

# MA5701: Statistical Methods

Course Introduction

Kui Zhang, Mathematical Sciences

# Contents

- Chapter 1: Data and Statistics
- Chapter 2: Probability and Sampling Distributions
- Chapter 3: Principles of Inference
- Chapter 4: Inferences on a Single Population
- Chapter 5: Inferences on Two Populations
- Chapter 6: Inferences for Two or More Means
- Chapter 7: Linear Regression
- Chapter 8: Multiple Regression

# Data and R

- **Data file**
  - Canvas Files Tab, then “Data” Folder
- **Software - R**
  - R Website - <https://www.r-project.org>

# Instructor Information

- **Office Location:** 211 Fisher Hall
- Email: [kuiz@mtu.edu](mailto:kuiz@mtu.edu)
- Office Hours: MWF 2:00pm – 2:50pm; or by appointment

# Grading Schemes

<b>Letter Grade</b>	<b>Percentage</b>	<b>Grade points/credit</b>	<b>Rating</b>
A	[90, 100]	4.00	Excellent
AB	[85, 90)	3.50	Very good
B	[80, 85)	3.00	Good
BC	[75, 80)	2.50	Above average
C	[70, 75)	2.00	Average
CD	[65, 70)	1.50	Below average
D	[60, 65)	1.00	Inferior
F	[0, 60)	0.00	Failure

# Grading Policy

Course Component	Percentage
Attendance	5%
Weekly Homework	50%
Mid-Term Exam (Open book/note)	15%
Final Exam (Open book/note)	30%
<b>Total</b>	<b>100%</b>

# Grading Policies

- **Late Assignments**
  - Not allowed unless approved by the instructor in advance.
- **No Round Up**
  - Calculated by CANVAS automatically and will not be rounded up or down.
  - For example, if your final grade in CANVAS is 89.75., then your letter grade will be B.

# Course Policies

- **Attendance (5%)**
  - Get approval in advance if you cannot attend a class.
- **Homework (50%)**
  - Weekly. Final week HW has bonus points.
  - Homework can be written or typed.
- **Mid-term Exam (15%)**
  - In-class, Open book, open notes.
  - Class on Friday 2/21, last Friday before the spring break.
- **Final Exam (30%)**
  - Open book, open notes

# Course Policies

- **Communication is the key.**
  - Canvas Inbox for individual correspondence.
- Let me know if you need any special accommodations for the class, the homework, and the exams as soon as possible.
- Please feel free to ask any question during class; your questions are an important part of this course.
- Please make sure to bring a calculator with you to class, so you can be appropriately prepared for assignments and/or exams.

# MA5701: Statistical Methods

Chapter 1 : Data and Statistics

Kui Zhang, Mathematical Sciences

## Exercise – Tree Data

- The discipline of forest science is a frequent user of statistics. An important activity is to gather data on the physical characteristics of a random sample of trees in a forest. The resulting data may be used to estimate the potential yield of the forest, to obtain information on the genetic composition of a particular species, or to investigate the effect of environmental conditions. The following data set consists of measurements of three characteristics of 64 sample trees of a particular species.

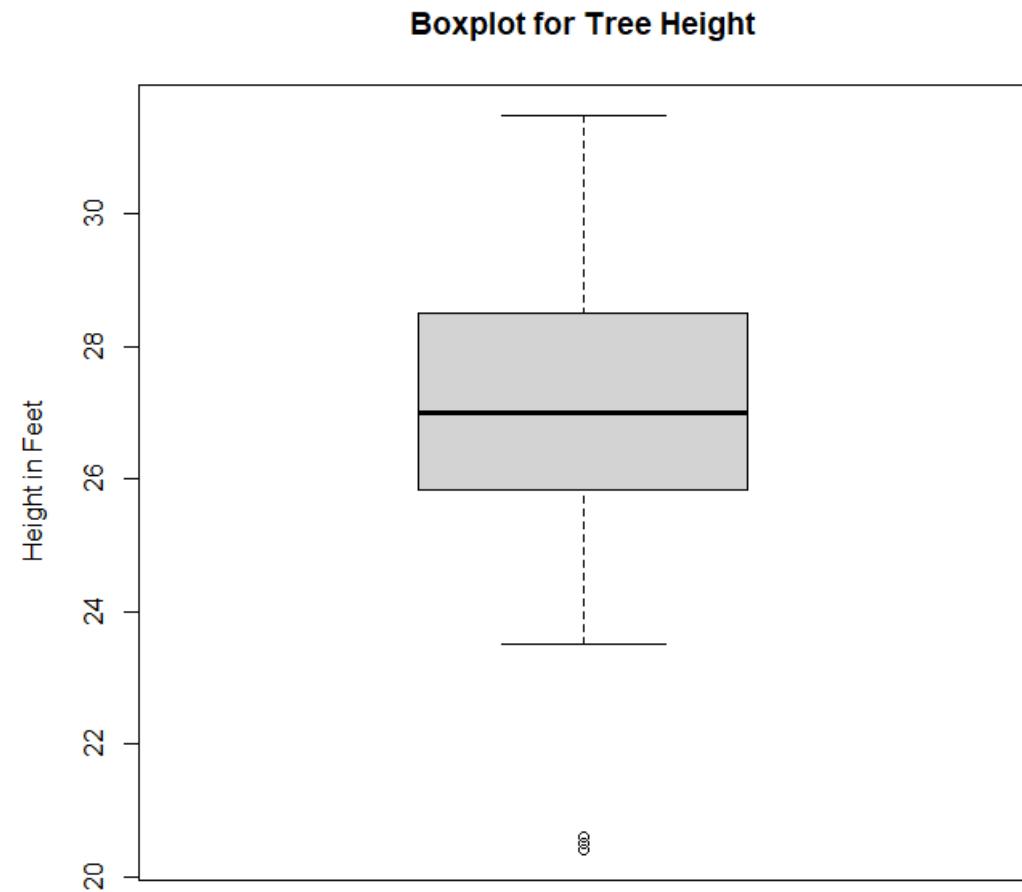
## Exercise – Tree Data

- The data look like this:

<b>obs</b>	<b>dfoot</b>	<b>hcrn</b>	<b>ht</b>		<b>obs</b>	<b>dfoot</b>	<b>hcrn</b>	<b>Ht</b>
1	4.1	1.5	24.5		23	4.3	2.0	25.6
2	3.4	4.7	25.0		24	2.7	3.0	20.4
3	4.4	2.8	29.0		25	4.3	2.0	25.0

- dfoot:** the diameter of the tree at one foot above ground level, measured in inches
- hcrn:** the height to the base of the crown measured in feet
- ht:** the total height of the tree measured in feet

# Boxplot for Height – Tree Data



## Exercise - Boxplot for Height from Tree Data

- Mean: 26.96
- Median: 27.00
- Maximum (largest) observation: 31.50
- Minimum (smallest) Observation: 20.40
- First Quartile ( $Q_1$ ): 25.875
- Third Quartile ( $Q_3$ ): 28.50

## Exercise - Boxplot for Height from Tree Data

- Interquartile Range?
- Step?
- Upper Inner Fence (UIF)?
- Lower Inner Fence (LIF)?
- Upper Outer Fence (UOF)?
- Lower Outer Fence (LOF)?
- Are smallest or largest observations here outliers?

## Exercise - Boxplot for Height from Tree Data

- **Interquartile Range:**  $28.50 - 25.875 = 2.625$
- Step:  $2.625 * 1.5 = 3.9375$
- **Upper Inner Fence (UIF):**  $28.50 + 3.9375 = 32.4375$
- **Lower Inner Fence (LIF):**  $25.875 - 3.9375 = 21.9375$
- **Upper Outer Fence (UOF):**  $28.50 + 2 * 3.9375 = 36.375$
- **Lower Outer Fence (LOF):**  $25.875 - 2 * 3.9375 = 18.0$
- **Smallest observation:**  $20.40 < \text{LIF}$  but  $> \text{LOF}$ , mild outlier
- **Largest observation:**  $31.50 < \text{UIF}$ , not outlier

# Introduction

- **Definition** - A set of **data** is a collection of observed values representing one or more characteristics of some objects or units.
- **Definition** - A **population** is a data set representing the entire entity of interest.

# Example – NORC Survey Data

<b>Respondent</b>	<b>AGE</b>	<b>SEX</b>	<b>HAPPY</b>	<b>TVHOURS</b>
1	41	1	2	0
9	31	2	1	0
22	64	2	3	0
36	26	1	2	2

This is a survey data from National Opinion Research Center.  
This is a survey conducted in 1996 for 2904 households from  
over 70 questions. This is only part of it.

# Data Resource and Format

- **Definition** - A **sample** is a data set consisting of a portion of a population. (Obtained in a way to represent the population)
- **Definition** – A **census** is the collection of the data from everyone on a population.
- Where we can obtain the data:
  - **Primary** data are collected as the part of study.
  - **Secondary** data are obtained from other resources.

# Example - NORC Survey Data

<b>Respondent</b>	<b>AGE</b>	<b>SEX</b>	<b>HAPPY</b>	<b>TVHOURS</b>
1	41	1	2	0
9	31	2	1	0
22	64	2	3	0
36	26	1	2	2

This is a survey data from National Opinion Research Center.  
This is a survey conducted in 1996 for 2904 households from  
over 70 questions. This is only part of it.

# Questions from NORC Survey Data Data

- Is this data a census data or sample data?
  - Sample data
- What is the population here?
  - All household in United States
- What is the relationship between the population and the sample?
  - Sample is the subset of the population
- Why can we not collect census data in this situation and in most of other situations?
  - It is too expensive and/or time consuming, may not be possible

# Observations and Variables

- **Data format**
  - Observation(s) – a row in the data file
  - Variables(s) – a column in the data file
- Two types of variables – Qualitative (Categorical) variables and Quantitative variables
- **Qualitative (Categorical) variable** - is a variable that is not numerical. It describes data that fits into categories.
- **Quantitative Variable** – is a variable that is measured on a numeric scale for which meaningful arithmetic operations make sense.
- **Remark – Numbers can be qualitative!**

# Variables – Texas House Data

Obs	Zip	Age	Bed	Bath	Size	Lot	Exter	garage	fp	Price
1	3	21	3	2	951	64904	Other	0	0	30000
3	4	7	1	1	676	54450	Other	2	0	46500
5	1	51	3	1	1186	10857	Other	1	0	51500
7	3	8	3	2	1368	.	Frame	0	0	56990
9	1	51	2	1	1176	6259	Frame	1	1	65500

Data: 69 families in a midsized city in east Texas. This is only part of it.

# Questions from Texas House Data

- The zip code of a house: qualitative or quantitative?
  - Qualitative
- The square footage of a house: qualitative or quantitative?
  - Quantitative
- Today's snowfall at Houghton: qualitative or quantitative?
  - Quantitative
- Time spent for MA5701: qualitative or quantitative?
  - Quantitative
- Your desire to study MA5701: qualitative or quantitative?
  - Qualitative

# Types of Quantitative Variables

- **Definition** - A **discrete** variable can assume only accountable number of values.
- **Definition** – A **continuous** variable is one that can take any one of an uncountable number of values in an interval.
- **Continuous Variable** – in real world, they may appear discrete but conceptually take any value in an interval. For example, AGE, generally, we calculate it according to your birthdate not in which minute which hour you were born. Other examples include height, blood pressure.

# Types of Qualitative Variables

- **Definition** – The **ordinal scale** distinguishes between measurements. Generally, the relative amounts of some characteristic they process.
- **Definition** – The **nominal scale** identifies observed values by name or classification.
  - Generally, for categorical or qualitative variables
  - Weakest scale

# Example - NORC Survey Data

Respondent	AGE	SEX	HAPPY	TVHOURS
1	41	1	2	0
9	31	2	1	0
22	64	2	3	0
36	26	1	2	2

**Respondent:**

- A. Qualitative
- B. Quantitative

- C. Ordinal
- D. Nominal

# Example - NORC Survey Data

Respondent	AGE	SEX	HAPPY	TVHOURS
1	41	1	2	0
9	31	2	1	0
22	64	2	3	0
36	26	1	2	2

**Age:**

- A. Qualitative
- B. Quantitative

- C. Continuous
- D. Discrete

# Example - NORC Survey Data

Respondent	AGE	SEX	HAPPY	TVHOURS
1	41	1	2	0
9	31	2	1	0
22	64	2	3	0
36	26	1	2	2

**Sex:**

- A. Qualitative
- B. Quantitative

- C. Ordinal
- D. Nominal

# Example - NORC Survey Data

Respondent	AGE	SEX	HAPPY	TVHOURS
1	41	1	2	0
9	31	2	1	0
22	64	2	3	0
36	26	1	2	2

**HAPPY:**

- A. Qualitative
- B. Quantitative
- C. Ordinal
- D. Nominal

# Example - NORC Survey Data

Respondent	AGE	SEX	HAPPY	TVHOURS
1	41	1	2	0
9	31	2	1	0
22	64	2	3	0
36	26	1	2	2

**TVHOURS:**

- A. Qualitative
- B. Quantitative

- C. Continuous
- D. Discrete

# Variables – Texas House Data

Obs	Zip	Age	Bed	Bath	Size	Lot	Exter	garage	fp	Price
1	3	21	3	2	951	64904	Other	0	0	30000
3	4	7	1	1	676	54450	Other	2	0	46500
5	1	51	3	1	1186	10857	Other	1	0	51500
7	3	8	3	2	1368	.	Frame	0	0	56990
9	1	51	2	1	1176	6259	Frame	1	1	65500

**Zip Code:**

- A. Qualitative
- B. Quantitative

- C. Ordinal
- D. Nominal

# Variables – Texas House Data

Obs	Zip	Age	Bed	Bath	Size	Lot	Exter	garage	fp	Price
1	3	21	3	2	951	64904	Other	0	0	30000
3	4	7	1	1	676	54450	Other	2	0	46500
5	1	51	3	1	1186	10857	Other	1	0	51500
7	3	8	3	2	1368	.	Frame	0	0	56990
9	1	51	2	1	1176	6259	Frame	1	1	65500

**Age:**

- A. Qualitative
- B. Quantitative
- C. Discrete
- D. Continuous

# Variables – Texas House Data

Obs	Zip	Age	Bed	Bath	Size	Lot	Exter	garage	fp	Price
1	3	21	3	2	951	64904	Other	0	0	30000
3	4	7	1	1	676	54450	Other	2	0	46500
5	1	51	3	1	1186	10857	Other	1	0	51500
7	3	8	3	2	1368	.	Frame	0	0	56990
9	1	51	2	1	1176	6259	Frame	1	1	65500

**Bed or Bath or Garage or Fire Place:**

- A. Qualitative
- C. Discrete
- B. Quantitative
- D. Continuous

# Variables – Texas House Data

Obs	Zip	Age	Bed	Bath	Size	Lot	Exter	garage	fp	Price
1	3	21	3	2	951	64904	Other	0	0	30000
3	4	7	1	1	676	54450	Other	2	0	46500
5	1	51	3	1	1186	10857	Other	1	0	51500
7	3	8	3	2	1368	.	Frame	0	0	56990
9	1	51	2	1	1176	6259	Frame	1	1	65500

**Size or Lot or Price:**

- A. Qualitative
- B. Quantitative
- C. Discrete
- D. Continuous

# Variables – Texas House Data

Obs	Zip	Age	Bed	Bath	Size	Lot	Exter	garage	fp	Price
1	3	21	3	2	951	64904	Other	0	0	30000
3	4	7	1	1	676	54450	Other	2	0	46500
5	1	51	3	1	1186	10857	Other	1	0	51500
7	3	8	3	2	1368	.	Frame	0	0	56990
9	1	51	2	1	1176	6259	Frame	1	1	65500

**Exterior Type:**

- A. Qualitative
- B. Quantitative
- C. Ordinal
- D. Nominal

# Data – Texas House Data

Obs	Zip	Age	Bed	Bath	Size	Lot	Exter	garage	fp	Price
1	3	21	3	2	951	64904	Other	0	0	30000
3	4	7	1	1	676	54450	Other	2	0	46500
5	1	51	3	1	1186	10857	Other	1	0	51500
7	3	8	3	2	1368	.	Frame	0	0	56990
9	1	51	2	1	1176	6259	Frame	1	1	65500

Data: 69 families in a midsized city in east Texas. This is only part of it.

# Distributions

- **Why we need distributions?**
  - Same as other statistics, to use them to summarize data to draw conclusions.
- **Definition** – A **frequency distribution** is a listing of frequencies (counts) of all categories of the observed values of variable.
- **Definition** – A **relative frequency distribution** consists of the relative frequencies, or proportions (percentages), of observations belong to each category.

# Data – Texas House Data

Obs	Zip	Age	Bed	Bath	Size	Lot	Exter	garage	fp	Price
1	3	21	3	2	951	64904	Other	0	0	30000
3	4	7	1	1	676	54450	Other	2	0	46500
5	1	51	3	1	1186	10857	Other	1	0	51500
7	3	8	3	2	1368	.	Frame	0	0	56990
9	1	51	2	1	1176	6259	Frame	1	1	65500

Data: 69 families in a midsized city in east Texas. This is only part of it.

# Frequency Table – Discrete Variable

bed	bed			
bed	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1	1.45	1	1.45
2	3	4.35	4	5.80
3	46	66.67	50	72.46
4	16	23.19	66	95.65
5	3	4.35	69	100

# Frequency Table – Continuous Variables

price	Frequency	Percent	Cumulative Frequency	Cumulative Percent
[ 0, 50k)	4	5.80	4	5.80
[ 50k, 100k)	22	31.88	26	37.68
[100k, 150k)	23	33.33	49	71.01
[150k, 200k)	10	14.49	59	85.51
[200k, 250k)	2	2.90	61	88.41
[250k, 300k)	1	1.45	62	89.86
[300k, 350k)	4	5.80	66	95.65
[350k, 400k)	3	4.35	69	100.00

# Statistics in Chapter 1

- Sample size for each variable
- For a single qualitative variable,
  - Frequency and Relative frequency
- For a single quantitative variable,
  - Sample mean/variance/standard deviation, median, quantiles
- For two qualitative variables
  - Two-dimensional contingency table with frequency
- For a qualitative variable and a quantitative variable
  - Sample mean/variance for each level of qualitative variable

# Importance of Data Displays

- It is a part of **Descriptive Statistics** – organizing and summarizing data through graphics.
- Clever plots of data provide a powerful way for looking at data – sometimes provide more than enough information to answer questions.
- Formal statistical analysis depends on underlying models and assumptions – plots provide a quick, easy, and effective way to check and verify those models and assumptions.
- Generally used Statistical Software can produce a correctly constructed chart and graph with some adjustments

# Plots in Chapter 1

- For a single qualitative variable
  - Barplot
- For a single quantitative variable
  - Boxplot – for small to moderate data
  - Histogram - for moderate and large data
- For a qualitative variable and a quantitative variable
  - Boxplots – boxplot of the quantitative variable for each level of the qualitative variable
- For two quantitative variables
  - Scatter plot

# Introduction of R

- R web site – download and install R
  - <https://www.r-project.org/>
- Open R
  - Create a data or input a data from a data file to R
  - Perform analysis
- Put R scripts to a file so they can be used later
  - Copy and paste to R

# Basic Data Types in R

- Numeric
  - Such as 1, 10, -0.24, 3.5
- Integer
  - 1L, 55L, -3L
- Complex – not used in our course
- Character (String)
  - “10”, “A”, “Car”
- Logical (Boolean)
  - TRUE or FALSE

# Basic Data Types in R

- Quantitative variables
  - Numeric
  - Integer
- Qualitative variables
  - Factor (a data structure and can be ordered or unordered)
  - Numeric – need to change to factor
  - Character – is considered as factor automatically

# Basic Data Structure in R - Vector

- Vector is an **ordered collection of same types** of data with a given length.
- Numeric vector: `x <- c(1, 3, 7, 8)`
- Character vector: `y <- c("A", "B", "MY")`
- Here, `x <- c(1, 3, 7, 8)`
  - A vector (data) named `x` is created. Equivalently, you can also use `x = c(1, 3, 7, 8)`,
  - `c()` is a R function that combines its arguments.
  - This vector (`x`) has 4 elements: 1 (first element), 3 (second element), 7 (third element), 8 (fourth element).

# Names in R

- Can be letters, numbers, ., and \_.
  - For example, A1, b2, c.s, c\_s
- Cannot start with a number
  - Better start with a letter
- Lower cases and upper cases are different
  - For example, A1 and a2 refer to different data/functions/etc.

# Subscripting with Vector

- $x <- 2 * (1:5)$  or  $x <- c(2, 4, 6, 8, 10)$
- Positive integers
  - $x[2]$  returns 4 (a scalar, which is a vector with length of 1)
  - $x[c(1, 3, 5)]$  returns a vector with three elements: 2, 6, 10
  - $x[c(5, 1, 3)]$  returns a vector with two elements: 10, 2, 6
- Negative integers
  - $x[-2]$  is equivalent to  $x[c(1, 3, 4, 5)]$
  - $x[c(-2, -4)]$  is equivalent to  $x[c(1, 3, 5)]$
  - How about  $x[c(-4, -2)]$ ?

# Subscripting with Vector

- Cannot combine positive and negative integers
- More on subscripting:  $x <- c(2, 4, 6, 8, 10)$ 
  - Results of  $x[c(1, 1)]$ ?
  - Results of  $x[c(1, 2, 1)]$ ?
- Logical
  - $x[c(\text{TRUE}, \text{FALSE}, \text{FALSE}, \text{TRUE}, \text{FALSE})]$  is equivalent to  $x[c(1, 4)]$
  - How about  $x[\text{TRUE}]$

# Element-wise Operation

- Arithmetic operation: +, -, \*, and /
- Element-wise: it performs each operation on each element of a vector independently of the other elements.
- Two vectors with same length:
  - $c(1, 3, 5) + c(2, 4, 6)$  is  $c(1 + 2, 3 + 4, 5 + 6)$
- One vector and one scalar:
  - $c(1, 3, 5) + 10$  is equivalent to  $c(1, 3, 5) + c(10, 10, 10)$
- Two vectors with different lengths (not covered here)

# Data Structure – Matrix and Data Frame

- Matrix and data frame are a two-dimensional of data in row and columns.
- All rows have the same length
- All columns have the same length
- Matrix
  - All columns must be the same type,
  - If one column is numeric, then all the other columns must be numeric
- Data frame
  - Columns can have different types of data
  - One column is numeric, the other column can be character
- Vector is used to store data with one variable
- Data frame is used to store data with more than one variable

# Subscripting with Matrix/Data Frame

- $x <- 2 * (1:5)$  or  $x <- c(2, 4, 6, 8, 10)$
- Positive integers
  - $x[2]$  returns 4 (a scalar, which is a vector with length of 1)
  - $x[c(1, 3, 5)]$  returns a vector with three elements: 2, 6, 10
  - $x[c(5, 1, 3)]$  returns a vector with two elements: 10, 2, 6
- Negative integers
  - $x[-2]$  is equivalent to  $x[c(1, 3, 4, 5)]$
  - $x[c(-2, -4)]$  is equivalent to  $x[c(1, 3, 5)]$
  - How about  $x[c(-4, -2)]$ ?

# Input Data from a Data File

- Set up the working directory – the file folder contains your data files
  - Put all data files to a single folder
  - Use the full path and /
  - `setwd("G:/My Drive/Zkui/Teaching/DataSets/MA5701")`
- Use R function `read.table` and/or `read.csv`
  - `norc <- read.csv(file = "norc.csv", stringsAsFactors = FALSE)`

# Basic Functions to Look at a Data Frame

- Print first or last few rows of a data frame
  - `head(norc)`
  - `tail(norc)`
- Name of columns
  - `names(norc)`
- Structure of a data frame
  - `str(norc)`

# Subscripting with Matrix/Data Frame

- Subscripting is done with [ for rows, for columns]
- Each part can be positive/negative integers and logical
  - Blank means all rows/all columns
  - First row and all columns: `norc[1, ]`
  - First and third row and all columns: `nocr[c(1, 3), ]`
  - Second column and all rows: `norc[ , 2]`
  - Second and third columns and all rows: `norc[ , c(2, 3)]`
  - First & third rows and first & second columns: `norc[c(1, 3), c(1, 2)]`
- Column names for one or more columns
  - Second column: `norc$age`

# Frequency Table – Discrete Variable

bed	bed			
bed	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1	1.45	1	1.45
2	3	4.35	4	5.80
3	46	66.67	50	72.46
4	16	23.19	66	95.65
5	3	4.35	69	100

# Construct Frequency Table with R

- Frequency table: `table()` function
  - `table(texas$bed)`
- Relative frequency table: `prop.table()`
  - Input is the output from `table()`
  - `prop.table( table(texas$bed) )`
- Use appropriate decimal digits: `round()`
  - `round(b, digits = 4)`

# Frequency Table – Continuous Variables

price	Frequency	Percent	Cumulative Frequency	Cumulative Percent
[ 0, 50k)	4	5.80	4	5.80
[ 50k, 100k)	22	31.88	26	37.68
[100k, 150k)	23	33.33	49	71.01
[150k, 200k)	10	14.49	59	85.51
[200k, 250k)	2	2.90	61	88.41
[250k, 300k)	1	1.45	62	89.86
[300k, 350k)	4	5.80	66	95.65
[350k, 400k)	3	4.35	69	100.00

# Construct Frequency Table with R

- For continuous variables: `cut()` and `table()`
- Function `cut()`:
  - Divides the range of data into intervals and codes the values according to which interval they fall.
  - ```
a <- cut(x = texas$price,  
           breaks = c(0, 50, 100, 150, 200, 250, 300, 350, 400) * 1000,  
           include.lowest = TRUE, right = FALSE)
```
  - Then use `table(a)` etc.

# Bar plot (bar chart)

- A **bar plot** is a plot that uses the height of rectangles (bars) to represent the frequency of each value.
- Look for differences in the heights of the bars.
- R function: **barplot()**
  - Use the output of **table()** as the input
  - For example, `barplot( table(texas$exter) )`

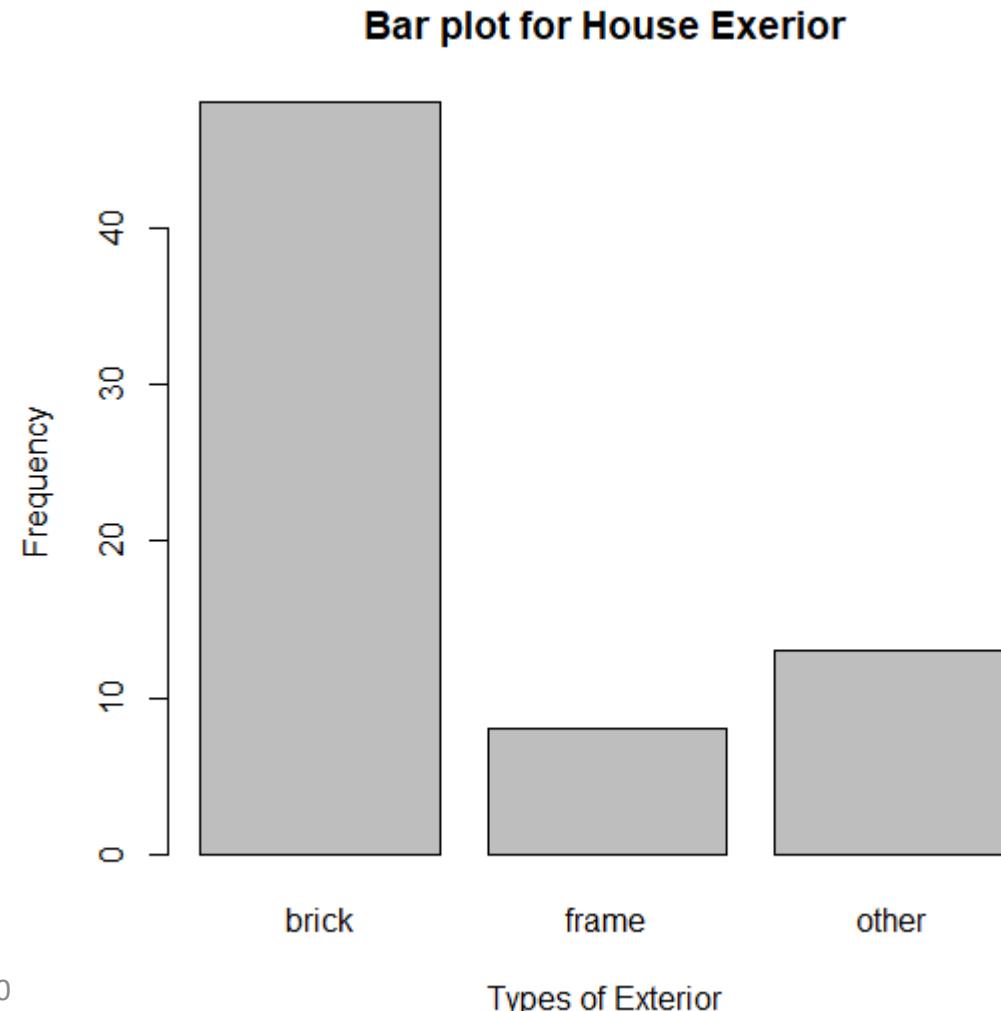
# Bar plot (bar chart)

- A **bar plot** is a plot that uses the height of rectangles (bars) to represent the frequency of each value.
- Look for differences in the heights of the bars.
- R function: **barplot()**
  - Use the output of **table()** as the input
  - For example, `barplot(table(texas$exter))`

# Bar plot (bar chart)

- **Conclusions**

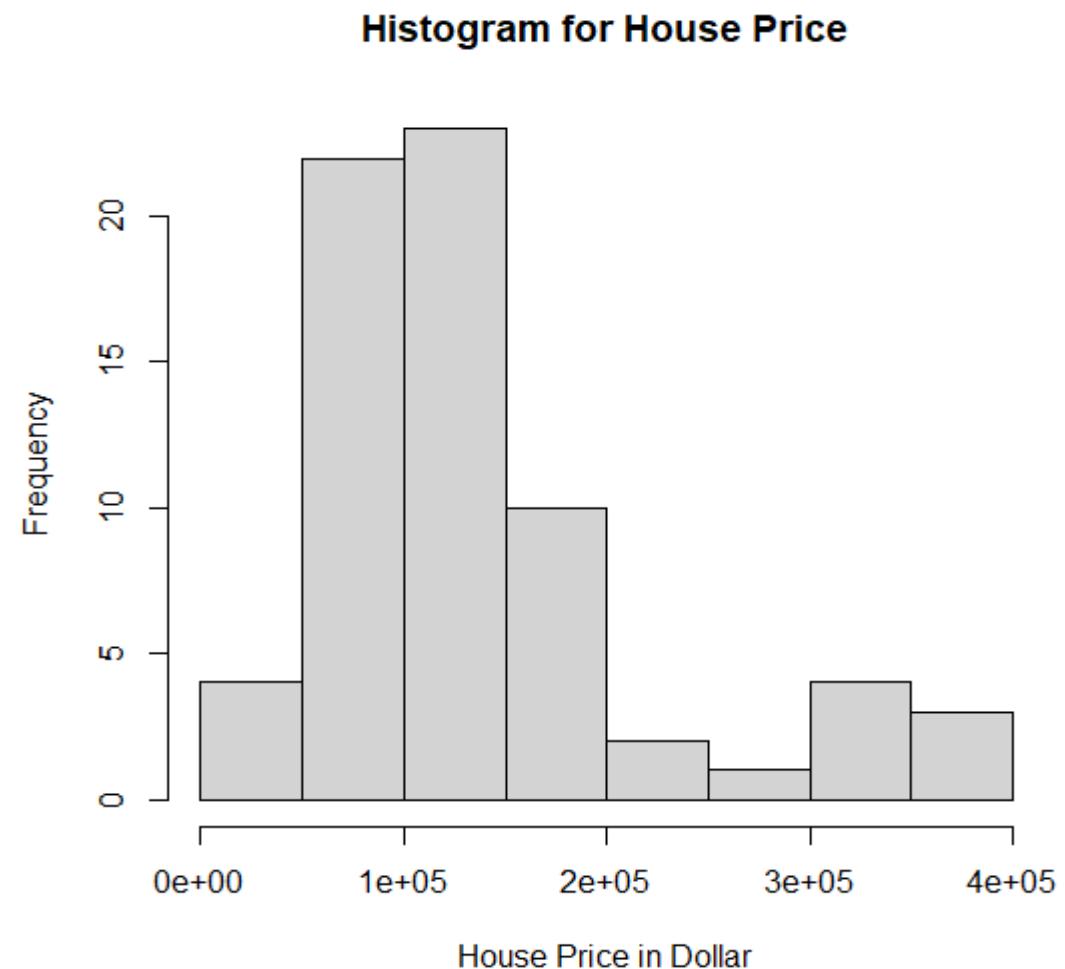
- Most of houses have the brick exterior
- About the same number of houses have the frame or other exteriors.



# Histogram

- A **histogram** is constructed by drawing rectangles for each class of data. The height of each rectangle is the frequency or relative frequency of the class. The width of each rectangle is the same and the rectangles touch each other.
- We will rely on R function `hist()` to construct the histogram.

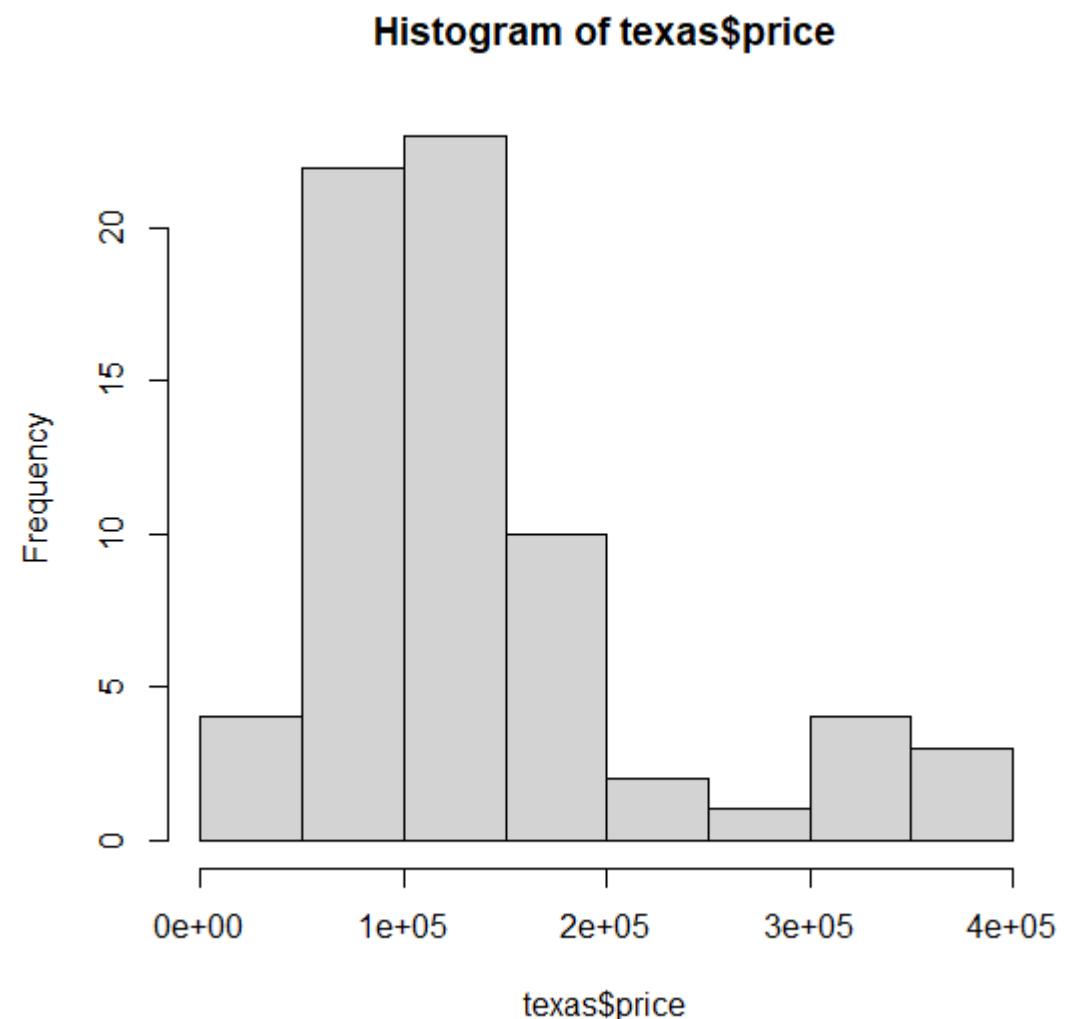
# Histogram from Texas House Data



# Histogram

```
hist(texas$price,  
      breaks = c(0, 50, 100, 150, 200, 250, 300, 350, 400) *  
      1000,  
      right = FALSE,  
      main = "Histogram for House Price",  
      xlab = "House Price in Dollar",  
      ylab = "Frequency")
```

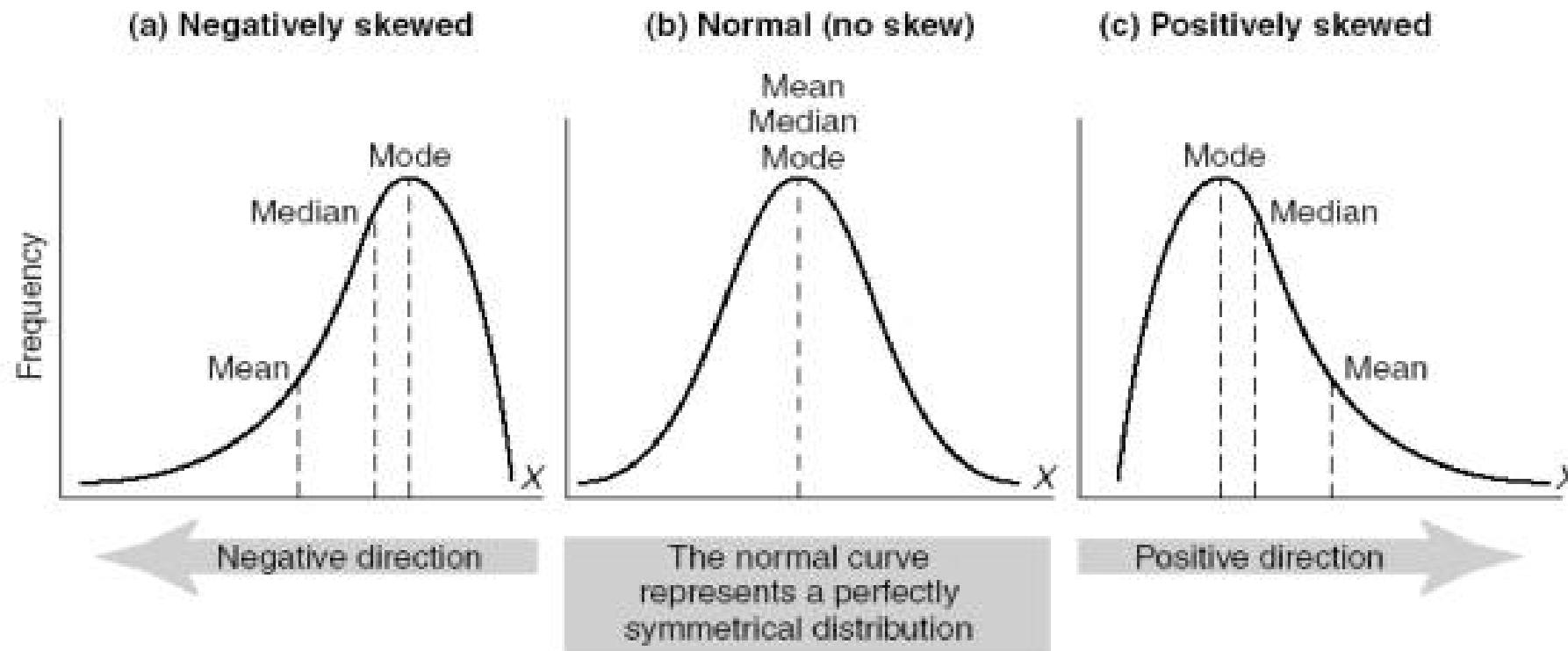
# Histogram from Texas House Data



# Histogram - Interpretation

- We describe the data by the “center” and the “spread”.
  - What is a typical value (“center”) of the data?
  - What is the variability (“spread”) of the data?
- Do the data follow some **patterns**?
  - Symmetric
  - Skewed-left (left-skewed, negatively skewed, or left-tailed)
    - More values on the right side and the tail on the left side is longer
  - Skewed-right (right-skewed, positively skewed, or right-tailed)
    - More values on the left side and the tail on the right side is longer
- Are there **multiple peaks** – multiple peaks suggest a mixture of populations

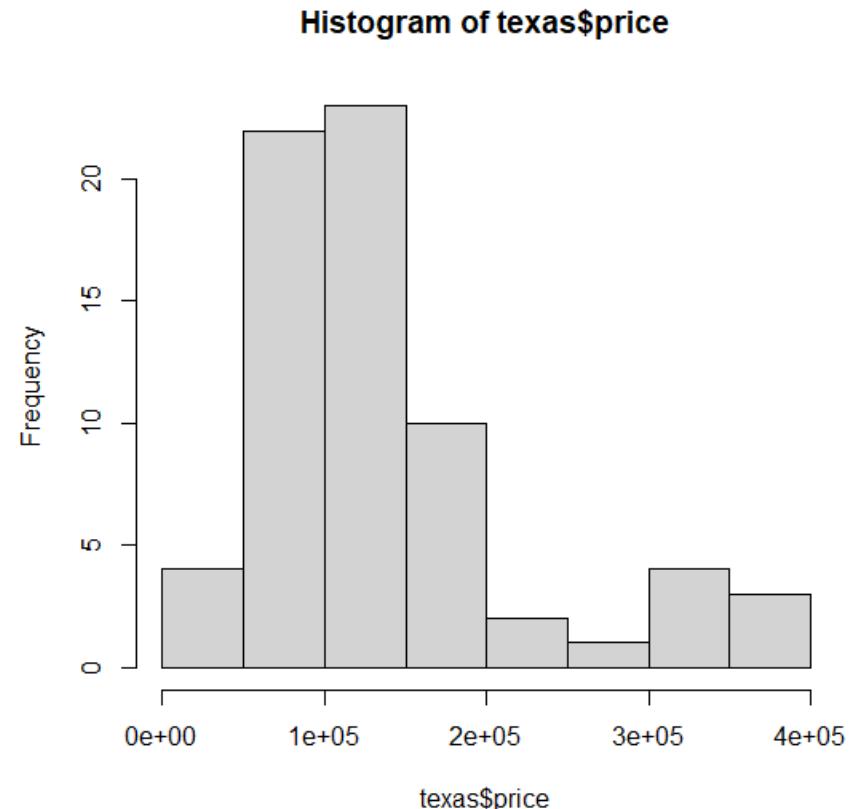
# Shape of the Data



■ FIGURE 15.6 Examples of normal and skewed distributions

# Histogram – House Price

- **Center:** Most of houses have the price between 50k and 250k
- **Spread:** The house price ranges from 30k to around 400k.
- **Patterns:** The data is not symmetric. The data is right-skewed.
  - This meaning that most houses have a lower price (less than 250k) but a few houses are much more expensive (greater than 350k).
- **Multiple peaks:** There is one peak around 200k.



# Descriptive Statistics - Mean

- **Definition** – The **mean (sample mean)** is the sum of all the observed values divided by the number of values.
- Let  $y_1, \dots, y_n$  denote a sample of interest. The mean (sample mean), denoted by  $\bar{y}$ , is given by
$$\bar{y} = \frac{1}{n}(y_1 + y_2 + \dots + y_n).$$
- Mean is a **Measure of Location or Center Tendency**.

# Descriptive Statistics – Sample Variance

- **Definition** – The **sample variance**, denoted by  $s^2$  is defined by

$$s^2 = \frac{1}{n-1} [(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2]$$

- It looks like an “average” of the squared deviations from the sample mean.
- Note that we use  $n - 1$  instead of  $n$ .
- Sample variance is a measure of **dispersion** which is the extent to which a distribution is stretched or squeezed.

# Descriptive Statistics – Sample Variance

- Another formula for sample variance,  $s^2$ , is

$$\begin{aligned}s^2 &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \\ &= \frac{1}{n-1} (\sum_{i=1}^n y_i^2 - n\bar{y}^2).\end{aligned}$$

- **Definition** – The **standard deviation** of a set of observed values is defined to be the positive square root of the variance. In other words, the sample standard deviation is:  
 $s = \sqrt{s^2}$ .

# Mean and Standard Deviation

- **Usefulness of the Mean and Standard Deviation**
  - Interval ( $\text{mean} \pm 1^*\text{SD}$ ) contains approximately 68% of observations
  - Interval ( $\text{mean} \pm 2^*\text{SD}$ ) contains approximately 95% of observations
  - Interval ( $\text{mean} \pm 3^*\text{SD}$ ) contains virtually all of the observations

# Sample Mean/Variance – A Toy Example

- Find sample mean/variance for data: {1, 5, 4, 9, 6}.

$$\sum y_i = 1 + 5 + 4 + 9 + 6 = 25$$

$$\sum y_i^2 = 1^2 + 5^2 + 4^2 + 9^2 + 6^2 = 159$$

$$\sum (y_i - \bar{y})^2 = (-4)^2 + 0^2 + (-1)^2 + 4^2 + 1^2 = 34$$

- Sample mean is:  $\frac{1}{n} \sum y_i = \frac{25}{5} = 5$

- Sample variance is:  $\frac{1}{n-1} (\sum y_i^2 - n\bar{y}^2) = \frac{159 - 5*5*5}{4} = 8.5$

- Sample variance is:  $\frac{1}{n-1} \sum (y_i - \bar{y})^2 = \frac{34}{4} = 8.5$

# Sample Mean/Variance – Packaged Weights

- **Example** – King (1992) discusses the net weights of a nominally 16-oz packaged product. An inspector collected 20 packages and measured their net contents.

|      |      |      |      |      |
|------|------|------|------|------|
| 16.4 | 16.4 | 16.5 | 16.5 | 16.6 |
| 16.7 | 16.2 | 16.4 | 16.4 | 16.5 |
| 16.6 | 16.6 | 16.8 | 16.3 | 16.4 |
| 16.5 | 16.5 | 16.6 | 16.7 | 16.8 |

# Sample Mean/Variance – Packaged Weights

- $\sum_{i=1}^{20} y_i = 16.4 + 16.4 + \cdots + 16.8 = 330.4$
- So sample mean is:  $\bar{y} = 330.4/20 = 16.52$
- $\sum_{i=1}^{20} y_i^2 = 16.4^2 + 16.4^2 + \cdots + 16.8^2 = 5458.68$
- So sample variance is
- $s^2 = (5458.68 - 20 * 16.52 * 16.52)/19 = 0.02484$
- $\sum_{i=1}^{20} (y_i - \bar{y})^2 = (16.4 - 16.52)^2 + \cdots + (16.8 - 16.52)^2 = 0.472$
- So sample variance is  $s^2 = \frac{0.472}{19} = 0.02484$

# Descriptive Statistics – Percentile (Quantile)

- **Definition** – The **median** of a set of observed values is defined to be the middle value when the measurement are arranged from lowest to the highest.
- **Definition** – The  **$p$ th percentile (quantile)** is defined to be that value for which at most  $(p)\%$  of the measurement are less and at most  $(100-p)\%$  of the measurement are greater.

