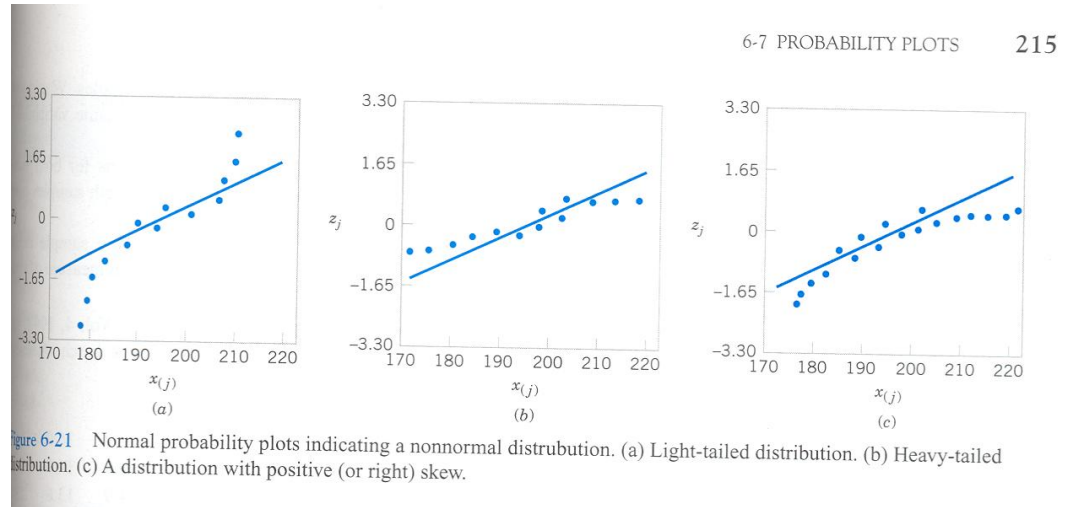
For the normal distribution find z_j that satisfies