# Descriptive Statistics - Percentile

- **Quartiles, 25%, 50%, 75% percentile**
  - 25% percentile – lower quartile, first quartile ( $Q_1$ )
  - 50% percentile – median, second quartile
  - 75% percentile – upper quartile, third quartile ( $Q_3$ )
- Question: what are 100% percentile and 0% percentile?
  - 100% percentile – maximum or largest
  - 0% percentile – minimum or smallest
- **Definition** – The **interquartile range** is the length of the interval between the 25<sup>th</sup> and 75<sup>th</sup> percentiles.

# Descriptive Statistics - Median

- **Definition** – The **median** of a set of observed values is defined to be the middle value when the measurement are arranged from lowest to the highest.
- Let  $y_1, \dots, y_n$  denote the data and  $\tilde{y}$  denote the median.
- Arrange the data in ascending order:  $y_{(1)} \leq \dots \leq y_{(n)}$
- Median  $\tilde{y} = y_{(\frac{n+1}{2})}$  if  $n$  is odd;  $\tilde{y} = \frac{y_{(\frac{n}{2})} + y_{(\frac{n}{2}+1)}}{2}$  if  $n$  is even.

# A Toy Example

- Suppose we have a data ( $n = 6$ ):  $3, 1, 5, 20, 3, 12$ , then  
 $y_1 = 3, y_2 = 1, y_3 = 5, y_4 = 20, y_5 = 3, y_6 = 12$
- We arrange the data in ascending order:  $1, 3, 3, 5, 12, 20$ , then  
 $y_{(1)} = 1, y_{(2)} = 3, y_{(3)} = 3, y_{(4)} = 5, y_{(5)} = 12, y_{(6)} = 20$
- So  $y_{(1)}$  is the smallest while  $y_{(n)}$  is the largest.
- The median is  $\tilde{y} = \frac{y_{(3)} + y_{(4)}}{2} = \frac{3+5}{2} = 4$

# Example – Wall Thickness of Aircraft Parts

- **Example** – Eck Industries, Inc. Manufacturers cast aluminum cylinder heads that used for liquid-cooled aircraft engines. The thicknesses (in inches) for 18 cylinder heads are given below:

|       |       |       |       |       |       |
|-------|-------|-------|-------|-------|-------|
| 0.223 | 0.193 | 0.218 | 0.201 | 0.231 | 0.204 |
| 0.228 | 0.223 | 0.215 | 0.223 | 0.237 | 0.226 |
| 0.214 | 0.213 | 0.233 | 0.224 | 0.217 | 0.210 |

- Find the median of this data.

# Example – Wall Thickness of Aircraft Parts

- First, sort the data to the ascending order:

|       |       |       |       |       |       |
|-------|-------|-------|-------|-------|-------|
| 0.193 | 0.201 | 0.204 | 0.210 | 0.213 | 0.214 |
| 0.215 | 0.217 | 0.218 | 0.223 | 0.223 | 0.223 |
| 0.224 | 0.226 | 0.228 | 0.231 | 0.233 | 0.237 |

- Since  $n = 18$ , the median is

$$\tilde{y} = \frac{y_{(9)} + y_{(10)}}{2} = \frac{0.218 + 0.223}{2} = 0.2205$$

# Descriptive Statistics - Quartiles

- We will rely on R to calculate quartiles.

# Descriptive Statistics with R

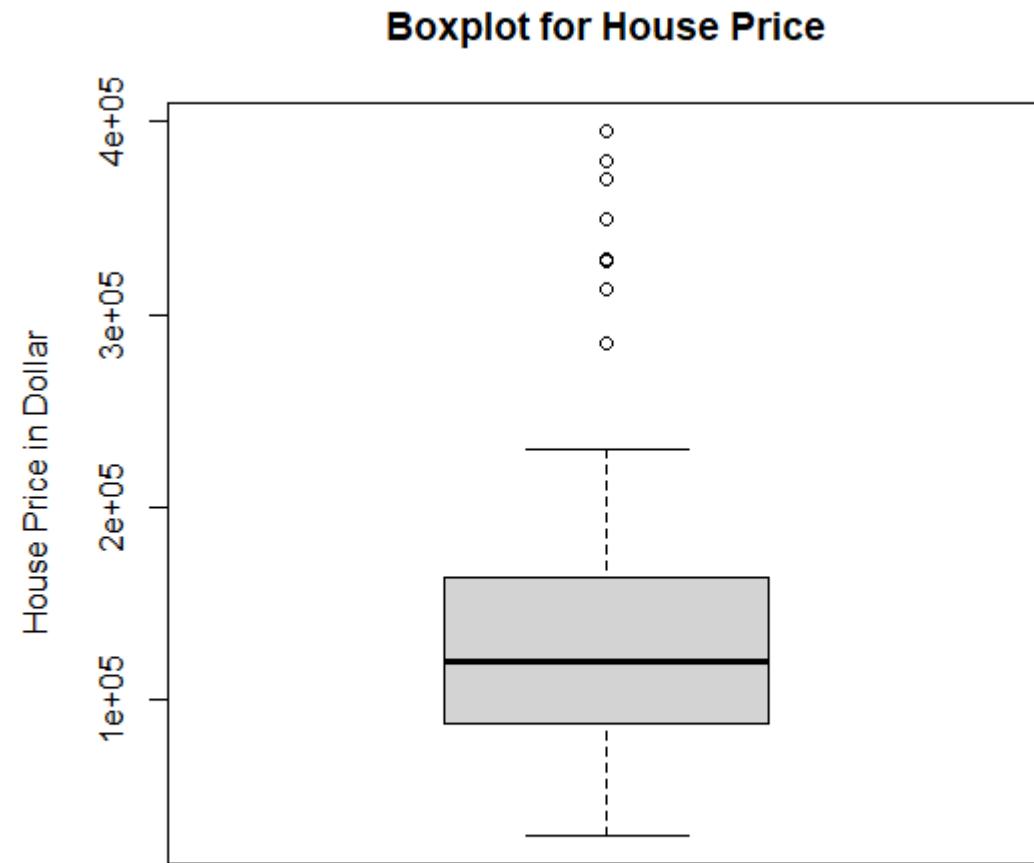
- Let use the data for the age from NORC
- Smallest and largest: `min()` and `max()`
- Sample mean: `mean()`
- Sample variance/standard deviation: `var()` and `sd()`
- Median: `median()`
- Quantile: `quantile()`

# Boxplots

The boxplot provides the analysis:

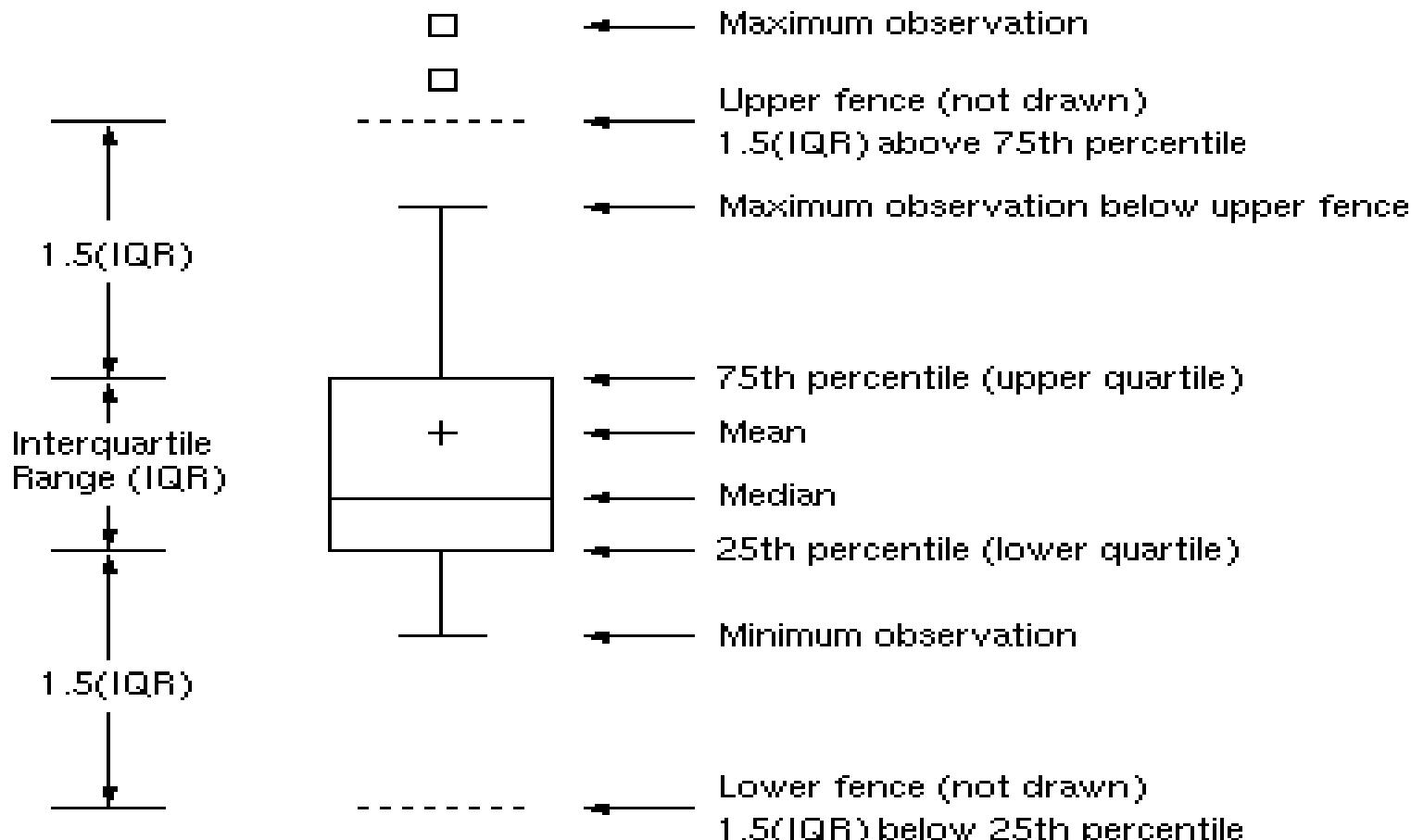
- The center of the data set;
- Where most of the data fall;
- The spread of the unquestionably ‘good’ data;
- Possible outliers;
- And more.....
- Again, we introduce steps to draw boxplots but rely on R.

# A Boxplot from Texas House Data



# Schematics of Boxplot

- Schematic Box-and-Whiskers Plot



# Seven Steps for Boxplots

1. Construct a vertical scale, marked clearly, that covers at least of the data.
2. Find the median and the quartiles ( $Q_1$  and  $Q_3$ ).
3. Find the step size:
  - Step =  $1.5 * (Q_3 - Q_1)$  ( $Q_3 - Q_1$  is called **inter-quartile range**)
4. Find the inner fences, which define the bounds for questionably good data. **Upper Inner Fence (UIF)** and **Lower Inner Fence(LIF)** are given by: UIF =  $Q_3 + \text{Step}$  and LIF =  $Q_1 - \text{Step}$ .

# Seven Steps for Boxplots

5. Locate the most extreme data values on or within the inner fence, draw vertical lines at these points and then draw whiskers to connect these points.
6. Find the outer fences for discriminating between mild and extreme outliers. **Upper Outer Fence (UOF)** and **Lower Outer Fence(LOF)** are:

$$UOF = Q_3 + 2 * \text{Step}$$

$$LOF = Q_1 - 2 * \text{Step}$$

# Seven Steps for Boxplots

## 7. Mark possible outliers.

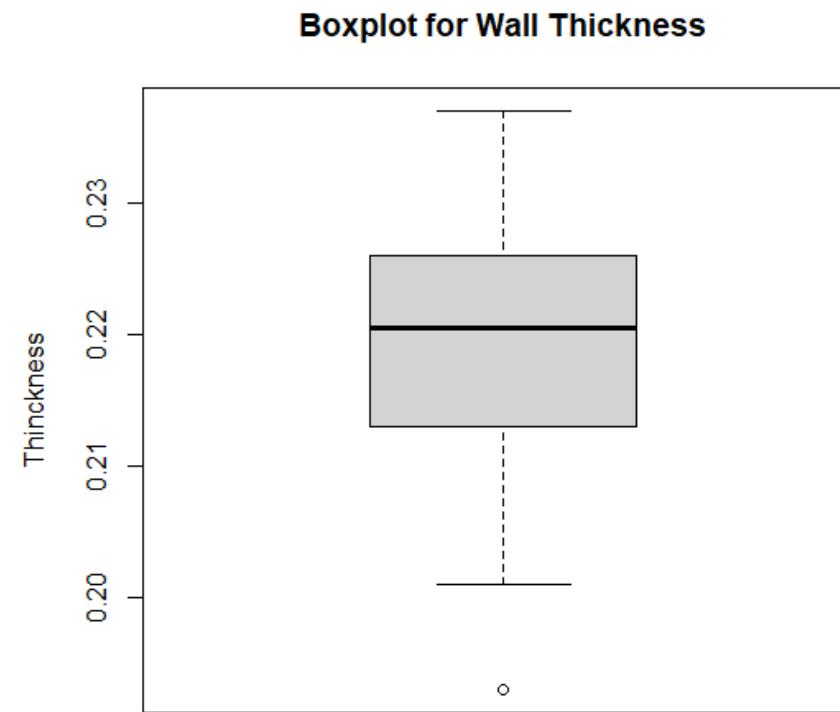
- Use a ‘o’ to denote mild outliers, which are those data points between inner and the outer fences.
- Use a ‘●’ to denote extreme outliers, which are those points on or beyond the outer fences.
- You can use other symbols too.

# Example – Wall Thickness of Aircraft Parts

- **Example** – Eck Industries, Inc. Manufacturers cast aluminum cylinder heads that used for liquid-cooled aircraft engines. The thicknesses (in inches) for 18 cylinder heads are given below:

|       |       |       |       |       |       |
|-------|-------|-------|-------|-------|-------|
| 0.223 | 0.193 | 0.218 | 0.201 | 0.231 | 0.204 |
| 0.228 | 0.223 | 0.215 | 0.223 | 0.237 | 0.226 |
| 0.214 | 0.213 | 0.233 | 0.224 | 0.217 | 0.210 |

# Boxplot – Wall Thickness of Aircraft Parts



# Boxplot – Texas House Data



# Frequency Table for Two Variables

- A two-dimensional contingency table should be used for two qualitative/discrete variables.
- Use R function `table(var1, var2)` for frequency
  - A row is for a level of var1
  - A column is for a level of var 2
- Use R function `prop.table(x, margin)` for frequency
  - The input is the output of `table()`
  - margin: default - all data; = 1 – for rows; = 2 – for columns
  - margin of 1 or 2 is preferred

# Frequency Table for Two Variables

- **Interpretation:** Look at different patterns between rows/columns
- **Frequency Table:** `table(texas$exter, texas$bed)`

| Exterior | Number of Bedrooms |   |    |    |   |
|----------|--------------------|---|----|----|---|
|          | 1                  | 2 | 3  | 4  | 5 |
| Brick    | 0                  | 1 | 30 | 14 | 3 |
| Frame    | 0                  | 1 | 7  | 0  | 0 |
| Others   | 1                  | 1 | 9  | 2  | 0 |

- **Interpretation:** Brick houses seem to have greater number of bedrooms.

# Frequency Table for Two Variables

- **Interpretation:** Look at different patterns between rows/columns
- **Frequency Table:** `prop.table(, margin = 1)`

| Exterior | Number of Bedrooms |       |       |       |       |
|----------|--------------------|-------|-------|-------|-------|
|          | 1                  | 2     | 3     | 4     | 5     |
| Brick    | 0.000              | 0.014 | 0.438 | 0.203 | 0.043 |
| Frame    | 0.000              | 0.014 | 0.101 | 0.000 | 0.000 |
| Others   | 0.014              | 0.014 | 0.130 | 0.029 | 0.000 |

- **Frequency Table:** `prop.table(, margin = 2)`

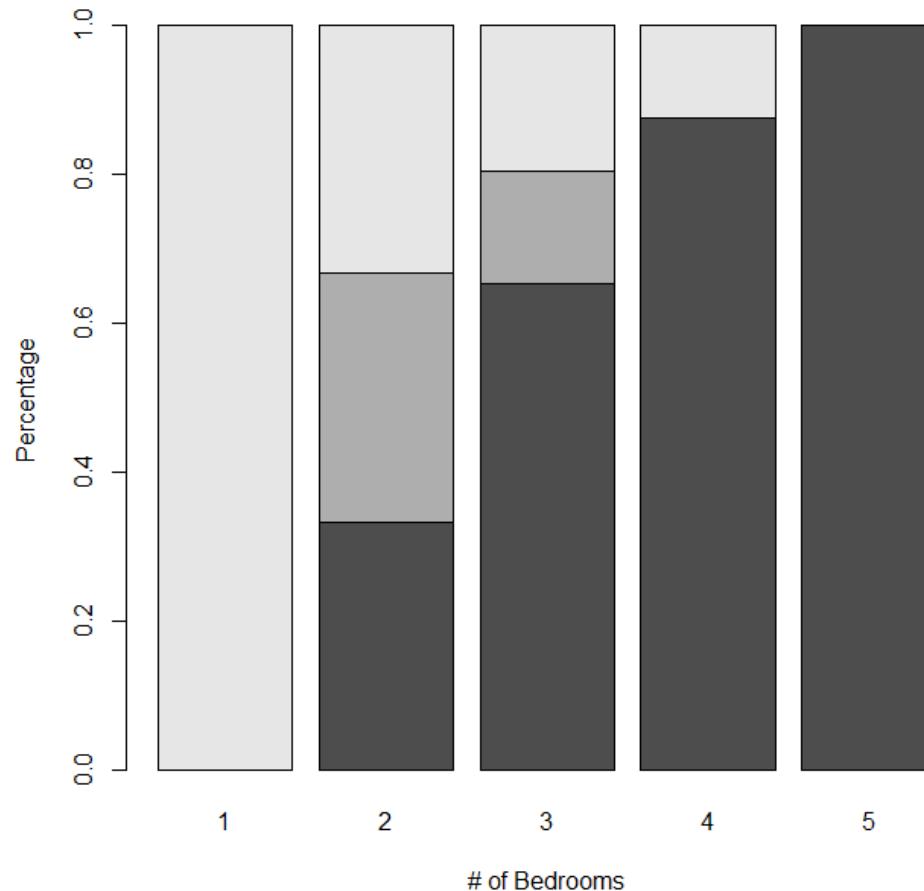
| Exterior | Number of Bedrooms |       |       |       |       |
|----------|--------------------|-------|-------|-------|-------|
|          | 1                  | 2     | 3     | 4     | 5     |
| Brick    | 0.000              | 0.333 | 0.652 | 0.875 | 1.000 |
| Frame    | 0.000              | 0.333 | 0.152 | 0.000 | 0.000 |
| Others   | 1.000              | 0.333 | 0.197 | 0.125 | 0.000 |

# Bar Plot for Two Variables

- **Interpretation:** Look at different patterns between rows/columns
- Use R function `barplot()`
  - Input should be a two-dimensional table from `prop.table()`
  - y-axis: each level of the column variable
  - Bars are stacked according to levels of the row column
  - For `prop.table()`, `margin = 2` should be used
- R program:

```
a <- table(texas$exter, texas$bed)
b <- prop.table(a, margin = 2)
barplot(b)
```

# Bar Plot for Two Variables

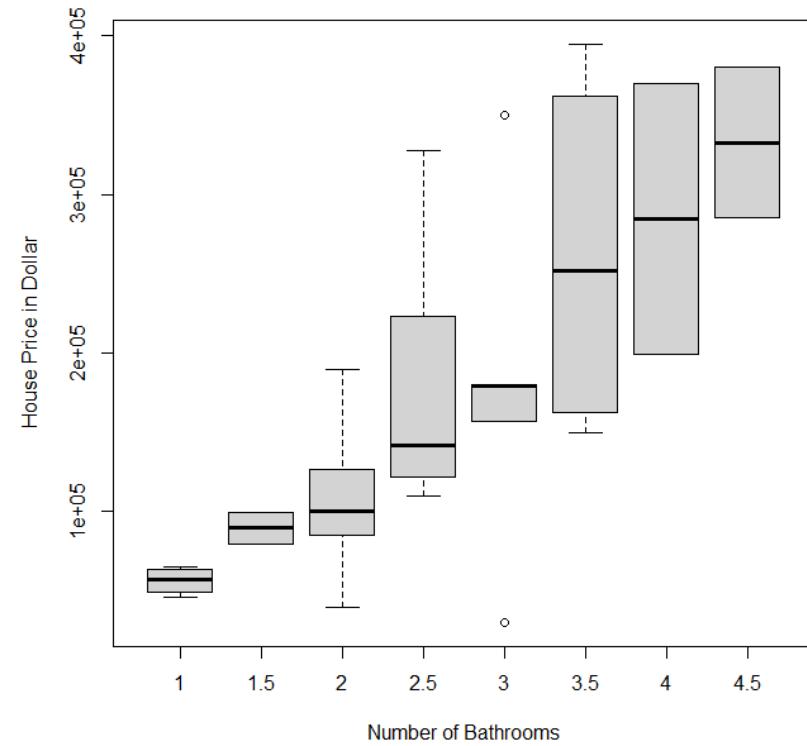


# Boxplots and Scatter Plots

- Boxplots for one qualitative/discrete variable and one quantitative variable
  - Use R function `boxplot()`
  - **Interpretation:** focus on difference/trend of means and difference of variability
- Scatter for two quantitative variables
  - Use R function `plot()`
  - **Interpretation:** focus on patterns/trends/variabilities

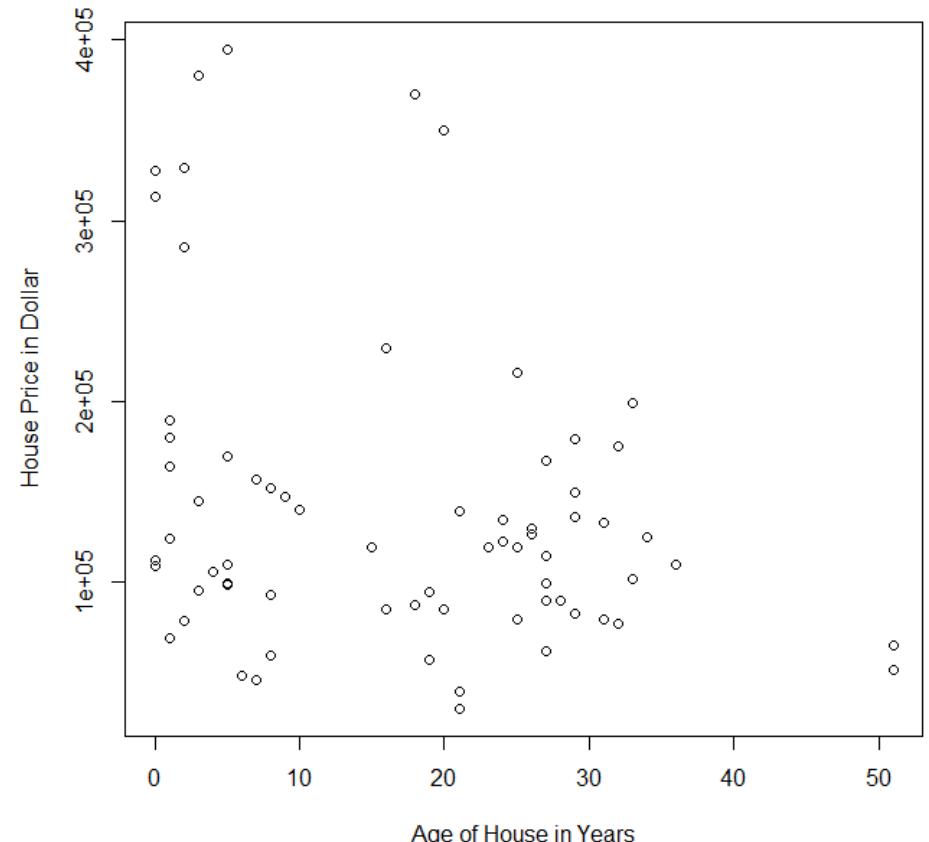
# Boxplots for Two Variables

- R code: `boxplot(price ~ bath, data = texas)`
  - First variable should be quantitative (y-axis)
  - Second variable should be qualitative or discrete (x-axis)
  - Use `data` argument for data frame used in the plot
  - **Interpretation**
    - Price increases with number of bathrooms.
    - The price for house with more than 3 bathrooms shows more variability.



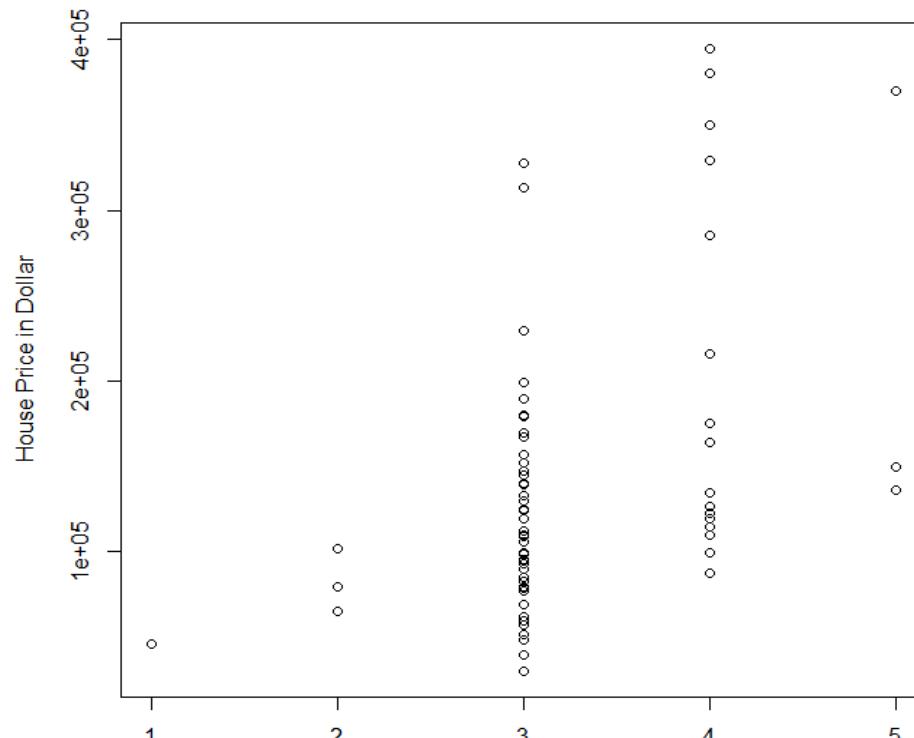
# Scatter Plots for Two Variables

- R code: `plot(price ~ age, data = texas)`
  - First variable (y-axis) and second variable (x-axis) should be chosen carefully
  - Use `data` argument for data frame used in the plot
  - **Interpretation**
    - Some new houses have quite high prices
    - Prices for most houses are not related to their ages
    - Price for newer houses seem to have a wider range

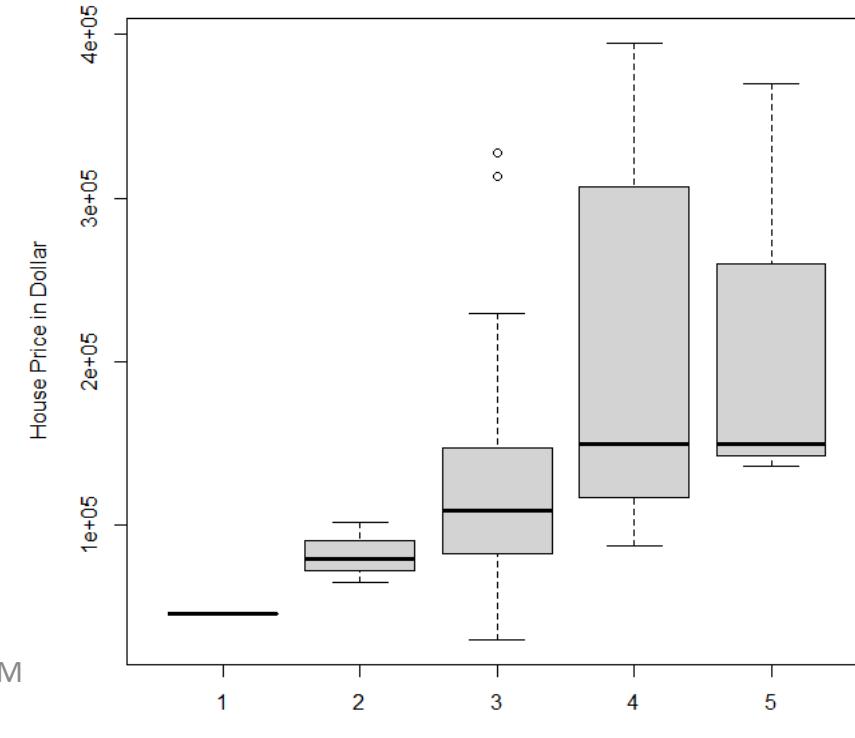


# Scatter Plots for Two Variables

- R code: `plot(price ~ bed, data = texas)` (boxplot is better here)
- House price increases with the number of bedrooms.
- Price for houses with more bedroom seem to have a wider range



A5701 – Statistical M



# Preview – Populations, Samples, Statistical Inference

- **Definition** – The **population** is the values of one or more variables for the entire collection of units relevant to a particular study
- **Population Parameters** – mean and variance, unknown, must be estimated from samples
- **Estimates** – the descriptive measures from samples, can reflect the population parameters, different from different sets of samples, how good an estimate is measured by “sampling error”
- **Statistics** – In this book, it is considered as the same as the estimate. In statistical theory, it refers to the function of a sample, which is either a random variable or random vector

# Data Collection

- **Goal** – make statements about population according to samples
- Random sampling or some more advanced probability sampling is the appropriate way to collect data. In this book, we assume all samples are from “simple random sampling”
- **Definition** – The **simple random sampling** is a sampling scheme that each possible sample of the specified size has an equal chance of occurring
- Random sampling can be difficult to implement in practice
- Convenience samples are dangerous (Be careful)
- Sample size and power calculation

# Chapter Summary

- **Statistics, Data, Sample, Population**
- **Variable** – quantitative, qualitative
- **Variable** - nominal, ordinal, discrete, continuous
- **Table** – Frequency table for one or two variables
- **Statistics** – mean, variance, standard deviation, largest, smallest, median, quartile
- **Graphic** – boxplot, histogram, scatter plot, etc.

# Writing Report

- Use appropriate tables and figures to summarize data
- Do not directly copy tables or output from R output unless you are instructed to do so, do some edits (e.g., add descriptions, appropriate row and/or column names, effective digit)
- Discrete variables – report frequency and percentage
- Continuous variables – mean, variance or standard deviation

# MA5701: Statistical Methods

Chapter 2 : Probability and Sampling Distributions

Kui Zhang, Mathematical Sciences

# Examples for Probability

- **Example 2.1 (Estimate the cost involved in replacing parts)** – defective screws can be produced at two points in a production line, which must be removed and replaced. We would like to estimate the total cost involved with about 1000 parts manufactured.

**Table 2.1 Summary of Defective Screws**

| Point in the Production Line | Proportion of Parts Having Defective Screws | Cost of Replacement |
|------------------------------|---------------------------------------------|---------------------|
| 1                            | 0.008                                       | \$0.23              |
| 2                            | 0.004                                       | \$0.69              |

# Examples for Probability

- **Example:** Estimate the number of fishes in a pond (Capturing and Re-capturing method) – interest in total number of fish ( $N$ ), captured  $M$  fishes, mark them and put them back, capture  $K$  fishes and find  $L$  marked. How to use  $K$ ,  $L$ , and  $M$  to estimate  $N$ ?
- **Example:** Lottery – what is the probability to win lottery if you buy one, or two, or ten tickets?

# Population and Sample

- **Definition 1.2** - A **population** is a data set representing the entire entity of interest.
- **Definition 1.3** - A **sample** is a data set consisting of a portion of a population. (Obtained in a way to represent the population)

# Variables – Texas House Data

| Obs | Zip | Age | Bed | Bath | Size | Lot   | Exter | garage | fp | Price |
|-----|-----|-----|-----|------|------|-------|-------|--------|----|-------|
| 1   | 3   | 21  | 3   | 2    | 951  | 64904 | Other | 0      | 0  | 30000 |
| 3   | 4   | 7   | 1   | 1    | 676  | 54450 | Other | 2      | 0  | 46500 |
| 5   | 1   | 51  | 3   | 1    | 1186 | 10857 | Other | 1      | 0  | 51500 |
| 7   | 3   | 8   | 3   | 2    | 1368 | .     | Frame | 0      | 0  | 56990 |
| 9   | 1   | 51  | 2   | 1    | 1176 | 6259  | Frame | 1      | 1  | 65500 |

Data: 69 families in a midsized city in east Texas. This is only part of it.

# Population and Sample

- **Population** – its characteristic is generally described by the parameter.
  - Usually denoted by Greek letters, such as  $\alpha, \beta, \mu$ , etc.
  - It is generally unknown.
  - It is primary of interest of statistical inference and is estimated from sample.
- **Sample** – its characteristic is generally described by the statistics.
  - Usually denoted by alphabetic letters, such as  $x, y, z, p$ , etc.
  - It can be calculated from the data and is considered as known once the data is collected.
  - It is used to estimate the population parameter.

# Parameter and Statistics

- **Definition 2.1** – A **parameter** is a quantity describes a particular characteristic of the distribution of a variable from a population.
- **Definition 2.2** – A **statistic** is a quantity calculated from data that described a particular characteristic of the sample.
- A statistic and a parameter are very similar. They are both descriptions of a characteristic. For example, “In average, more than 5% of MTU graduate students in School of Forestry take MA5701 – Statistical Methods, in last three years”.

# Parameter and Statistics

- The difference between them is that a statistic describes a **sample** while a parameter describes an entire **population**.
  - **Parameter** – a numerical value describing some characteristic of a *population*.
  - **Statistic** – a numerical value describing some characteristic of a *sample*.
- For the example on the last slide, is 5% a statistic or parameter?
- The answer of this depends on the context – need to know the population first.

# Parameter or Statistic?

Determine whether the underlined value is a parameter or a statistic.

- In a survey conducted in the town of Atherton, 25% of adult respondents reported that they had been involved in at least one car accident in the past ten years. Here 25% is a \_\_\_\_\_

– Statistic

– Parameter

# Parameter or Statistic?

Determine whether the underlined value is a parameter or a statistic.

- 27.2% of the mayors of cities in a certain state are from minority groups. Here 27.2% is a \_\_\_\_\_

- Statistic
- Parameter

# Variables – Texas House Data

| Obs | Zip | Age | Bed | Bath | Size | Lot   | Exter | garage | fp | Price |
|-----|-----|-----|-----|------|------|-------|-------|--------|----|-------|
| 1   | 3   | 21  | 3   | 2    | 951  | 64904 | Other | 0      | 0  | 30000 |
| 3   | 4   | 7   | 1   | 1    | 676  | 54450 | Other | 2      | 0  | 46500 |
| 5   | 1   | 51  | 3   | 1    | 1186 | 10857 | Other | 1      | 0  | 51500 |
| 7   | 3   | 8   | 3   | 2    | 1368 | .     | Frame | 0      | 0  | 56990 |
| 9   | 1   | 51  | 2   | 1    | 1176 | 6259  | Frame | 1      | 1  | 65500 |

Data: 69 families in a midsized city in east Texas. This is only part of it.

# Parameter or Statistic?

Determine whether the underlined value is a parameter or a statistic. From the Texas House data,

- Among 69 houses selected, about 33.3% of houses have a price from \$100000 to \$150000. Here the percentage 33.3% is a \_\_\_\_\_

– Statistic

– Parameter

# Statistical Inference

- **Definition 2.3 - Statistical inference** is the process of using sample statistics to make decisions about population parameters (probability distribution).
- **Examples of Statistical Inference**
  - The use of sample mean to estimate the population mean. For example, what is the average GRE score of MTU graduate students?
  - Decide if population mean is greater than a certain value (Hypothesis Testing using sample mean and sample variance). For example, is the average GRE score of MTU students greater than the national average?

# Sample Space and Event

- **Definition 2.4** – An **experiment** is any process that yields an observation.
- **Definition 2.5** – An **outcome** is a specific result of an experiment.
- Definition – The **sample space** is the combination of all possible outcomes of an experiment, denoted by  $S$ .
- **Definition 2.6** – An **event** is a combination of outcomes having some special characteristic of interest. In other words, the event is a subset of the sample space.

# Sample Space and Event

- **Example:** Toss a coin.
  - The sample space is  $S = \{H, T\}$ .
- **Example:** The experiment consists two steps. First a coin is flipped. If it is a tail, then a die is tossed. If the outcome is head, then the coin is flipped again.
  - The sample space is  $S = \{T1, T2, T3, T4, T5, T6, HT, HH\}$
- **Example:** The life of a light bulb.
  - The sample space is  $S = \{x: x \geq 0\} = [0, \infty)$
- **Example:** we toss a coin and stop until we get a head. The outcome is the number of tosses.
  - The sample space is  $S = \{1, 2, \dots\}$

# Sample Space and Event

- **Example:** Toss two coins.
  - The sample space is  $S = \{HH, HT, TH, HH\}$ .
  - An event can be  $A = \{HH, HT, TH\}$  – at least one head.
- **Example:** Toss two coins. The number of heads is the outcome.
  - The sample space is  $S = \{0,1,2\}$ .
- **Example:** Toss two dice.
  - The sample space is  $S = \{(1,1), (1,2), (1,3), \dots, (6,6)\}$
  - An event can be  $A = \{(5,6), (6,5), (6,6)\}$  – sum is at least 11
- **Example:** A bus arrives at between 10:00am to 11:00am.
  - The sample space is  $S = (10:00\text{am}, 11:00\text{am})$

# Probability

- **Definition – Probability** is the *measure of chance (likelihood)* that an event will occur.
- **Definition of Probability** - A rigorous definition here is difficult, can be considered as a “long-rang relative frequency” or “percentage”.
- **Example:** if we want to know the probability of getting a head when a coin is tossed, we can repeat the experiments many times, record the number and calculate percentage of heads obtained.
- **Example:** we can use the percentage of light blubs last more than 200 hours as probability of that a light bulb lasts more than 200 hours.

# Properties of Probability

- **Probability** is the *measure of chance (likelihood)* that an event will occur.
  - Can its value be negative? **No**
  - Can its value be equal to 0? **Yes**
  - Can its value be equal to 1? **Yes**
  - Can its value be greater than 1? **No**
- Its value will be in the range from 0 to 1 (including 0 and 1).
- If  $A$  is an event, then the probability of  $A$  is denoted by  $\Pr(A)$  or  $P(A)$ . And we have  $0 \leq \Pr(A) \leq 1$ .

# Probability of Equally Likely Sample Space

- If  $A$  is an event, then the probability of  $A$  is denoted by  $\Pr(A)$ . We have

$$\Pr(A) = \frac{\text{Size of the event } A}{\text{Size of the sample space } S}$$

- In this case, size refers to the appropriate measure of chance. Essentially,  $\Pr(A)$  represents size of the event  $A$  relative to size of sample space  $S$ .
- Size can be calculated using counting methods. This formula assumes that all the elements in  $S$  have the same probability (equally likely to happen).

# A Simple Example – Fair Die

- **Example** – Roll a *fair* die and record the number obtained.
- What are the possible outcomes?  $1, 2, 3, 4, 5, 6$
- What is the sample space?  $S = \{1, 2, 3, 4, 5, 6\}$
- Is  $A = \{1, 3\}$  an event and what is  $\Pr(A)$ ? Yes.  $\Pr(A) = 1/3$
- Is  $B = \{1, 3, 5\}$  an event and what is  $\Pr(B)$ ? Yes.  $\Pr(B) = 1/2$
- Is  $C = \{1\}$  an event and what is  $\Pr(C)$ ? Yes.  $\Pr(C) = 1/6$
- Is  $D = \{\}$  an event and what is  $\Pr(D)$ ? Yes.  $\Pr(D) = 0$
- Is  $E = \{1, 2, 3, 4, 5, 6\}$  an event and what is  $\Pr(E)$ ? Yes.  $\Pr(E) = 1$
- Define an event  $F$  = “the number is even” –  $F = \{2, 4, 6\}$ .  $\Pr(F) = 1/2$
- How many different events do we have for this experiment?  $64 = 2^6$

# Example – Maintenance of Spinning Machines

- **Example** – A major manufacturer of textile fibers has several spinning “plants” at a single location. The central maintenance shop provides support for all major repairs and overhauls. Plant 2A has 60 spinning parts while plant 3 has 18 spinning parts. Assume that each part (in plant either 2A or 3) is equally likely to require maintenance (have problem).
- What is the probability that a service request is from Plant 3?

# Example – Maintenance of Spinning Machines

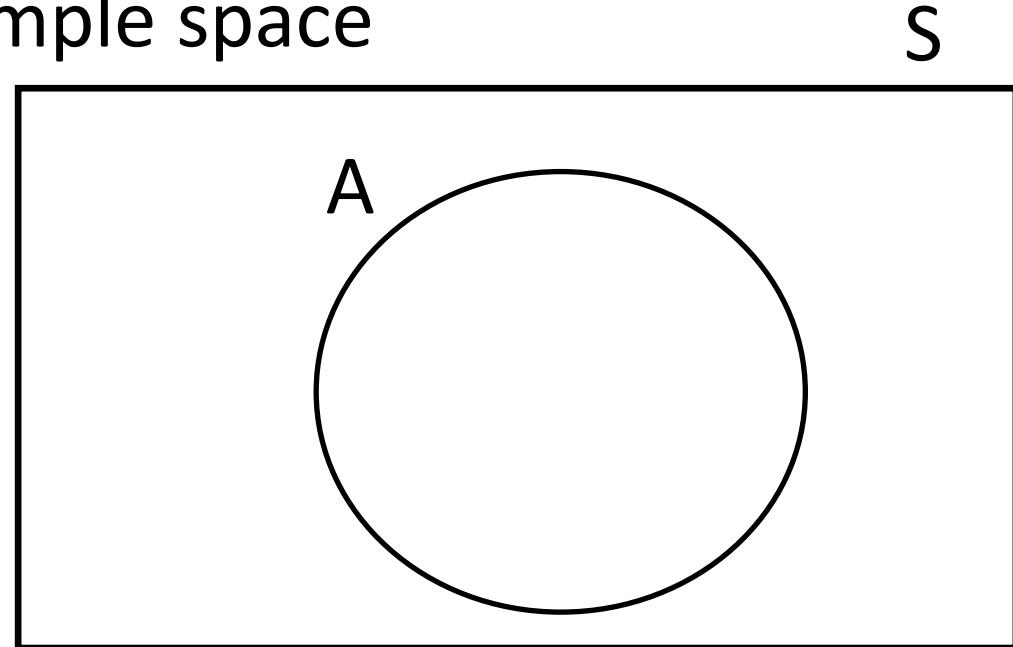
- Sample Space  $S = \{1, 2, 3, \dots, 78\}$
- Define the event  $A$  as the spinning parts of Plant 3 need the service. Then  $A = \{61, \dots, 78\}$ .

$$\Pr(A) = \frac{\text{Size of } A}{\text{Size of } S} = \frac{18}{78} = \frac{3}{13}$$

- How to calculate the probability if we do not think each spinning part is equally likely to require maintenance?
  - Best way is to use basic rules of probability.

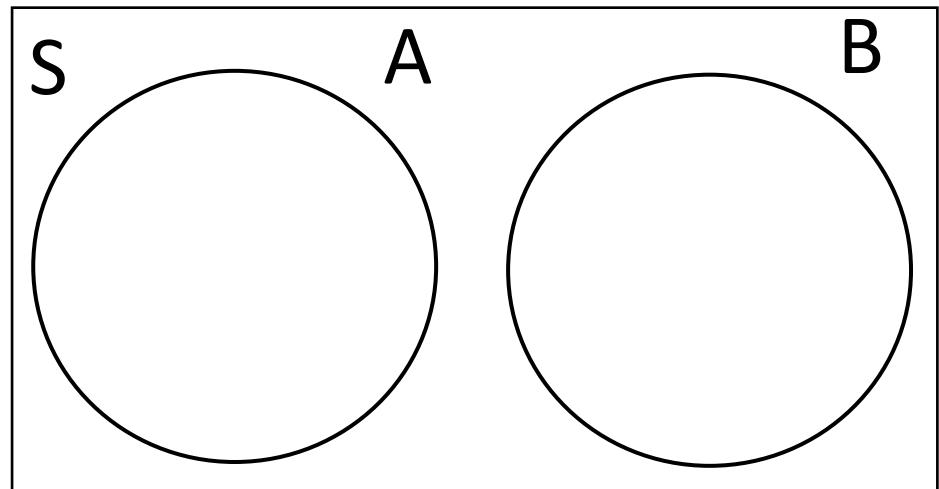
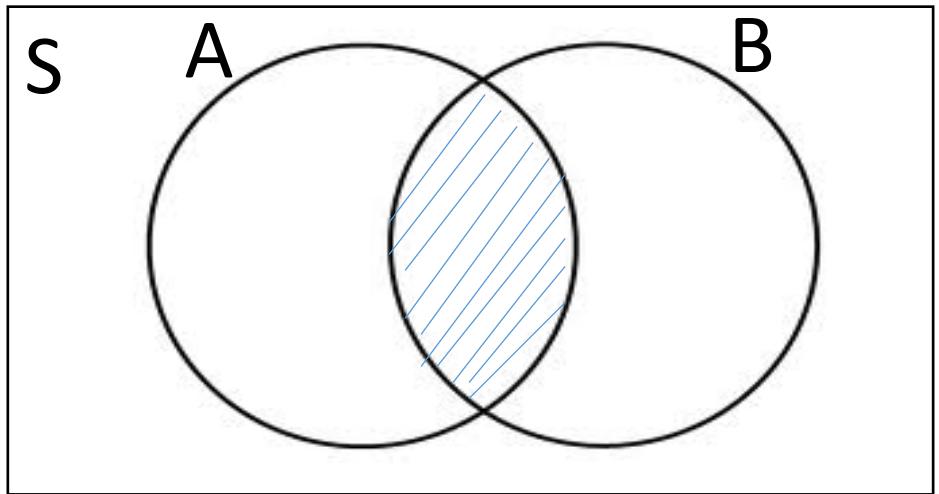
# Events and Venn Diagram

- We can use *Venn diagram* to:
  - Illustrate the relationship between multiple events.
  - Help us to calculate the corresponding probability.
  - Rectangle: represent the sample space
  - Other shapes inside: event
- An example of Venn diagram:



# Event Relation: Intersection

- **Intersection:**  $A \cap B$  (Both  $A$  and  $B$  or simply  $A$  and  $B$ )
- If the intersection of  $A$  and  $B$  is empty,  $A \cap B = \emptyset$ , then  $A$  and  $B$  are called **mutually exclusive (Definition 2.8)**.

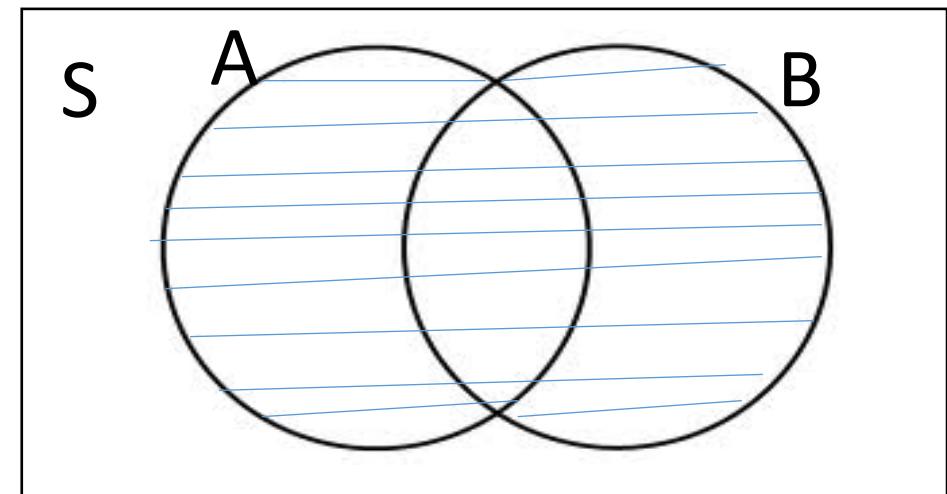


# Event Relationship – Mutually Exclusive

- **Definition 2.8** – If two events cannot occur simultaneously, that is, one “excludes” the other, then the two events are said to be **mutually exclusive** ( in other words, if two events are mutually exclusion, then the intersection of them is an empty set).
- **Example:** If we toss a coin, the outcome {Head} and {Tail} are mutually exclusive since you can not get both a head and tail at the same time.
- **Example:** Raining at Chicago today and raining at Houghton today are NOT mutually exclusive.

# Event Operation: Union

- **Union:**  $A \cup B$  ( $A$  or  $B$ )
- **Union:** in other words, is  $A$  alone or  $B$  alone or  $A$  and  $B$  both



# Event Operation: Complement

- **Complement** of  $A$ :  $A^c$  or  $\bar{A}$  (Not  $A$ )

$$A \cap A^c = \emptyset$$

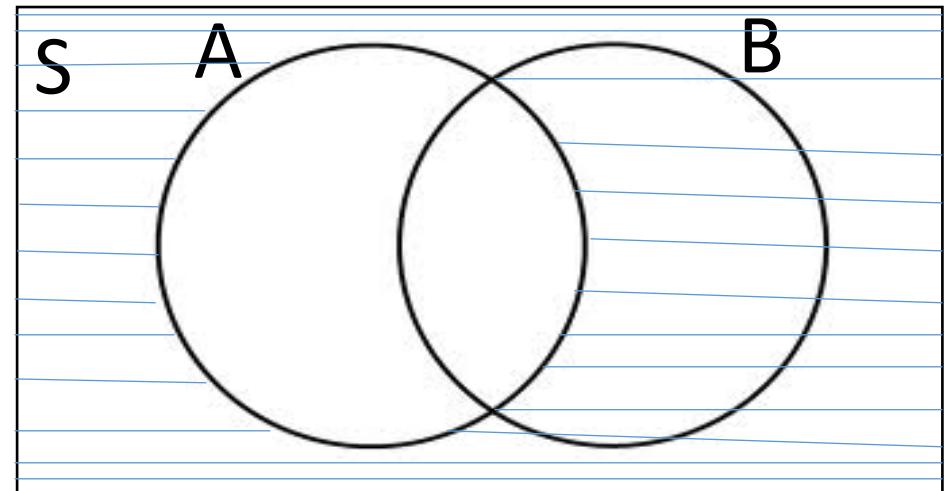
$$A \cup A^c = S$$

- For complement of intersection:

$$(A \cap B)^c = A^c \cup B^c$$

- For complement of union:

$$(A \cup B)^c = A^c \cap B^c$$



# Event Relationship – Complement

- **Definition 2.9** – The **complement** of an outcome or event  $A$  is the occurrence of any event or outcome that precludes  $A$  from happening (also, the union of an event and its complement is an event with all possible outcomes).
- **Example:** There are at least two students in my class have the same birthday. Its **complement** is
  - All students have different birthday.

# Event Relationship - Independent

- **Definition 2.10** – Two events  $A$  and  $B$  are said to be **independent** if the probability of  $A$  occurring is in no way affected by event  $B$  having occurred or vice versa.
- **Example:** Toss two coins. The outcome from the first toss is independent of the outcome from the second toss.
- **Example:** Are a person's study attitude to study MA5701 and his grade in MA5701 independent?

# Rules for Probability Calculation

- **Probability of Empty Set:**  $\Pr(\emptyset) = 0$
- **Probability of An Event with All Outcomes:**  $\Pr(S) = 1$
- **Probability of Intersection of Two Events (Independent):**  
$$\Pr(A \text{ and } B) = \Pr(A \cap B) = \Pr(A) * \Pr(B)$$
- **Probability of Union of Two Events (General Rule):**  
$$\Pr(A \text{ or } B) = \Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$
- **Probability of Union of Two Events (Mutually Exclusive):**  
$$\Pr(A \text{ or } B) = \Pr(A \cup B) = \Pr(A) + \Pr(B)$$
  
$$\Pr(A \cup A^c) = 1 = \Pr(A) + \Pr(A^c)$$

# A Simple Example – Die Example

- **Example** – Roll a fair die and record the number obtained.
- The sample space is  $S = \{1,2,3,4,5,6\}$
- Is  $A = \{2, 4, 6\}$  and  $\Pr(A) = 1/2$ , why?
- Is  $B = \{1, 2, 3, 4\}$  and  $\Pr(B) = 2/3$ , why?
- Actually,  $A$  – the event that an outcome is an even number
- Actually,  $B$  – the event that the number is at most 4.
- Find the following probability (**Exercise**):  
 $\Pr(A \cap B)$ ;  $\Pr(A \cup B)$

# Calculation of Probability – Die Example

- The sample space is equally likely.
- $\Pr(A \cap B) = \Pr(\{4,6\}) = 1/3$
- $\Pr(A \cup B) = \Pr(\{1,2,3,4,6\}) = 5/6$
- $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) = \frac{1}{2} + \frac{2}{3} - \frac{1}{3} = 5/6$

# Calculation of Probability – More Examples

- **Example:** Suppose that 75% all investors invest traditional annuities and 45% of them invest in the stock market. If 85% invest in the stock market and/or traditional annuities, what percent invest in both?
- $A$  = invest in traditional annuities;  $B$  = invest in stock market;
- $\Pr(A) = 0.75$ ;  $\Pr(B) = 0.45$ ;  $\Pr(A \cup B) = 0.85$
- Invest in both:  $\Pr(A \cap B) = \Pr(A) + \Pr(B) - \Pr(A \cup B) = 0.35$
- Invest only in annuities :  $\Pr(A \cap B^c) = \Pr(A) - \Pr(A \cap B) = 0.40$

# Calculation of Probability – Exercises

- **Exercises:** Suppose that in a community of 400 adults, 300 bike or swim or do both, 160 swim, and 120 swim and bike, how many of them bike? For an adult selected at random from this community, what is the probability that he/she bikes?
- **Details:** skipped. Discussed in the class.
- **Solution:** 260 of them bike.
- **Solution:** the probability is  $260/400 = 0.65$ .

# Probability of Intersection

- If  $A$  and  $B$  are independent, then  $\Pr(A \cap B) = \Pr(A) \Pr(B)$
- If  $A$  and  $B$  are not independent, then the calculation of  $\Pr(A \cap B)$  can be difficult and tricky – we may need to calculate probability of the event  $A \cap B$  directly or rely on  $\Pr(A)$ ,  $\Pr(B)$ , and  $\Pr(A \cup B)$ .

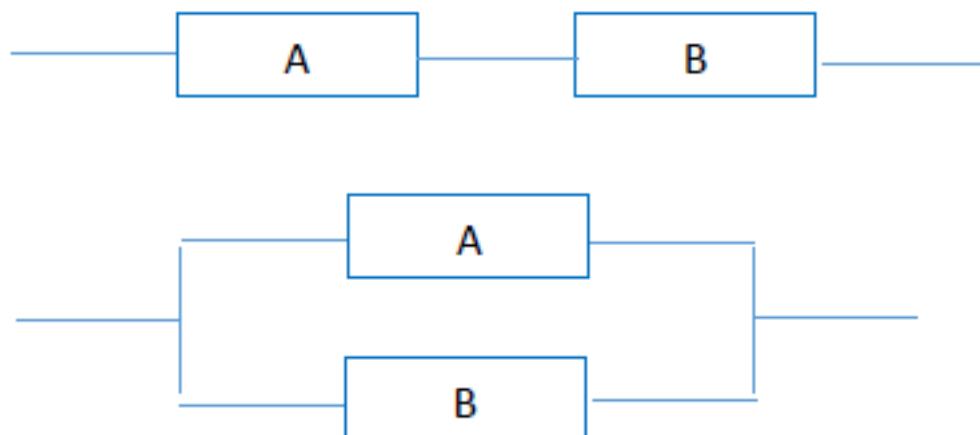
# Independence – Dice Example

- Roll a fair dice two times, what is the probability getting two 6s?
- **Solution:**

$$\begin{aligned} & \Pr(\text{First Dice is 6 and Second Dice is 6}) \\ &= \Pr(\text{First Dice is 6} \cap \text{Second Dice is 6}) \\ &= \Pr(\text{First Dice is 6}) * \Pr(\text{Second Dice is 6}) \\ &= \frac{1}{6} * \frac{1}{6} = \frac{1}{36} \end{aligned}$$

# Probability Calculation – System Reliability

- **Reliability of System:**  $A$  failing is independent of  $B$  failing  
 $\Pr(A) = \Pr(A \text{ failing}) = 0.01$  and  $\Pr(B) = \Pr(B \text{ failing}) = 0.02$
- **Serial System:**  $\Pr = \Pr(A \text{ or } B) = \Pr(A) + \Pr(B) - \Pr(A \text{ and } B) = 0.0298$
- **Parallel System:**  $\Pr = \Pr(A \text{ and } B) = \Pr(A) * \Pr(B) = 0.0002$



# Exercise - Engineering Design Project Bids

- **Example** – Suppose a project manager of an engineering design firm bids on two projects. The chance that a bid will be accepted is:  $\Pr(A) = 0.3$ ;  $\Pr(B) = 0.8$ . We can reasonably assume that bids are *independent*.
  1. What is the probability that at least one of bids is accepted?
  2. What is the probability that only bid  $A$  is accepted?

# Solution - Engineering Design Project Bids

1. First, we have  $\Pr(A \cap B) = \Pr(A) * \Pr(B) = 0.3 * 0.8 = 0.24$
2. What is the probability that at least one of bids is accepted?
  - $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) = 0.3 + 0.8 - 0.24 = 0.86$
3. What is the probability that only bid A is accepted?
  - $\Pr(A \cap B^c) = \Pr(A) - \Pr(A \cap B) = 0.3 - 0.24 = 0.06$

# Exercise – Number of Defects

- **Exercise:** Two factories monitor their production line and find the following distributions of the number of defects:

|           | 0    | 1    | 2    | $\geq 3$ |
|-----------|------|------|------|----------|
| Factory 1 | 0.92 | 0.04 | 0.03 | 0.01     |
| Factory 2 | 0.95 | 0.03 | 0.02 | 0.00     |

- We further assume that the number of defects from factory 1 is independent of the number of defects from factory 2.
- $A$  = at most two defects from factory 1;  $B$  = 1 defect from factory 2;
- $C$  = total of 4 defects from Factory 1 and Factory 2
- Calculate:  $\Pr(A)$ ;  $\Pr(B)$ ;  $\Pr(A \cap B)$ ;  $\Pr(A \cup B)$ ;  $\Pr(A \cup C)$ ;  $\Pr(A \cap C)$

# Random Variables

- **Definition 2.11** – A **random variable** is a rule that assigns a numerical value to an outcome of interest (in other words, a function from outcomes to real numbers).
- If  $Y$  is a random variable, then the **cumulative distribution function (cdf)**, denoted by  $F(y)$ , is given by

$$F(y) = \Pr(Y \leq y)$$

for all real numbers  $y$  ( $-\infty < y < \infty$ ).

- Properties of  $F(y)$ :  
 $0 \leq F(y) \leq 1$ ; non-decreasing;  $F(-\infty) = 0$ ;  $F(\infty) = 1$ .

# Random Variable - Example

- **Example 2.4** - Distribution of number of heads of tossing two fair coins.
  - We know the sample space is  $\{HH, HT, TH, TT\}$ .
  - We assign 0 to  $\{TT\}$
  - We assign 1 to  $\{HT\}$  or  $\{TH\}$
  - We assign 2 to  $\{HH\}$ .
  - Now the random variable take values 0, 1, 2.
- What is the *cdf* of this random variable?
  - $\Pr(Y \leq 1.5) = \Pr(Y = 0) + \Pr(Y = 1) = 0.75$
  - $\Pr(Y \leq 1) = \Pr(Y = 0) + \Pr(Y = 1) = 0.75$
  - $\Pr(Y < 1) = \Pr(Y = 0) = 0.25$

# Random Variables – Why We Use Them?

- Why we use the random variable instead of the original sample space?
- Answer: data reduction and a better tool to describe the experiments.
- **Example:** suppose that we toss 100 fair coins.
  - The original sample space has  $2^{100}$  elements.
  - If we define the number of heads as the random variable, then we have 101 possible values. Why?
  - Actually, we are more (or only) interested in knowing the number of heads than knowing which tosses result in a head.

# Random Variable – Another Example

- **Example:** The random variable is the life of the light bulb in years.
- For this example, the sample space is  $S = [0, \infty)$ .
- We assign the same value to each value in  $S$ .
- There are many possible choices of cumulative probability function (*cdf*) for this random variable.
- For example, we can use  $F(y) = \Pr(Y \leq y) = 1 - \exp(-\frac{y}{3})$ .
- Based on this *cdf*, we can calculate

$$\Pr(Y > 10) = 1 - \Pr(Y \leq 10) = \exp\left(-\frac{10}{3}\right) = 0.037$$

# Discrete and Continuous Random Variables

- **Definition 2.5** – A **discrete random variable** is one that can take on only a countable number of values.
- **Definition 2.6** – A **continuous random variable** is one that can take on any value in an interval.
- A more formal definition based on the cumulative probability function  $F(y)$ :  $Y$  is a discrete random variable if  **$F(y)$  is a step function** while  $Y$  is a continuous random variable if  **$F(y)$  is a continuous function**.

# Discrete and Continuous Random Variables

- **Example:** Toss two dice and the random variable is the sum of two numbers. This is a discrete random variable.
- **Example:** The lifetime of a light bulb. This is a continuous random variable.
- **Note:** Some random variables can be neither discrete nor continuous.

# Exercise – Types of Random Variable

- $Y$  is the aluminum contamination from an area.
  - Continuous random variable.
- $Y$  is the number obtained when we roll a dice.
  - Discrete random variable.
- $Y$  is the number of car accidents in Houghton.
  - Discrete random variable.
- $Y$  is the strength of yarns from a manufacturer.
  - Continuous random variable.

# Discrete Random Variable

- The **probability mass function (pmf)** for a discrete random variable, denoted by  $f(y)$ , is defined by  $f(y) = \Pr(Y = y)$ .
- Two properties of  $f(y)$  are:
  1. Non-negative:  $0 \leq f(y) \leq 1$ .
  2.  $\sum_y f(y) = 1$ .
- Is  $f(y)$  a monotone function?
  - May not be a monotone function.
- The *cdf* can be calculated through *pmf*:  $F(y) = \sum_{t \leq y} f(t)$

# Discrete Random Variable - Example

**Example** – Roll a fair dice, let  $Y$  be the number obtained.

- Is  $Y$  a discrete random variable?
- What are the possible values of  $Y$ ?
- What are the *cdf* and *pmf* of  $Y$ ?
- The *pmf* of  $Y$ :  $f(y) = \frac{1}{6}, y = 1, \dots, 6$
- The *cdf* is a step function. For example,  $F(1) = \Pr(Y = 1) = 1/6$ ,  
 $F(2.5) = \Pr(Y = 1) + \Pr(Y = 2) = \frac{1}{3}$ .
- How about  $F(-1)$  and  $F(10)$ ?  $F(-1) = 0; F(10) = 1$

# Expected Values and Population Mean

- **Definition** – The **expected value** of the discrete random variable is the probability-weighted average of all possible values.
- This value is also called population mean and denoted by  $\mu$ .
- The mathematical formula is based on the possible values and their probability mass function.

$$\mu = E[Y] = \sum_y f(y) * y$$

- Intuitively it is the long-run average value of repetitions of the experiment it represents.

# Discrete Random Variable - Example

- **Example 2.4** - Distribution of number of heads of tossing two fair coins. We know that  $f(0) = 0.25$ ;  $f(1) = 0.50$ ;  $f(2) = 0.25$ .
- The expected value is  $\mu = 0.25 * 0 + 0.50 * 1 + 0.25 * 2 = 1$ .
- This value can be considered as the long-run average value of repetitions of the experiment it represents, what does it mean?
- That means we conduct such experiment (tossing two fair coins) many times, the average number of heads (the sample mean) is (approximately) 1. If we conduct such experiment infinite times, then the average number of heads is 1.

# Population Variance

- **Definition** – The **variance** of the discrete random variable is the probability-weighted average of squared distance of all possible values and its expected value.
- Again, this variance is considered as population variance and can be estimated by sample variance.
- **Population variance** of  $Y$ , denoted by  $\sigma^2$ , is
$$\sigma^2 = \text{var}(Y) = \sum_y f(y) * (y - \mu)^2.$$
- **Population standard deviation** of  $Y$ , denoted by  $\sigma$ , is  $\sigma = \sqrt{\sigma^2}$ , the square root of the population variance.

# Population Variance - Example

- **Example** – Roll a fair dice. let  $Y = 0$  if the number is not greater than 4 and  $Y = 1$  otherwise.
- We know that  $f(0) = 2/3$  and  $f(1) = 1/3$ , why?
- Population mean is:

$$\mu = 0 * f(0) + 1 * f(1) = 1/3$$

- Population variance is:

$$\sigma^2 = (0 - \frac{1}{3})^2 * f(0) + \left(1 - \frac{1}{3}\right)^2 * f(1) = \frac{1}{9} * \frac{2}{3} + \frac{4}{9} * \frac{1}{3} = \frac{2}{9}$$

# Exercise – Mean and Variance

- **Exercise:** A factory monitor their production line and find the following distributions of the number of defects:

| # Defects   | 0    | 1    | 2    | 3    |
|-------------|------|------|------|------|
| Probability | 0.92 | 0.04 | 0.03 | 0.01 |
|             |      |      |      |      |

- Let  $Y$  be the number defects from this factory, then  $Y$  is a discrete random variable.
- Find  $\Pr(Y \geq 1)$ ,  $\Pr(Y \leq 2)$ , and  $\Pr(Y \leq 10)$ .
- Find the mean and variance of  $Y$ .
- **Solution:** in class.

# Common Discrete Distributions

What important things do you need to memorize for a discrete distribution (random variable)?

- The **possible values** that the random variable can take.
- The **probability mass function**.
- The **population mean** and **population variance**.

You also want to understand -

- What type of experiments can be described by such a random variable.

# Discrete Uniform Distribution

- A random variable  $Y$  has a *discrete uniform*  $(1, N)$  distribution if  $\Pr(Y = k) = \frac{1}{N}$ ,  $k = 1, \dots, N$ .
- Population mean:  $\mu = E[Y] = \frac{N+1}{2}$ .
- Population variance:  $\sigma^2 = \text{var}(Y) = \frac{(N+1)*(N-1)}{12}$
- Examples:  $Y$  = number obtained from rolling a fair dice

$$Y \sim U(1, 6), E[Y] = 3.5, \text{var}(Y) = \frac{35}{12} = 2.917$$

# Bernoulli Distribution

- A random variable  $Y$  has a  $Bernoulli(p)$  distribution if  $f(1) = \Pr(Y = 1) = p$  and  $f(0) = \Pr(Y = 0) = 1 - p$ .
- Population mean:  $\mu = E[Y] = p$ .
- Population variance:  $\sigma^2 = \text{var}(Y) = p(1 - p)$ .
- Bernoulli distribution is used to describe a trial (an experiment) with two outcomes: success and failure. And  $p$  is the probability of success.

# Bernoulli Distribution - Examples

- Toss a coin, define  $Y = 1$  if a head obtained and  $Y = 0$  otherwise. Now  $p$  is the probability to get a head.
- In an election poll, define  $Y = 1$  if candidate A gets a vote and  $Y = 0$  otherwise. Now  $p$  is the probability that candidate A gets a vote.
- In a manufacturer, define  $Y = 1$  if a battery plate meets specifications and  $Y = 0$  otherwise. Now  $p$  is the probability that a battery plate meets specifications.

# Binomial Distribution

- A random variable  $Y$  has a *Binomial distribution*,  $\text{Binomial}(n, p)$ , if its *pmf* is

$$f(y) = \Pr(Y = y) = \binom{n}{y} p^y (1 - p)^{n-y}, y = 0, \dots, n.$$

- If  $Y$  is the number of successes from  $n$  independent identical Bernoulli trials, then  $Y$  has  $\text{Binomial}(n, p)$ .
- If  $Y \sim \text{Bernoulli}(p)$  then  $Y \sim \text{Binomial}(1, p)$ .
- Population mean:  $\mu = E[Y] = np$ .
- Population variance:  $\sigma^2 = \text{var}(Y) = np(1 - p)$ .

# Binomial Distribution

Consider an experiment satisfies these four conditions:

1. The number of experiments  $n$  must be fixed.
2. The trials are independent.
3. Each trial has two mutually exclusive outcomes, success or failure.
4. The probability of success ( $p$ ) is fixed for each trial of the experiment.

Then  $Y$  = the number of successes has  $\text{Binomial}(n, p)$ .

# Binomial Distribution - Examples

Determine if the following distributions are binomial:

1. Toss a coin 100 times,  $Y$  = number of heads.
2. Roll a fair dice 10 times,  $Y$  = the number of rolls getting number greater than 4.
3. Suppose our class has 55 students and 30 of them are female students. I randomly select 5 students.  $Y$  = number of females.
4. The probability to get a defected bulb is 3%. There are 1000 bulbs produced in one day.  $Y$  = number of defected bulbs.

**Answer:** 1, 2, and 4 have a binomial distribution, while 3 does not. The distribution of 3 is hypergeometric distribution.

# Example – Nonconforming Brick

- **Example** - Marcucci (1985) reports that a brick manufacturer classifies the product into conforming and nonconforming groups. Historically, nonconforming bricks counts 5% of all bricks. A facility makes 25 bricks per hour. Let  $Y$  = number of nonconforming bricks. Then  $Y$  has Binomial(25, 0.05).
- Probability of two non-conforming bricks is:

$$\Pr(Y = 2) = \binom{25}{2} 0.05^2 (1 - 0.05)^{25-2} = 0.2305.$$

- Probability of at least one non-conforming brick is:

$$\Pr(Y \geq 1) = 1 - \Pr(Y = 0) = 1 - \binom{n}{0} 0.05^0 (1 - 0.05)^{25-0} = 0.7226.$$

# Example – Nonconforming Brick

- Consider the population mean,  $\mu = np = 25 * 0.05 = 1.25$ .
- Consider the population variance:
$$\sigma^2 = np(1 - p) = 25 * 0.05 * 0.95 = 1.1875$$
- This analysis assumes that:
  - Each brick is independent of the other.
  - The historical probability of a brick being defective holds true for the particular hour.

# Exercises – Find Distribution and Probability

- **Exercise:** A typesetter, on the average, makes one error in every 500 words typeset. A typical page contains 300 words. What is the probability that there will be no more than two errors in five pages?
- **Solution:** Let  $Y$  be the number of errors, then  
$$Y \sim Binomial(1500, 0.002).$$

$$\Pr(Y \leq 2) = \sum_{y=0}^2 \binom{1500}{y} 0.002^y (1 - 0.002)^{1500-y} = 0.4230$$

# Exercises – Find Distribution

- A manufacturer receives a lot of 100 parts from a vendor. The probability of the part is defective is 0.01. Let  $Y$  be the number of parts that are defective.
  - $Y \sim \text{Binomial}(100, 0.01)$
- A manufacturer of water filters for refrigerators monitors the process for defective filters (the filter leaks). Historically, this process averages 5% defective filters. 20 filters are randomly chosen for testing and let  $Y$  be the number of defective filters.
  - $Y \sim \text{Binomial}(20, 0.05)$

# Exercises – Find Distribution and Probability

- A standard drug is known to be effective in 80% of the cases in which it is used. The drug is tested on 100 patients and found to be effective in 85 cases. What distribution can be used to calculate the probability that we can observe at least 85 effective cases among 100 tested patients?
  - $Y \sim \text{Binomial}(100, 0.80)$
  - $\Pr(Y \geq 85) = \binom{100}{85} 0.80^{85} 0.20^{100-85} + \dots + \binom{100}{100} 0.80^{100} 0.20^{100-100}$ 
$$= \sum_{i=85}^{100} \binom{100}{i} 0.80^i 0.20^{100-i} = 0.1285$$

# Continuous Random Variable

- A **continuous random variable** is one that can be any possible real value over some interval.
- For any random variable (continuous, discrete, or other types), we have the cumulative probability function  $F(y) = \Pr(Y \leq y)$ .
- Properties of  $F(y)$ :  
$$0 \leq F(y) \leq 1; \text{ increasing}; F(-\infty) = 0; F(\infty) = 1$$
- A more formal definition based on  $F(y)$ :  $Y$  is a discrete random variable if  **$F(y)$  is a step function** while  $Y$  is a continuous random variable if  **$F(y)$  is a continuous function**.

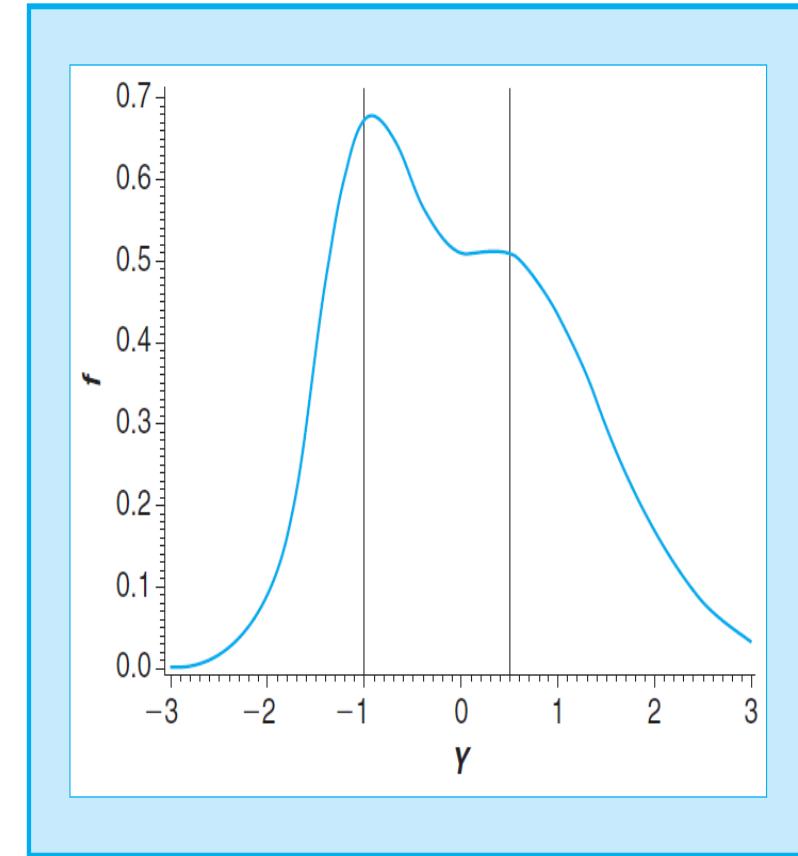
# Continuous Random Variables

- Recall for a discrete random variable, we have the **probability mass function (pmf)**  $f(y) = \Pr(Y = y)$  and  $F(y) = \sum_{t \leq y} f(t)$ .
- For a continuous random variable, we have  $\Pr(Y = y) = 0$ .
- The **probability density function (pdf)**  $f(y)$  for a continuous random variable satisfies:

$$F(y) = \Pr(Y \leq y) = \int_{-\infty}^y f(t)dt$$

# Continuous Probability Distribution

- The graph of the distribution is a smooth curve.
- The total area under the curve is 1.
- The area between the curve and horizontal axis from  $a$  to  $b$  is the probability of the random variable taking a value in the interval  $(a, b)$ 
  - The probability of taking a specific value is zero and for continuous random variables, therefore
$$\Pr(a \leq X \leq b) = \Pr(a < X < b)$$
- Finding areas under curves (probability) can be difficult – involving the use of calculus.  
Sometimes, there is no analytical solution.



# Probability Density Function

Properties of *pdf*:

1.  $f(y) \geq 0.$
2.  $\int_{-\infty}^{\infty} f(t)dt = 1 = F(\infty).$
3.  $F(y) = \int_{-\infty}^y f(t)dt$

- Actually, any function that satisfies conditions (1) and (2) can be considered as a valid pdf.
- **Question:** do we have  $f(y) \leq 1$  for all  $y$ ?
  - No. For *pdf*,  $f(y)$  can be greater than 1 for some  $y$ . For *pmf*,  $f(y) \leq 1$ .

# Common Continuous Distributions

What important things do you need to memorize for a continuous distribution (random variable)?

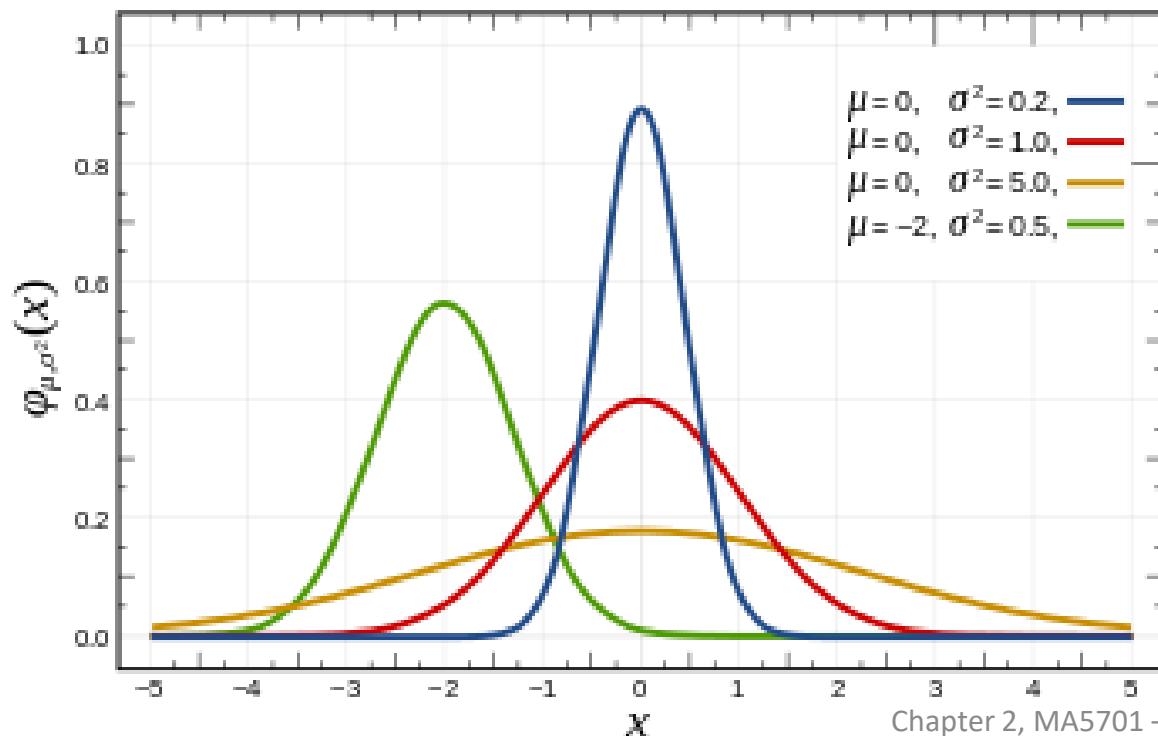
- The **possible values** that the random variable can take.
- The **probability density function**.
- The **population mean** and **population variance**.

You also want to understand -

- What type of experiments can be described by such a random variable.

# Normal Distribution

- The most often commonly continuous probability function.
- Probability density function:  $f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right), -\infty < y < \infty$



# Normal Distribution

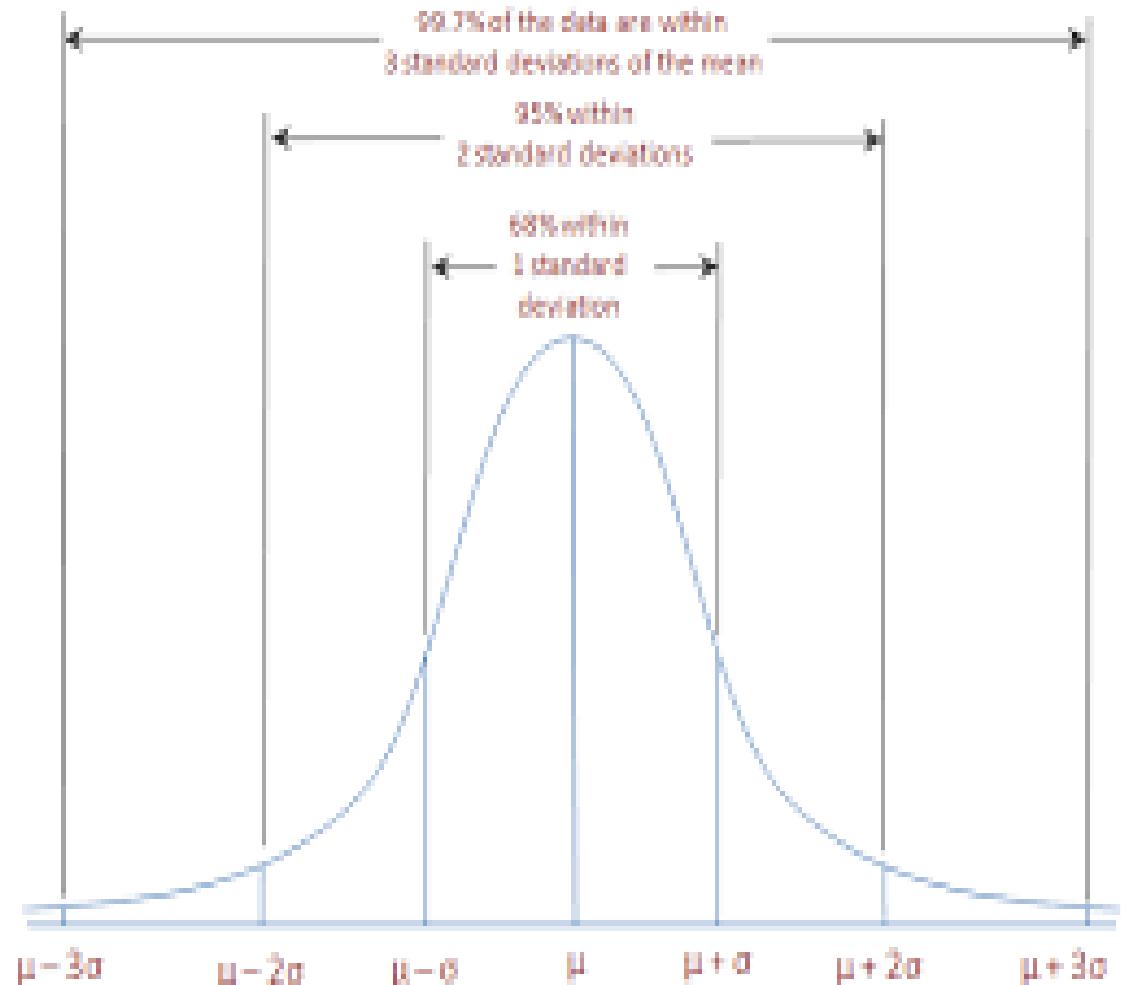
- It is the single most important distribution, which is often described as “bell-shaped” distribution or curve.
- It can be used to model the behavior of many phenomena.
- Under certain conditions, it can be used to model the behavior of averages.
- Many times, we use the simple notation:  $Y \sim N(\mu, \sigma^2)$
- Population mean is:  $\mu$ .
- Population variance and standard deviation are:  $\sigma^2$  and  $\sigma$ .

# Normal Distribution

- It is symmetric at its mean  $\mu$ .
- Mean = median = mode, this is a single peak and the highest point occurs at  $y = \mu$ .
- The area under the curve is 1.
- The area under the curve to the left (right) of  $\mu$  is 0.5.
- The curve never reaches the horizontal axis.

# Normal Distribution – Empirical Rule

- Approximately 68% of data fall within  $\mu \pm \sigma$
- Approximately 95% of data fall within  $\mu \pm 2\sigma$
- Approximately 99.7% of data fall within  $\mu \pm 3\sigma$



# Standard Normal Distribution

- If  $Y \sim N(\mu, \sigma^2)$  and  $\mu = 0$  and  $\sigma^2 = \sigma = 1$ , then  $Y$  has a standard normal distribution.
- In other words, the standard normal distribution is a normal distribution with mean of 0 and population variance (standard deviation) of 1.
- The *cdf* for the standard normal distribution can be found in a table.

# Standard Normal Distribution

- It is the most important normal distribution. Why?
- If  $Y \sim N(\mu, \sigma^2)$ , then we can not analytically integrate to get *cdf*  $F(y)$  - we must rely on numerical integration.
- Fortunately, any normal distribution can be converted to a standard normal distribution.
- So we can summarize the *cdf* of any normal distribution with the *cdf* of the standard normal distribution.

# Normal Distribution: Calculations

- We generally use  $Z \sim N(0,1)$  to represent the standard normal random variable.

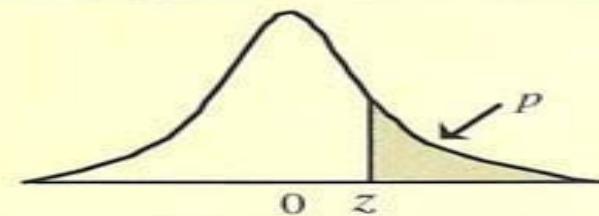
$$\Pr(Z > z) = \Pr(Z < -z) = 1 - \Pr(Z < z)$$

$$\Pr(a < Z < b) = \Pr(Z < b) - \Pr(Z < a)$$

$$\Pr(Z > 0) = \Pr(Z < 0) = 0.5$$

- We rely on normal tables, which has many formats.