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

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

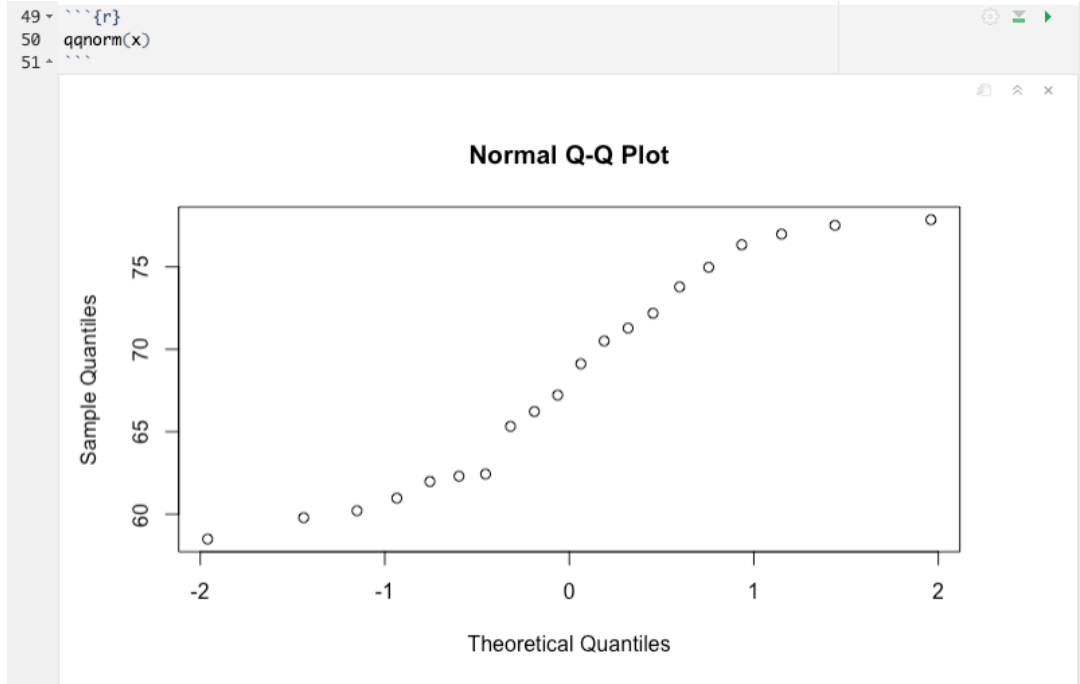
Usage

If the data plots as a straight line, the assumed distribution is correct



normal probability plots from textbook, figure 6.21 on page 215

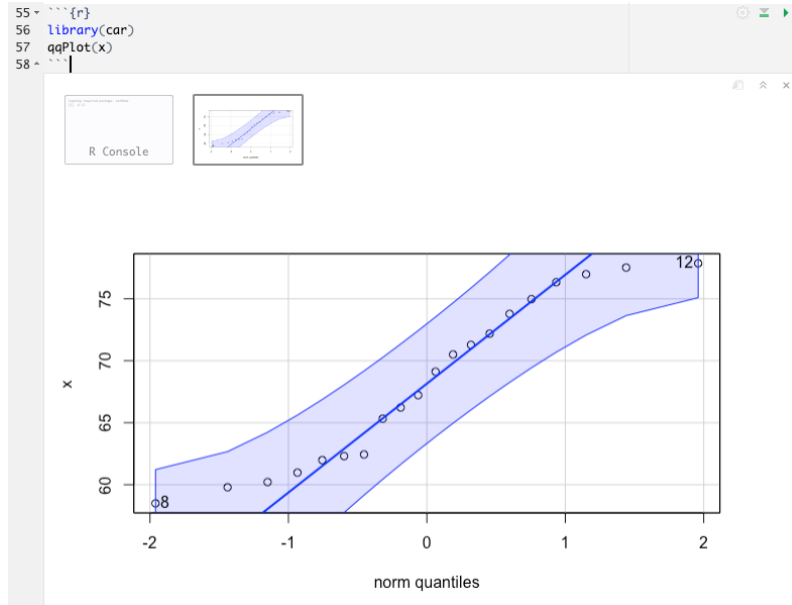
Probability Plot Example 1 in R



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

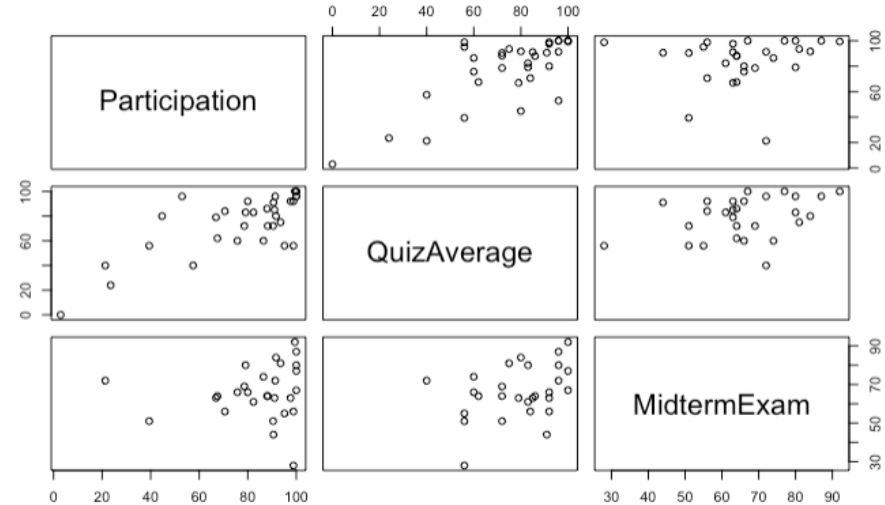


QQ Plot with band

Multivariate Data

Matrix of Scatter Plot in R

```
62 > {r}  
63 plot(midterm)  
64 ^ {r}
```



Scatter Plots

Covariance in R

```
67 ~ ```{r}  
68 midterm_NA <- na.omit(midterm)  
69 print(cov(midterm_NA))  
70 ^ ```
```

	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

```
74 ~~~{r}  
75 print(cor(midterm_NA))  
76 ~~~
```

	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