# Standard Normal Distribution



| z   | Second decimal place of $z$ |       |       |       |       |       |       |       |       |       |  |
|-----|-----------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|--|
|     | .00                         | .01   | .02   | .03   | .04   | .05   | .06   | .07   | .08   | .09   |  |
| 0.0 | .5000                       | .4960 | .4920 | .4880 | .4840 | .4801 | .4761 | .4721 | .4681 | .4641 |  |
| 0.1 | .4602                       | .4562 | .4522 | .4483 | .4443 | .4404 | .4364 | .4325 | .4286 | .4247 |  |
| 0.2 | .4207                       | .4168 | .4129 | .4090 | .4052 | .4013 | .3974 | .3936 | .3897 | .3859 |  |
| 0.3 | .3821                       | .3783 | .3745 | .3707 | .3669 | .3632 | .3594 | .3557 | .3520 | .3483 |  |
| 0.4 | .3446                       | .3409 | .3372 | .3336 | .3300 | .3264 | .3228 | .3192 | .3156 | .3121 |  |
| 0.5 | .3085                       | .3050 | .3015 | .2981 | .2946 | .2912 | .2877 | .2843 | .2810 | .2776 |  |
| 0.6 | .2743                       | .2709 | .2676 | .2643 | .2611 | .2578 | .2546 | .2514 | .2483 | .2451 |  |
| 0.7 | .2420                       | .2389 | .2358 | .2327 | .2297 | .2266 | .2236 | .2206 | .2177 | .2148 |  |
| 0.8 | .2119                       | .2090 | .2061 | .2033 | .2005 | .1977 | .1949 | .1922 | .1894 | .1867 |  |
| 0.9 | .1841                       | .1814 | .1788 | .1762 | .1736 | .1711 | .1685 | .1660 | .1635 | .1611 |  |
| 1.0 | .1587                       | .1562 | .1539 | .1515 | .1492 | .1469 | .1446 | .1423 | .1401 | .1379 |  |
| 1.1 | .1357                       | .1335 | .1314 | .1292 | .1271 | .1251 | .1230 | .1210 | .1190 | .1170 |  |
| 1.2 | .1151                       | .1131 | .1112 | .1093 | .1075 | .1056 | .1038 | .1020 | .1003 | .0985 |  |
| 1.3 | .0968                       | .0951 | .0934 | .0918 | .0901 | .0885 | .0869 | .0853 | .0838 | .0823 |  |
| 1.4 | .0808                       | .0793 | .0778 | .0764 | .0749 | .0735 | .0721 | .0708 | .0694 | .0681 |  |
| 1.5 | .0668                       | .0655 | .0643 | .0630 | .0618 | .0606 | .0594 | .0582 | .0571 | .0559 |  |
| 1.6 | .0548                       | .0537 | .0526 | .0516 | .0505 | .0495 | .0485 | .0475 | .0465 | .0455 |  |
| 1.7 | .0446                       | .0436 | .0427 | .0418 | .0409 | .0401 | .0392 | .0384 | .0375 | .0367 |  |
| 1.8 | .0359                       | .0351 | .0344 | .0336 | .0329 | .0322 | .0314 | .0307 | .0301 | .0294 |  |
| 1.9 | .0287                       | .0281 | .0274 | .0268 | .0262 | .0256 | .0250 | .0244 | .0239 | .0233 |  |
| 2.0 | .0228                       | .0222 | .0217 | .0212 | .0207 | .0202 | .0197 | .0192 | .0188 | .0183 |  |
| 2.1 | .0179                       | .0174 | .0170 | .0166 | .0162 | .0158 | .0154 | .0150 | .0146 | .0143 |  |
| 2.2 | .0139                       | .0136 | .0132 | .0129 | .0125 | .0122 | .0119 | .0116 | .0113 | .0110 |  |
| 2.3 | .0107                       | .0104 | .0102 | .0099 | .0096 | .0094 | .0091 | .0089 | .0087 | .0084 |  |
| 2.4 | .0082                       | .0080 | .0078 | .0075 | .0073 | .0071 | .0069 | .0068 | .0066 | .0064 |  |
| 2.5 | .0062                       | .0060 | .0059 | .0057 | .0055 | .0054 | .0052 | .0051 | .0049 | .0048 |  |
| 2.6 | .0047                       | .0045 | .0044 | .0043 | .0041 | .0040 | .0039 | .0038 | .0037 | .0036 |  |
| 2.7 | .0035                       | .0034 | .0033 | .0032 | .0031 | .0030 | .0029 | .0028 | .0027 | .0026 |  |

# Standard Normal Distribution

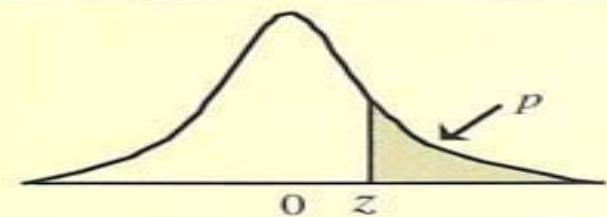
Table 1: Table of the Standard Normal Cumulative Distribution Function  $\Phi(z)$

| $z$  | 0.00   | 0.01   | 0.02   | 0.03   | 0.04   | 0.05   | 0.06   | 0.07   | 0.08   | 0.09   |
|------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| -3.4 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0002 |
| -3.3 | 0.0005 | 0.0005 | 0.0005 | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0003 |
| -3.2 | 0.0007 | 0.0007 | 0.0006 | 0.0006 | 0.0006 | 0.0006 | 0.0006 | 0.0005 | 0.0005 | 0.0005 |
| -3.1 | 0.0010 | 0.0009 | 0.0009 | 0.0009 | 0.0008 | 0.0008 | 0.0008 | 0.0008 | 0.0007 | 0.0007 |
| -3.0 | 0.0013 | 0.0013 | 0.0013 | 0.0012 | 0.0012 | 0.0011 | 0.0011 | 0.0011 | 0.0010 | 0.0010 |
| -2.9 | 0.0019 | 0.0018 | 0.0018 | 0.0017 | 0.0016 | 0.0016 | 0.0015 | 0.0015 | 0.0014 | 0.0014 |
| -2.8 | 0.0026 | 0.0025 | 0.0024 | 0.0023 | 0.0023 | 0.0022 | 0.0021 | 0.0021 | 0.0020 | 0.0019 |
| -2.7 | 0.0035 | 0.0034 | 0.0033 | 0.0032 | 0.0031 | 0.0030 | 0.0029 | 0.0028 | 0.0027 | 0.0026 |
| -2.6 | 0.0047 | 0.0045 | 0.0044 | 0.0043 | 0.0041 | 0.0040 | 0.0039 | 0.0038 | 0.0037 | 0.0036 |
| -2.5 | 0.0062 | 0.0060 | 0.0059 | 0.0057 | 0.0055 | 0.0054 | 0.0052 | 0.0051 | 0.0049 | 0.0048 |
| -2.4 | 0.0082 | 0.0080 | 0.0078 | 0.0075 | 0.0073 | 0.0071 | 0.0069 | 0.0068 | 0.0066 | 0.0064 |
| -2.3 | 0.0107 | 0.0104 | 0.0102 | 0.0099 | 0.0096 | 0.0094 | 0.0091 | 0.0089 | 0.0087 | 0.0084 |
| -2.2 | 0.0139 | 0.0136 | 0.0132 | 0.0129 | 0.0125 | 0.0122 | 0.0119 | 0.0116 | 0.0113 | 0.0110 |
| -2.1 | 0.0179 | 0.0174 | 0.0170 | 0.0166 | 0.0162 | 0.0158 | 0.0154 | 0.0150 | 0.0146 | 0.0143 |
| -2.0 | 0.0228 | 0.0222 | 0.0217 | 0.0212 | 0.0207 | 0.0202 | 0.0197 | 0.0192 | 0.0188 | 0.0183 |
| -1.9 | 0.0287 | 0.0281 | 0.0274 | 0.0268 | 0.0262 | 0.0256 | 0.0250 | 0.0244 | 0.0239 | 0.0233 |
| -1.8 | 0.0359 | 0.0351 | 0.0344 | 0.0336 | 0.0329 | 0.0322 | 0.0314 | 0.0307 | 0.0301 | 0.0294 |
| -1.7 | 0.0446 | 0.0436 | 0.0427 | 0.0418 | 0.0409 | 0.0401 | 0.0392 | 0.0384 | 0.0375 | 0.0367 |

# Normal Distribution: Calculations

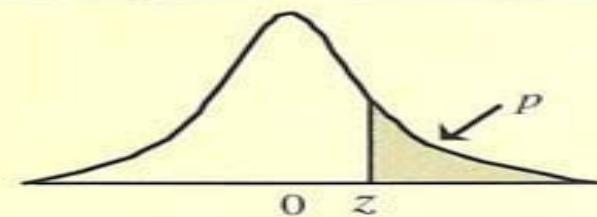
- $\Pr(Z > 2.0) = 0.0228$
- $\Pr(Z < -1) = \Pr(Z > 1) = 0.1587$
- $\Pr(Z < 1.53) = 1 - \Pr(Z > 1.53) = 1 - 0.0630 = 0.937$
- $\Pr(-1 < Z < 1.53) = \Pr(Z < 1.53) - \Pr(Z < -1) = 0.937 - 0.1587 = 0.7783$

# Normal Distribution: $\Pr(Z > 2.0)$



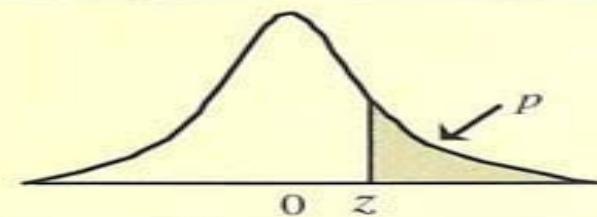
| z   | Second decimal place of z |       |       |       |       |       |       |       |       |       |  |
|-----|---------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|--|
|     | .00                       | .01   | .02   | .03   | .04   | .05   | .06   | .07   | .08   | .09   |  |
| 0.0 | 5000                      | .4960 | .4920 | .4880 | .4840 | .4801 | .4761 | .4721 | .4681 | .4641 |  |
| 0.1 | 4602                      | .4562 | .4522 | .4483 | .4443 | .4404 | .4364 | .4325 | .4286 | .4247 |  |
| 0.2 | 4207                      | .4168 | .4129 | .4090 | .4052 | .4013 | .3974 | .3936 | .3897 | .3859 |  |
| 0.3 | 3821                      | .3783 | .3745 | .3707 | .3669 | .3632 | .3594 | .3557 | .3520 | .3483 |  |
| 0.4 | 3446                      | .3409 | .3372 | .3336 | .3300 | .3264 | .3228 | .3192 | .3156 | .3121 |  |
| 0.5 | 3085                      | .3050 | .3015 | .2981 | .2946 | .2912 | .2877 | .2843 | .2810 | .2776 |  |
| 0.6 | 2743                      | .2709 | .2676 | .2643 | .2611 | .2578 | .2546 | .2514 | .2483 | .2451 |  |
| 0.7 | 2420                      | .2389 | .2358 | .2327 | .2297 | .2266 | .2236 | .2206 | .2177 | .2148 |  |
| 0.8 | 2119                      | .2090 | .2061 | .2033 | .2005 | .1977 | .1949 | .1922 | .1894 | .1867 |  |
| 0.9 | 1841                      | .1814 | .1788 | .1762 | .1736 | .1711 | .1685 | .1660 | .1635 | .1611 |  |
| 1.0 | 1587                      | .1562 | .1539 | .1515 | .1492 | .1469 | .1446 | .1423 | .1401 | .1379 |  |
| 1.1 | 1357                      | .1335 | .1314 | .1292 | .1271 | .1251 | .1230 | .1210 | .1190 | .1170 |  |
| 1.2 | 1151                      | .1131 | .1112 | .1093 | .1075 | .1056 | .1038 | .1020 | .1003 | .0985 |  |
| 1.3 | 0968                      | .0951 | .0934 | .0918 | .0901 | .0885 | .0869 | .0853 | .0838 | .0823 |  |
| 1.4 | 0808                      | .0793 | .0778 | .0764 | .0749 | .0735 | .0721 | .0708 | .0694 | .0681 |  |
| 1.5 | 0668                      | .0655 | .0643 | .0630 | .0618 | .0606 | .0594 | .0582 | .0571 | .0559 |  |
| 1.6 | 0548                      | .0537 | .0526 | .0516 | .0505 | .0495 | .0485 | .0475 | .0465 | .0455 |  |
| 1.7 | 0446                      | .0436 | .0427 | .0418 | .0409 | .0401 | .0392 | .0384 | .0375 | .0367 |  |
| 1.8 | 0359                      | .0351 | .0344 | .0336 | .0329 | .0322 | .0314 | .0307 | .0301 | .0294 |  |
| 1.9 | 0287                      | .0281 | .0274 | .0268 | .0262 | .0256 | .0250 | .0244 | .0239 | .0233 |  |
| 2.0 | 0228                      | .0222 | .0217 | .0212 | .0207 | .0202 | .0197 | .0192 | .0188 | .0183 |  |
| 2.1 | 0179                      | .0174 | .0170 | .0166 | .0162 | .0158 | .0154 | .0150 | .0146 | .0143 |  |
| 2.2 | 0139                      | .0136 | .0132 | .0129 | .0125 | .0122 | .0119 | .0116 | .0113 | .0110 |  |
| 2.3 | 0107                      | .0104 | .0102 | .0099 | .0096 | .0094 | .0091 | .0089 | .0087 | .0084 |  |
| 2.4 | 0082                      | .0080 | .0078 | .0075 | .0073 | .0071 | .0069 | .0068 | .0066 | .0064 |  |
| 2.5 | 0062                      | .0060 | .0059 | .0057 | .0055 | .0054 | .0052 | .0051 | .0049 | .0048 |  |
| 2.6 | 0047                      | .0045 | .0044 | .0043 | .0041 | .0040 | .0039 | .0038 | .0037 | .0036 |  |
| 2.7 | 0035                      | .0034 | .0033 | .0032 | .0031 | .0030 | .0029 | .0028 | .0027 | .0026 |  |

# Normal Distribution: $\Pr(Z < -1)$



| z   | Second decimal place of $z$ |       |       |       |       |       |       |       |       |       |  |
|-----|-----------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|--|
|     | .00                         | .01   | .02   | .03   | .04   | .05   | .06   | .07   | .08   | .09   |  |
| 0.0 | 5000                        | .4960 | .4920 | .4880 | .4840 | .4801 | .4761 | .4721 | .4681 | .4641 |  |
| 0.1 | 4602                        | .4562 | .4522 | .4483 | .4443 | .4404 | .4364 | .4325 | .4286 | .4247 |  |
| 0.2 | 4207                        | .4168 | .4129 | .4090 | .4052 | .4013 | .3974 | .3936 | .3897 | .3859 |  |
| 0.3 | 3821                        | .3783 | .3745 | .3707 | .3669 | .3632 | .3594 | .3557 | .3520 | .3483 |  |
| 0.4 | 3446                        | .3409 | .3372 | .3336 | .3300 | .3264 | .3228 | .3192 | .3156 | .3121 |  |
| 0.5 | 3085                        | .3050 | .3015 | .2981 | .2946 | .2912 | .2877 | .2843 | .2810 | .2776 |  |
| 0.6 | 2743                        | .2709 | .2676 | .2643 | .2611 | .2578 | .2546 | .2514 | .2483 | .2451 |  |
| 0.7 | 2420                        | .2389 | .2358 | .2327 | .2297 | .2266 | .2236 | .2206 | .2177 | .2148 |  |
| 0.8 | 2119                        | .2090 | .2061 | .2033 | .2005 | .1977 | .1949 | .1922 | .1894 | .1867 |  |
| 0.9 | 1841                        | .1814 | .1788 | .1762 | .1736 | .1711 | .1685 | .1660 | .1635 | .1611 |  |
| 1.0 | 1587                        | .1562 | .1539 | .1515 | .1492 | .1469 | .1446 | .1423 | .1401 | .1379 |  |
| 1.1 | 1357                        | .1335 | .1314 | .1292 | .1271 | .1251 | .1230 | .1210 | .1190 | .1170 |  |
| 1.2 | 1151                        | .1131 | .1112 | .1093 | .1075 | .1056 | .1038 | .1020 | .1003 | .0985 |  |
| 1.3 | 0968                        | .0951 | .0934 | .0918 | .0901 | .0885 | .0869 | .0853 | .0838 | .0823 |  |
| 1.4 | 0808                        | .0793 | .0778 | .0764 | .0749 | .0735 | .0721 | .0708 | .0694 | .0681 |  |
| 1.5 | 0668                        | .0655 | .0643 | .0630 | .0618 | .0606 | .0594 | .0582 | .0571 | .0559 |  |
| 1.6 | 0548                        | .0537 | .0526 | .0516 | .0505 | .0495 | .0485 | .0475 | .0465 | .0455 |  |
| 1.7 | 0446                        | .0436 | .0427 | .0418 | .0409 | .0401 | .0392 | .0384 | .0375 | .0367 |  |
| 1.8 | 0359                        | .0351 | .0344 | .0336 | .0329 | .0322 | .0314 | .0307 | .0301 | .0294 |  |
| 1.9 | 0287                        | .0281 | .0274 | .0268 | .0262 | .0256 | .0250 | .0244 | .0239 | .0233 |  |
| 2.0 | 0228                        | .0222 | .0217 | .0212 | .0207 | .0202 | .0197 | .0192 | .0188 | .0183 |  |
| 2.1 | 0179                        | .0174 | .0170 | .0166 | .0162 | .0158 | .0154 | .0150 | .0146 | .0143 |  |
| 2.2 | 0139                        | .0136 | .0132 | .0129 | .0125 | .0122 | .0119 | .0116 | .0113 | .0110 |  |
| 2.3 | 0107                        | .0104 | .0102 | .0099 | .0096 | .0094 | .0091 | .0089 | .0087 | .0084 |  |
| 2.4 | 0082                        | .0080 | .0078 | .0075 | .0073 | .0071 | .0069 | .0068 | .0066 | .0064 |  |
| 2.5 | 0062                        | .0060 | .0059 | .0057 | .0055 | .0054 | .0052 | .0051 | .0049 | .0048 |  |
| 2.6 | 0047                        | .0045 | .0044 | .0043 | .0041 | .0040 | .0039 | .0038 | .0037 | .0036 |  |
| 2.7 | 0035                        | .0034 | .0033 | .0032 | .0031 | .0030 | .0029 | .0028 | .0027 | .0026 |  |

# Normal Distribution: $\Pr(Z < 1.53)$



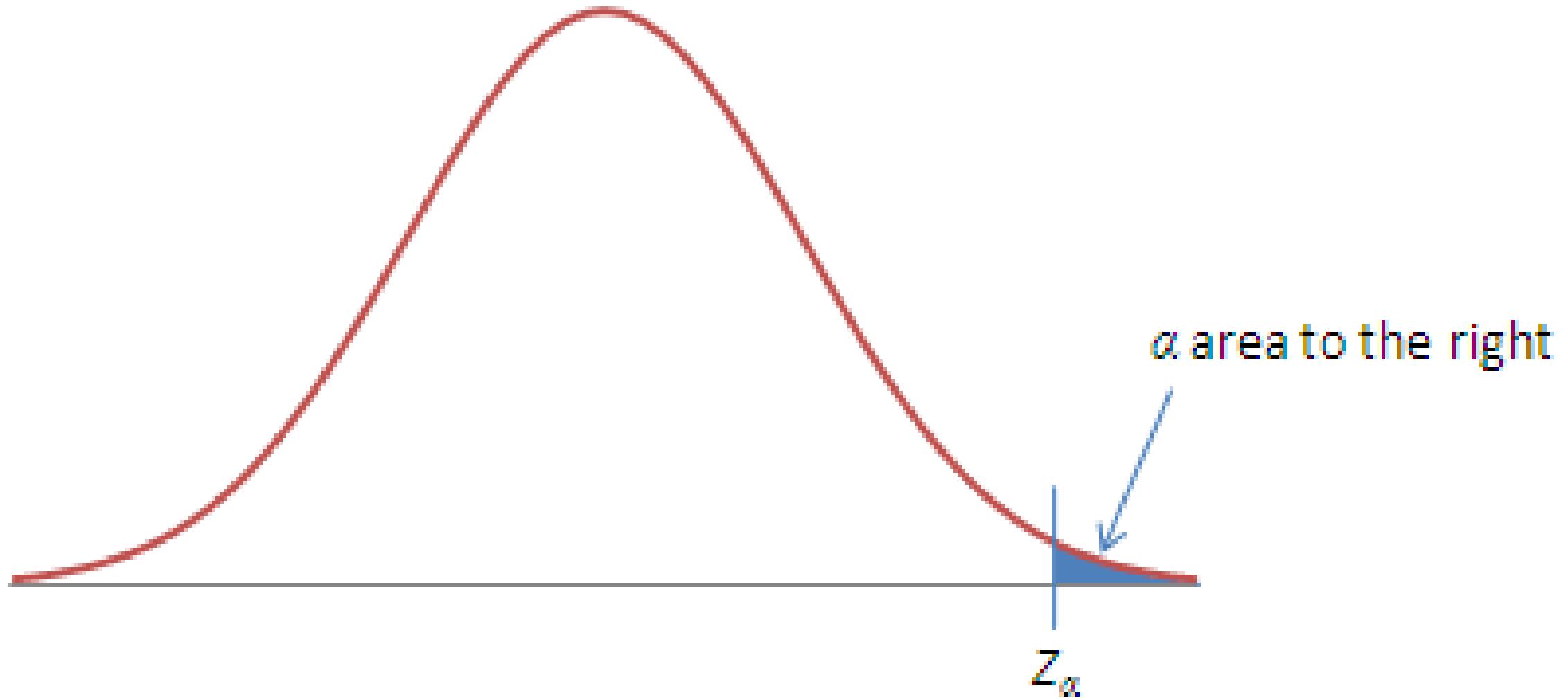
| z   | Second decimal place of z |       |       |      |       |       |       |       |       |       |
|-----|---------------------------|-------|-------|------|-------|-------|-------|-------|-------|-------|
|     | .00                       | .01   | .02   | .03  | .04   | .05   | .06   | .07   | .08   | .09   |
| 0.0 | .5000                     | .4960 | .4920 | 4880 | .4840 | .4801 | .4761 | .4721 | .4681 | .4641 |
| 0.1 | .4602                     | .4562 | .4522 | 4483 | .4443 | .4404 | .4364 | .4325 | .4286 | .4247 |
| 0.2 | .4207                     | .4168 | .4129 | 4090 | .4052 | .4013 | .3974 | .3936 | .3897 | .3859 |
| 0.3 | .3821                     | .3783 | .3745 | 3707 | .3669 | .3632 | .3594 | .3557 | .3520 | .3483 |
| 0.4 | .3446                     | .3409 | .3372 | 3336 | .3300 | .3264 | .3228 | .3192 | .3156 | .3121 |
| 0.5 | .3085                     | .3050 | .3015 | 2981 | .2946 | .2912 | .2877 | .2843 | .2810 | .2776 |
| 0.6 | .2743                     | .2709 | .2676 | 2643 | .2611 | .2578 | .2546 | .2514 | .2483 | .2451 |
| 0.7 | .2420                     | .2389 | .2358 | 2327 | .2297 | .2266 | .2236 | .2206 | .2177 | .2148 |
| 0.8 | .2119                     | .2090 | .2061 | 2033 | .2005 | .1977 | .1949 | .1922 | .1894 | .1867 |
| 0.9 | .1841                     | .1814 | .1788 | 1762 | .1736 | .1711 | .1685 | .1660 | .1635 | .1611 |
| 1.0 | .1587                     | .1562 | .1539 | 1515 | .1492 | .1469 | .1446 | .1423 | .1401 | .1379 |
| 1.1 | .1357                     | .1335 | .1314 | 1292 | .1271 | .1251 | .1230 | .1210 | .1190 | .1170 |
| 1.2 | .1151                     | .1131 | .1112 | 1093 | .1075 | .1056 | .1038 | .1020 | .1003 | .0985 |
| 1.3 | .0968                     | .0951 | .0934 | 0918 | .0901 | .0885 | .0869 | .0853 | .0838 | .0823 |
| 1.4 | .0808                     | .0793 | .0778 | 0764 | .0749 | .0735 | .0721 | .0708 | .0694 | .0681 |
| 1.5 | .0668                     | .0655 | .0643 | 0630 | .0618 | .0606 | .0594 | .0582 | .0571 | .0559 |
| 1.6 | .0548                     | .0537 | .0526 | 0516 | .0505 | .0495 | .0485 | .0475 | .0465 | .0455 |
| 1.7 | .0446                     | .0436 | .0427 | 0418 | .0409 | .0401 | .0392 | .0384 | .0375 | .0367 |
| 1.8 | .0359                     | .0351 | .0344 | 0336 | .0329 | .0322 | .0314 | .0307 | .0301 | .0294 |
| 1.9 | .0287                     | .0281 | .0274 | 0268 | .0262 | .0256 | .0250 | .0244 | .0239 | .0233 |
| 2.0 | .0228                     | .0222 | .0217 | 0212 | .0207 | .0202 | .0197 | .0192 | .0188 | .0183 |
| 2.1 | .0179                     | .0174 | .0170 | 0166 | .0162 | .0158 | .0154 | .0150 | .0146 | .0143 |
| 2.2 | .0139                     | .0136 | .0132 | 0129 | .0125 | .0122 | .0119 | .0116 | .0113 | .0110 |
| 2.3 | .0107                     | .0104 | .0102 | 0099 | .0096 | .0094 | .0091 | .0089 | .0087 | .0084 |
| 2.4 | .0082                     | .0080 | .0078 | 0075 | .0073 | .0071 | .0069 | .0068 | .0066 | .0064 |
| 2.5 | .0062                     | .0060 | .0059 | 0057 | .0055 | .0054 | .0052 | .0051 | .0049 | .0048 |
| 2.6 | .0047                     | .0045 | .0044 | 0043 | .0041 | .0040 | .0039 | .0038 | .0037 | .0036 |
| 2.7 | .0035                     | .0034 | .0033 | 0032 | .0031 | .0030 | .0029 | .0028 | .0027 | .0026 |

# Normal Distribution: $Z$ -value

- We often need to use the  $Z$ -value associated with specific “tail” areas of the standard normal distribution.
- Let  $z_\alpha$  represent the  $Z$ -value associated with a right-hand “tail area” of  $\alpha$ , such that

$$\Pr(Z > z_\alpha) = \alpha \text{ so } \Pr(Z < z_\alpha) = 1 - \alpha$$

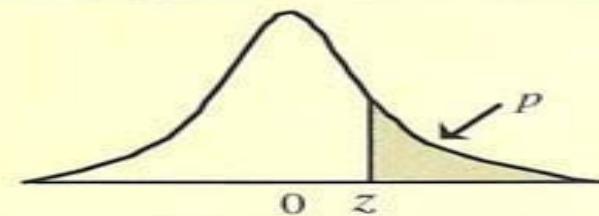
# Normal Distribution: Z-value



# Normal Distribution: Calculations

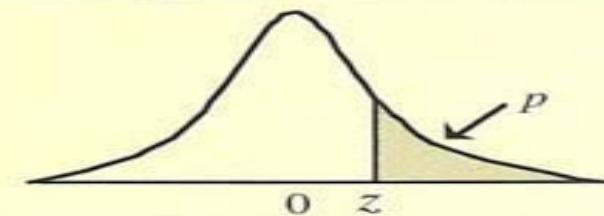
- Find  $z$  such that  $\Pr(|Z| > z) = 0.10$
- $\Pr(|Z| > z) = 2 \Pr(Z > z)$
- $\Pr(Z > z) = 0.05$ , so  $z = 1.64$  or  $1.65$  ( $1.6448$ )
- Therefore,  $z_{0.05} = 1.6448$  or  $1.64$  or  $1.65$ .
- How about  $z_{0.025}$  and  $z_{0.085}$ ?
- From the table  $z_{0.025} = 1.96$  and  $z_{0.085} = 1.37$

# Normal Distribution: Z<sub>0.05</sub>



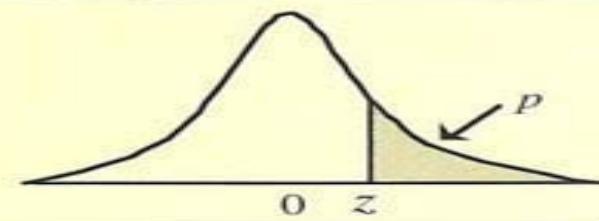
| z   | Second decimal place of z |       |       |       |       |      |       |       |       |       |
|-----|---------------------------|-------|-------|-------|-------|------|-------|-------|-------|-------|
|     | .00                       | .01   | .02   | .03   | .04   | .05  | .06   | .07   | .08   | .09   |
| 0.0 | .5000                     | .4960 | .4920 | .4880 | .4840 | 4801 | .4761 | .4721 | .4681 | .4641 |
| 0.1 | .4602                     | .4562 | .4522 | .4483 | .4443 | 4404 | .4364 | .4325 | .4286 | .4247 |
| 0.2 | .4207                     | .4168 | .4129 | .4090 | .4052 | 4013 | .3974 | .3936 | .3897 | .3859 |
| 0.3 | .3821                     | .3783 | .3745 | .3707 | .3669 | 3632 | .3594 | .3557 | .3520 | .3483 |
| 0.4 | .3446                     | .3409 | .3372 | .3336 | .3300 | 3264 | .3228 | .3192 | .3156 | .3121 |
| 0.5 | .3085                     | .3050 | .3015 | .2981 | .2946 | 2912 | .2877 | .2843 | .2810 | .2776 |
| 0.6 | .2743                     | .2709 | .2676 | .2643 | .2611 | 2578 | .2546 | .2514 | .2483 | .2451 |
| 0.7 | .2420                     | .2389 | .2358 | .2327 | .2297 | 2266 | .2236 | .2206 | .2177 | .2148 |
| 0.8 | .2119                     | .2090 | .2061 | .2033 | .2005 | 1977 | .1949 | .1922 | .1894 | .1867 |
| 0.9 | .1841                     | .1814 | .1788 | .1762 | .1736 | 1711 | .1685 | .1660 | .1635 | .1611 |
| 1.0 | .1587                     | .1562 | .1539 | .1515 | .1492 | 1469 | .1446 | .1423 | .1401 | .1379 |
| 1.1 | .1357                     | .1335 | .1314 | .1292 | .1271 | 1251 | .1230 | .1210 | .1190 | .1170 |
| 1.2 | .1151                     | .1131 | .1112 | .1093 | .1075 | 1056 | .1038 | .1020 | .1003 | .0985 |
| 1.3 | .0968                     | .0951 | .0934 | .0918 | .0901 | 0885 | .0869 | .0853 | .0838 | .0823 |
| 1.4 | .0808                     | .0793 | .0778 | .0764 | .0749 | 0735 | .0721 | .0708 | .0694 | .0681 |
| 1.5 | .0668                     | .0655 | .0643 | .0630 | .0618 | 0606 | .0594 | .0582 | .0571 | .0559 |
| 1.6 | .0548                     | .0537 | .0526 | .0516 | .0505 | 0495 | .0485 | .0475 | .0465 | .0455 |
| 1.7 | .0446                     | .0436 | .0427 | .0418 | .0409 | 0401 | .0392 | .0384 | .0375 | .0367 |
| 1.8 | .0359                     | .0351 | .0344 | .0336 | .0329 | 0322 | .0314 | .0307 | .0301 | .0294 |
| 1.9 | .0287                     | .0281 | .0274 | .0268 | .0262 | 0256 | .0250 | .0244 | .0239 | .0233 |
| 2.0 | .0228                     | .0222 | .0217 | .0212 | .0207 | 0202 | .0197 | .0192 | .0188 | .0183 |
| 2.1 | .0179                     | .0174 | .0170 | .0166 | .0162 | 0158 | .0154 | .0150 | .0146 | .0143 |
| 2.2 | .0139                     | .0136 | .0132 | .0129 | .0125 | 0122 | .0119 | .0116 | .0113 | .0110 |
| 2.3 | .0107                     | .0104 | .0102 | .0099 | .0096 | 0094 | .0091 | .0089 | .0087 | .0084 |
| 2.4 | .0082                     | .0080 | .0078 | .0075 | .0073 | 0071 | .0069 | .0068 | .0066 | .0064 |
| 2.5 | .0062                     | .0060 | .0059 | .0057 | .0055 | 0054 | .0052 | .0051 | .0049 | .0048 |
| 2.6 | .0047                     | .0045 | .0044 | .0043 | .0041 | 0040 | .0039 | .0038 | .0037 | .0036 |
| 2.7 | .0035                     | .0034 | .0033 | .0032 | .0031 | 0030 | .0029 | .0028 | .0027 | .0026 |

# Normal Distribution: Z<sub>0.025</sub>



| z   | Second decimal place of $z$ |       |       |       |       |       |      |       |       |       |
|-----|-----------------------------|-------|-------|-------|-------|-------|------|-------|-------|-------|
|     | .00                         | .01   | .02   | .03   | .04   | .05   | .06  | .07   | .08   | .09   |
| 0.0 | .5000                       | .4960 | .4920 | .4880 | .4840 | .4801 | 4761 | .4721 | .4681 | .4641 |
| 0.1 | .4602                       | .4562 | .4522 | .4483 | .4443 | .4404 | 4364 | .4325 | .4286 | .4247 |
| 0.2 | .4207                       | .4168 | .4129 | .4090 | .4052 | .4013 | 3974 | .3936 | .3897 | .3859 |
| 0.3 | .3821                       | .3783 | .3745 | .3707 | .3669 | .3632 | 3594 | .3557 | .3520 | .3483 |
| 0.4 | .3446                       | .3409 | .3372 | .3336 | .3300 | .3264 | 3228 | .3192 | .3156 | .3121 |
| 0.5 | .3085                       | .3050 | .3015 | .2981 | .2946 | .2912 | 2877 | .2843 | .2810 | .2776 |
| 0.6 | .2743                       | .2709 | .2676 | .2643 | .2611 | .2578 | 2546 | .2514 | .2483 | .2451 |
| 0.7 | .2420                       | .2389 | .2358 | .2327 | .2297 | .2266 | 2236 | .2206 | .2177 | .2148 |
| 0.8 | .2119                       | .2090 | .2061 | .2033 | .2005 | .1977 | 1949 | .1922 | .1894 | .1867 |
| 0.9 | .1841                       | .1814 | .1788 | .1762 | .1736 | .1711 | 1685 | .1660 | .1635 | .1611 |
| 1.0 | .1587                       | .1562 | .1539 | .1515 | .1492 | .1469 | 1446 | .1423 | .1401 | .1379 |
| 1.1 | .1357                       | .1335 | .1314 | .1292 | .1271 | .1251 | 1230 | .1210 | .1190 | .1170 |
| 1.2 | .1151                       | .1131 | .1112 | .1093 | .1075 | .1056 | 1038 | .1020 | .1003 | .0985 |
| 1.3 | .0968                       | .0951 | .0934 | .0918 | .0901 | .0885 | 0869 | .0853 | .0838 | .0823 |
| 1.4 | .0808                       | .0793 | .0778 | .0764 | .0749 | .0735 | 0721 | .0708 | .0694 | .0681 |
| 1.5 | .0668                       | .0655 | .0643 | .0630 | .0618 | .0606 | 0594 | .0582 | .0571 | .0559 |
| 1.6 | .0548                       | .0537 | .0526 | .0516 | .0505 | .0495 | 0485 | .0475 | .0465 | .0455 |
| 1.7 | .0446                       | .0436 | .0427 | .0418 | .0409 | .0401 | 0392 | .0384 | .0375 | .0367 |
| 1.8 | .0350                       | .0351 | .0344 | .0336 | .0329 | .0322 | 0314 | .0307 | .0301 | .0294 |
| 1.9 | .0287                       | .0281 | .0274 | .0268 | .0262 | .0256 | 0250 | .0244 | .0239 | .0233 |
| 2.0 | .0228                       | .0222 | .0217 | .0212 | .0207 | .0202 | 0197 | .0192 | .0188 | .0183 |
| 2.1 | .0179                       | .0174 | .0170 | .0166 | .0162 | .0158 | 0154 | .0150 | .0146 | .0143 |
| 2.2 | .0139                       | .0136 | .0132 | .0129 | .0125 | .0122 | 0119 | .0116 | .0113 | .0110 |
| 2.3 | .0107                       | .0104 | .0102 | .0099 | .0096 | .0094 | 0091 | .0089 | .0087 | .0084 |
| 2.4 | .0082                       | .0080 | .0078 | .0075 | .0073 | .0071 | 0069 | .0068 | .0066 | .0064 |
| 2.5 | .0062                       | .0060 | .0059 | .0057 | .0055 | .0054 | 0052 | .0051 | .0049 | .0048 |
| 2.6 | .0047                       | .0045 | .0044 | .0043 | .0041 | .0040 | 0039 | .0038 | .0037 | .0036 |
| 2.7 | .0035                       | .0034 | .0033 | .0032 | .0031 | .0030 | 0029 | .0028 | .0027 | .0026 |

# Normal Distribution: Z<sub>0.085</sub>

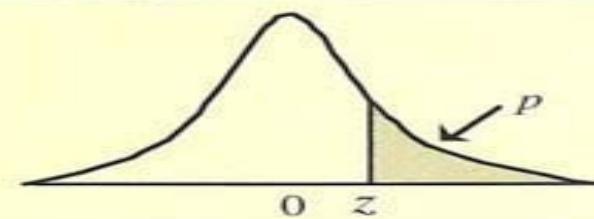


| z   | Second decimal place of $z$ |       |       |       |       |       |       |       |       |       |
|-----|-----------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
|     | .00                         | .01   | .02   | .03   | .04   | .05   | .06   | .07   | .08   | .09   |
| 0.0 | .5000                       | .4960 | .4920 | .4880 | .4840 | .4801 | .4761 | .4721 | .4681 | .4641 |
| 0.1 | .4602                       | .4562 | .4522 | .4483 | .4443 | .4404 | .4364 | .4325 | .4286 | .4247 |
| 0.2 | .4207                       | .4168 | .4129 | .4090 | .4052 | .4013 | .3974 | .3936 | .3897 | .3859 |
| 0.3 | .3821                       | .3783 | .3745 | .3707 | .3669 | .3632 | .3594 | .3557 | .3520 | .3483 |
| 0.4 | .3446                       | .3409 | .3372 | .3336 | .3300 | .3264 | .3228 | .3192 | .3156 | .3121 |
| 0.5 | .3085                       | .3050 | .3015 | .2981 | .2946 | .2912 | .2877 | .2843 | .2810 | .2776 |
| 0.6 | .2743                       | .2709 | .2676 | .2643 | .2611 | .2578 | .2546 | .2514 | .2483 | .2451 |
| 0.7 | .2420                       | .2389 | .2358 | .2327 | .2297 | .2266 | .2236 | .2206 | .2177 | .2148 |
| 0.8 | .2119                       | .2090 | .2061 | .2033 | .2005 | .1977 | .1949 | .1922 | .1894 | .1867 |
| 0.9 | .1841                       | .1814 | .1788 | .1762 | .1736 | .1711 | .1685 | .1660 | .1635 | .1611 |
| 1.0 | .1587                       | .1562 | .1539 | .1515 | .1492 | .1469 | .1446 | .1423 | .1401 | .1379 |
| 1.1 | .1357                       | .1335 | .1314 | .1292 | .1271 | .1251 | .1230 | .1210 | .1190 | .1170 |
| 1.2 | .1151                       | .1131 | .1112 | .1093 | .1075 | .1056 | .1038 | .1020 | .1003 | .0985 |
| 1.3 | .0968                       | .0951 | .0934 | .0918 | .0901 | .0885 | .0869 | .0853 | .0838 | .0823 |
| 1.4 | .0808                       | .0793 | .0778 | .0764 | .0749 | .0735 | .0721 | .0708 | .0694 | .0681 |
| 1.5 | .0668                       | .0655 | .0643 | .0630 | .0618 | .0606 | .0594 | .0582 | .0571 | .0559 |
| 1.6 | .0548                       | .0537 | .0526 | .0516 | .0505 | .0495 | .0485 | .0475 | .0465 | .0455 |
| 1.7 | .0446                       | .0436 | .0427 | .0418 | .0409 | .0401 | .0392 | .0384 | .0375 | .0367 |
| 1.8 | .0359                       | .0351 | .0344 | .0336 | .0329 | .0322 | .0314 | .0307 | .0301 | .0294 |
| 1.9 | .0287                       | .0281 | .0274 | .0268 | .0262 | .0256 | .0250 | .0244 | .0239 | .0233 |
| 2.0 | .0228                       | .0222 | .0217 | .0212 | .0207 | .0202 | .0197 | .0192 | .0188 | .0183 |
| 2.1 | .0179                       | .0174 | .0170 | .0166 | .0162 | .0158 | .0154 | .0150 | .0146 | .0143 |
| 2.2 | .0139                       | .0136 | .0132 | .0129 | .0125 | .0122 | .0119 | .0116 | .0113 | .0110 |
| 2.3 | .0107                       | .0104 | .0102 | .0099 | .0096 | .0094 | .0091 | .0089 | .0087 | .0084 |
| 2.4 | .0082                       | .0080 | .0078 | .0075 | .0073 | .0071 | .0069 | .0068 | .0066 | .0064 |
| 2.5 | .0062                       | .0060 | .0059 | .0057 | .0055 | .0054 | .0052 | .0051 | .0049 | .0048 |
| 2.6 | .0047                       | .0045 | .0044 | .0043 | .0041 | .0040 | .0039 | .0038 | .0037 | .0036 |
| 2.7 | .0035                       | .0034 | .0033 | .0032 | .0031 | .0030 | .0029 | .0028 | .0027 | .0026 |

# Normal Distribution: Calculations

- How about  $z_{0.70}$ ?
- We know that  $\Pr(Z > z_{0.70}) = 0.70$  so  
 $z_{0.70} < 0.$
- So  $\Pr(Z > -z_{0.70}) = 0.30$
- We have  $z_{0.70} = -z_{0.30} = -0.52.$

# Normal Distribution: $Z_{0.30} = Z_{-0.70}$



| z   | Second decimal place of z |       |       |       |       |       |       |       |       |       |
|-----|---------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
|     | .00                       | .01   | .02   | .03   | .04   | .05   | .06   | .07   | .08   | .09   |
| 0.0 | .5000                     | .4960 | .4920 | .4880 | .4840 | .4801 | .4761 | .4721 | .4681 | .4641 |
| 0.1 | .4602                     | .4562 | .4522 | .4483 | .4443 | .4404 | .4364 | .4325 | .4286 | .4247 |
| 0.2 | .4207                     | .4168 | .4129 | .4090 | .4052 | .4013 | .3974 | .3936 | .3897 | .3859 |
| 0.3 | .3821                     | .3783 | .3745 | .3707 | .3669 | .3632 | .3594 | .3557 | .3520 | .3483 |
| 0.4 | .3446                     | .3409 | .3372 | .3336 | .3300 | .3264 | .3228 | .3192 | .3156 | .3121 |
| 0.5 | .3085                     | .3050 | .3015 | .2981 | .2946 | .2912 | .2877 | .2843 | .2810 | .2776 |
| 0.6 | .2743                     | .2709 | .2676 | .2643 | .2611 | .2578 | .2546 | .2514 | .2483 | .2451 |
| 0.7 | .2420                     | .2389 | .2358 | .2327 | .2297 | .2266 | .2236 | .2206 | .2177 | .2148 |
| 0.8 | .2119                     | .2090 | .2061 | .2033 | .2005 | .1977 | .1949 | .1922 | .1894 | .1867 |
| 0.9 | .1841                     | .1814 | .1788 | .1762 | .1736 | .1711 | .1685 | .1660 | .1635 | .1611 |
| 1.0 | .1587                     | .1562 | .1539 | .1515 | .1492 | .1469 | .1446 | .1423 | .1401 | .1379 |
| 1.1 | .1357                     | .1335 | .1314 | .1292 | .1271 | .1251 | .1230 | .1210 | .1190 | .1170 |
| 1.2 | .1151                     | .1131 | .1112 | .1093 | .1075 | .1056 | .1038 | .1020 | .1003 | .0985 |
| 1.3 | .0968                     | .0951 | .0934 | .0918 | .0901 | .0885 | .0869 | .0853 | .0838 | .0823 |
| 1.4 | .0808                     | .0793 | .0778 | .0764 | .0749 | .0735 | .0721 | .0708 | .0694 | .0681 |
| 1.5 | .0668                     | .0655 | .0643 | .0630 | .0618 | .0606 | .0594 | .0582 | .0571 | .0559 |
| 1.6 | .0548                     | .0537 | .0526 | .0516 | .0505 | .0495 | .0485 | .0475 | .0465 | .0455 |
| 1.7 | .0446                     | .0436 | .0427 | .0418 | .0409 | .0401 | .0392 | .0384 | .0375 | .0367 |
| 1.8 | .0359                     | .0351 | .0344 | .0336 | .0329 | .0322 | .0314 | .0307 | .0301 | .0294 |
| 1.9 | .0287                     | .0281 | .0274 | .0268 | .0262 | .0256 | .0250 | .0244 | .0239 | .0233 |
| 2.0 | .0228                     | .0222 | .0217 | .0212 | .0207 | .0202 | .0197 | .0192 | .0188 | .0183 |
| 2.1 | .0179                     | .0174 | .0170 | .0166 | .0162 | .0158 | .0154 | .0150 | .0146 | .0143 |
| 2.2 | .0139                     | .0136 | .0132 | .0129 | .0125 | .0122 | .0119 | .0116 | .0113 | .0110 |
| 2.3 | .0107                     | .0104 | .0102 | .0099 | .0096 | .0094 | .0091 | .0089 | .0087 | .0084 |
| 2.4 | .0082                     | .0080 | .0078 | .0075 | .0073 | .0071 | .0069 | .0068 | .0066 | .0064 |
| 2.5 | .0062                     | .0060 | .0059 | .0057 | .0055 | .0054 | .0052 | .0051 | .0049 | .0048 |
| 2.6 | .0047                     | .0045 | .0044 | .0043 | .0041 | .0040 | .0039 | .0038 | .0037 | .0036 |
| 2.7 | .0035                     | .0034 | .0033 | .0032 | .0031 | .0030 | .0029 | .0028 | .0027 | .0026 |

# Calculating Probabilities of Normal Distribution

- Characteristics of standard normal distribution. If  $Z \sim N(0,1)$ , then

$$\Pr(Z > z) = \Pr(Z < -z) = 1 - \Pr(Z < z)$$

$$\Pr(a < Z < b) = \Pr(Z < b) - \Pr(Z < a)$$

- Exercise** – find the following probability:

$$\Pr(Z > 1.02)$$

$$\Pr(Z < -0.80)$$

$$\Pr(Z < 1.35)$$

$$\Pr(-0.80 < Z < 1.35)$$

# Calculating Probabilities of Normal Distribution

- Sometimes you want to find a value  $z_\alpha$  ( $0 < \alpha < 1$ ):

$$\Pr(Z > z_\alpha) = \Pr(Z < -z_\alpha) = \alpha$$

$$\Pr(Z < z_\alpha) = 1 - \alpha$$

$$\Pr\left(-\frac{z_\alpha}{2} < Z < \frac{z_\alpha}{2}\right) = 1 - \alpha$$

- **Exercise:** find a value  $z$  such that

$$\Pr(|Z| > z) = 0.08$$

# Normal Distribution: Transformation

- If  $Y \sim N(\mu, \sigma^2)$ , then  $E(Y) = \mu$  and  $Var(Y) = \sigma^2$
- Define  $Z = \frac{Y-\mu}{\sigma}$  - First re-center  $Y$  with population mean  $\mu$  then rescale it with  $\sigma$ . Then  $Z \sim N(0,1)$ .
- We have

$$\Pr(Y < y) = \Pr\left(\frac{Y-\mu}{\sigma} < \frac{y-\mu}{\sigma}\right) = \Pr(Z < \frac{y-\mu}{\sigma})$$

# Steps for Calculations with Normal Distribution

1. Determine if  $Y$  has a normal distribution.
2. Find the population mean and variance of  $Y$ .
3. Represent probability in terms of  $Y$ .
4. Transfer  $Y$  to the standard normal  $Z$ .
5. Find the probability with a normal table.

# Calculating Probabilities of Normal Distribution

- Suppose that  $Y \sim N(10,20)$ , find the following probability:

$$\Pr(Y > 15) = \Pr\left(\frac{Y - 10}{\sqrt{20}} > \frac{15 - 10}{\sqrt{20}}\right) = \Pr(Z > 1.12) = 0.1314$$

- **Exercises:** for  $Y \sim N(10,10)$ , find:

$$\Pr(Y < 5)$$

$$\Pr(5 < Y < 10)$$

# Example – Production at Titanium Dioxide Facility

- **Example** – A major titanium dioxide facility has a designed capacity of 600 tons. Historically, the daily production approximately follows a normal distribution with mean 500 tons and a standard deviation of 50 tons.

## Example – Production at Titanium Dioxide Facility

- The company can sell anything this facility can make. So management really would like to know the probability that it can make more than 600 tons (exceeding the capacity) of product.
- Let  $Y$  be the total production, then  $Y \sim N(500, 50)$ .

$$\begin{aligned}\Pr(Y > 600) &= \Pr\left(\frac{Y - 500}{50} > \frac{600 - 500}{50}\right) \\ &= \Pr(Z > 2) = 0.0228.\end{aligned}$$

- So the facility should exceed the capacity only 2% of time.

# Example – Setting Mean for Dairy Packing

**Example** – 8 oz of milk should weigh 245 grams. Federal inspectors require that the mean amount of milk packaged must be significantly greater than 245 grams in order to minimize any underage. Plant manager would like the mean amount to be no more than necessary to meet standards and argues that a reasonable mean weight should produce less than 1% of cartons underweight. Historically, the weight approximately follows a normal distribution with a standard deviation of 1.65 grams.

# Example – Setting Mean for Dairy Packing

- Let  $Y$  be the milk weight and  $Y \sim N(\mu, 1.65^2)$

$$\begin{aligned}\Pr(Y < 245) &= \Pr\left(\frac{Y-\mu}{1.65} < \frac{245-\mu}{1.65}\right) \\ &= \Pr\left(Z < \frac{245 - \mu}{1.65}\right) < 0.01\end{aligned}$$

- So  $\frac{245-\mu}{1.65} < z_{0.99}$  and

$$\mu > 245 - 1.65 * z_{0.99} = 245 - 1.65 * (-2.33) = 248.83$$

- So we need an average of 248.83 grams.

# Rational Behind Boxplot

- Recall the boxplot from Chapter 1, we call an observation  $y$ :
  - An outlier, if  $y > \text{Upper Inner fence}$  or  $y < \text{Lower Inner Fence}$
  - An extreme outlier, if  $y > \text{Upper Outer fence}$  or  $y < \text{Lower Outer Fence}$
- Interquartile Range (IQR) =  $Q_3 - Q_1$  and Step =  $1.5 * IQR$
- Upper Inner Fence =  $Q_3 + Step$
- Lower Inner Fence =  $Q_1 - Step$
- Upper Outer Fence =  $Q_3 + 2 * Step$ ;
- Lower Outer Fence =  $Q_1 - 2 * Step$ ;

# Exercise - Rational Behind Boxplot

Suppose that the data is from the standard normal. Find

- $Q_3; Q_1$
- $IQR = Q_3 - Q_1$
- $Step = 2 * IQR$
- $UIF = Q_3 + Step; LIF = Q_3 - Step$
- $\Pr(Z > UIF)$  and  $\Pr(Z < LIF)$
- $UOF = Q_3 + 2 * Step; LOF = Q_1 - 2 * Step$
- $\Pr(Z > UOF)$  and  $\Pr(Z < LOF)$

# Solution - Rational Behind Boxplot

Suppose that the data is from the standard normal. Find

- $Q_3 = z_{0.25} = 0.6749; Q_1 = -z_{0.25} = -0.6749$
- $IQR = Q_3 - Q_1 = 2 * 0.6749 = 1.3490$
- $Step = 2 * IQR = 1.5 * 1.3490 = 2.0234$
- $UIF = Q_3 + Step = 2.6980; LIF = -UIF = -2.6980$
- $\Pr(Z > UIF) = \Pr(Z < LIF) = 0.0035$
- $UOF = Q_3 + 2 * Step = 4.7214; LOF = -UOF = -4.7214$
- $\Pr(Z > UOF) = \Pr(Z < LOF) \approx 0$

# Sampling Distribution

- **Statistical Inference** - making inferences on population parameters using sample statistic.
- **Definition 2.15** - The **sampling distribution** of a statistic is the probability distribution of that statistic.
- Statistic is a random variable.

# Sample Mean

- Sample mean – let  $y_1, \dots, y_n$  denote a sample of interest. The sample mean, denoted by  $\bar{y}$ , is given by

$$\bar{y} = \frac{1}{n}(y_1 + \dots + y_n) = \frac{1}{n}\sum_{i=1}^n y_i.$$

- Sample mean is a random variable!

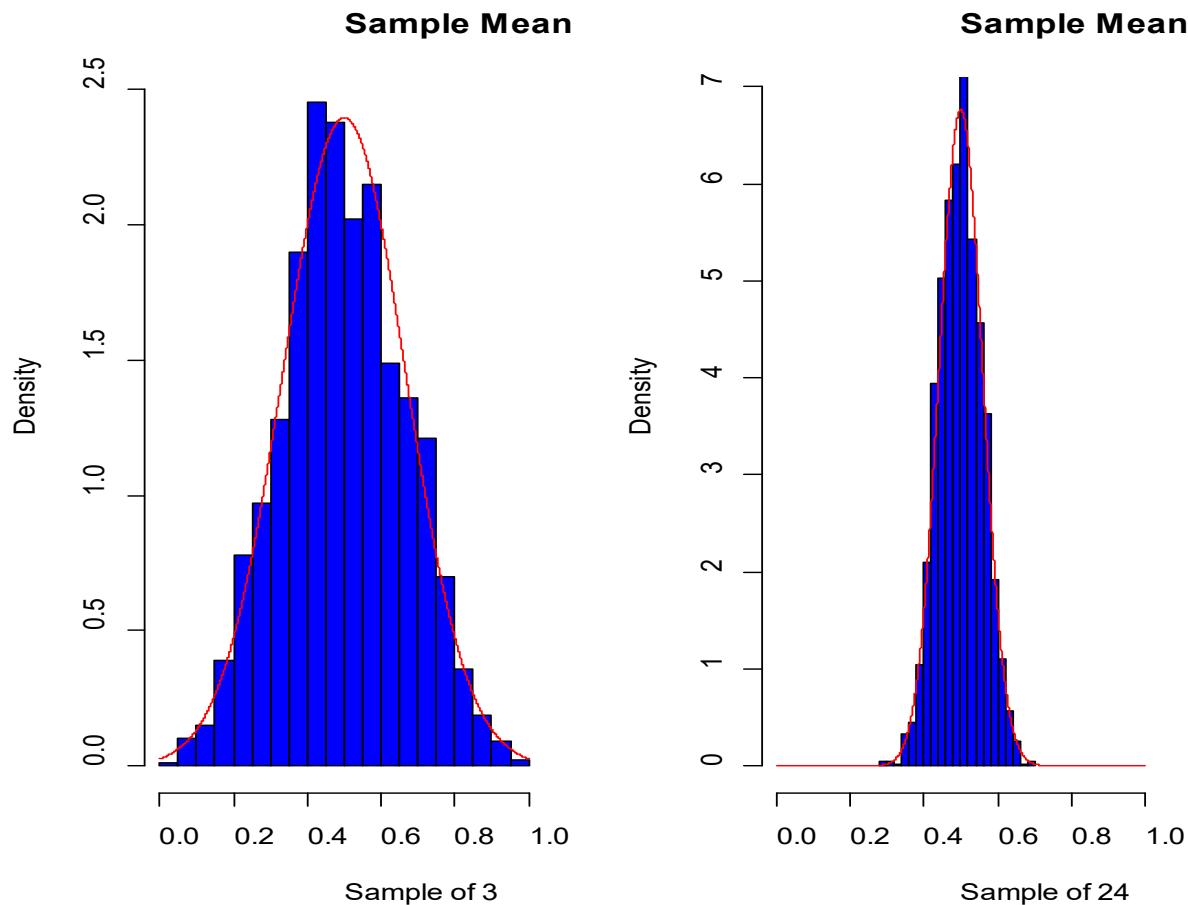
# Sampling Distribution

- **Theorem 2.5.1 Sampling distribution of the mean** - The sampling distribution of the mean from a random sample of size  $n$  from a population with mean  $\mu$  and variance  $\sigma^2$  will have mean  $\mu$  and variance  $\sigma^2/n$ .
- This is true for a random sample from any distribution.

# Sampling Distribution: An example

- For a uniform distribution from 0 to 1
  - Population mean and variance are 0.5 and 0.08333
- Consider the sample size of 3
  - Mean and variance of sample mean is 0.5 and  $0.08333/3=0.0278$
- The empirical estimate from 1000 samples
  - Mean and variance of sample mean is 0.5 and  $0.08333/3=0.0278$
- How about the sample size of 24?
  - Mean and variance of sample mean is 0.5 and  $0.08333/24=0.00347$

# Sampling Distribution: An example



# Usefulness of Sampling Distribution

- Mean of sampling distribution of the sample mean is the population mean - implies that “on the average” the sample mean is the same as the population mean. We therefore say that the sample mean is an **unbiased estimate** of the population mean.
- Variance of the distribution of the sample mean is  $\sigma^2/n$ . This implies the variability of sample mean decreases with the increasing sample size.

# Sampling Distribution of Sample Mean

- But what is the exact distribution of sample mean,  $\bar{y}$  ( $\bar{Y}$ )?
- Its distribution depends on
  - The distribution of sample.
  - The samples themselves.
- In this book, we focus on the ***random sample*** – A random sample is one in which all the observations are statistically independent and follow exactly the same distribution.

# Sampling Distribution of Sample Mean

- If the sample is a random sample from the normal distribution  $N(\mu, \sigma^2)$ , then sample mean  $\sim N\left(\mu, \frac{\sigma^2}{n}\right)$
- If the sample is a random sample from another distribution, then the sampling distribution of sample mean is generally unknown. However, the normal distribution can be used as an approximation.

# Central Limit Theorem

- **Central Limit Theorem** - If random samples of size  $n$  are taken from any distribution with mean  $\mu$  and variance  $\sigma^2$ , sample mean will have a distribution approximately normal with mean  $\mu$  and variance  $\sigma^2/n$ .
- How large  $n$  should be for the Central Limit Theorem:  
$$n > 30 \text{ in general.}$$
- Much smaller  $n$  is fine if data is approximately normal.

# Central Limit Theorem - Example

- **Example** - An aptitude test for high school students is designed so that scores on the test have  $\mu = 90$  and  $\sigma = 20$ . In a section of 100 students the mean score is 86. What is the probability of getting a mean of 86 or lower on test?
- We have  $\bar{Y} \sim N(90, \frac{20^2}{100})$  and

$$\begin{aligned}\Pr(\bar{Y} < 86) &= \Pr\left(\frac{\bar{Y} - 90}{2} < \frac{86 - 90}{2}\right) \\ &= Pr(Z < -2) = 0.0228\end{aligned}$$

# CLT – Thickness of Silicon Wafers

- **Example** – Hurwitz and Spagon (1993) analyzed the performance of a planarization device that polishes silicon wafers to a high degree of smoothness. Historically, the thickness of the wafer has a mean of 3200 angstroms with a standard deviation 80 angstroms. The following thicknesses are from 23 wafers.

## CLT – Thickness of Silicon Wafers

|      |      |      |      |      |      |
|------|------|------|------|------|------|
| 3240 | 3200 | 3220 | 3210 | 3250 | 3220 |
| 3190 | 3190 | 3150 | 3160 | 3270 | 3180 |
| 3200 | 3270 | 3180 | 3300 | 3250 | 3330 |
| 3300 | 3280 | 3270 | 3270 | 3200 |      |

- Sample mean is  $\bar{y} = 3232$ , which is larger than 3200.
- Production management would like to know whether there is evidence to suggest that this lot is thicker than usual.

# Exercise - Thickness of Silicon Wafers

- How to formulate “whether there is evidence to suggest that this lot is thicker than usual”?
- If the probability of the mean thickness of 23 wafers greater than the observed mean (3232) is small, we have an evidence of “unusual”.
- We need to calculate:

$$\Pr(\bar{Y} > 3232)$$

# Solution - Thickness of Silicon Wafers

- We have

$$\begin{aligned}\Pr(\bar{Y} > 3232) &= \Pr\left(\frac{\bar{Y} - 3200}{\frac{80}{\sqrt{23}}} > \frac{3232 - 3200}{\frac{80}{\sqrt{23}}}\right) \\ &= \Pr(Z > 1.92) = 0.0274\end{aligned}$$

- Yes. We have some evidence to suggest that this lot is thicker than usual.

# Central Limit Theorem

- When do we use  $N(\mu, \sigma^2)$  and when do we use  $N(\mu, \sigma^2/n)$  ?
- One is for an individual observation and one is for the sample mean.
- Example (Thickness of Silicon Wafer) – What is the probability of the mean thickness of 100 wafers great than 3232?  $N(\mu, \sigma^2/n)$
- Example (Thickness of Silicon Wafer) – What is the probability of thickness of a random selected wafer great than 3232?  $N(\mu, \sigma^2)$

# Normal Approximation to Binomial

- If  $Y \sim \text{Binomial}(n, p)$ , then  $Y = \sum_{i=1}^n Y_i$  and  $Y_i (i = 1, \dots, n)$  is a random sample from Bernoulli distribution. So  $\frac{Y}{n}$  is considered as a sample mean.
- By Central Limit Theorem (CLT):

$$\frac{\frac{Y}{n} - p}{\sqrt{p(1-p)/n}} = \frac{Y - np}{\sqrt{np(1-p)}}$$

can be approximated by  $N(0,1)$  for large  $n$ .

# Normal Approximation to Binomial

- By Central Limit Theorem (CLT):  $\frac{\frac{Y}{n} - p}{\sqrt{p(1-p)/n}} = \frac{Y - np}{\sqrt{np(1-p)}}$  can be approximated by  $N(0,1)$  for large  $n$ .
- If we want to use normal distribution to approximate the sample proportion: use  $N(p, \frac{p(1-p)}{n})$ .
- If we want to use normal to approximate the number of success (which has a binomial distribution): use  $N(np, np(1 - p))$ .

# Normal Approximation to Binomial

- In other words,  $\frac{\frac{Y}{n} - p}{\sqrt{p(1-p)/n}}$  or  $\frac{Y - np}{\sqrt{np(1-p)}}$  can be approximated by  $N(0,1)$  for sufficiently large  $n$ .
- We consider  $n$  to be sufficiently large if
$$np \geq 5 \text{ and } n(1 - p) \geq 5.$$
- The approximation works even better if
$$np \geq 10 \text{ and } n(1 - p) \geq 10.$$

## Example 2.15 - Election

- **Example 2.15:** Suppose a random sample of 100 voters show 61 with a preference for candidate A. If the election were in fact a toss-up (that is,  $p = 0.5$ ) what is the probability of obtaining that (or a more extreme value)?
- Let  $Y$  be the number of voters with a preference for candidate A, then  $Y \sim \text{Binomial}(100, 0.5)$  and we want to calculate  $\Pr(Y \geq 61)$
- Can we use normal approximation?
- Check:  $100 * 0.5 = 50$ . Yes. We can.

# Exercise – Normal Approximation to Binomial

- **Exercise:** Based on data from 2007 National Health Interview Survey, it is estimated that “10% of adults experienced feelings of sadness for all, most, or some of the time” during the 30 days prior to the interview. You interview a random sample of 68 people who have recently filed for unemployment benefits in your county, and ask this same question in your survey. If the proportion of your population with these feelings is the same as the 10% nationally, what is the probability that your sample will have 12 or more people with these feelings?

# Exercise – Normal Approximation to Binomial

- **Exercise:** An insurance company wishes to keep the error rates in medical claims at or below 10%. If there is evidence of an error rate greater than this, they will need to introduce new quality procedures. They randomly select 60 independent claims and audit them for errors and use the following rule: *Decide error rate is acceptable if there are eight or fewer errors in the sample of 60.*
- **Question:** If the probability of error is truly 10%, what is the chance they will decide their error rate is acceptable?

# Sample Variance

- To use the central limit theorem, we need to know the population variance. If we want to use data to make inference about the population, we generally do not know the population variance.
- In this situation, we can use sample variance in our calculation.

# Descriptive Statistics – Sample Variance

- **Definition 1.14** – The **sample variance**, denoted by  $s^2$  is defined by

$$s^2 = \frac{1}{n-1} [(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2]$$

- It looks like an “average” of the squared deviations from the sample mean.
- Note that we use  $n - 1$  instead of  $n$ .
- Sample variance is a measure of **dispersion** which is the extent to which a distribution is stretched or squeezed.

# Descriptive Statistics – Sample Variance

- Another formula for sample variance,  $s^2$ , is

$$\begin{aligned}s^2 &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \\ &= \frac{1}{n-1} (\sum_{i=1}^n y_i^2 - n\bar{y}^2).\end{aligned}$$

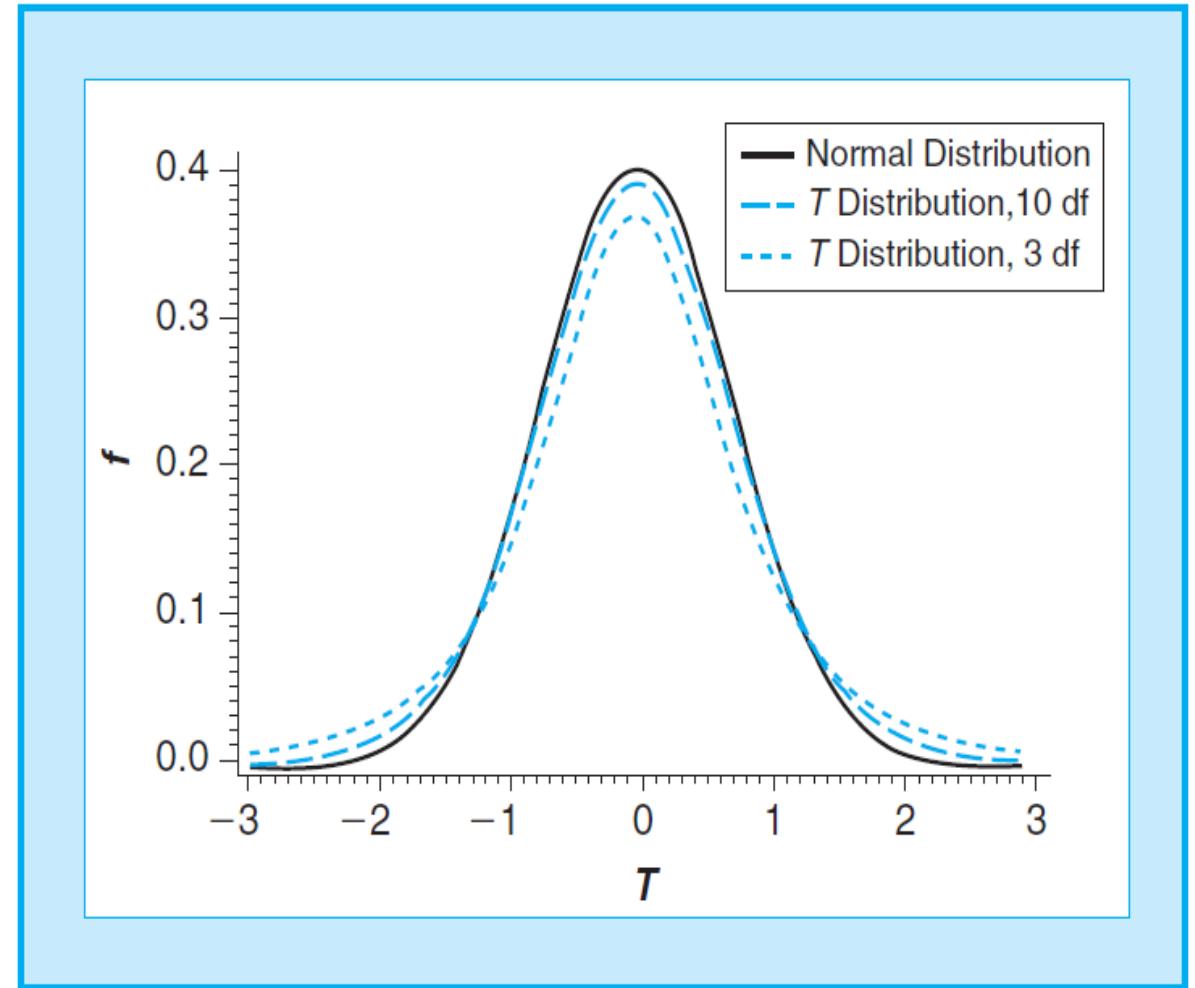
- **Definition 1.15** – The **standard deviation** of a set of observed values is defined to be the positive square root of the variance. In other words, the sample standard deviation is:  $s = \sqrt{s^2}$ .

# Random Behavior of Means with Unknown Variance

- For a random sample size of  $n$  from  $N(\mu, \sigma^2)$ , we can
- Use  $Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$  if  $\sigma$  is known.
- Use  $t = \frac{\bar{Y} - \mu}{s/\sqrt{n}} \sim t_{n-1}$  -  $t$ -distribution with  $n - 1$  degrees of freedom if  $\sigma$  is unknown.
- In general, degrees of freedom = number of observations – number of parameters estimated

# *t*-distribution

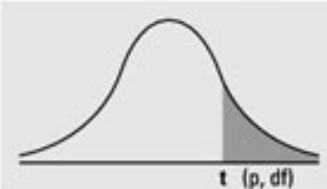
- *t*-distribution has a similar shape with standard normal but with a “fatter” tail.
- When  $n \rightarrow \infty$ ,  $t_{n-1}$  becomes  $N(0, 1)$  since the sample variance converges to population variance.



# *t*-distribution

- Similarly with standard normal, we can use *t*-table to calculate the probability.
- May need to use symmetric property of *t*-distribution.
- We can find  $t_{n,\alpha}$  such that  $\Pr(T_n > t_{n,\alpha}) = \alpha$ .
- **Exercise:** Find  $\Pr(T_{10} > 0.5)$ .
- **Example:** Find  $t_{10,0.05}$ .
- **Example:** Find  $t_{26,0.05}$ .

Numbers in each row of the table are values on a  $t$ -distribution with  
( $df$ ) degrees of freedom for selected right-tail (greater-than) probabilities ( $p$ ).



| <b>df/p</b> | <b>0.40</b> | <b>0.25</b> | <b>0.10</b> | <b>0.05</b> | <b>0.025</b> | <b>0.01</b> | <b>0.005</b> | <b>0.0005</b> |
|-------------|-------------|-------------|-------------|-------------|--------------|-------------|--------------|---------------|
| <b>1</b>    | 0.324920    | 1.000000    | 3.077684    | 6.313752    | 12.70620     | 31.82052    | 63.65674     | 636.6192      |
| <b>2</b>    | 0.288675    | 0.816497    | 1.885618    | 2.919986    | 4.30265      | 6.96456     | 9.92484      | 31.5991       |
| <b>3</b>    | 0.276671    | 0.764892    | 1.637744    | 2.353363    | 3.18245      | 4.54070     | 5.84091      | 12.9240       |
| <b>4</b>    | 0.270722    | 0.740697    | 1.533206    | 2.131847    | 2.77645      | 3.74695     | 4.60409      | 8.6103        |
| <b>5</b>    | 0.267181    | 0.726687    | 1.475884    | 2.015048    | 2.57058      | 3.36493     | 4.03214      | 6.8688        |
| <b>6</b>    | 0.264835    | 0.717558    | 1.439756    | 1.943180    | 2.44691      | 3.14267     | 3.70743      | 5.9588        |
| <b>7</b>    | 0.263167    | 0.711142    | 1.414924    | 1.894579    | 2.36462      | 2.99795     | 3.49948      | 5.4079        |
| <b>8</b>    | 0.261921    | 0.706387    | 1.396815    | 1.859548    | 2.30600      | 2.89646     | 3.35539      | 5.0413        |
| <b>9</b>    | 0.260955    | 0.702722    | 1.383029    | 1.833113    | 2.26216      | 2.82144     | 3.24984      | 4.7809        |
| <b>10</b>   | 0.260185    | 0.699812    | 1.372184    | 1.812461    | 2.22814      | 2.76377     | 3.16927      | 4.5869        |
| <b>11</b>   | 0.259556    | 0.697445    | 1.363430    | 1.795885    | 2.20099      | 2.71808     | 3.10581      | 4.4370        |
| <b>12</b>   | 0.259033    | 0.695483    | 1.356217    | 1.782288    | 2.17881      | 2.68100     | 3.05454      | 4.3178        |
| <b>13</b>   | 0.258591    | 0.693829    | 1.350171    | 1.770933    | 2.16037      | 2.65031     | 3.01228      | 4.2208        |
| <b>14</b>   | 0.258213    | 0.692417    | 1.345030    | 1.761310    | 2.14479      | 2.62449     | 2.97684      | 4.1405        |
| <b>15</b>   | 0.257885    | 0.691197    | 1.340606    | 1.753050    | 2.13145      | 2.60248     | 2.94671      | 4.0728        |
| <b>16</b>   | 0.257599    | 0.690132    | 1.336757    | 1.745884    | 2.11991      | 2.58349     | 2.92078      | 4.0150        |
| <b>17</b>   | 0.257347    | 0.689195    | 1.333379    | 1.739607    | 2.10982      | 2.56693     | 2.89823      | 3.9651        |
| <b>18</b>   | 0.257123    | 0.688364    | 1.330391    | 1.734064    | 2.10092      | 2.55238     | 2.87844      | 3.9216        |
| <b>19</b>   | 0.256923    | 0.687621    | 1.327728    | 1.729133    | 2.09302      | 2.53948     | 2.86093      | 3.8834        |
| <b>20</b>   | 0.256743    | 0.686954    | 1.325341    | 1.724718    | 2.08596      | 2.52798     | 2.84534      | 3.8495        |
| <b>21</b>   | 0.256580    | 0.686352    | 1.323188    | 1.720743    | 2.07961      | 2.51765     | 2.83136      | 3.8193        |
| <b>22</b>   | 0.256432    | 0.685805    | 1.321237    | 1.717144    | 2.07387      | 2.50832     | 2.81876      | 3.7921        |
| <b>23</b>   | 0.256297    | 0.685306    | 1.319460    | 1.713872    | 2.06866      | 2.49987     | 2.80734      | 3.7676        |

## Example 2.18 – Grade Point Ratio Study

- **Example 2.18:** Grade point ratios (GPRs) have been recorded for a random sample of 16 from the entering freshman class at a major university. It can be assumed that the distribution of GPR values is approximately normal. The sample yielded a mean,  $\bar{y} = 3.1$ , and standard deviation,  $s = 0.8$ . The nationwide mean GPR of entering freshmen is  $\mu = 2.7$ . We want to know the probability of getting this sample mean (or higher) if the mean GPR of this university is the same as the nationwide population of students.
- **What do we want to know?** We want the probability of getting a  $\bar{y}$  that is greater than or equal to 3.1 from a population whose mean is 2.7.

## Example 2.18 – Grade Point Ratio Study

- **Statistically**, we want to know  $\Pr(\bar{Y} \geq 3.1)$ , where  $\bar{Y}$  is the sample mean from 16 samples with  $N(2.7, \sigma^2)$ . If we know  $\sigma^2$ , we can use central limit theorem to calculate it. Unfortunately, we do not know the population variance  $\sigma^2$ .
- **Solution:** we can use the *t*-distribution.

$$\Pr(\bar{Y} \geq 3.1) = \Pr\left(\frac{\bar{Y} - 2.7}{\frac{s}{\sqrt{16}}} \geq \frac{3.1 - 2.1}{\frac{s}{\sqrt{16}}}\right) = \Pr\left(T_{15} \geq \frac{3.1 - 2.1}{\frac{0.8}{\sqrt{16}}}\right) = \Pr(T_{15} \geq 2.0)$$

- From Appendix Table on Slide 133, we can find the range of this probability, which is between 0.025 and 0.05. Therefore, we can say that the probability of obtaining a sample mean this large or larger is between 0.025 and 0.05.
- The exact probability from SAS is 0.032.

# Exercise – Filling Bottles

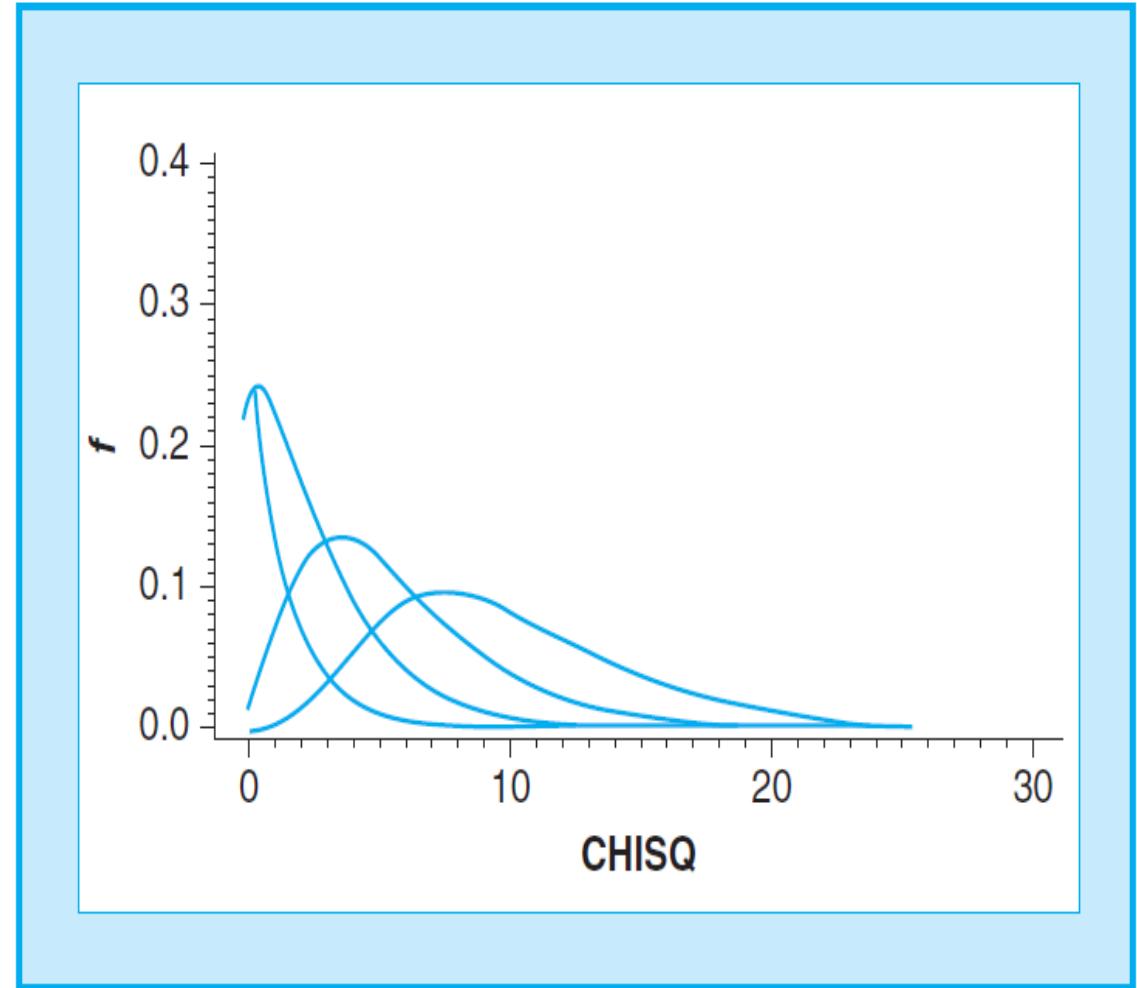
- Consider the filling operation for 20-oz bottles of a popular soft drink. Historically, this operation average 20.2 oz. A recent random sample of 12 bottles yielded these volumes:

|      |      |      |      |      |      |
|------|------|------|------|------|------|
| 20.1 | 20.1 | 20.0 | 19.9 | 20.5 | 20.9 |
| 20.1 | 20.4 | 20.2 | 19.1 | 20.1 | 20.0 |

- From this data, we have  $\bar{y} = 20.12$ ;  $s^2 = 0.1779$
- Exercise:** assume that the historical average is true, find the probability that you can observe a sample mean from 12 random samples not greater than 20.12? **Why we want to do this?**

# Other Distributions

- Chi-square distribution describes the distribution of sample variance.
- We use  $\chi_n$  to represent a random variable with  $n$  degrees of freedom.
- $F$  distribution describes the distribution of the ratio of two estimates of population variance.



# How to Check Normality

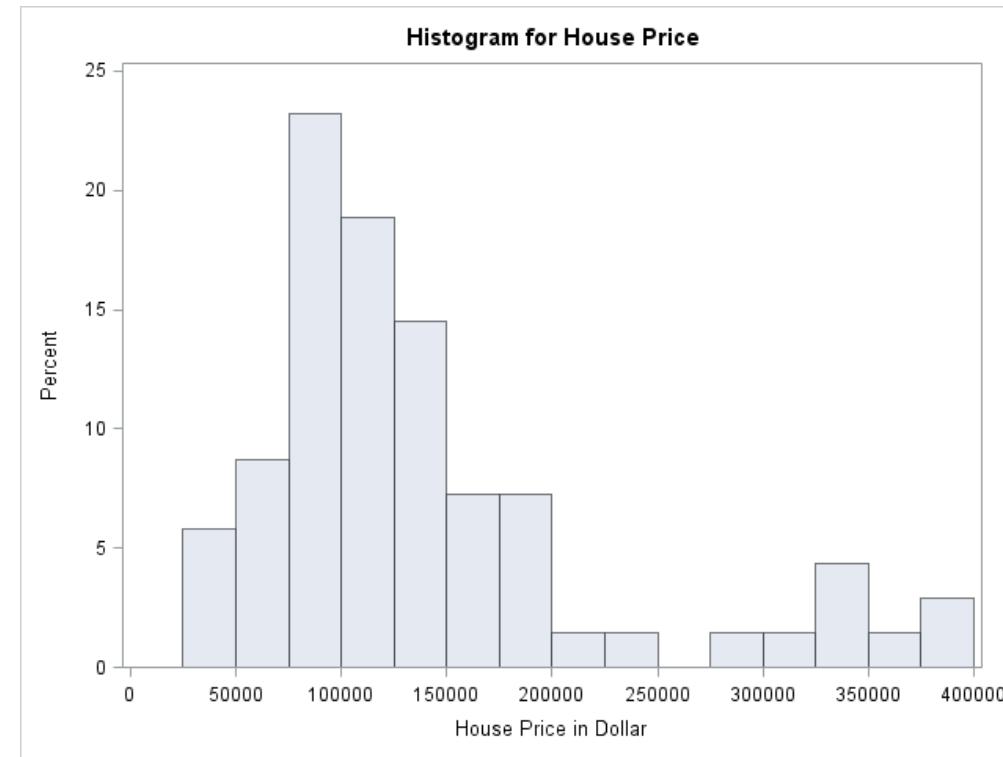
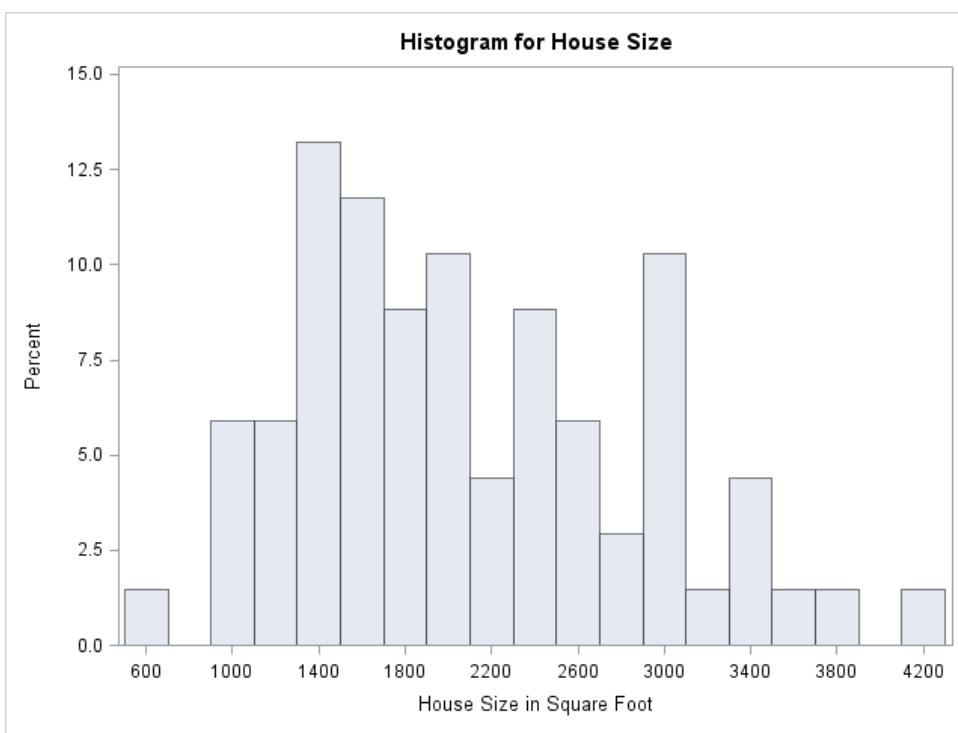
- We can see that we require that the data have a normal distribution in many of our calculations and models.
- How do I know if the data have a normal distribution?
- Construct either a histogram or stem-and-leaf display for the data and check the shape of the graph. If the data are approximately normal, the shape of the histogram or stem-and-leaf display should be bell-shaped.

# Data – Texas House Data

| Obs | Zip | Age | Bed | Bath | Size | Lot   | Exter | garage | fp | Price |
|-----|-----|-----|-----|------|------|-------|-------|--------|----|-------|
| 1   | 3   | 21  | 3   | 2    | 951  | 64904 | Other | 0      | 0  | 30000 |
| 3   | 4   | 7   | 1   | 1    | 676  | 54450 | Other | 2      | 0  | 46500 |
| 5   | 1   | 51  | 3   | 1    | 1186 | 10857 | Other | 1      | 0  | 51500 |
| 7   | 3   | 8   | 3   | 2    | 1368 | .     | Frame | 0      | 0  | 56990 |
| 9   | 1   | 51  | 2   | 1    | 1176 | 6259  | Frame | 1      | 1  | 65500 |

Data: 69 families in a midsized city in east Texas. This is only part of it.

# Histogram - Texas House Data



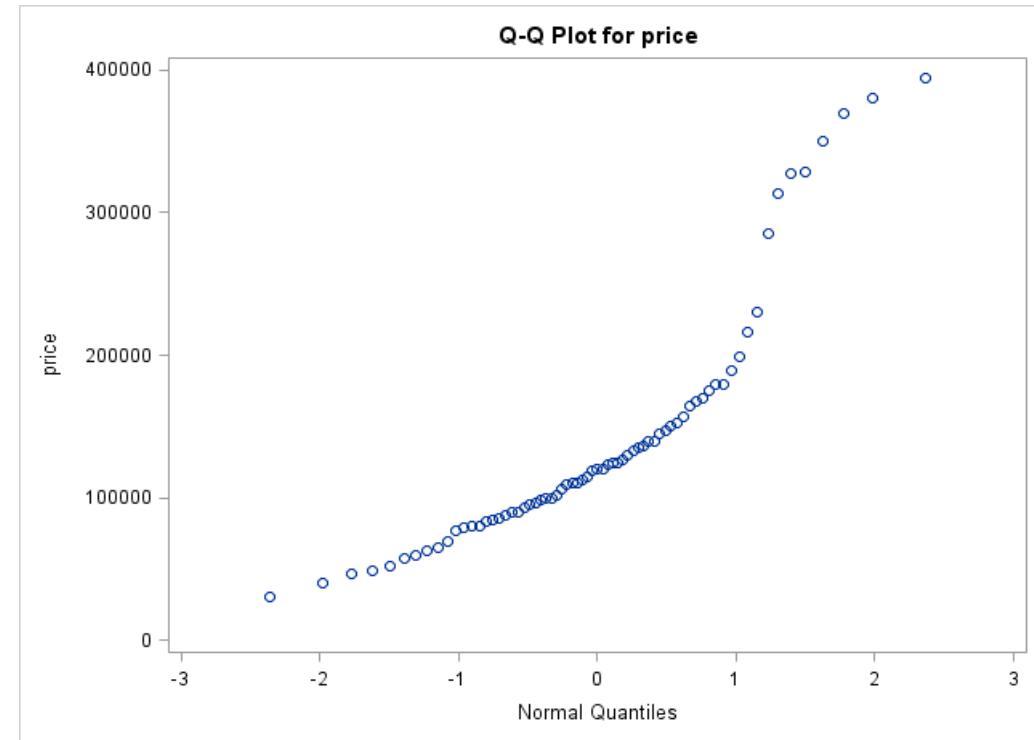
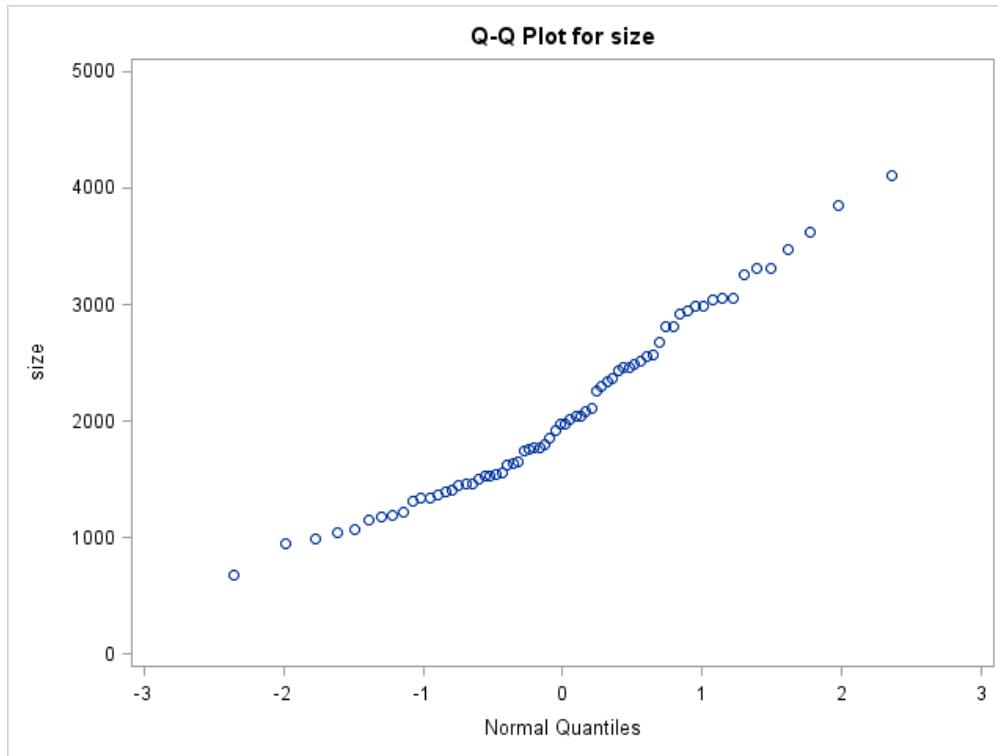
# How to Check Normality

- We can use Quantile-Quantile (Q-Q) plot to check the normality.
- The Q-Q plot is a graphical technique for determining if a set of data plausibly came from some theoretical distribution such as a **Normal** or other distribution.
- A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

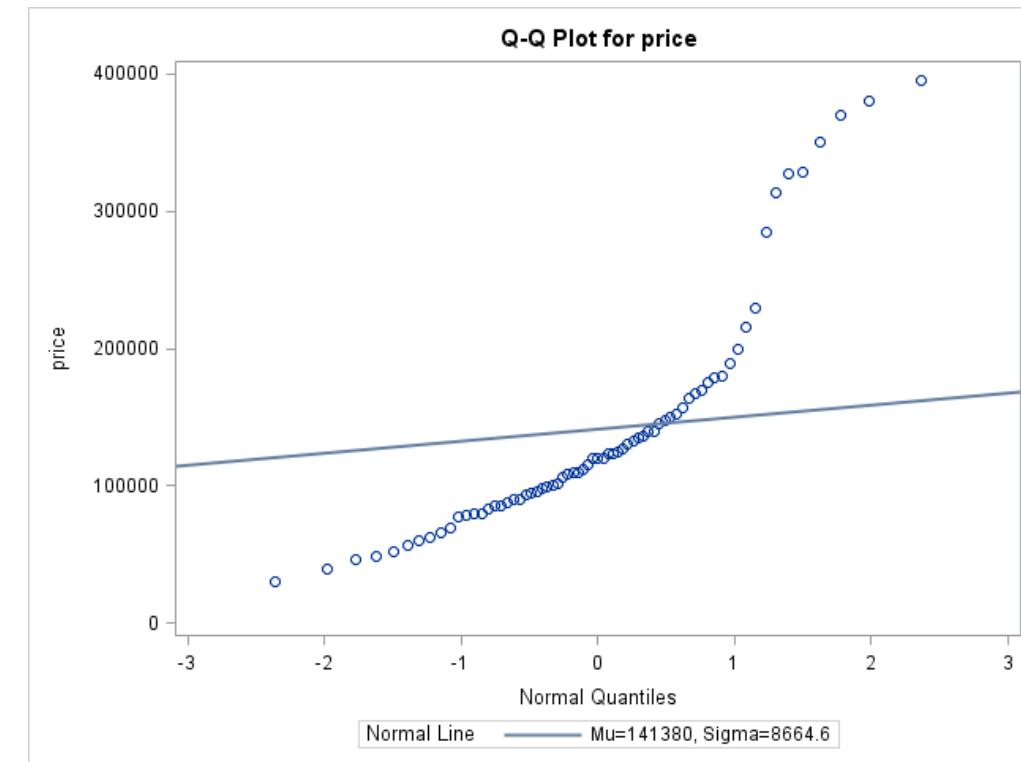
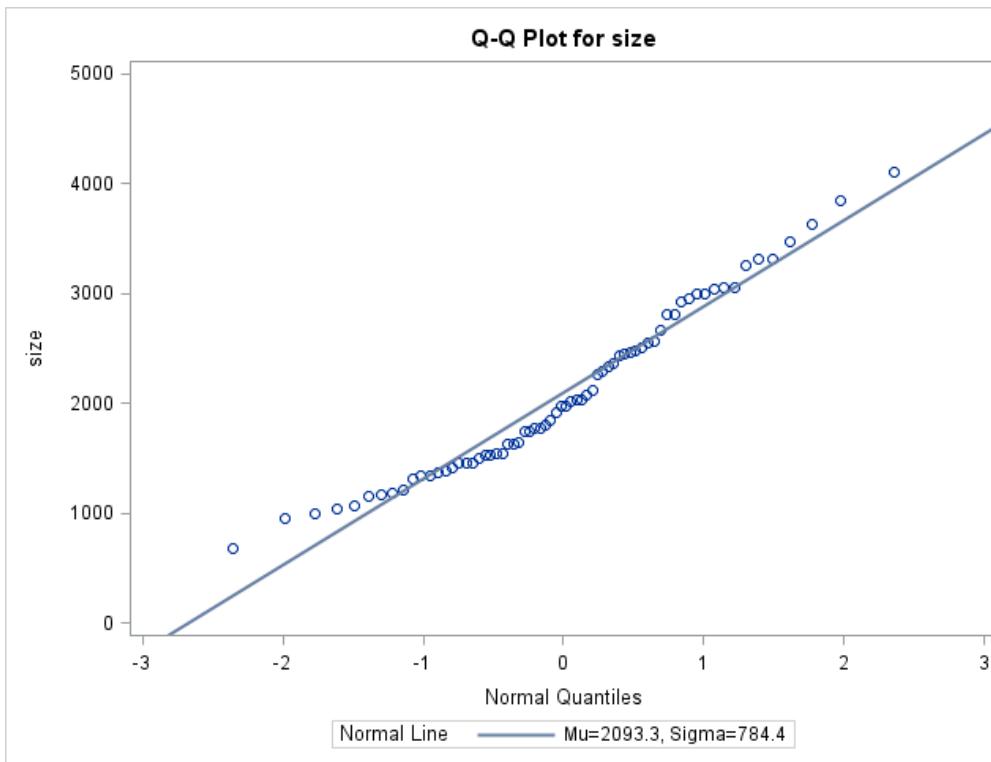
# Normal Q-Q Plot

- The “quantile” are often referred to as “percentiles”.
- Q-Q plots take your sample data, sort it in ascending order, and then plot them versus quantiles calculated from a theoretical distribution.
- The number of quantiles is selected to match the size of your sample data.
- For normal Q-Q plot, the theoretical distribution is normal.
- For example, you can calculate  $Q_1, Q_2, Q_3$  from the data, then correspond quantiles from the standard normal distribution are  $z_{0.75} = -0.6745$ ,  $z_{0.5} = 0$ , and  $z_{0.25} = 0.6745$ , respectively.

# Q-Q Plot - Texas House Data



# Q-Q Plot - Texas House Data



# MA5701: Statistical Methods

Chapter 3 : Principles of Inference

Kui Zhang, Mathematical Sciences



## Example 3.1

- **Example 3.1 (see book)** - The National Center for Education Statistics reports that the year 2007 reading scores for fourth graders had a mean of **220.99** and a standard deviation of **35.73** (from 191,000 fourth grades). You believe that your school district is doing a great job of teaching reading that want to show that mean scores in your district would be higher than this national mean. You randomly select 50 fourth grader in your district, give the same exam, get 230.2 as the sample mean.
- Since your mean is **higher**, this seems to vindicate your belief.
- A critic points out that **you simply may have been lucky** in your sample.
- You can only afford to have this sample; how can you **take sampling variability into account** to explain your high score in your data?

# Introduction

- There are two parts of statistical inference:
  - A **statement** about the value of that parameter.
  - A measure of the **reliability** (probability) of that statement.
- **Two different but related objectives of statistical inference**
  - **Tests of hypotheses** - hypothesize that parameters have some specific values or relationships and make decisions about that. Reliability is the probability that the decision is incorrect.
  - **Estimating parameters** - which is usually done in the form of an interval, and reliability is expressed as probability of true value is covered.
- In this chapter, we present basic principles of statistic inference.

# Populations and Samples

- **Population** – its characteristic is generally described by the parameter.
  - Usually denoted by Greek letters, such as  $\alpha, \beta, \mu$ , etc.
  - It is generally unknown.
  - It is primary of interest of statistical inference and is estimated from sample.
  - In chapter 3, we developed models to describe “populations”.
- **Sample** – its characteristic is generally described by the statistics.
  - Usually denoted by alphabetic letters, such as  $x, y, z, p$ , etc.
  - It can be calculated from the data and is considered as known once the data is collected.
  - It is used to estimate the population parameter.

# Examples

| Population                                                         | Samples                                                                                                                                                    | Statistical Inference                                                                                                                                                                              |
|--------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Difference in elastic strengths of polymer yarn from two machines  | Measure the strength of 10 yarns from each machine                                                                                                         | Use statistics from the sample to infer the unknown parameters                                                                                                                                     |
| Parameters                                                         | Statistics                                                                                                                                                 |                                                                                                                                                                                                    |
| $Y_1 \sim N(\mu_1, \sigma_1^2)$<br>$Y_2 \sim N(\mu_2, \sigma_2^2)$ | $\bar{y}_1, s_1^2$<br>$\bar{y}_2, s_2^2$                                                                                                                   | For example, use $\bar{y}_1$ as an estimate of $\mu_1$                                                                                                                                             |
| Parameters are constants, but unknown.                             | Statistics are random variables, for example,<br>$\bar{Y}_1 \sim N(\mu_1, \frac{\sigma_1^2}{n_1})$ .<br>Statistics are fixed once the sample is collected. | <b>Point Estimation:</b> $\hat{\mu}_1 = \bar{y}_1$<br><b>Interval Estimation:</b> $\bar{y}_1 \pm t_{n-1, \alpha/2} \sqrt{\frac{s_1^2}{n_1}}$<br><b>Hypothesis Testing:</b><br>$H_0: \mu_1 = \mu_2$ |

# Hypothesis Testing - Overview

- **Statistical Hypothesis** – a statement of population parameters.
- **Test of hypothesis** – A procedure that enables us to agree and disagree with hypothesis using data from a sample.
- Hypothesis testing starts by making a set of two statements about the parameter. These two statements are exclusive and exhaustive, which means that one or the other statement must be true, but they cannot both be true.

## Example 3.3 - Filling Peanuts Jars

- **Example 3.3 - Filling Peanuts Jars.** A company that packages salted peanuts in 8-oz. jars is interested in maintaining control on the amount of peanuts put in jars by one of its machines. Control is defined as averaging 8 oz. per jar and not consistently over or under filling the jars. To monitor this control, a sample of 16 jars is taken from the line at random time intervals and their contents weighed. The mean weight of peanuts in these 16 jars will be used to test the hypothesis that the machine is indeed working properly.

## Example 3.3 - Filling Peanuts Jars

- **Example 3.3 – Filling Peanuts Jars.** From the sample of 16 jars, we find the sample mean is

$$\bar{y} = 7.89 \text{ oz.}$$

- We further assume that
  - The population standard deviation,  $\sigma$ , is known and  $\sigma = 0.2$ .
  - In this chapter, we assume that the data is either normally distributed or the sample size is large enough so the Central Limit Theorem can be applied.

# Hypothesis Testing - Overview

- **Example 3.3 – Filling Peanuts Jars.** The mean weight of peanuts in these 16 jars will be used to test the hypothesis that the machine is indeed working properly.
- We can solve this by the hypothesis testing.
  - If the machine is not working properly, we need to take some appropriate actions, which can be costly.
  - If the machine is working properly, then we can leave the process alone.

# Hypothesis Testing - Overview

## Example 3.3 – Filling Peanuts Jars.

- We make two complementary statements:
  - The average weight of peanuts in the Jar is 8 oz.
  - The average weight of peanuts in the Jar is different from 8 oz.
- We need to use data (here are weights from 16 samples) to make a decision about two statements.

# Null Hypotheses

- **Definition 3.1 – Null hypothesis ( $H_0$ )** - a statement about the values of one or more parameters. This hypothesis represents the status quo and is usually not rejected unless the sample results strongly imply that it is false.

# Alternative Hypotheses

- **Definition 3.2 - Alternative hypothesis ( $H_a$ )** - a statement that contradicts the null hypothesis. This hypothesis is accepted if the null hypothesis is rejected. The alternative hypothesis is often called the *research hypothesis* because it usually implies that some action is to be performed, some money spent, or some established theory overturned.

## Hypotheses – Example 3.3

- The null hypothesis is:

$$H_0: \mu = 8$$

- If we fail to reject it, we do not need to do anything.
- The alternative hypothesis is:

$$H_a: \mu \neq 8$$

- If we reject the null hypothesis, we need to correct the process – thus we need strong evidence to do so.

# Fail to Reject Null Hypothesis

- When we reject the null hypothesis, we have strong evidence to support that the alternative hypothesis is true – so we accept the alternative hypothesis.
- When we fail to reject the null hypothesis, we just do not have strong evidence to support that the alternative hypothesis is true. In this case, we do not (and never) state to “accept the null hypothesis”.

# Possible Errors in Hypothesis Testing

- A **type I error** occurs when we incorrectly reject  $H_0$ , that is, when  $H_0$  is true, and our sample-based inference procedure rejects it.
- A **type II error** occurs when we incorrectly fail to reject  $H_0$ , that is, when  $H_0$  is not true, and our inference procedure fails to detect this fact.

|                       |              | In the Population |                   |
|-----------------------|--------------|-------------------|-------------------|
| The Decision          |              | $H_0$ is True     | $H_0$ is Not True |
| $H_0$ is Not Rejected | Correct      | Type II Error     |                   |
| $H_0$ is Rejected     | Type I Error | Correct           |                   |

# A Few More Definitions

- The **rejection region** (also called the **critical region**) is the range of values of a sample statistic that will lead to rejection of the null hypothesis.
- $\alpha$ : denotes the probability of making a type I error;
- $\beta$  : denotes the probability of making a type II error.
- $1 - \beta$ : the power of test.
- In the hypothesis testing, we use a fixed small type I error and try to decrease the type II error (increase the power).

## Example 3.3 – Filling Peanuts Jars

- **Example 3.3 – Filling Peanuts Jars.** We would like to use 16 samples to test if the weight is 8 oz.
- **Rejection Region of Example 3.3** – the sample mean is much larger or less than 8 oz. But the exact rejection depends on  $\alpha$ , the probability of making a type I error and variability of the data (or population).
- Here we assume that we reject the null hypothesis when  $\bar{y} > 8.1$  or  $\bar{y} < 7.9$ .

## Example 3.3 – Filling Peanuts Jars

- When the mean is 8 oz, there is a probability that  $\bar{Y} > 8.1$  or  $\bar{Y} < 7.9$ , this probability is  $\alpha$ , the probability of type I error. Here  $\alpha = 0.0455$ .
- When the mean is 8.15 oz (not 8 oz), there is a probability that  $\bar{Y} < 8.1$  and  $\bar{Y} > 7.9$ , this probability is  $\beta$ , the probability of type II error. Here  $\beta = 0.1587$ .
- When the mean is 8.15 oz (not 8 oz),  $1 - \beta$ , is the probability that  $\bar{Y} > 8.1$  or  $\bar{Y} < 7.9$ , this probability is the power. Here the power is  $1 - \beta = 1 - 0.1587 = 0.8413$ .

# Type I Error, Type II Error, Sample Size

- For any fixed  $\alpha$ , an increase in the sample size will cause a decrease in  $\beta$ .
- For a fixed sample size  $n$ , a decrease in  $\alpha$  will cause an increase in  $\beta$ . Conversely, an increase in  $\alpha$  will cause a decrease in  $\beta$ .
- An increase in the sample size will cause a decrease in both  $\alpha$  and  $\beta$ .

# Fail to Reject Null Hypothesis

- As we have mentioned, we do not “accept” the null hypothesis when we fail to reject the null hypothesis.
- We do not know the population parameter in most situations.
- Also, we use small  $\alpha$  to control type I error. In most cases, we use  $\alpha = 0.05$ . We use smaller  $\alpha$  if the rejection of null hypothesis has serious consequences.

# Type I Error - Example

- **Example** – A drug company tests a new drug and must consider
  - Toxicity (side effects) – null hypothesis here is the drug is toxic, so what  $\alpha$  should be used?
  - For toxicity, since the consequence is so severe if we reject the null hypothesis while the drug is toxic, we use a very small  $\alpha$  such as  $\alpha = 0.0001$  or less is common.

# Type I Error – Example 3.3

- **Example 3.3** – We would like to know if the weight of peanuts is 8 oz.
- **Example 3.3** – Should we use  $\alpha = 0.05$  or  $\alpha = 0.01$ ?
- **Conclusion:** we should use a smaller one,  $\alpha = 0.01$ .
- **Why?**

# Type I Error – Example 3.3

- **Conclusion:** we should use a smaller one,  $\alpha = 0.01$ .
- Since we conduct the test often, we really are more concerned about the type I error than the type II error.
- If we make a type I error (reject the null when it is true), we need to search for the cause of a change when none is present. We are conducting this test so frequently, we run the risk of constantly searching for problems that do not exist. This can be costly.
- Since we are conducting this test so frequently, if we do not detect a true change on any sample (a type II error), we should pick it up later.

# Type I Error, Type II Error, Rejection Region

If the null hypothesis is false, which of these statements characterizes a situation where the value of the test statistic falls in the rejection region?

1. A type I error has been committed.
2. A type II error has been committed.
3. Insufficient information has been given to make a decision.
4. The decision is correct.
5. None of the above is correct.

# Type I Error, Type II Error, Rejection Region

If the value of the test statistic does not fall in the rejection region, the decision is:

1. Reject the null hypothesis.
2. Reject the alternative hypothesis.
3. Fail to reject the null hypothesis.
4. Fail to reject the alternative hypothesis.
5. There is insufficient information to make a decision.

# Type I Error, Type II Error, Rejection Region

If the value of the test statistic falls in the rejection region, then:

1. We cannot commit a type I error.
2. We cannot commit a type II error.
3. We have proven that the null hypothesis is true.
4. We have proven that the null hypothesis is false.
5. None of the above is correct.



# Five-Step Procedure for Hypothesis Testing

- **Step 1:** Specify appropriate  $H_0$ ,  $H_a$ , and an acceptable level of  $\alpha$ .
- **Step 2:** Define a sample-based test statistic based on  $H_0$ .
- **Step 3:** Find the rejection region for the specified  $H_a$  and  $\alpha$ .
- **Step 4:** Collect the sample data and calculate the test statistic.
- **Step 5:** Make a decision to either reject or fail to reject  $H_0$ . This decision will normally result in a recommendation for action. Then interpret the results in the language of the problem.
- **Note:** It is imperative that the results be usable by the practitioner. Since  $H_a$  is of primary interest, this conclusion should be stated in terms of whether there was or was not evidence for the alternative hypothesis.

# Step 1 - Specify $H_0$ , $H_a$ , and Choose $\alpha$

- Note that  $\alpha$  and  $\beta$  are inversely related – for a fixed sample size, we can reduce  $\alpha$  only at the cost of increasing  $\beta$ .
- **Example 3.3 Filling Peanuts Jars –**

$$H_0: \mu = 8$$

$$H_a: \mu \neq 8$$

$$\alpha = 0.01$$

- We call this alternative as a two-sided alternative.



# Three Different Alternatives

- In our course, the null hypothesis is always:

$$H_0: \text{parameter} = \text{a value.}$$

- For alternative hypothesis, we have

- Not equal (two-sided alternative, two-tailed test) -

$$H_a: \text{parameter} \neq \text{a value}$$

- Greater (one-sided alternative, right-tailed test):

$$H_a: \text{parameter} > \text{a value}$$

- Less (one-sided alternative, left-tailed test):

$$H_a: \text{parameter} < \text{a value}$$

# Three Different Alternatives

|                  | Two-Tailed                                  | Right-Tailed<br>(Greater)                | Left-Tailed<br>(Less)                    |
|------------------|---------------------------------------------|------------------------------------------|------------------------------------------|
| Hypotheses       | $H_0: \mu = \mu_0$<br>$H_a: \mu \neq \mu_0$ | $H_0: \mu = \mu_0$<br>$H_a: \mu > \mu_0$ | $H_0: \mu = \mu_0$<br>$H_a: \mu < \mu_0$ |
| Key Words        | Changed<br>Different                        | Improved<br>Greater<br>Increased         | Reduced<br>Less<br>Decreased             |
| Rejection Region | Both Tails                                  | Right Tails                              | Left Tails                               |

## Step 2 – Define Test Statistic

- **Definition 3.8** - The **test statistic** is a statistic whose sampling distribution can be specified for both the null and alternative hypothesis (although sampling distribution when the alternative hypothesis is true may often be quite complex).
- For example, to test the hypothesis about the population mean and assume that the variance is known, we can use the sample mean.
- The sample mean:  $\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$  if the random sample from a normal distribution  $N(\mu, \sigma^2)$  or by the central limit theorem.



## Step 2 – Define Test Statistic

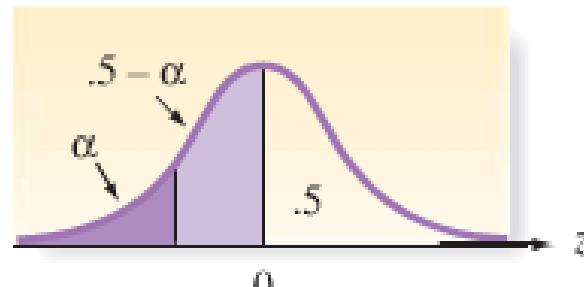
- **Example 3.3 – Filling Peanuts Jars.** We have 16 samples and further assume that the population standard deviation is 0.2 and the data follows a normal distribution.
- The test statistic is  $\frac{\bar{Y}-\mu_0}{\sigma/\sqrt{n}} = \frac{\bar{Y}-8}{0.2/\sqrt{16}} \sim N(0,1)$  when  $H_0: \mu = 8$  is true.
- The test statistic is  $\frac{\bar{Y}-\mu_0}{\sigma/\sqrt{n}} = \frac{\bar{Y}-8}{0.2/\sqrt{16}}$ , what is its distribution when  $H_a: \mu = 8.2$  is true?

## Step 3 – Determine Rejection Region

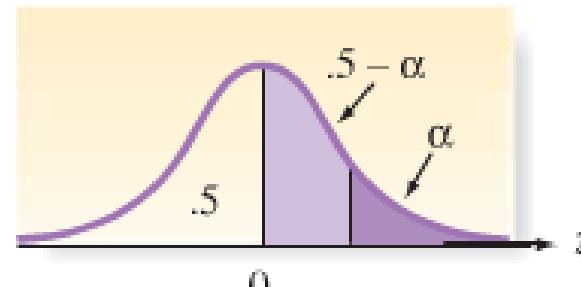
- **Definition 3.9** - The **rejection region** comprises the values of the test statistic for which the researchers reject the null hypothesis ( $H_0$ ).
- The probability of the test statistic falling in the rejection region is the specified  $\alpha$  when the null hypothesis is true.
- **Decision Rule:**

We reject the null hypothesis when the test statistic falls in the rejection region.

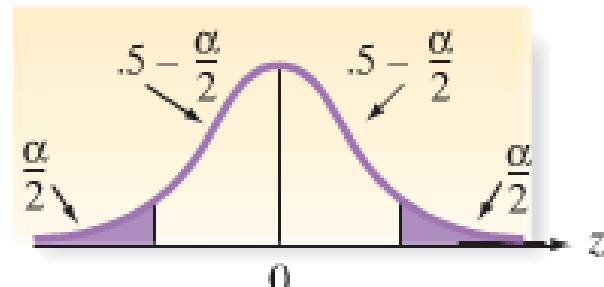
## Step 3 – Determine Rejection Region



a. Form of  $H_a: <$



b. Form of  $H_a: >$



c. Form of  $H_a: \neq$

**Table 6.2** Rejection Regions for Common Values of  $\alpha$

|                | Alternative Hypotheses |              |                             |
|----------------|------------------------|--------------|-----------------------------|
|                | Lower-Tailed           | Upper-Tailed | Two-Tailed                  |
| $\alpha = .10$ | $z < -1.28$            | $z > 1.28$   | $z < -1.645$ or $z > 1.645$ |
| $\alpha = .05$ | $z < -1.645$           | $z > 1.645$  | $z < -1.96$ or $z > 1.96$   |
| $\alpha = .01$ | $z < -2.33$            | $z > 2.33$   | $z < -2.575$ or $z > 2.575$ |

## Step 3 – Determine Rejection Region

**Example 3.3 - Filling Peanuts Jars –**

$$H_0: \mu = 8$$

$$H_a: \mu \neq 8$$

$$\alpha = 0.01$$

- This is a two-sided (two-tailed) test.
- The rejection region is:

$$\left| \frac{\bar{y} - \mu_0}{\sigma / \sqrt{n}} \right| = \left| \frac{\bar{y} - 8}{0.2 / \sqrt{16}} \right| > z_{0.005} = 2.575$$



## Step 4 and Step 5

- **Step 4:** Collect the sample data and calculate the test statistic.
- **Example 3.3 – Filling Peanuts Jars,** we have

$$z = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} = \frac{7.89 - 8.0}{0.2/\sqrt{16}} = -2.2$$

- **Step 5:** Make a decision to either reject or fail to reject  $H_0$ .
- **Example 3.3,** we have  $|z| = 2.2 < 2.575$ , we fail to reject the null hypothesis.



## Step 5 - Interpretation

- **Step 5:** Then interpret the results in the language of the problem.
- If we test statistic falls in rejection region, we reject the null hypothesis, we state that:

At the ( )% significance level, the test statistic ( $z =$ ) falls in the rejection region. Therefore, we reject the null hypothesis. The data provides sufficient evidence that (state the problem specified in the problem).



## Step 5 - Interpretation

- **Step 5:** Then interpret the results in the language of the problem.
- If we test statistic does not fall in rejection region, we fail to reject the null hypothesis, we state that:

At the ( )% significance level, the test statistic ( $z =$ ) does not fall in the rejection region. Therefore, we fail to reject the null hypothesis. The data does not provide sufficient evidence that (state the problem specified in the problem).



## Step 5 - Interpretation

- **Step 5:** Then interpret the results in the language of the problem.
- **Example 3.3 – Filling Peanuts Jars,** we state like this way:

At the 1% significance level, the test statistic  $z = -2.2$  does not fall in the rejection region. Therefore, we fail to reject the null hypothesis. The data dose not provide sufficient evidence that the filling process is out of control (or the machine is not working properly).



## Step 5 – Interpretation

- When we reject the null hypothesis at the  $\alpha$  level, we can also state that “that is statistically significant at the  $\alpha$  level”.
- When we fail to reject the null hypothesis at the  $\alpha$  level, we can also state that “that is not statistically significant at the  $\alpha$  level”.
- For **Example 3.3 – Filling Peanuts Jars**, we can state that “the difference between the weight of filling process and the weight of 8 oz is not statistically significant at the 0.01 level”.
- For **Example 3.3 – Filling Peanuts Jars**, we can also state that “the weight of filling process is not statistically significantly different from the weight of 8 oz at the 0.01 level”.



## Step 5 – Interpretation

- Therefore, when we state that “that is statistically significant at the  $\alpha$  level”, we just mean that we reject the null hypothesis at the  $\alpha$  level.
- Therefore, the true meaning of “that is statistically significant at the  $\alpha$  level” is that if the null hypothesis is true, the corresponding statistic observed in the sample would occur by chance with probability of no more than  $\alpha$ .

# Hypothesis Testing - Interpretation

A research report states: The differences between public and private school seventh graders' attitudes toward minority groups was statistically significant at the  $\alpha = 0.05$  level. This means that:

1. It has been proven that the two groups are different.
2. There is a probability of 0.05 that the attitudes of the two groups are different.
3. There is a probability of 0.95 that the attitudes of the two groups are different.
4. If there is no difference between the groups, the difference observed in the sample would occur by chance with probability of no more than 0.05.
5. None of the above is correct.

# *P*-Values – Observed Significance Level

- Problems with significance level and rejection region:
  - Users would have to specify an alpha and determine the corresponding rejection region for every test being requested.
  - The conclusion may be affected by very minor changes in sample statistics.
  - Only give a “discrete” but not “continuous” scale for rejecting the null hypothesis.
  - In many situations, we may need want to use our own “alpha”.
- Most statistical packages provide *p*-value, since what alpha that a researcher would like to use is generally unknown.

# *P*-Values – Observed Significance Level

- **Definition 3.10** - The ***p*-value** is the probability of committing a type I error if the actual sample value of the statistic is used as the boundary of the rejection region.
- The ***p*-value** is also the probability of observing a test statistic or more extreme.
- **Example 3.3 – Filling Peanut Jars** – The test statistic is -2.2, we have

$$p\text{-value} = 2\Pr(Z > |-2.2|) = 0.0278$$

# *P*-Values – Observed Significance Level

- If the  $p$ -value  $< \alpha$ , we reject the null hypothesis.
- If the  $p$ -value  $> \alpha$ , we failed to reject the null hypothesis.
- Therefore, the  $p$ -value is therefore the smallest level of significance for which we would reject the null hypothesis with that sample.
- Consequently, the  $p$ -value is often called the “attained” or the “observed” significance level. It is also interpreted as an indicator of the weight of evidence against the null hypothesis.
- Smaller  $p$ -value indicates stronger evidence to reject the null hypothesis.

# *P*-value Approach

To use the *p*-value to test hypothesis, you can:

- In Step 3, skip the calculation of rejection region.
- In Step 4, calculate the corresponding *p*-value based on  $H_a$  in Step 1.
- In Step 5, make a decision based on the *p*-value value and the significance level -  $\alpha$  in Step 1. Interpret the results.



# Calculation of $p$ -value

- For two-sided test ( $H_a: \mu \neq \mu_0$ ),

$$p\text{-value} = 2\Pr(Z > |z|)$$

- For upper-tailed (right-tailed) test ( $H_a: \mu > \mu_0$ ),

$$p\text{-value} = \Pr(Z > z)$$

- For lower-tailed (left-tailed) test ( $H_a: \mu < \mu_0$ ),

$$p\text{-value} = \Pr(Z < z)$$

- When the  $p$ -value  $> \alpha$ , we fail to reject the null hypothesis.
- When the  $p$ -value  $< \alpha$ , we reject the null hypothesis.

# *P*-value Approach

In a hypothesis test the *p* value is 0.043. This means that we can find statistical significance at:

1. both the 0.05 and 0.01 levels
2. the 0.05 but not at the 0.01 level
3. the 0.01 but not at the 0.05 level
4. neither the 0.05 or 0.01 levels

# *P*-value Approach

You are reading a research article that states that there is no significant evidence that the median income in the two groups differs, at  $\alpha = 0.05$ . You are interested in this conclusion but prefer to use  $\alpha = 0.01$ .

1. You would also say there is no significant evidence that the medians differ.
2. You would say there is significant evidence that the medians differ.
3. You do not know whether there is significant evidence or not, until you know the *p*-value.

# *P*-value Approach – Example 3.3

- **Example 3.3 – Filling Peanuts Jars.** Use the *p*-value approach to test

$$H_0: \mu = 8 \text{ versus } H_a: \mu \neq 8; \alpha = 0.01$$

- The *p*-value is:

$$p\text{-value} = 2 * \Pr(Z > |-2.2|) = 0.0278.$$

- The *p*-value = 0.0278 is greater than the significance level  $\alpha = 0.01$ . Therefore, we fail to reject the null hypothesis. The data does not provide sufficient evidence that the filling process is out of control (or the machine is not working properly).

## Example 3.4 – Aptitude Test

- **Example 3.4 – Aptitude Test.** An aptitude test has been used to test the ability of fourth graders to reason quantitatively. The test is constructed so that the scores are normally distributed with **a mean of 50 and standard deviation of 10**. It is suspected that, with increasing exposure to computer-assisted learning, the test has become obsolete. That is, **it is suspected that the mean score is no longer 50, although  $\sigma$  remains the same**. This suspicion may be tested based on a sample of students who have been exposed to a certain amount of computer-assisted learning. A sample of **500 students** was collected and the **sample mean was 51.07**.



## Example 3.4 – Aptitude Test

**Example 3.4 – Aptitude Test.** What information can we get?

- Sample size:  $n = 500$ , which is large enough so the central limit theorem can be applied to the sample mean.
- The sample mean is 51.07.
- A standard deviation of 10 is given – this should be used as the population standard deviation.
- We are interested in if the score is **different** from 50.
- Here we need the significance level,  $\alpha$ , which is set as 0.05.

# Uniformly Most Powerful Tests

- The test with higher power is more desirable so the test with the highest power should be used.
- For any specified alternative hypothesis, sample size, and level of significance, the test with the highest power is called a “**uniformly most powerful**” (**UMP**) test
- The construction of a power curve and the UMP test are not simple, and it becomes increasingly difficult for the applications in subsequent chapters.
- In our course, virtually all of the procedures we will be using provide uniformly most powerful tests, assuming that basic assumptions are met.
- Power calculations for more complex applications can be made easier through the use of computer programs. While there is no single program that calculates power for all hypothesis tests, some programs either have the option of calculating power for specific situations or can be adapted to do so.

# Example 3.1 – As An Exercise

- **Example 3.1 (see book)** - The National Center for Education Statistics reports that the year 2007 reading scores for fourth graders had a mean of **220.99** and a standard deviation of **35.73** (from 191,000 fourth grades). You believe that your school district is doing a great job of teaching reading and want to show that mean scores in your district **would be higher than** this national mean. You randomly select **50 fourth graders** in your district, give the same exam, get **230.2** as the sample mean.



## Example 3.1 – As An Exercise

**Example 3.1 (see book)** – Get to know some basics:

- A sample of 50 fourth grader is selected, so  $n = 50$ , which is large enough so the sample mean is has an approximate normal distribution even the score does not have a normal distribution.
- The sample mean is 230.2.
- A standard deviation of **35.73** is given – this should be used as the population standard deviation.
- We are interested in if the score in our district is **greater** than the national average.
- Here we also need the significance level,  $\alpha$ , which is set as 0.05.

# Point Estimator

- **Estimation** is a inferential procedure to use data from a sample to estimate the value of a parameter of the population.
- An ***estimator*** is a statistic used to estimate an unknown parameter of a population. We generally use  $\hat{\theta}$  to represent an estimator of the arbitrary parameter.
- Examples of estimator:
  - Sample mean for population mean:  $\hat{\mu}_1 = \bar{y}_1$ .
  - Sample variance for population variance:  $\hat{\sigma}_1^2 = s_1^2$ .

# Point Estimator - Example 3.3

**Example 3.3 – Filling Peanuts Jars.** A sample of 16 jars with peanuts was collected to see if the filling process is normal. So we are interested in two questions:

- What is the true current mean weight of peanuts in jars?
- Is the current mean weight different from the mean weight of 8 oz?

# Point Estimator - Example 3.3

**Example 3.3 -Filling Peanuts Jars.** Suppose that the weight of peanuts follows a normal distribution with mean  $\mu$  and variance  $\sigma^2$ .

- We may estimate the population mean using either sample mean or sample median.
- We can also estimate the population variance using either sample variance or some constant multiple of the interquartile range.
- The question here is: **which estimator should be used?**
- There are statistical theories about how to choose the best estimator. The estimators used in our book are generally “optimal”.

# Unbiased Estimator

- An ***unbiased estimator*** of an unknown parameter is one whose expected value is equal to the parameter of interest. In other words, if  $E(\hat{\theta}) = \theta$ , then  $\hat{\theta}$  is unbiased estimator of  $\theta$ .
- For example,  $\bar{y}$  and  $s^2$  from a random sample from  $N(\mu, \sigma^2)$  are unbiased, since

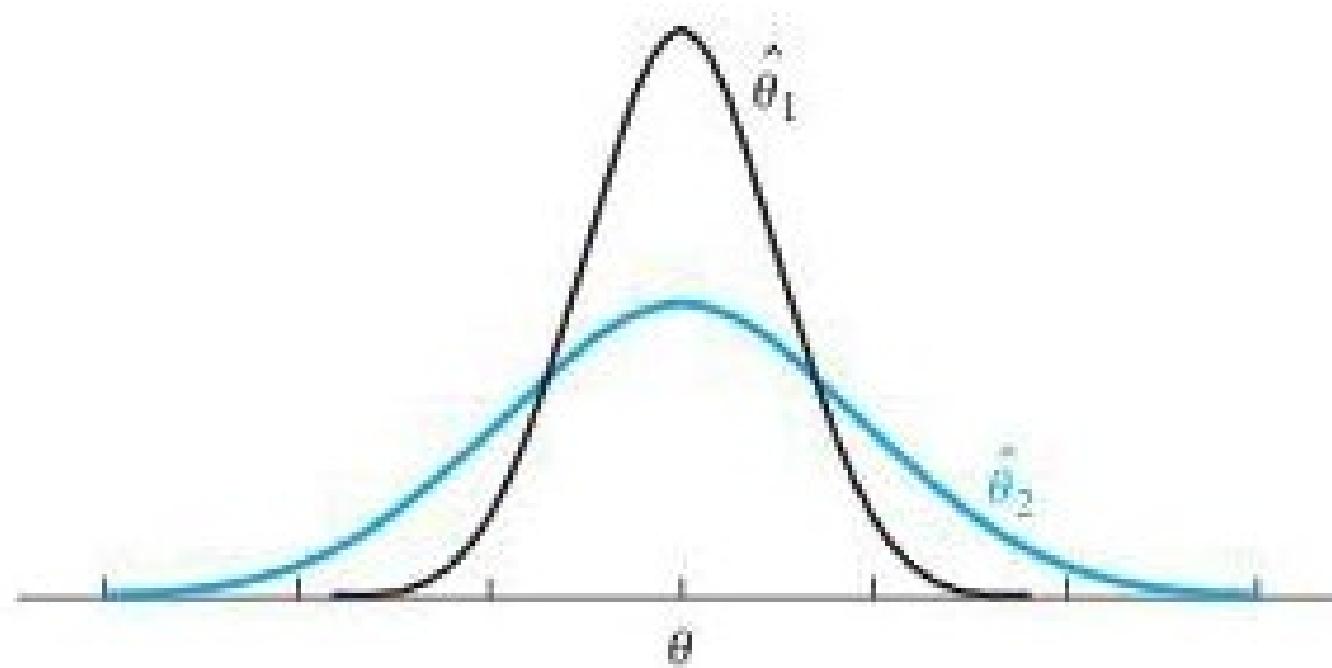
$$E(\bar{Y}) = \mu \text{ and } E(S^2) = \sigma^2.$$

- However, sample standard deviation is a biased estimator of population standard deviation, because

$$E(S) \neq \sigma$$

# Precision of Unbiased Estimators

- An estimator is *more precise* if its sampling distribution has a smaller standard error (or variance).



# Sample Mean from a Normal Population

For a random sample from  $N(\mu, \sigma^2)$ , the sample mean is the “best” estimator for the population mean  $\mu$ , because:

1. The sample mean is an unbiased estimator of the population mean.
2. Among all unbiased estimator of the population mean, the sample mean has the smallest variance (Need some sophisticated statistical theory to prove this statement).

# Point Estimators

- The sample mean is a *point estimator* since we use this specific value to estimate the parameter. If the random sample is from  $N(\mu, \sigma^2)$ , then  $\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ .
- We have  $\Pr(\bar{Y} = \mu) = 0$ .
- Thus, we know that this **point estimate has no chance of being correct**.
- Thus, we prefer to use interval estimators.

# Interval Estimators – Variance Known

- We assume the random sample is from a normal distribution, or the central limit theorem holds, we have  $\bar{Y} \sim N(\mu, \frac{\sigma^2}{n})$ .
- The two-sided interval estimate of the population mean is given by

$$\left( \bar{y} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{y} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)$$

# Interval Estimators – Variance Known

For interval estimate:  $\left(\bar{y} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{y} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right)$

- Here,  $1 - \alpha$  is the **coverage probability (confidence level)**.
- We call it as  $100(1 - \alpha)\%$  confidence interval for  $\mu$ .
- We can use any  $\alpha$ . But in general, we use  $\alpha = 0.05$  or  $\alpha = 0.01$ .

# Interval Estimators – Variance Known

For interval estimate:  $\left(\bar{y} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{y} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right)$

- $z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$  is the **margin of error (Definition 3.13)** which is defined as one-half the width of a confidence interval.
  - It increases with increased confidence coefficient (decreased  $\alpha$ ).
  - It increases with decreased sample size.
  - It increases with increased population variance.

# Interval Estimators – Variance Known

- We have:

$$\begin{aligned}1 - \alpha &= \Pr\left(-z_{\alpha/2} < \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}} < z_{\alpha/2}\right) \\&= \Pr\left(-z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \bar{Y} - \mu < z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) \\&= \Pr\left(-z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu - \bar{Y} < z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) \\&= \Pr\left(\bar{Y} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{Y} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right)\end{aligned}$$

# Interval Estimators – Interpretation

- The  $100(1 - \alpha)\%$  (or just use  $1 - \alpha$ ) confidence interval of  $\mu$  is

$$\left( \bar{y} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{y} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)$$

- Therefore, **we are  $100(1 - \alpha)\%$  confident that the population mean  $\mu$  is inside of this interval.**
- We can state: **the probability that the interval generated by this way covers the true population mean  $\mu$  is  $1 - \alpha$ .**

# Interval Estimators – Interpretation

- The  $100(1 - \alpha)\%$  confidence interval of  $\mu$  is
$$\left(\bar{y} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{y} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right)$$
- You **can not** state: **the probability that the population mean  $\mu$  is inside of this interval is  $1 - \alpha$  since  $\mu$  is a constant, not a random variable.**
- For a given sample, you **can not** state that: **the probability that the interval covers the true  $\mu$  is  $1 - \alpha$  since after the confidence interval is calculated, it either covers or does not cover the true  $\mu$ .**

# Interval Estimators – Examples/Exercises

- **Example 3.3 – Filling Peanuts Jars.** From the sample of 16 jars, we find the sample mean is  $\bar{y} = 7.89$ . The population standard deviation,  $\sigma$ , is known and  $\sigma = 0.2$ .
- Find 99% confidence interval of the population mean.
- **Solution:** The 99% confidence interval of population mean is:

$$\left( \bar{y} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right) = \left( 7.89 \pm 2.5758 * \frac{0.2}{\sqrt{16}} \right) = (7.761, 8.019)$$

# Interval Estimators – Examples/Exercises

- **Example 3.3 – Filling Peanuts Jars.** From the sample of 16 jars, we find the sample mean is  $\bar{y} = 7.89$ . The population standard deviation,  $\sigma$ , is known and  $\sigma = 0.2$ .
- Find 95% confidence interval of the population mean.
- **Solution:** The 95% confidence interval of population mean is:

$$\left( \bar{y} \pm z_{0.05} \frac{\sigma}{\sqrt{n}} \right) = \left( 7.89 \pm 1.96 * \frac{0.2}{\sqrt{16}} \right) = (7.792, 7.988)$$

# Interval Estimators – Interpretation

Interpretation of 95% confidence interval,  $(7.792, 7.988)$ . Which statement is correct:

1.  $\Pr(7.792 < \mu < 7.988) = 0.95$ .
2. 95% of all weights between 7.792 and 7.988.
3. We sampled 95% of all weights.
4. We know that  $7.792 < \mu < 7.988$ .
5. We are 95% confident that the true population mean is between 7.792 and 7.988.

# Interval Estimators – Interpretation

The blood pressure of 100 patients from hospitable is collected. Based on the data, the 95% confidence interval of mean blood pressure of patients in a hospital is from 95.1 to 143.6, which of the following statements is correct:

1. The true mean blood pressure of patients in that hospital is between 95.1 and 143.6.
2. About 95% of patients in that hospital will have the blood pressure in the interval from 95.1 to 143.6.
3. The probability of true mean blood pressure of patients in that hospital in the interval from 95.1 to 143.6 is 0.95.
4. None of the above is correct.



# Interval Estimators – Exercises

- **Example 3.4 – Aptitude Test.** A sample of 500 students was collected and the sample mean was 51.07. The population standard deviation is 10.
- Find 99% confidence interval of the population mean.
- **Solution:** In Class Exercise.

# Hypothesis Testing and Confidence Interval

- There is a direct relationship between hypothesis testing and confidence interval estimation (for two-tailed hypothesis tests)
  - A hypothesis test for  $H_0: \mu = \mu_0$  against  $H_1: \mu \neq \mu_0$  will be rejected at a significance level of  $\alpha$  if  $\mu_0$  is not in the  $(1 - \alpha)$  confidence interval for  $\mu$ .
  - Any value of  $\mu$  inside the  $(1 - \alpha)$  confidence interval will not be rejected by an  $\alpha$  -level significance test.



# Hypothesis Testing and Confidence Interval

- **Example 3.3** – 95% CI is (7.792, 7.988), we reject  $H_0: \mu = 8$  at the 0.05 level of significance.
- **Example 3.4** – 99% CI is (49.92, 52.22), we do not reject  $H_0: \mu = 50$  at the 0.01 level of significance.
- For one-tailed test, we need one-sided confidence interval.

# One-Sided Interval Estimators When Variance Known

- The  $100(1 - \alpha)\%$  upper (one-sided) confidence interval of  $\mu$  is given by

$$\left(-\infty, \bar{y} + z_\alpha \frac{\sigma}{\sqrt{n}}\right)$$

- The  $100(1 - \alpha)\%$  lower (one-sided) confidence interval of  $\mu$  is given by

$$\left(\bar{y} - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty\right)$$

# One-Sided Interval Estimators - Example 3.1

**Example 3.1 – Some basics:**

- A sample of 50 fourth grader is selected with a sample mean of 230.2.
- The population standard deviation is **35.73**.
- We are interested in if the score in our district is **greater** than the national average, which is 220.99, at  $\alpha = 0.05$ .
- We need to calculate 95% lower confidence interval.
- **Solution:**  $\left(\bar{y} - z_{\alpha} \frac{\sigma}{\sqrt{n}}, \infty\right) = \left(230.2 - 1.645 * \frac{35.73}{\sqrt{50}}, \infty\right) = (221.89, \infty)$

# One-Sided Interval Estimators - Example 3.3

- **Example 3.3 – Filling Peanuts Jars.** From the sample of 16 jars, we find the sample mean is

$$\bar{y} = 7.89 \text{ oz.}$$

- We further assume that
  - The population standard deviation,  $\sigma$ , is known and  $\sigma = 0.2$ .
  - Now we are more interested in knowing if  $H_a: \mu < 8$ .



# One-Sided Interval Estimators - Example 3.3

- **Example 3.3 – Filling Peanuts Jars.**
- Five steps to perform the test. Assume  $\alpha = 0.05$ .
- Calculate the  $p$ -value.
- Calculate the appropriate one-sided confidence interval.

# One-Sided Interval Estimators - Example 3.3

- **Step 5:** Then interpret the results in the language of the problem.
- **Example 3.3 – Filling Peanuts Jars (Slide 78)** , we state like this way:

At the 5% significance level, the test statistic  $z = -2.2$  falls in the rejection region. Therefore, we reject the null hypothesis. The data provides sufficient evidence that the weight of peanuts from the filling process is less than the weight of 8 oz.

# One-Sided Interval Estimators - Example 3.3

- **Step 5:** Then interpret the results in the language of the problem.
- **Example 3.3 – Filling Peanuts Jars (Slide 78)** , we state like this way:

At the 5% significance level, the p-value is 0.0139 and less than 0.05. Therefore, we reject the null hypothesis. The data provides sufficient evidence that the weight of peanuts from the filling process is less than the weight of 8 oz.

# One-Sided Interval Estimators - Example 3.3

- **Step 5:** Then interpret the results in the language of the problem.
- **Example 3.3 – Filling Peanuts Jars (Slide 78)** , we state like this way:

The 95% upper confidence interval of the mean weight of peanuts in jars is  $(-\infty, 7.972)$ , which does not contain the weight of 8 oz. Therefore, we reject the null hypothesis at 5% significance level. The data provides sufficient evidence that the weight of peanuts from the filling process is less than the weight of 8 oz.



# Sample Size for Interval Estimation

- We will skip this part. Please refer to the textbook if you want to learn this part.

# Probability of Type II Error

- There are many reasons for concerning the probability of the type II error, for example:
  - The probability of making a type II error may be so large that the test may not be useful.
  - Because of the trade-off between  $\alpha$  and  $\beta$ , we may find that we may need to increase  $\alpha$  in order to have a reasonable value for  $\beta$ .
  - Sometimes we have a choice of testing procedures where we may get different values of  $\beta$  for a given  $\alpha$ .

# Power and Sample Size

- **Definition 3.11** - The **power** of a test is the probability of correctly rejecting the null hypothesis when it is false.
- Therefore, we have power =  $1 - \beta$ .
- More reasons for concerning the probability of the type II error ( or the power):
  - Sometimes we have a choice of testing procedures where we may get different values of  $\beta$  for a given  $\alpha$ . So we want a test with the largest power.
  - We may need to calculate sample size to ensure the study will have enough power.
- The calculation of power and sample size can be very difficult in some situations.



# Power Calculation

Use the following steps to calculate the power for  $\mu = \mu_1 (\mu_1 \neq \mu_0)$ :

- **Step 1** – Find the rejection region,  $R$ . We have  $\Pr\left(\frac{\bar{Y} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \in R\right) = \alpha$ .
- **Step 2** – Rewrite the formula -  $\Pr(\bar{Y} \in R^*) = \alpha$ . Note that under the null hypothesis  $\mu = \mu_0$ ,  $\bar{Y} \sim N(\mu_0, \frac{\sigma^2}{n})$ .
- **Step 3** – The power is  $\Pr(\bar{Y} \in R^*)$ . Note that under the alternative hypothesis  $\mu = \mu_1$ ,  $\bar{Y} \sim N(\mu_1, \frac{\sigma^2}{n})$ .

# Power Calculation – Example

- **Example:** Assume that a random sample of size 25 is to be taken from a normal population with  $\mu = 10$  and  $\sigma = 2$ . The value of  $\mu$ , however, is not known by the person taking the sample.
- **Problem 1 (Used as An Example)** Suppose the person wanted to test  $H_0: \mu = 10.6$  against  $H_a: \mu < 10.6$ . Compute the power for  $\alpha = 0.05$  and  $\mu = 10$ .
- **Problem 2 (Used as An Exercise):** Suppose the person wanted to test  $H_0: \mu = 10.6$  against  $H_a: \mu \neq 10.6$ . Compute the power for  $\alpha = 0.05$  and  $\mu = 10$ .



# Sample Size for Hypothesis Testing

- Sample size determination must satisfy:
  - Required Confidence - level of significance ( $\alpha$ ) for hypothesis testing
  - Error – the probability of type II error ( $\beta$ ) or power for hypothesis testing
  - The difference, called  $\delta$  (delta), between the hypothesized value and the specified value ( $\delta = \mu_a - \mu_0$ )
  - The population variance or standard deviation.
- If the population standard deviation is unknown, we can use the following empirical rule:
  - Standard deviation is estimated by the range divided by 4.



# Sample Size for Hypothesis Testing

- One-tailed Tests -  $n = \sigma^2(z_\alpha + z_\beta)^2/\delta^2$
- Two-tailed Tests -  $n = \sigma^2(z_{\alpha/2} + z_\beta)^2/\delta^2$
- **Example 3.6** -  $H_0: \mu = 35; H_1: \mu > 35; \alpha = 0.05; \beta = 0.10; \delta = 37 - 35 = 2$

# Sample Size for Hypothesis Testing - Example

- **Example 3.6** - In a study of the effect of a certain drug on the behavior of laboratory animals, a research psychologist needed to determine the appropriate sample size. The study was to estimate the time necessary for the animal to travel through a maze under the influence of this drug.
- We know: (1)  $\alpha = 0.05$ ; (2)  $\beta = 0.10$ ; (3)  $\mu_0 = 35$ ; (4)  $\mu_1 = 37$ ; (5) an anticipated range of times of 15 to 60 seconds.
- **Solution:**  $\delta = 37 - 35 = 2$  and  $\sigma = \frac{60-15}{\sqrt{4}} = 11.25$ . So sample size is  
$$n = \frac{11.25^2 * (z_{0.05} + z_{0.10})^2}{2^2} = \frac{11.25^2 * (1.64485 + 1.28155)^2}{2^2} = 271$$



# Assumptions about Normality

- Generally, are based on normal distribution
- Robust methods have been developed when normality is not satisfied
  - Generally they have wider confidence intervals and/or lower power
- Two principles to develop robust methods
  - Trimming, which consists of discarding a small pre-specified portion of the most extreme observations and making appropriate adjustments to the test statistics.
  - Nonparametric methods, which avoid dependence on the sampling distribution by making strictly probabilistic arguments (often referred to as distribution-free methods).



# Statistical Significance versus Practical Significance

- We can have a statistically significant result that has no practical implications.

## Example 3.7

- **Example 3.7.** In the January/February 1992 *International Contact Lens Clinic* publication, there is an article that presented the results of a clinical trial designed to determine the effect of defective disposable contact lenses on ocular integrity (Efron and Veys, 1992).
- This is double blind experiment: 29 samples who wore a defective lens in one eye and a nondefective one in the other. Neither the research officer nor the subject knew which eye wore the defective lens.

## Example 3.7 – Statistical Significance

- The study indicated that a significantly greater ocular response was observed in eyes wearing defective lenses in the form of corneal epithelial microcysts (among other results) -  $p$  value is 0.04. With a level of significance of 0.05, the conclusion would be that the defective lenses resulted in more microcysts being measured.

## Example 3.7 – No Practical Significance

- The study reported a mean number of microcysts for the eyes wearing defective lenses as 3.3 and the mean for eyes wearing the nondefective lenses as 1.6.
- In an invited commentary, Dr. X points out that the observation of fewer than 50 microcysts per eye requires no clinical action other than regular patient follow up.
- We may use the following hypothesis test:  $H_a: |d| \geq 50$ .



# Chapter Summary

- Steps to conduct hypothesis tests:
  - State the hypotheses and the significance level
  - Collect data and compute test statistics
  - Make a decision to confirm or deny hypothesis.
- Steps to calculate confidence interval:
  - Identify the parameter and the confidence level
  - Collect data and compute the statistics for the confidence interval
  - Interpret the interval in the context of the situation.
- The distinction between null and alternative hypotheses can be difficult in some situations.

# MA5701: Statistical Methods

Chapter 4 : Inferences on a Single Population

Kui Zhang, Mathematical Sciences

# Introduction

- The examples used in Chapter 3 to introduce the concepts of statistical inference were not very practical, because they required outside knowledge of the population variance. We did this to avoid distractions from issues that were irrelevant to the principles we were introducing.

# Introduction

- In this chapter, we will present procedures for:
  - Making inferences on the mean of a normally distributed population where the variance is unknown.
  - Making inferences on the proportion of successes in a binomial population.

# Inference on Population Mean

- In Chapter 3, we use the statistic  $z = \frac{(\bar{y}-\mu)}{\sigma/\sqrt{n}}$ , which has the standard normal distribution, to make inferences about the population mean.
- This statistic has limited practical value because, if the population mean is unknown, it is also likely that the variance of the population is unknown.
- The idea is to use the estimate of population variance in the statistic.

# One Sample $t$ -test – Test Statistic

- The test statistic is:

$$T = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}}$$

- Where  $S$  is the sample standard deviation.
- When the null hypothesis is true ( $\mu = \mu_0$ ),

$$T = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

- **Question:** when  $\mu \neq \mu_0$ , what is the distribution of  $T$ ?

# Hypothesis Test on Population Mean

- The test is called ***t*-test**.
- To test the hypothesis  $H_0: \mu = \mu_0$  with the significance level of  $\alpha$ .
  - The test statistic is:  $t = \frac{(\bar{y} - \mu_0)}{s/\sqrt{n}}$ .
  - Here we use the sample standard deviation instead of the population standard deviation.
  - Again, the rejection region and  $p$ -values depends on  $H_a$ .

# One Sample $t$ -test – Rejection Region

- For two-sided test ( $H_a: \mu \neq \mu_0$ ), the rejection region is:

$$|t| > t_{n-1, \alpha/2}$$

- For upper-tailed (right-tailed) test ( $H_a: \mu > \mu_0$ ), the rejection region:

$$t > t_{n-1, \alpha}$$

- For lower-tailed (left-tailed) test ( $H_a: \mu < \mu_0$ ), the rejection region is:

$$t < -t_{n-1, \alpha}$$

- Here the statistic is  $t = \frac{(\bar{y} - \mu_0)}{s/\sqrt{n}}$ .

# One Sample $t$ -test – $p$ -value Approach

- For two-sided test ( $H_a: \mu \neq \mu_0$ ):

$$p\text{-value} = 2\Pr(T_{n-1} > |t|)$$

- For upper-tailed (right-tailed) test ( $H_a: \mu > \mu_0$ ):

$$p\text{-value} = \Pr(T_{n-1} > t)$$

- For lower-tailed (left-tailed) test ( $H_a: \mu < \mu_0$ ):

$$p\text{-value} = \Pr(T_{n-1} < t)$$

- Here  $T_{n-1}$  represents a random variable with  $t$ -distribution of  $n - 1$  degrees of freedom.

# One Sample $t$ -test – Confidence Interval

- For two-sided test ( $H_a: \mu \neq \mu_0$ ), construct a  $100(1 - \alpha)\%$  two-sided confidence interval  $\left(\bar{y} - t_{n-1, \frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{y} + t_{n-1, \frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right)$ .
- For upper-tailed (right-tailed) test ( $H_a: \mu > \mu_0$ ), construct a  $100(1 - \alpha)\%$  lower confidence interval  $\left(\bar{y} - t_{n-1, \alpha} \frac{s}{\sqrt{n}}, \infty\right)$ .
- For lower-tailed (left-tailed) test ( $H_a: \mu < \mu_0$ ), construct a  $100(1 - \alpha)\%$  upper confidence interval  $\left(-\infty, \bar{y} + t_{n-1, \alpha} \frac{s}{\sqrt{n}}\right)$ .

# One Sample t-test - Example 4.2

- **Example 4.2** - In **Example 3.3** we presented a quality control problem in which we tested the hypothesis that the mean weight of peanuts being put in jars was the required 8 oz. We assumed that we knew the population standard deviation, possibly from experience. We now relax that assumption and estimate both mean and variance from the sample. Data are presented on the next slide.

# One Sample t-test - Example 4.2

**Example 4.2 – Some basics:**

- Test if  $H_a: \mu \neq 8$ .
- Sample size:  $n = 16$
- Sample mean:  $\bar{y} = 7.8925$ ;
- Sample variance:  $s^2 = 0.03174$ ;  $s = 0.1782$
- Still, follow 5-steps of hypothesis testing.
- You can use either rejection region,  $p$ -value, or confidence interval approach.

# One Sample $t$ -test – Example 4.2

- **Step 1:** Specify appropriate  $H_0$ ,  $H_a$ , and an acceptable level of  $\alpha$ .

$$H_0: \mu = 8; H_a: \mu \neq 8; \alpha = 0.05$$

- **Step 2:** Define a sample-based test statistic:

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} = \frac{\bar{y} - 8}{0.1782/\sqrt{16}}$$

- **Step 3:** Find the rejection region for the specified  $H_a$  and  $\alpha$ .

The rejection is  $|t| > t_{n-1,\alpha/2} = t_{15,0.025} = 2.1315$ .

# One Sample $t$ -test – Example 4.2

- **Step 4:** Collect the sample data and calculate the test statistic.

$$t = \frac{\bar{y} - 8}{s/\sqrt{16}} = \frac{7.8925 - 8}{0.1782/\sqrt{16}} = -2.4130$$

- **Step 5:** Decide to either reject or fail to reject  $H_0$ .
  - At the 5% significance level, the test statistic  $t = -2.4130$  falls in the rejection region. Therefore, we reject the null hypothesis. The data provides sufficient evidence that the mean weight from filling process is different from the targeted weight of 8 oz.

# One Sample $t$ -test – Example 4.2

- **Step 4:** we have  $t = -2.4130$ , so the  $p$ -value is

$$p\text{-value} = 2\Pr(T_{n-1} > |t|) = 2\Pr(T_{15} > 2.4130) = 0.0291$$

- **Step 5:** Make a decision to either reject or fail to reject  $H_0$ .
  - The  $p$ -value is 0.0291 and is less than the significance level of 0.05. Therefore, we reject the null hypothesis. The data provides sufficient evidence that the mean weight from filling process is different from the targeted weight of 8 oz.

# One Sample $t$ -test – Example 4.2

- **Step 4:** The 95% confidence interval of mean concentricity is

$$\left( \bar{y} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}} \right) = \left( 7.8912 \pm 2.1315 * \frac{0.1782}{\sqrt{16}} \right) = (7.7962, 7.9862)$$

- **Step 5:** Make a decision to either reject or fail to reject  $H_0$ .

- The targeted mean weight of 8 oz does not fall in the 95% confidence interval of mean weight (7.7962, 7.9862). Therefore, we reject the null hypothesis. The data provides sufficient evidence that the mean weight from filling process is different from the targeted weight of 8 oz.

## Example 4.2 – R Program and Output

```
t.test(x = peanuts$weight, mu = 8)  
t.test(x = peanuts$weight, , mu = 8,  
       alternative = "two.sided", conf.level = 1 - alpha)
```

```
alternative = c("two.sided", "less", "greater")
```

# Example 4.2 – R Program and Output

## One Sample t-test

```
data: peanuts$weight  
t = -2.4136, df = 15, p-value = 0.02904  
alternative hypothesis: true mean is not equal to 8  
95 percent confidence interval:  
 7.797567 7.987433  
sample estimates:  
mean of x  
 7.8925
```

# Example 4.2 – R Program and Output

## One Sample t-test

data: peanuts\$weight

t = -2.4136, df = 15, p-value = 0.01452

alternative hypothesis: true mean is less than 8

95 percent confidence interval:

-Inf 7.97058

sample estimates:

mean of x

7.8925

# One Sample $t$ -test – Exercise

- **Exercise (Porosity of Battery)** – Nickel-hydrogen (Ni-H) batteries use a nickel plate as the anode. A critical quality characteristic is the plate's porosity. The manufacturer has set a target porosity of 80%. The production have felt that there is no sufficient porous due to the equipment. They collected 10 samples to investigate the problem.

|      |      |      |      |      |
|------|------|------|------|------|
| 79.1 | 79.5 | 79.3 | 79.3 | 78.8 |
| 79.0 | 79.2 | 79.7 | 79.0 | 79.2 |

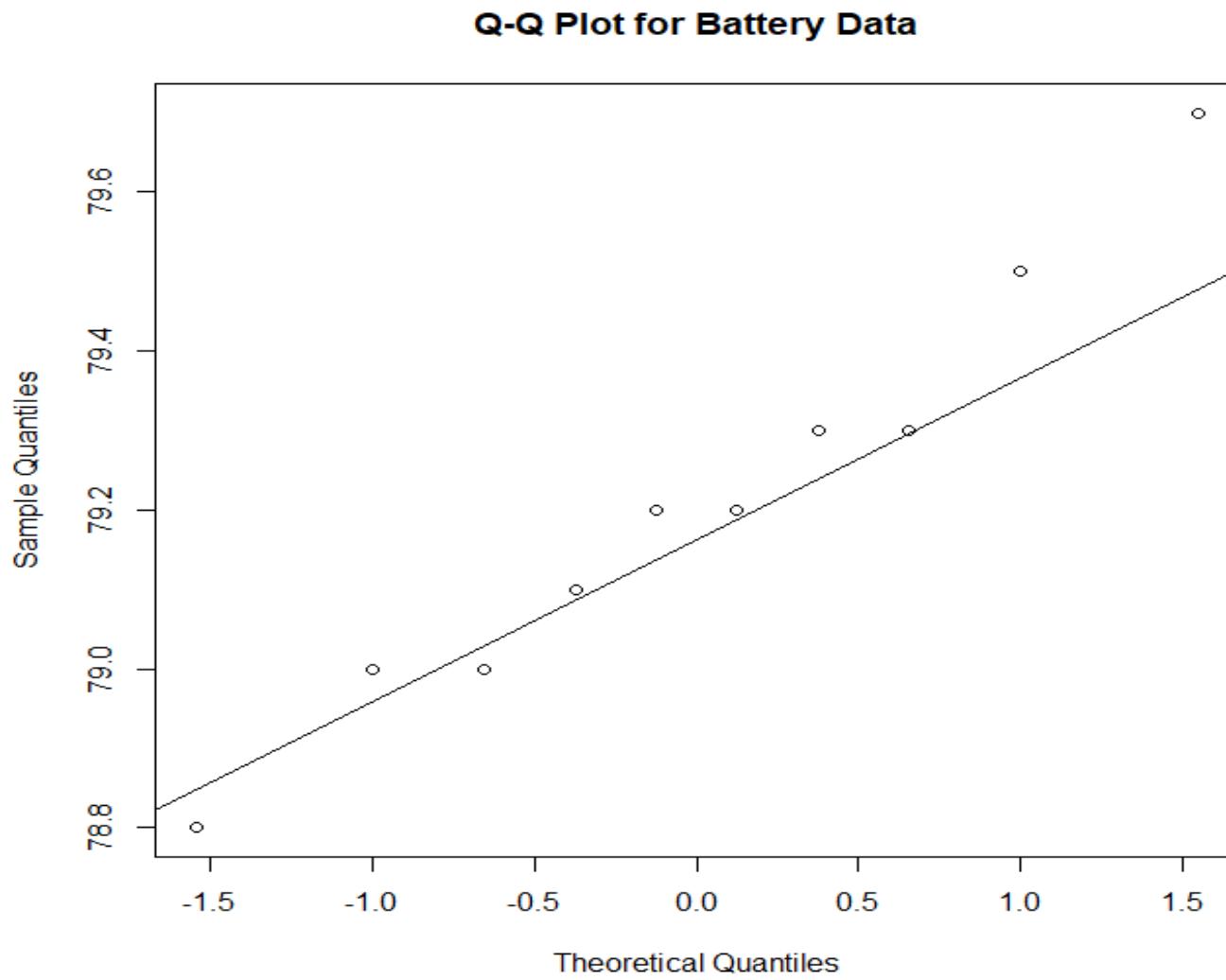
- Use 0.05 significance level to perform your test.

# One Sample $t$ -test – Exercise

## Exercise (Porosity of Battery):

- Sample mean:  $\bar{y} = 79.21$
- Sample Size:  $n = 10$
- Sample variance:  $s^2 = 0.0677$
- Sample standard deviation:  $s = 0.2601$
- Significance level:  $\alpha = 0.05$
- Check the assumption for the normal approximation.

# Porosity of Battery – Q-Q Plot



# Degrees of Freedom

- **Degrees of freedom:** It is important to remember that the degrees of freedom of the  $t$  statistic are always those used to estimate the variance, which may not be sample size minus 1.
  - For example, suppose that we take 100 stones to estimate the average size of stones produced by a gravel crusher. We do not have time to weigh each stone individually. So we weigh the entire 100 in one weighing and choose and weigh 10 stones individually. The test statistic is  $t = \frac{\bar{y}_{100} - \mu_0}{s/\sqrt{100}}$  where  $s^2 = \sum(y_i - \bar{y}_{10})^2/9$ . It has a  $t$ -distribution with 9 degrees of freedom.

# Inference On Proportions

- If  $Y \sim Binomial(n, p)$ , then  $Y = \sum_{i=1}^n Y_i$  and  $Y_i (i = 1, \dots, n)$  is a random sample from Bernoulli distribution. So  $\frac{Y}{n}$  (the proportion) is considered as a sample mean.
- By Central Limit Theorem (CLT):  $\frac{\frac{Y}{n} - p}{\sqrt{p(1-p)/n}} = \frac{Y - np}{\sqrt{np(1-p)}}$  can be approximated by  $N(0,1)$ .
- Also,  $E\left(\frac{Y}{n}\right) = p$ ; and  $\text{Var}\left(\frac{Y}{n}\right) = \frac{p(1-p)}{n}$
- This statistic can be used for hypothesis testing.

# Test for Proportion

- The null hypothesis:  $H_0: p = p_0$ .
- Denote  $\hat{p} = \frac{y}{n}$ . The test statistic is:

$$Z = \frac{\frac{Y}{n} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{Y - np_0}{\sqrt{np_0(1-p_0)}}$$

- Assumptions for this test are:
  - $Y \sim Binomial(n, p)$ .
  - Sample size is large enough:  $np_0 \geq 5$  and  $n(1 - p_0) \geq 5$ .

# Test for Proportion – Rejection Region

- For two-sided test ( $H_a: p \neq p_0$ ), the rejection region is:

$$|z| > z_{\alpha/2}$$

- For upper-tailed (right-tailed) test ( $H_a: p > p_0$ ), the rejection region:

$$z > z_\alpha$$

- For lower-tailed (left-tailed) test ( $H_a: p < p_0$ ), the rejection region is:

$$z < -z_\alpha$$

# Test for Proportion – Confidence Interval

- For two-sided test ( $H_a: p \neq p_0$ ), construct a  $100(1 - \alpha)\%$  two-sided confidence interval  $\left(\hat{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$ .
- For upper-tailed (right-tailed) test ( $H_a: p > p_0$ ), construct a  $100(1 - \alpha)\%$  lower confidence interval  $\left(\hat{p} - z_{\alpha} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, 1\right)$ .
- For lower-tailed (left-tailed) test ( $H_a: p < p_0$ ), construct a  $100(1 - \alpha)\%$  upper confidence interval  $\left(0, \hat{p} + z_{\alpha} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$ .

# Test for Proportion – $p$ -value Approach

- For two-sided test ( $H_a: p \neq p_0$ ):

$$p\text{-value} = 2\Pr(Z > |z|)$$

- For upper-tailed (right-tailed) test ( $H_a: p > p_0$ ):

$$p\text{-value} = \Pr(Z > z)$$

- For lower-tailed (left-tailed) test ( $H_a: p < p_0$ ):

$$p\text{-value} = \Pr(Z < z)$$

# Inference on a Proportion – Example 4.4

- **Example 4.4** - An advertisement claims that more than 60% of doctors prefer a particular brand of painkiller. An agency established to monitor truth in advertising conducts a survey consisting of a random sample of 120 doctors. Of the 120 questioned, 82 indicated a preference for the particular brand. Is the advertisement justified?
- **Example 4.4** – Use five steps for hypothesis testing. The significance level is set as 0.05.

# Example 4.4 – R Program and Output

1-sample proportions test without continuity correction

data: y out of n, null probability 0.6

X-squared = 3.4722, df = 1, p-value = 0.0312

alternative hypothesis: true p is greater than 0.6

95 percent confidence interval:

0.6100991 1.0000000

sample estimates:

p

0.6833333

# Test for Proportion – Exercise

- **Exercise (Breaking Strengths Carbon Fiber)** – Padgett and Spurrier (1990) analyzed the breaking strengths of carbon fiber. Suppose that historically, the proportion of nonconforming product is 10%. To test if the true proportion of nonconforming is 10%, 100 fibers are collected and 6 are nonconforming.
- **Question:** Is the true proportion of nonconforming 10%? Use 0.01 significance level to perform your test.

# Test for a Proportion - Example 4.5

- **Example 4.5** – A pre-election poll using a random sample of 150 voters indicated that 84 favored candidate Smith, that is,  $\hat{p} = 0.56$ . We would like to construct a 0.99 confidence interval on the true proportion of voters favoring Smith.
- This will be used as an exercise problem in the class. You can look at the R program.

# MA5701: Statistical Methods

Chapter 5 : Inferences for Two Populations

Kui Zhang, Mathematical Sciences

# Inference for Two Parameters

- So far we have performed the statistical inference for one parameter.
- The case of two populations (two parameters) is important too.
  - Many interesting applications involve only two populations, for example, any comparisons involving differences between the two sexes, comparing a drug with a placebo, comparing old versus new, or before and after some events.
  - Some of the concepts underlying comparing several populations are more easily introduced for the two-population case.
  - The comparison of two populations results in a single easily understood statistic: the difference between two sample means.

# Methods for Collecting Data

- **Independent Samples, for example,**
  - A sample of migraine sufferers is randomly divided into two groups. The first group is given remedy A while the other is given remedy B, both to be taken at the onset of a migraine attack. The pills are not identified, so patients do not know which pill they are taking.
- **Dependent or Paired Samples, for example,**
  - Each person in a group of migraine sufferers is given two pills, one of which is red and the other is green. The group is randomly split into two subgroups and one is told to take the green pill the first time a migraine attack occurs and the red pill for the next one. The other group is told to take the red pill first and the green pill next.

# Introduction

- In this chapter, we will present procedures for:
  - Making inferences on the difference of means of two normally distributed populations where the variances are unknown.
  - Making inferences on the difference of proportions of successes in two binomial populations.

# Inferences on the Difference of Means

- We are interested in comparing two populations whose means are  $\mu_1$  and  $\mu_2$  and variances are  $\sigma_1^2$  and  $\sigma_2^2$ .
- To compare means, we have  $H_0: \mu_1 - \mu_2 = \delta_0$  versus  $H_a: \mu_1 - \mu_2 \neq \delta_0$ .
- In many situations, we set  $\delta_0 = 0$ .

# Inference For Two Independent Samples

- **Example (Packaging of Ground Beef)** – Maxcy and Lowry (1984) looked a packaging process for ground beef over a series of days. An interesting question is whether the true mean amount delivered by this process changes from day to day.
- Denote the population mean and variance of beef packed on one day are  $\mu_1$  and  $\sigma_1^2$  and on another day are  $\mu_2$  and  $\sigma_2^2$ , respectively.
- We are interested in the difference of means,  $\mu_1 - \mu_2$ .

# Inference For Two Independent Samples

- If  $\mu_1 - \mu_2 > 0$  then the true mean amount of beef packed on one day is larger than the true mean amount of beef packed on the other day.
- If  $\mu_1 - \mu_2 = 0$  then the true mean amount of beef packed on one day is no different from the true mean amount of beef packed on the other day.
- If  $\mu_1 - \mu_2 < 0$ , then the true mean amount of beef packed on one day is less than the true mean amount of beef packed on the other day.

# Distribution of A Linear Function of Random Variables

For a set of  $n$  independent random variables  $Y_1, \dots, Y_n$ , whose means are  $\mu_1, \dots, \mu_n$  and whose variances are  $\sigma_1^2, \dots, \sigma_n^2$ . A linear function of these random variables is defined as  $L = \sum_{i=1}^n (a_i Y_i + b_i)$ , where the  $a_i$  and  $b_i$  are arbitrary constants, then

1.  $E(L) = \sum_{i=1}^n (a_i \mu_i + b_i)$
2.  $Var(L) = \sum_{i=1}^n a_i^2 \sigma_i^2$ .
3. If  $Y_i$  are normally distributed, then  $L$  is normally distributed.

# Sampling Distribution of Difference of Means

- Since sample means are random variables, the difference between two sample means is a linear function of two random variables.
- First,  $E(\bar{Y}_1) = \mu_1$ ,  $Var(\bar{Y}_1) = \sigma_1^2/n_1$ ,  $E(\bar{Y}_2) = \mu_2$ ,  $Var(\bar{Y}_2) = \sigma_2^2/n_2$
- Then  $L = \bar{Y}_1 - \bar{Y}_2$ , we can get

$$E(L) = \mu_1 - \mu_2; Var(L) = \sigma_1^2/n_1 + \sigma_2^2/n_2$$

- When  $\sigma_1^2 = \sigma_2^2 = \sigma^2$ , then  $Var(L) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}$
- Finally, the central limit theorem states that if the sample sizes are sufficiently large, the sample means are approximately normally distributed; hence generally  $\bar{Y}_1 - \bar{Y}_2$  is also normally distributed too.

# Inferences with Known Variance

- If the variances are known, we can use the following random variable to make the inference on  $\delta = \mu_1 - \mu_2$ :  $(\bar{Y}_1 - \bar{Y}_2)$  and we know that  $Z = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$  has a standard normal distribution.
- In this situation, you can consider  $(\bar{Y}_1 - \bar{Y}_2)$  as a random variable with  $N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$ , then all formulas introduced in chapter 3 can be used here.
- Again, this method has little practical use since we generally do not know the populations variances.

# Variances Unknown but Assumed Equal

- **Solution** to unknown variance - Assume that the two population variances are equal and find an estimate of that variance. The equal variance assumption is actually quite reasonable since in many studies, a focus on means implies that the populations are similar in many respects. However, if the assumption of equal variances cannot be made, then other methods must be used.
- Again, we will use the point estimate of that difference ( $\bar{y}_1 - \bar{y}_2$ ).

# Two Sample $t$ -test – Equal Variance

- Two sample  $t$ -test is used for the inference of a difference of two means with unknown population variance.
- Assumptions for the two sample  $t$ -test (equal variance) are:
  - Each random sample is selected from the target population.
  - Two samples are independent.
  - The sample mean,  $\bar{Y}_1$  and  $\bar{Y}_2$  are normal or approximate normal.
  - The variances of two populations are same:  $\sigma_1^2 = \sigma_2^2 = \sigma^2$ .
  - Note that two sample sizes may not be equal.

# Two Sample $t$ -test – Equal Variance

- The null hypothesis is:  $H_0: \mu_1 - \mu_2 = \delta_0$
- $\delta_0$  is the nominal difference two means. In many situations,  $\delta_0 = 0$
- The test statistic is:

$$T = \frac{(\bar{y}_1 - \bar{y}_2) - \delta_0}{\text{Estimated Standard Deviation of } (\bar{y}_1 - \bar{y}_2)}$$

- How to estimate the standard deviation of  $(\bar{Y}_1 - \bar{Y}_2)$ ?
- Note that the variance of  $\bar{Y}_1 - \bar{Y}_2$  is  $\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}$  here.
- What is the degrees of freedom of the  $t$ -distribution here?

# Pooled Estimate of $\sigma^2$

- $n_1$ : sample size from the first sample;
- $n_2$ : sample size from the second sample;
- $s_1^2$ : sample variance from the first sample;
- $s_2^2$ : sample variance from the second sample;
- The pooled estimate of  $\sigma^2$  is given by:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{SS_1 + SS_2}{(n_1 - 1) + (n_2 - 1)}$$

# Two Sample $t$ -test – Equal Variance

- The null hypothesis is:  $H_0: \mu_1 - \mu_2 = \delta_0$
- The test statistic is:

$$T = \frac{(\bar{y}_1 - \bar{y}_2) - \delta_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1 + n_2 - 2} \text{ when } H_0 \text{ is true}$$

- Where  $s_p = \sqrt{s_p^2}$  and  $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{SS_1 + SS_2}{n_1 + n_2 - 2}$ .
- The degrees of freedom is  $n_1 + n_2 - 2$ .

# Two Sample $t$ -test – Rejection Region

- For two-sided test ( $H_a: \mu_1 - \mu_2 \neq \delta_0$ ), the rejection region is:  
$$|t| > t_{n_1+n_2-2, \alpha/2}$$
- For upper-tailed (right-tailed) test ( $H_a: \mu_1 - \mu_2 > \delta_0$ ), the rejection region is:  
$$t > t_{n_1+n_2-2, \alpha}$$
- For lower-tailed (left-tailed) test ( $H_a: \mu_1 - \mu_2 < \delta_0$ ), the rejection region is:  
$$t < -t_{n_1+n_2-2, \alpha}$$

# Two Sample $t$ -test – $p$ -value Approach

- For two-sided test ( $H_a: \mu_1 - \mu_2 \neq \delta_0$ ):

$$p\text{-value} = 2\Pr(T_{n_1+n_2-2} > |t|)$$

- For upper-tailed (right-tailed) test ( $H_a: \mu_1 - \mu_2 > \delta_0$ ):

$$p\text{-value} = \Pr(T_{n_1+n_2-2} > t)$$

- For lower-tailed (left-tailed) test ( $H_a: \mu_1 - \mu_2 < \delta_0$ ):

$$p\text{-value} = \Pr(T_{n_1+n_2-2} < t)$$

- Here  $T_{n_1+n_2-2}$  represents a random variable with  $t$ -distribution of  $n_1 + n_2 - 2$  degrees of freedom.

# Two Sample $t$ -test – Confidence Interval

- For two-sided test ( $H_a: \mu_1 - \mu_2 \neq \delta_0$ ), construct a  $100(1 - \alpha)\%$  two-sided confidence interval

$$(\bar{y}_1 - \bar{y}_2 \pm t_{n_1+n_2-2,\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}).$$

- For upper-tailed test ( $H_a: \mu_1 - \mu_2 > \delta_0$ ), construct a  $100(1 - \alpha)\%$  lower confidence interval  $(\bar{y}_1 - \bar{y}_2 - t_{n_1+n_2-2,\alpha} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \infty)$ .
- For lower-tailed test ( $H_a: \mu_1 - \mu_2 < \delta_0$ ), construct a  $100(1 - \alpha)\%$  upper confidence interval  $(-\infty, \bar{y}_1 - \bar{y}_2 + t_{n_1+n_2-2,\alpha} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}})$ .

# Two Sample $t$ -test – Packing of Ground Beef

- **Example (Packing of Ground Beef)** – The data are shown in Tables.

|            |        |        |        |        |        |
|------------|--------|--------|--------|--------|--------|
| First Day  | 1397.8 | 1394.8 | 1391.7 | 1400.0 | 1393.5 |
| First Day  | 1391.2 | 1384.0 | 1391.0 | 1385.7 | 1385.3 |
| Second Day | 1410.0 | 1393.9 | 1405.9 | 1404.2 | 1387.3 |
| Second Day | 1398.5 | 1399.9 | 1392.5 | 1402.5 | 1391.8 |

- **Question:** Is the mean amount of beef delivered on the first day different from the mean amount of beef delivered on the second day? Use 0.05 significance level to perform your test.

# Two sample $t$ -test – Packing of Ground Beef

- **Step 1:** Specify appropriate  $H_0$ ,  $H_a$ , and an acceptable level of  $\alpha$ .

$$H_0: \mu_1 - \mu_2 = 0; H_a: \mu_1 - \mu_2 \neq 0; \alpha = 0.05$$

- **Step 2:** Define a sample-based test statistic:

$$t = \frac{\bar{y}_1 - \bar{y}_2 - 0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- **Step 3:** Find the rejection region for the specified  $H_a$  and  $\alpha$ .

The rejection region is  $|t| > t_{n_1+n_2-2, \alpha/2} = t_{18, 0.025} = 2.1009$ .

# Two Sample $t$ -test – Packing of Ground Beef

- **Step 4:** Collect the sample data and calculate the test statistic.

$$\bar{y}_1 = 1391.50; s_1^2 = 28.3933; \bar{y}_2 = 1398.65; s_2^2 = 51.6361$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{9 * 28.3933 + 9 * 51.6361}{10 + 10 - 2} \\ = 40.0147$$

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{1391.50 - 1398.65}{\sqrt{40.0147} \sqrt{\frac{1}{10} + \frac{1}{10}}} = -2.5274$$

# Two Sample $t$ -test – Packing of Ground Beef

- **Step 5:** Make a decision to either reject or fail to reject  $H_0$ .
  - We have  $|t| = |-2.5274| > 2.1009$ . At the 5% significance level, the test statistic  $t = -2.5274$  falls in the rejection region. Therefore, we reject the null hypothesis. The data provides sufficient evidence that the mean amount of beef delivered on the first day is different from the mean amount of beef delivered on the second day.

# Packing of Ground Beef – Confidence Interval

- **Step 4:** The 95% confidence interval of mean amount of beef is

$$\begin{aligned} & \left( \bar{y}_1 - \bar{y}_2 \pm t_{n_1+n_2-2, \alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) \\ &= \left( 1391.50 - 1398.65 \pm 2.1009 * \sqrt{40.1047} \sqrt{\frac{1}{10} + \frac{1}{10}} \right) \\ &= (-13.1000, -1.2000) \end{aligned}$$

# Packing of Ground Beef – Confidence Interval

- **Step 5:** Make a decision to either reject or fail to reject  $H_0$ .
  - The nominal difference of mean amount of beef 0 does not fall in the 95% confidence interval  $(-13.1000, -1.2000)$ . Therefore, we reject the null hypothesis. The data provides sufficient evidence that the mean amount of beef delivered on the first day is different from the mean amount of beef delivered on the second day.

# Packing of Ground Beef – $p$ -value Approach

- **Step 4:** we have  $t = -2.5274$ , so the  $p$ -value is

$$p\text{-value} = 2\Pr(T_{18} > |-2.5274|) = 0.0211$$

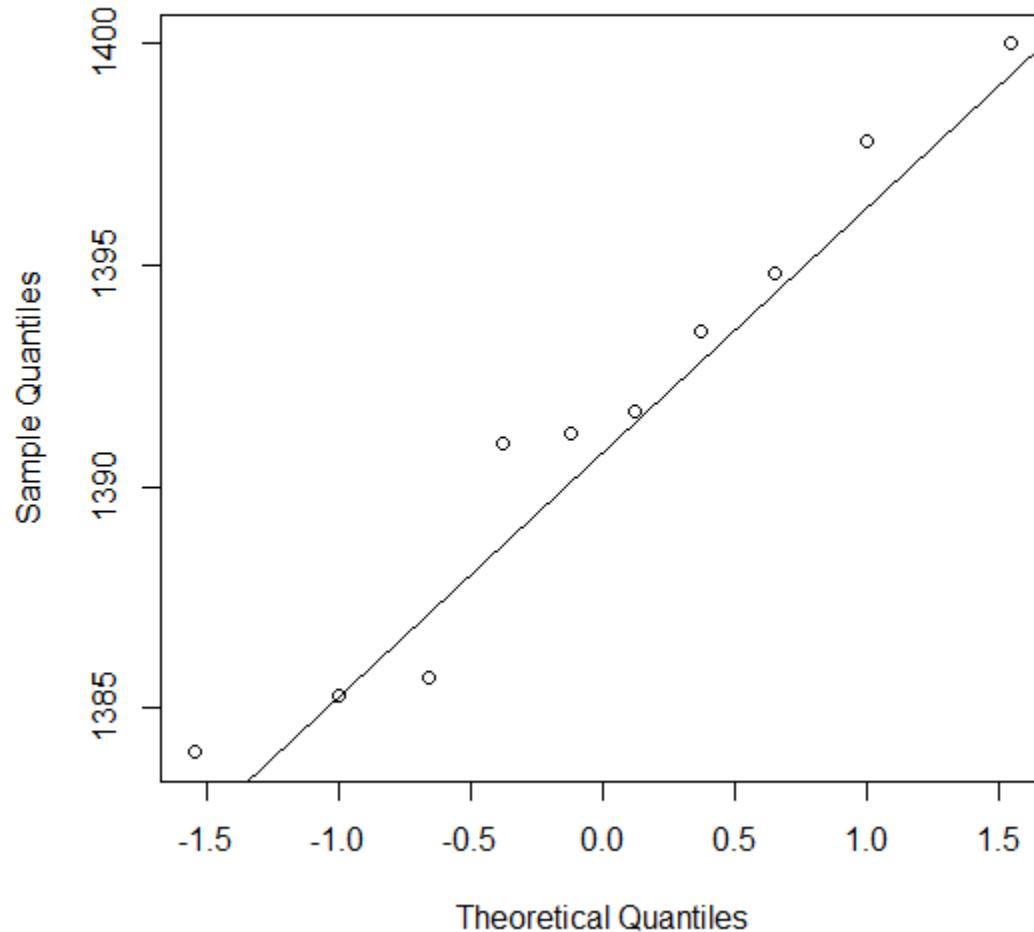
- **Step 5:** Make a decision to either reject or fail to reject  $H_0$ .
  - The  $p$ -value is 0.0211 and is less than the significance level of 0.05. Therefore, we reject the null hypothesis. The data provides sufficient evidence that the mean amount of beef delivered on the first day is different from the mean amount of beef delivered on the second day.

# Packing of Ground Beef – Check Assumptions

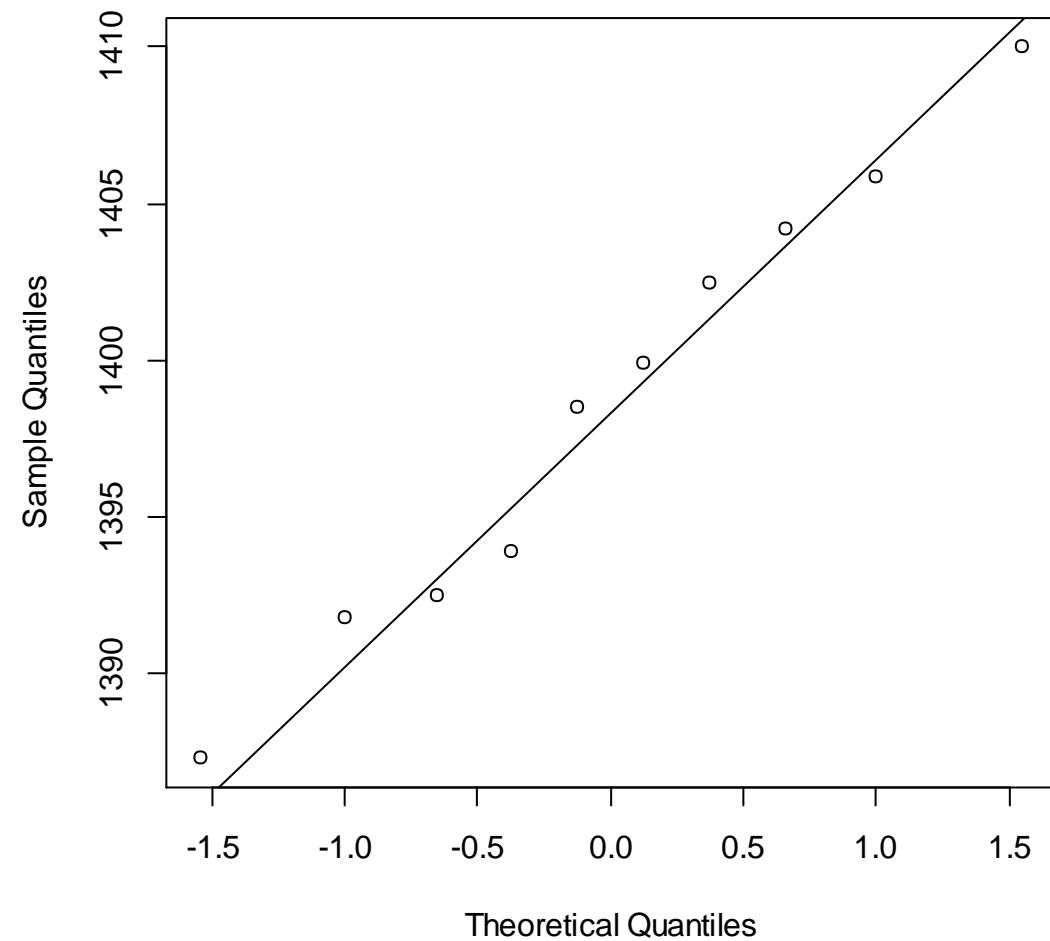
- When the sample size is small, you should check the normality assumption with the stem-and-leaf display and/or Q-Q plot.
- Use a box plot to check the equal variance assumption. There are formal tests for the equal variance assumption. However, they will not be covered in our course.
- From the figures from the next two slides, it seems the data are normally distributed and two samples have the similar variance.

# Packing of Ground Beef – Q-Q Plot

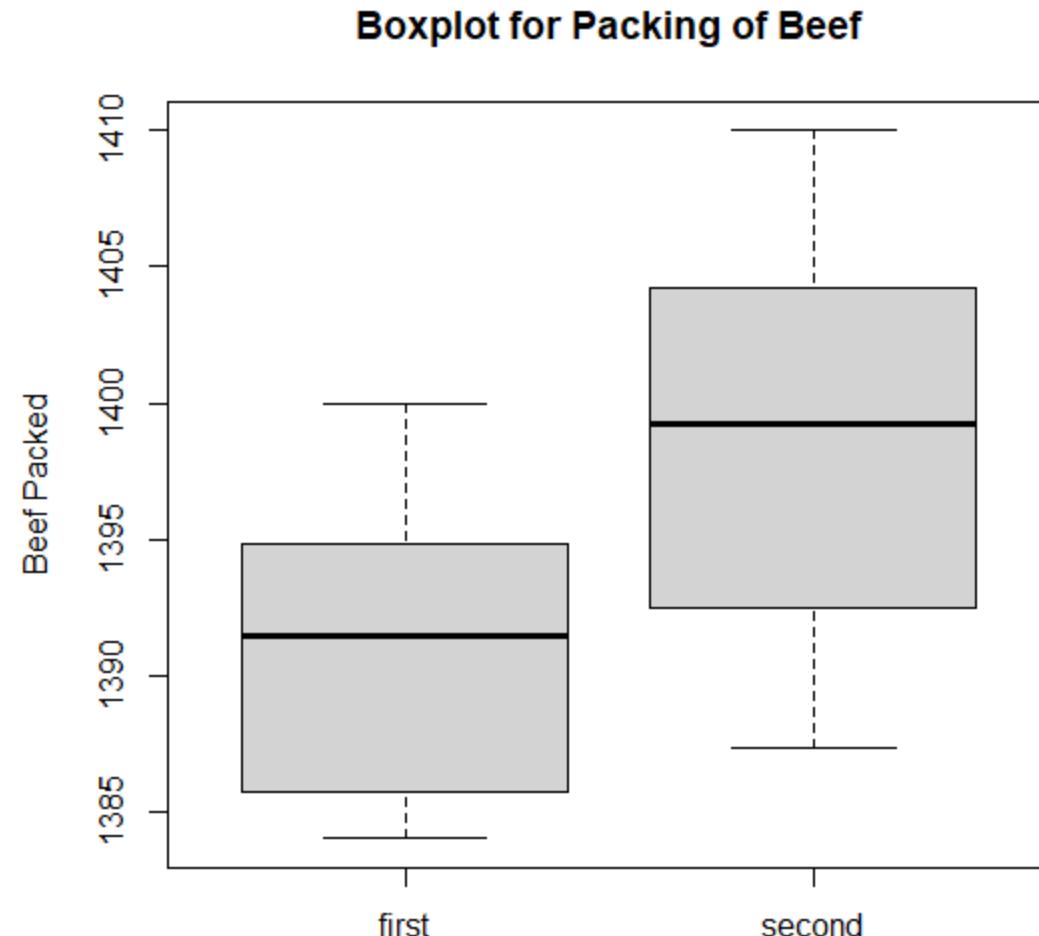
Normal Q-Q Plot for First Day



Normal Q-Q Plot for Second Day



# Packing of Ground Beef – Boxplot



# R Function – t.test

```
t.test( x = beef$first,  
        y = beef$second,  
        alternative = "two.sided", # two sided is the default  
        mu = 0,  
        paired = FALSE, # default is FALSE  
        var.equal = TRUE, # default is FALSE  
        conf.level = 1 - alpha)
```

# R Function – t.test

Two Sample t-test

data: beef\$first and beef\$second

t = -2.5274, df = 18, p-value = 0.02107

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-13.093398 -1.206602

sample estimates:

mean of x mean of y

1391.50 1398.65

# Exercise – Diet Formulation

- **Exercise (Die Example).** To assess the effectiveness of a new diet formulation, a sample of 8 steers is fed a regular diet and another sample of 10 steers is fed a new diet. The weights of the steers at 1 year are given. Do these results imply that the new diet results in higher weights? (Use  $\alpha = 0.05$ )
- **Weights from Regular Diet:** 831, 858, 833, 860, 922, 875, 797, 788
- **Weights from New Diet:** 870, 882, 896, 925, 842, 908, 944, 927, 965, 887

# Exercise – Diet Formulation

**Exercise (Die Example).** Some Basic Statistics.

- Regular Diet:

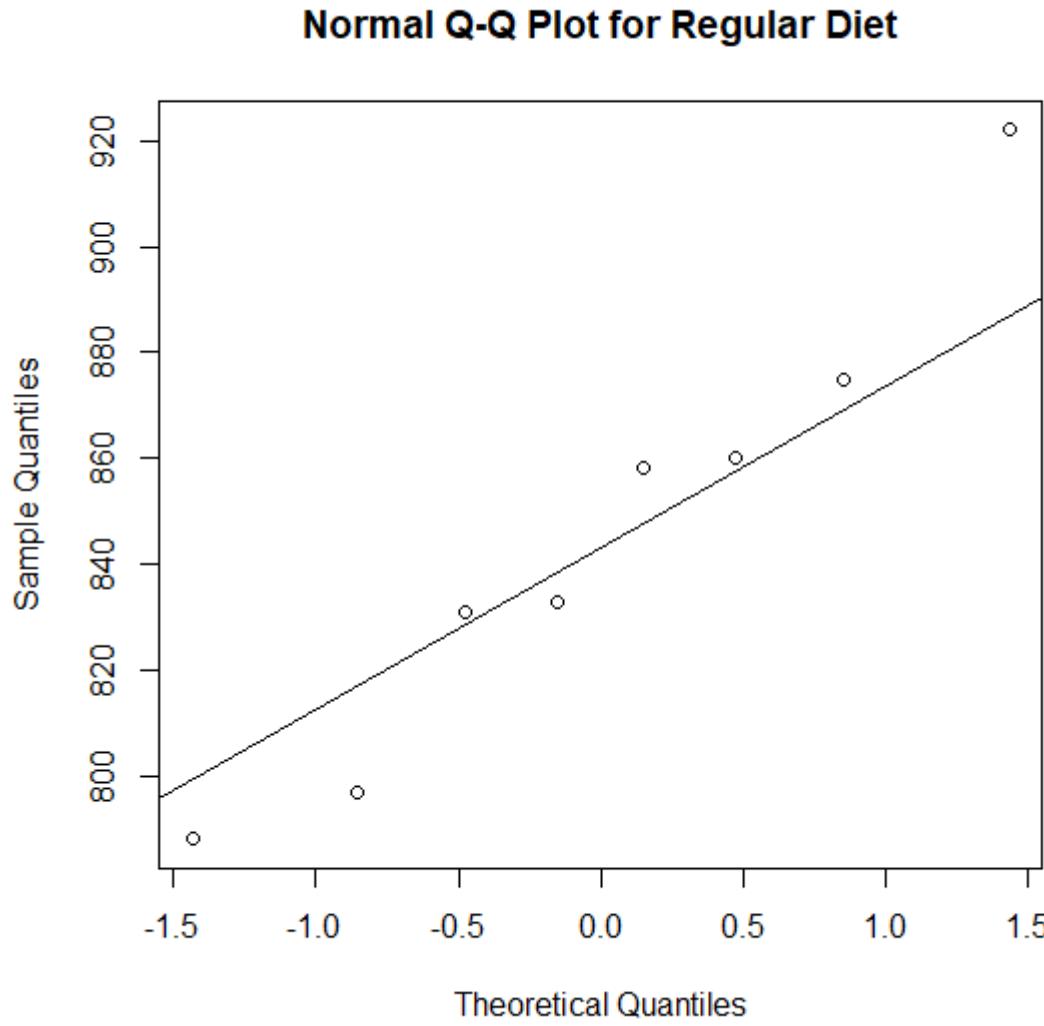
$$n_1 = 8; \bar{y}_1 = 845.5; s_1^2 = 1873.4286$$

- New Diet:

$$n_2 = 10; \bar{y}_2 = 904.6; s_2^2 = 1348.9333$$

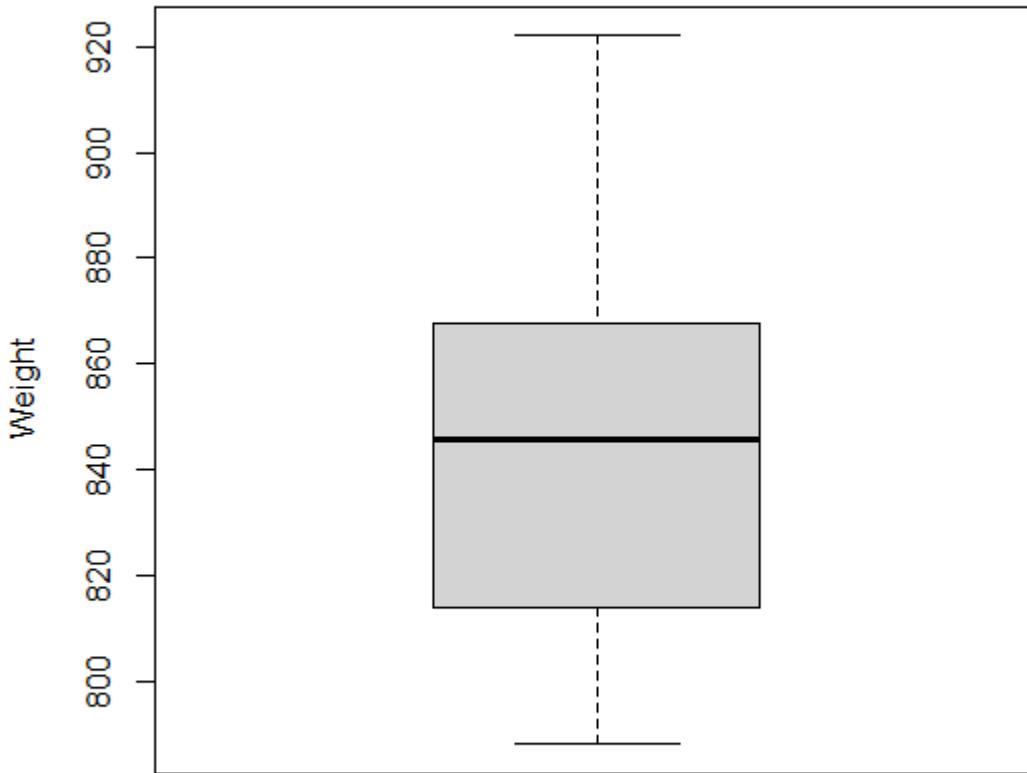
- Now use five steps for hypothesis testing.
- Use the significance level of 0.05.

# Exercise – Diet Formulation

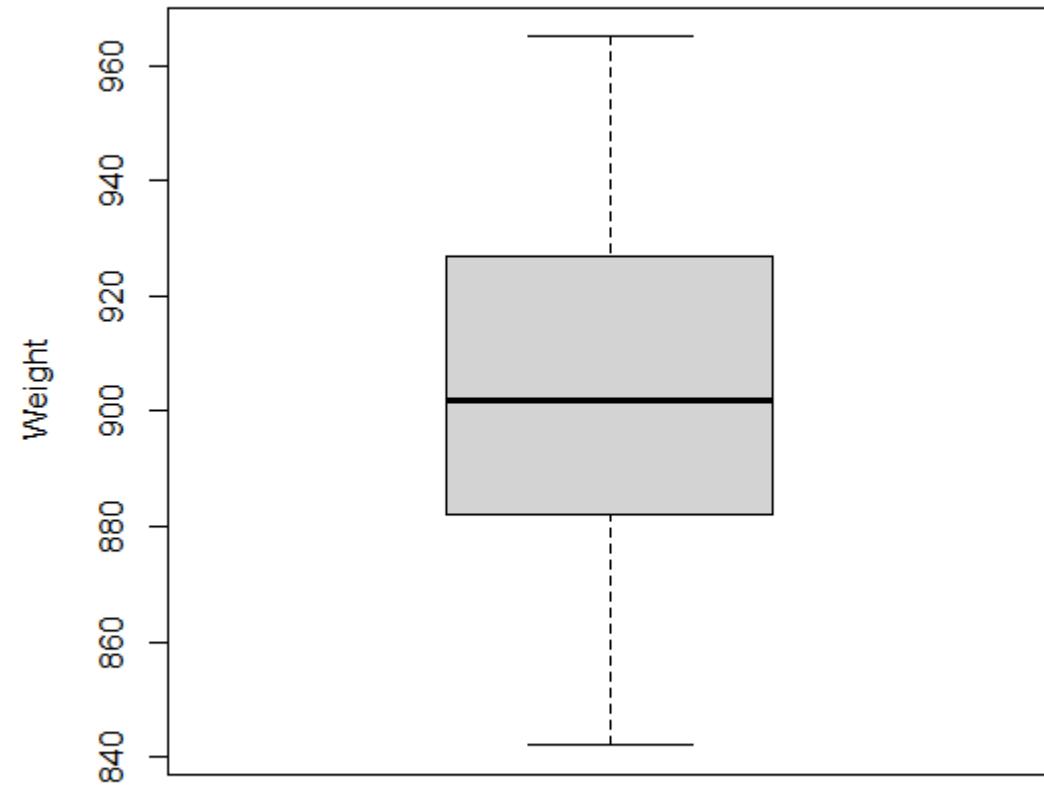


# Exercise – Diet Formulation

Boxplot for Regular Drug



Boxplot for New Drug



# Two Sample $t$ -test – Unequal Variance

- In **Packing of Ground Beef Example**, we saw that the sample variance of the weights of second day is almost twice that of first day. Therefore we may need to provide a method for comparing means that does not assume equal variances. (A test for equality of variances is presented in Section 5.3 and according to this test these two variances are not significantly different.)
- A test statistic similar to previous ones (next slide) can be used.
- Transformation to make variance equal (e.g., log transformation).

# Two Sample $t$ -test - Unequal Variance

- We can use the following statistic:  $t' = \frac{(\bar{y}_1 - \bar{y}_2) - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$
- If both sample size are large (both over 30) we can assume a normal distribution and compute the test statistic.

# Two Sample $t$ -test - Unequal Variance

- We can use the following statistic:  $t' = \frac{(\bar{y}_1 - \bar{y}_2) - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$
- If either sample size is not large but the data come from approximately normally distributed populations, this statistic does have an approximate Student's  $t$  distribution, but the degrees of freedom cannot be precisely determined.
- A reasonable (and conservative) approximation is to use the degrees of freedom for the smaller sample.
- More precise but complex approximations are available. One such approximation, called Satterthwaite's approximation, is implemented in many statistical packages.

# Paired *t*-Test

- **Paired Samples** - pairs of observed values.
- Why paired samples? Using the weight loss of a special diet as an example:
  - Divide samples to two groups – one group with the general diet and the other group with the special diet and then compare the weights of samples. The drawback is that the estimate of the variance is based on the differences in weights among individuals in each sample, and these differences are probably larger than those induced by the special diet. Thus a huge sample size may be needed.
  - Give all samples the special diet but compare their weights before and after the special diet.
- Methods for paired samples – use differences between paired values.

# Paired Versus Independent *t*-test

- Paired *t*-test can remove the sampling unit to sampling unit variability, increase the test statistic by decreasing the estimated variance.
- Paired *t*-test reduces the degrees of freedom thus results in the larger critical value, especially for small sample size.
- We should obtain paired data whenever we know the sampling unit to sampling unit variability is large.
- We can use two independent samples when the sampling unit to sampling unit variability is not an issue and the available sample size is small.

# Paired $t$ -test

- The paired  $t$ -test is essentially the one sample  $t$ -test that is applied to the difference of paired data.
- Let  $d_i = y_{1,i} - y_{2,i}$  ( $i = 1, \dots, n$ ) be the difference of the paired data.
- We perform the one sample  $t$ -test to  $d_i$  ( $i = 1, \dots, n$ ).

# Paired $t$ -test - Notations

We have the following notations:

- $n$ : number of *pairs* of data
- $\mu_d$ : population mean of difference.
- $\bar{d}$ : sample mean of difference  $d_i$  ( $i = 1, \dots, n$ ).
- $s_d$ : sample standard deviation of difference  $d_i$  ( $i = 1, \dots, n$ ).

# Paired $t$ -test

- The null hypothesis is:  $H_0: \mu_d = \delta_0$
- $\delta_0$  is the nominal difference two means. In many situations,  $\delta_0 = 0$ .
- The test statistic is:

$$T = \frac{\bar{d} - \delta_0}{s_d / \sqrt{n}} \sim t_{n-1} \text{ when } H_0 \text{ is true.}$$

- You can make a decision based on the rejection region, the confidence interval, or the  $p$ -value. The procedure is exactly same as the one sample  $t$ -test.

# Paired *t*-Test – Example 5.6

- **Example 5.6 (Baseball Teams)** - For the first 60 years major league baseball consisted of 16 teams, eight each in the National and the American leagues. In 1961 the Los Angeles and the Washington Senators became the first expansion teams in baseball history. It is conjectured that the main reason that the league allowed expansion teams was the fact that total attendance dropped from 20 million in 1960 to slightly over 17 million in 1961.

# Data from Example 5.6

Table 5.7 Baseball Attendance  
(Thousands)

| Team | 1960 | 1961 | Diff. |
|------|------|------|-------|
| 1    | 809  | 673  | -136  |
| 2    | 663  | 1123 | 460   |
| 3    | 2253 | 1813 | -440  |
| 4    | 1497 | 1100 | -397  |
| 5    | 862  | 584  | -278  |
| 6    | 1705 | 1199 | -506  |
| 7    | 1096 | 855  | -241  |
| 8    | 1795 | 1391 | -404  |
| 9    | 1187 | 951  | -236  |
| 10   | 1129 | 850  | -279  |
| 11   | 1644 | 1151 | -493  |
| 12   | 950  | 735  | -215  |
| 13   | 1167 | 1606 | 439   |
| 14   | 774  | 683  | -91   |
| 15   | 1627 | 1747 | 120   |
| 16   | 743  | 597  | -146  |

# Paired $t$ -Test - Example 5.6

## Example 5.6 (Baseball Teams)

- Sample Size:  $n = 16$
- 1960 Samples:  $\bar{y}_1 = 1243.8125$ ;  $s_1^2 = 212016.9625$
- 1961 Samples:  $\bar{y}_2 = 1066.1250$ ;  $s_2^2 = 161046.6500$
- Difference:  $\bar{d} = -177.6875$ ;  $s_d^2 = 86019.0292$
- Note that:  $\bar{d} = \bar{y}_2 - \bar{y}_1$  but  $s_d^2 \neq s_2^2 - s_1^2$

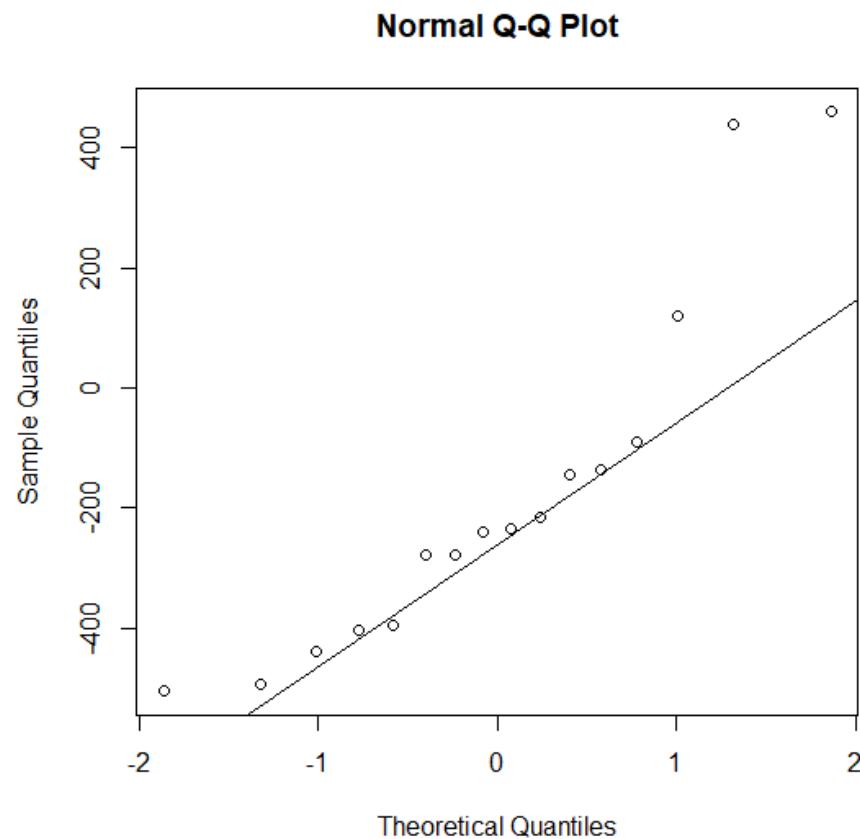
## Paired t-Test - Example 5.6

**Example 5.6 (Baseball Teams) – Key Results** (Again, some steps are skipped here)

- Test statistic:  $t = -177.6875/\sqrt{86019.0292/16} = -2.4233$
- Rejection region:  $t < -t_{15,0.05} = -1.7531$
- The  $p$ -value is:  $\Pr(T_{15} < -2.4233) = 0.0142$
- The upper 95% confidence interval is:

$$\left( -\infty, -177.6875 + t_{15,0.05} * \sqrt{\frac{86019.0292}{16}} \right) = (-\infty, -49.1495)$$

# Paired t-Test - Example 5.6



# Paired *t*-test – Baseball Teams

```
t.test(x = baseball$F3, y = baseball$F2, alternative = "less",
       mu = 0, paired = TRUE, conf.level = 1 - alpha)
```

Paired t-test

data: baseball\$F3 and baseball\$F2

**t = -2.4234, df = 15, p-value = 0.01425**

alternative hypothesis: true mean difference is less than 0

95 percent confidence interval:

**-Inf -49.14946**

sample estimates:

mean difference

**-177.6875**

# Paired *t*-test – Baseball Teams

```
t.test(x = baseball$diff, alternative = "less",
       mu = 0, conf.level = 1 - alpha)
```

One Sample t-test

data: baseball\$diff

**t = -2.4234, df = 15, p-value = 0.01425**

alternative hypothesis: true mean is less than 0

95 percent confidence interval:

**-Inf -49.14946**

sample estimates:

mean of x

**-177.6875**

# Paired *t*-test – Exercise

- **Exercise (BUN in Cat)** - Elevated levels of blood urea nitrogen (BUN) denote poor kidney function. Five elderly cats are placed on a standard high-protein diet. Their BUN is measured both initially and three months after they are placed on a standard high-protein diet.

| Cat         | 1  | 2  | 3  | 4  | 5  | Sample Mean | Sample Variance |
|-------------|----|----|----|----|----|-------------|-----------------|
| Initial BUN | 52 | 41 | 49 | 62 | 39 | 48.6        | 85.30           |
| Final BUN   | 58 | 41 | 58 | 75 | 44 | 55.2        | 183.70          |
| Difference  | 6  | 0  | 9  | 13 | 5  | 6.6         | 23.30           |

- **Question:** Is there a significant increase in mean BUN? Use 0.05 as the significance level.

# Inference on Two Proportions

- **Purpose** – Compare the probability of success from two binomial distributions.
- Let  $p_1$  and  $p_2$  be the probabilities of success, respectively.
- Let  $n_1$  and  $n_2$  be the sample sizes, respectively.
- Let  $y_1$  and  $y_2$  be the samples (number of observed success), respectively.
- Estimates of proportion of success:

$$\hat{p}_1 = Y_1/n_1, \hat{p}_2 = Y_2/n_2$$

# Sampling Distribution of the Difference between Two Proportions

- Let  $\hat{p}_1 = Y_1/n_1$ ,  $\hat{p}_2 = Y_2/n_2$ . Similarly, we have

$$E(\hat{p}_1) = p_1, \text{Var}(\hat{p}_1) = p_1(1-p_1)/n_1$$

$$E(\hat{p}_2) = p_2, \text{Var}(\hat{p}_2) = p_2(1-p_2)/n_2$$

- Then  $L = \hat{p}_1 - \hat{p}_2$ , we can get

$$E(L) = p_1 - p_2$$

$$\text{Var}(L) = p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2$$

- Idea:** use  $L$  as the test statistic and use the normal approximation for the calculation.

# Inference for Two Proportions

- Assumptions for the test of two proportions are:
  - Each random sample is selected from the target population.
  - Two samples are independent.
  - The sample proportions,  $\hat{p}_1$  and  $\hat{p}_2$  are approximate normal. That means, the normal approximation to the binomial is appropriate:
$$n_1\hat{p}_1 \geq 5; n_1(1 - \hat{p}_1) \geq 5$$
$$n_2\hat{p}_2 \geq 5; n_2(1 - \hat{p}_2) \geq 5$$
- Note that two sample sizes may not be equal.

# Inference for Two proportions

- The null hypothesis is:  $H_0: p_1 - p_2 = \delta_0$
- $\delta_0$  is the nominal difference two means. In our course,  $\delta_0 = 0$
- The test statistic is:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2)}{\text{Estimated Standard Deviation of } (\hat{p}_1 - \hat{p}_2)}$$

- How to estimate the standard deviation of  $(\hat{p}_1 - \hat{p}_2)$ ?
- $Z$  has a standard normal distribution when  $H_0$  is true.

# Pooled Estimate of Proportion

- $n_1$ : sample size from the first sample
- $n_2$ : sample size from the second sample
- $\hat{p}_1 = y_1/n_1$ : sample proportion from the first sample
- $\hat{p}_2 = y_2/n_2$ : sample proportion from the second sample
- The pooled estimate of proportion is given by:

$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} = \frac{y_1 + y_2}{n_1 + n_2}$$

# Inference for Two Proportions

- The null hypothesis is:  $H_0: p_1 - p_2 = 0$
- The test statistic is:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1) \text{ when } H_0 \text{ is true}$$

- Where  $\hat{p} = \frac{y_1 + y_2}{n_1 + n_2}$  is the pooled estimate of proportion.

# Inference for Two Proportions – Rejection Region

- For two-sided test ( $H_a: p_1 - p_2 \neq 0$ ), the rejection region is:  
$$|z| > z_{\alpha/2}$$
- For upper-tailed (right-tailed) test ( $H_a: p_1 - p_2 > 0$ ), the rejection region:

$$z > z_\alpha$$

- For lower-tailed (left-tailed) test ( $H_a: p_1 - p_2 < 0$ ), the rejection region is:

$$z < -z_\alpha$$

# Inference for Two Proportions – Confidence Interval

- For two-sided test ( $H_a: p_1 - p_2 \neq 0$ ), construct a  $100(1 - \alpha)\%$  two-sided confidence interval  $\left( \hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right)$ .
- For upper-tailed test ( $H_a: p_1 - p_2 > 0$ ), construct a  $100(1 - \alpha)\%$  lower confidence interval  $\left( \hat{p}_1 - \hat{p}_2 - z_{\alpha} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, 1 \right)$ .
- For lower-tailed test ( $H_a: p_1 - p_2 < 0$ ), construct a  $100(1 - \alpha)\%$  upper confidence interval  $\left( -1, \hat{p}_1 - \hat{p}_2 + z_{\alpha} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right)$ .

# Inference for Two Proportions – Confidence Interval

- Note that when we construct the confidence interval for  $p_1 - p_2$ , we can not assume that  $p_1 = p_2$ .
- The estimate of the variance of  $\hat{p}_1 - \hat{p}_2$  is

$$\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2.$$

# Inference for Two Proportions – $p$ -value Approach

- For two-sided test ( $H_a: p_1 - p_2 \neq 0$ ):

$$p\text{-value} = 2\Pr(Z > |z|)$$

- For upper-tailed (right-tailed) test ( $H_a: p_1 - p_2 > 0$ ):

$$p\text{-value} = \Pr(Z > z)$$

- For lower-tailed (left-tailed) test ( $H_a: p_1 - p_2 < 0$ ):

$$p\text{-value} = \Pr(Z < z)$$

# Inference on Two Proportions – Example 5.8

- **Example 5.8** - A candidate for political office wants to determine whether he is more popular in women than in men. He conducts a sample survey of 250 men and 250 women, of which 105 men and 128 women favor his candidacy. Do these values indicate the candidate is more popular in women than in men?
- Use a significance level of 0.04.
- Basic Statistics:

$$n_1 = n_2 = 250; \hat{p}_1 = 0.42; \hat{p}_2 = 0.512$$

# Inference on Two Proportions – Example 5.8

**Example 5.8 – Key Results (Again, some steps are skipped)**

- Pooled estimate of  $\hat{p} = \frac{105+128}{250+250} = 0.4660$
- Test statistic:  $z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})/n_1 + \hat{p}(1-\hat{p})/n_2}}$
- Test statistic:  $z = \frac{0.42 - 0.512}{\sqrt{0.466*(1-0.466)/250 + 0.466*(1-0.466)/250}} = -2.0620$
- Rejection Region:  $z < -z_{0.04} = -1.7507$
- The  $p$ -value is:  $\Pr(Z < -2.0620) = 0.0196$
- The upper 96% CI is  $(-1, -0.0142)$

# Inference on Two Proportions – Example 5.8

```
prop.test(x = c(y1, y2), n = c(n1, n2),  
          alternative = "less", conf.level = 1 - alpha, correct = FALSE)
```

2-sample test for equality of proportions without continuity correction

data: c(y1, y2) out of c(n1, n2)

**X-squared = 4.2517, df = 1, p-value = 0.01961**

alternative hypothesis: less

96 percent confidence interval:

**-1.0000000 -0.01422098**

sample estimates:

prop 1 prop 2

**0.420 0.512**

# Inference Two Proportions - Exercise

- **Exercise (Newly Hatched Larvae)** – A researcher studied newly hatched larvae of the common carp for exposure to copper or lead during the embryonic development. They were interested in if the percentage of defective larvae differed for the two metal solutions. They put 100 eggs copper and 80 eggs in lead. The number of defective larvae was 38 with copper and 18 with lead.
- **Question:** Did the percentage of defective larvae differ for the two metal solutions? Use the significance of 0.04 for the test.

# Comparing Portions Using Paired Samples

- **Example 5.9** - In an experiment for evaluating a new headache remedy, 80 chronic headache sufferers are given a standard remedy and a new drug on different days, and the response is whether their headache was relieved. In the experiment 56, or 70%, were relieved by the standard remedy and 64, or 80%, by the new drug. Do the data suggest that the new drug is better to relieve the headache.

**Table 5.9** Data on Headache Remedy

|                 | STANDARD REMEDY |                    |  | <b>Totals</b> |
|-----------------|-----------------|--------------------|--|---------------|
|                 | <b>Headache</b> | <b>No Headache</b> |  |               |
| <i>New drug</i> |                 |                    |  |               |
| Headache        | 10              | 6                  |  | 16            |
| No Headache     | 14              | 50                 |  | 64            |
| Totals          | 24              | 56                 |  | 80            |

# Comparing Portions Using Paired Samples

- **Example 5.9** – The method here is to look at the pairs that are different and test if the proportion is 0.5.
- The two numbers are 6 and 14.
- 6: number of persons had headache after new drug but did not have headache after standard drug.
- 14: number of persons did not have headache after new drug but had headache after standard drug.
- Note these 20 samples are independent.
- Use 6 as the observed success to perform the test:

$$H_0: p = 0.5 \text{ versus } H_a: p < 0.5$$

# Comparing Portions Using Paired Samples

- **Example 5.9 – Key Results (Again, some steps are skipped)**
- $\hat{p} = \frac{6}{20} = 0.30$
- Test statistic:  $z = \frac{\hat{p}-0.50}{\sqrt{0.5*0.5/20}} = \frac{0.30-0.50}{\sqrt{0.5*0.5/20}} = -1.7889$
- Rejection region:  $z < -z_{0.05} = 1.6449$
- The  $p$ -value is:  $\Pr(Z < -1.7889) = -0.0368$
- The upper 95% CI is:
- $\left(0, \hat{p} + z_{0.05} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = \left(0, 0.30 + 1.6449 * \sqrt{\frac{0.3*0.7}{20}}\right) = (0, 0.4685)$

# Assumptions

- **Pooled t statistic:** (a) Two sets of samples are independent; (b) Distributions are normal or approximate normal (large sample size); (c) Variances are equal.
- **Paired t statistic:** (a) samples are independent; (b) Observations are paired; (c) Distributions are normal or approximate normal.
- **Inferences on binomial populations:** (a) Observations are independent (for McNemar's test pairs are independent); (b) Probability of success is constant for all observations; (c) Large sample sizes for normal approximation.

# Assumptions

- When assumptions are not fulfilled - analysis is not appropriate and/or the significance levels ( $p$ -values) are not as advertised.
- Violation of distributional assumptions may be detected by the exploratory data analysis methods described in Chapter 1, which should be routinely applied to all data.
- When assumptions are not fulfilled or not clear-cut.
  - For the  $t$  statistics, minor violations are not particularly serious because these statistics are relatively robust.
  - It will be necessary to investigate other analysis strategies.

# Chapter Summary

- **Inferences on means based on independent samples where the variances can be assumed equal** - use a single pooled estimate of the common variance.
- **Inferences on means based on independent samples where the variances cannot be assumed equal** - use the estimated variances for large samples. For small samples an approximation must be used.
- **Inferences on means based on dependent (paired) samples** - use differences between the pairs.

# Chapter Summary

- **Inferences on proportions from independent samples** - use the normal approximation of the binomial to compute a statistic similar to that for inferences on means when variances are assumed known.
- **Inferences on proportions from dependent samples** - use a statistic based on information only on pairs whose responses differ between the two groups.
- **If assumptions are violated** – use alternative methods.

# MA5701: Statistical Methods

Summary

Kui Zhang, Mathematical Sciences

# Final Exam

- **Time and Date:** 12:45pm, Wednesday, April 23, 2025
- **Room:** Rekhi 214
- **Requirements:**
  - Two Hours
  - Pens, One Calculator, 4 letter size one-sided note
  - Covers contents from Chapter 1 to Chapter 5
  - Problems will be similar with homework problems

# Chapter 1 – Variables

- **Qualitative (Categorical) variable** - is a variable that is not numerical. It describes data that fits into categories.
  - The **ordinal scale** distinguishes between measurements Generally, the relative amounts of some characteristic they process.
  - The **nominal scale** identifies observed values by name or classification.
- **Quantitative Variable** – is a variable that is measured on a numeric scale for which meaningful arithmetic operations make sense.
  - A **discrete** variable can assume only accountable number of values.
  - A **continuous** variable is one that can take any one of an uncountable number of values in an interval.

# Chapter 1 – Descriptive Statistics

- Let  $y_1, \dots, y_n$  denote a sample of interest.

- **Sample Mean:**  $\bar{y} = \frac{1}{n}(y_1 + y_2 + \dots + y_n)$

- **Sample Variance:**

$$s^2 = \frac{1}{n-1} [(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \dots + (y_n - \bar{y})^2] = \frac{1}{n-1} (\sum_{i=1}^n y_i^2 - n\bar{y}^2)$$

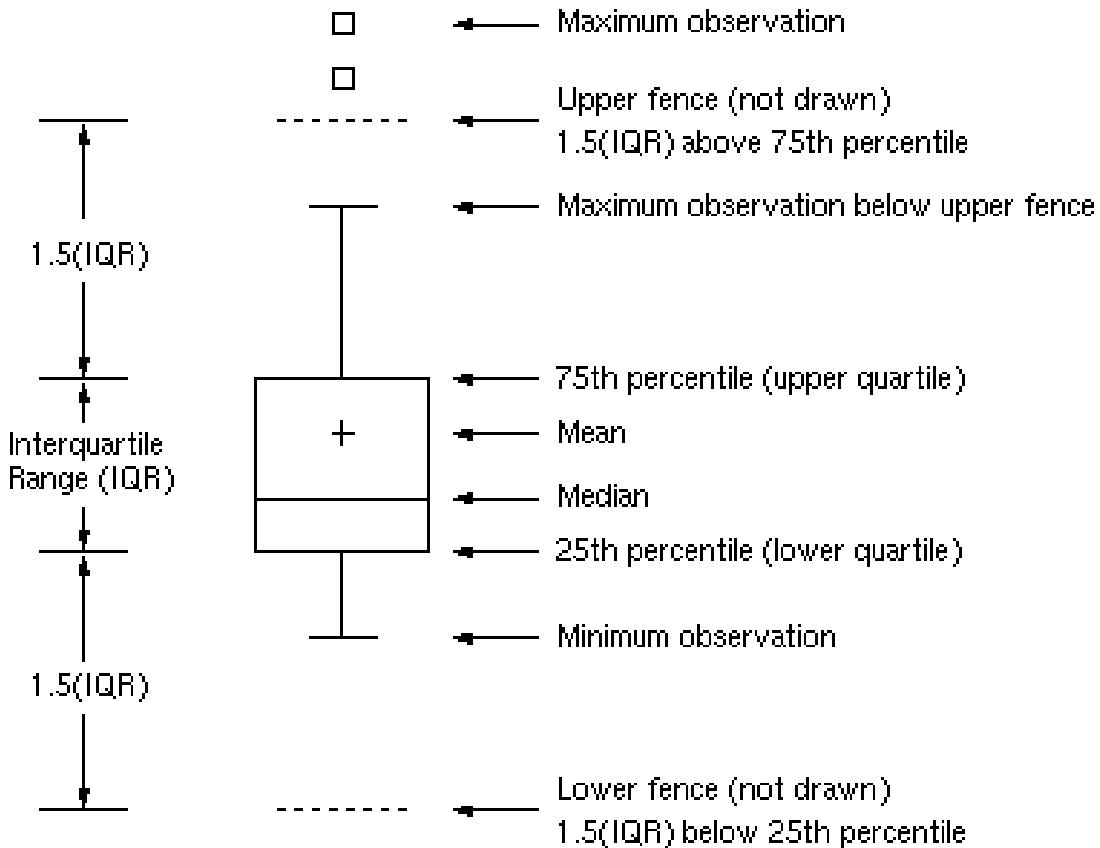
- Sample Median: The **median** of a set of observed values is defined to be the middle value when the measurements are arranged from lowest to the highest. **Need to know how to find it.**

- Median  $\tilde{y} = y_{(\frac{n+1}{2})}$  if  $n$  is odd;  $\tilde{y} = \frac{y_{(\frac{n}{2})} + y_{(\frac{n}{2}+1)}}{2}$  if  $n$  is even

# Chapter 1 – Descriptive Statistics

- **Quartiles**, 25%, 50%, 75% percentile
  - 25% percentile – lower quartile, first quartile ( $Q_1$ )
  - 50% percentile – median, second quartile
  - 75% percentile – upper quartile, third quartile ( $Q_3$ )
- The **interquartile range** is the length of the interval between the 25<sup>th</sup> and 75<sup>th</sup> percentiles.

# Chapter 1 – Schematics of Boxplot



# Chapter 2 – Probability Calculation

- **Always True:**

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$

- **If  $A$  and  $B$  are mutually exclusive**

$$\begin{aligned}\Pr(A \cup B) &= \Pr(A) + \Pr(B) \\ \Pr(A \cap B) &= 0\end{aligned}$$

- **If  $A$  and  $B$  are independent**

$$\Pr(A \cap B) = \Pr(A) * \Pr(B)$$

- **For any event  $A$ ,**

$$\Pr(A) + \Pr(A^c) = 1$$

# Chapter 2 – Probability Calculation

- If  $A_1, \dots, A_n$  are mutually exclusive

$$\Pr(A_1 \cup \dots \cup A_n) = \Pr(A_1) + \dots + \Pr(A_n)$$

- If  $A_1, \dots, A_n$  are independent

$$\Pr(A_1 \cap \dots \cap A_n) = \Pr(A_1) * \dots * \Pr(A_n)$$

- If  $A, B, C$  are independent, then

$$\Pr(A \cap B \cap C) = \Pr(A) * \Pr(B) * \Pr(C)$$

$$\Pr(A^c \cap B \cap C) = \Pr(A^c) * \Pr(B) * \Pr(C)$$

$$\Pr(A \cap B^c \cap C^c) = \Pr(A) * \Pr(B^c) * \Pr(C^c)$$

# Chapter 2 – Random Variables

- A **discrete random variable** is one that can take on only a countable number of values.
  - It has a probability mass function:  $f(y) = \Pr(Y = y)$
  - $\Pr(Y \leq y) = \sum_{x \leq y} f(x)$
- A **continuous random variable** is one that can take on any value in an interval.
  - It has a probability density function:  $f(y)$  (and  $\Pr(Y = y) = 0$ )
  - $\Pr(Y \leq y) = \int_{-\infty}^y f(x) dx$

# Chapter 2 – Discrete Random Variable

For a discrete random variable  $y$  with the pmf  $f(y)$

- **Expected Value**

$$\mu = E[Y] = \sum_y f(y) * y$$

- **Population variance** of  $Y$ , denoted by  $\sigma^2$ , is:

$$\sigma^2 = \text{var}(Y) = \sum_y f(y) * (y - \mu)^2$$

- **Population standard deviation** of  $Y$ , denoted by  $\sigma$ , is  $\sigma = \sqrt{\sigma^2}$ , the square root of the population variance.

# Chapter 2 – Bernoulli and Binomial R.V.s

- A random variable  $Y$  has a **Bernoulli( $p$ )** distribution if  $f(1) = \Pr(Y = 1) = p$  and  $f(0) = \Pr(Y = 0) = 1 - p$ .
  - Population mean:  $\mu = E[Y] = p$ .
  - Population variance:  $\sigma^2 = \text{var}(Y) = p(1 - p)$ .
- A random variable  $Y$  has a **Binomial distribution**,  $\text{Binomial}(n, p)$ , if its *pmf* is

$$f(y) = \Pr(Y = y) = \binom{n}{y} p^y (1 - p)^{n-y}, y = 0, \dots, n.$$

- If  $Y$  is the number of successes from  $n$  independent identical Bernoulli trials, then  $Y$  has  $\text{Binomial}(n, p)$ .
- Population mean:  $\mu = E[Y] = np$ .
- Population variance:  $\sigma^2 = \text{var}(Y) = np(1 - p)$ .

# Chapter 2 – Normal Distribution

- The random variable has the following pdf:

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right), -\infty < y < \infty$$

- Many times, we use the simple notation:  $Y \sim N(\mu, \sigma^2)$
- Population mean is:  $\mu$ .
- Population variance and standard deviation are:  $\sigma^2$  and  $\sigma$ .
- If  $Y \sim N(\mu, \sigma^2)$  and  $\mu = 0$  and  $\sigma^2 = \sigma = 1$ , then  $Y$  has a standard normal distribution.

# Chapter 2 – Normal Table

- Need to know how to use the normal table to find:
- If  $Z \sim N(0,1)$ ,  
 $\Pr(Z > a), \Pr(Z < b), \Pr(a < Z < b)$
- If  $Y \sim N(\mu, \sigma^2)$   
 $\Pr(Y > a), \Pr(Y < b), \Pr(a < Y < b)$
- Find  $Z_\alpha$  such as  $Z_{0.05}, Z_{0.02}$ , etc.

# Chapter 2 – Distribution of Sample Mean

- If the sample is a random sample from the normal distribution  $N(\mu, \sigma^2)$ , then sample mean  $\sim N\left(\mu, \frac{\sigma^2}{n}\right)$
- **Central Limit Theorem** - If random samples of size  $n$  are taken from any distribution with mean  $\mu$  and variance  $\sigma^2$ , sample mean will have a distribution approximately normal with mean  $\mu$  and variance  $\sigma^2/n$ .
- If  $Y \sim Binomial(n, p)$ , then approximately

$$\frac{\frac{Y}{n} - p}{\sqrt{p(1-p)/n}} = \frac{Y - np}{\sqrt{np(1-p)}} \sim N(0,1)$$

# Chapter 3 – Hypothesis Testing

- **Alternative hypothesis ( $H_a$ )** - a statement that contradicts the null hypothesis. This hypothesis is accepted if the null hypothesis is rejected. The alternative hypothesis is often called the ***research hypothesis*** because it usually implies that some action is to be performed, some money spent, or some established theory overturned.

# Chapter 3 - Possible Errors in Hypothesis Testing

- A **type I error** occurs when we incorrectly reject  $H_0$ , that is, when  $H_0$  is true, and our sample-based inference procedure rejects it.
- A **type II error** occurs when we incorrectly fail to reject  $H_0$ , that is, when  $H_0$  is not true, and our inference procedure fails to detect this fact.

|                       |              | In the Population |                   |
|-----------------------|--------------|-------------------|-------------------|
| The Decision          |              | $H_0$ is True     | $H_0$ is Not True |
| $H_0$ is Not Rejected | Correct      | Type II Error     |                   |
| $H_0$ is Rejected     | Type I Error | Correct           |                   |

# Chapter 3 – Rejection Region

- The **rejection region** (also called the **critical region**) is the range of values of a sample statistic that will lead to rejection of the null hypothesis.
- $R$ : rejection region and  $W$  is your test statistic
- Probability of making a type I error
$$\alpha = \Pr(W \in R | H_0 \text{ is true})$$
- Probability of making a type II error
$$\beta = \Pr(W \in R^c | H_a \text{ is true})$$
- $1 - \beta$ : the power of test.

# Chapter 3 – 5 Steps

- **Step 1:** Specify  $H_0$ ,  $H_a$ , and  $\alpha$
- **Step 2:** Define test statistic
- **Step 3:** Determine rejection region
- **Step 4:** Calculate test statistic based on data
- **Step 5:** State conclusions
- **Alternative approach:**  $p$ -value
- **Alternative approach:** confidence interval

# Chapter 3 – Interval Estimators

**How to interpret an interval estimator? For example:**

Interpretation of 95% confidence interval, (7.792, 7.988). Which statement is correct:

1.  $\Pr(7.792 < \mu < 7.988) = 0.95$ .
2. 95% of all weights between 7.792 and 7.988.
3. We sampled 95% of all weights.
4. We know that  $7.792 < \mu < 7.988$ .
5. We are 95% confident that the true population mean is between 7.792 and 7.988.

# Chapter 4 – One Sample $t$ -test

- $H_0: \mu = \mu_0$  versus  $H_a: \mu \neq \mu_0$
- **Test statistic is always:**  $T = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}}$
- **Rejection Region:**  $|t| > t_{n-1, \alpha/2}$
- **$p$ -value:**  $2\Pr(T_{n-1} > |t|)$
- **Two-sided CI:**  $\left(\bar{y} - t_{n-1, \frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{y} + t_{n-1, \frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right)$

# Chapter 4 – Test for One Proportion

- $H_0: p = p_0$  versus  $H_a: p > p_0$
- **Test statistic is always:**  $Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$
- **Rejection Region:**  $|z| > z_\alpha$
- **p-value:**  $\Pr(Z > z)$
- **Lower CI:**  $\left( \hat{p} - z_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, 1 \right)$

# Chapter 5 – Two Sample $t$ -test

- $H_0: \mu_1 - \mu_2 = \delta_0$  versus  $H_a: \mu_1 - \mu_2 \neq \delta_0$ .
- **Test statistic is always:**  $T = \frac{(\bar{y}_1 - \bar{y}_2) - \delta_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
- Where  $s_p = \sqrt{s_p^2}$  and  $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{SS_1 + SS_2}{n_1+n_2-2}$
- **Rejection Region:**  $|t| > t_{n_1+n_2-2, \alpha/2}$
- **p-value:**  $2\Pr(T_{n_1+n_2-2} > |t|)$
- **Two-sided CI:**  $(\bar{y}_1 - \bar{y}_2 \pm t_{n_1+n_2-2, \alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}})$

# Chapter 5 – Paired $t$ -test

- Paired  $t$ -test is just one sample  $t$ -test for difference of paired data.

# Chapter 5 – Inference for Two Proportions

- $H_0: p_1 - p_2 = 0$  versus  $H_a: p_1 - p_2 < 0$
- **Test statistic** is:  $Z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$
- Where  $\hat{p} = \frac{y_1 + y_2}{n_1 + n_2}$  is the pooled estimate of proportion.
- **Rejection region** is:  $z < -z_\alpha$
- **p-value** is:  $\Pr(Z < z)$
- **Upper CI:**  $\left(-1, \hat{p}_1 - \hat{p}_2 + z_\alpha \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right)$

# Chapter 5 – Assumptions

- Assumptions for t-test (one-sample, two-sample, paired)
  - Data is normal or approximate normal
  - Check with Q-Q plot
- Assumptions for inference of proportions
  - One proportion:  $np_0 \geq 5$  and  $n(1 - p_0) \geq 5$
  - Two proportions:

$$n_1\hat{p}_1 \geq 5; n_1(1 - \hat{p}_1) \geq 5$$
$$n_2\hat{p}_2 \geq 5; n_2(1 - \hat{p}_2) \geq 5$$

Attach all of your R code at the end of your solution file or in a separate text file.

### Problem 1 (40 points)

A measure of the time a drug stays in the blood system is given by the half-life of the drug. This measure is dependent on the type of drug, the weight of the patient, and the dose administered. To study the half-life of aminoglycosides in trauma patients, a pharmacy researcher recorded the data for patients in a critical care facility. The data is in the data file drug.csv which can be downloaded from the canvas. The data consists of measurements of dosage per kilogram of weight of the patient, type of drug, either Amikacin or Gentamicin, and the half-life measured 1 hour after administration.

- (1) **(3 points)** Practice of R. Use R to (a) create a data named “drug” from the data file drug.csv; (b) look at a few first lines of data; (c) the names of the variables in the data. For this part, you do not need to include any output from R.
- (2) **(3 points)** Describe the data set. You need to include the information about the purpose of the study, the variables (and their names), and the sample size in your description.
- (3) **(4 points)** Determine the type of each variable. First, determine if a variable is quantitative or qualitative. If a variable is quantitative, further determine if it is continuous or discrete. If a variable is qualitative, further determine if it is ordinal or nominal.
- (4) **(10 points)** For the type of drugs, use R to construct a frequency table. Your table should contain the frequency and the relative frequency (with 3 decimal digits). Present your table in your solution file. **Comment** on your findings about this table.
- (5) **(10 points)** For the Half-life, use R to construct a frequency table. Your table should contain the frequency and the relative frequency (with 3 decimal digits) based on the following intervals: [0.50, 1.00), [1.00, 1.50), [1.50, 2.00), [2.00, 2.50), [2.50, 3.00), and [3.00, 3.50). Present your table in the solution file. **Comment** on your findings about this table.
- (6) **(10 points)** For the dosage, use R to construct a frequency table. Your table should contain the frequency and the relative frequency (with 3 decimal digits) based on the following intervals: [0, 2.00), [2.00, 4.00), [4.00, 6.00), [6.00, 8.00), [8.00, 10.00) , and [10.00, 12.00). **Comment** on your findings about this table.

(1) (a)  
`drug <- read.csv(file = "drug.csv", stringsAsFactors = FALSE)`

(1) (b)  
`head(drug)`

(1) (c)  
`names(drug)`

40/40 great!

## (2)

The purpose of the study is to analyze the half-life of aminoglycosides in trauma patients, which is affected by the drug type, dosage, and patient weight. This is part of a study by pharmacy researchers to understand drug behavior in a critical care setting.

Variables in the dataset

1. patient: unique identifier for each patient
2. drug: type of drug administered to the patient
3. half.life: half-life of the drug in the patient's body
4. dosage: dosage of the drug administered to the patient

sample size: 43

`n <- length(drug$patient)`

## (3)

patient: qualitative and nominal data

drug: qualitative and nominal data

half.life: quantitative and continuous data

dosage: quantitative and continuous data

## (4)

Frequency table for type of drugs

| Drug | Frequency | Relative Frequency |
|------|-----------|--------------------|
| a    | 22        | 0.512              |
| b    | 21        | 0.488              |

There are two drug types in the dataset: Amikacin (a) and Gentamicin (g). Amikacin (a) is slightly more common, with a relative frequency of 0.512 (51.2%). Gentamicin (g) accounts for 0.488 (48.8%) of the observations. The near-equal distribution suggests that both drugs were used similarly frequently in the study which may help comparable results when analyzing the effects of the two drugs.

## (5)

Frequency table for the Half-life

| half. life | Frequency | Relative Frequency |
|------------|-----------|--------------------|
| [0.5,1)    | 1         | 0.023              |

|         |    |       |
|---------|----|-------|
| [1,1.5) | 7  | 0.163 |
| [1.5,2) | 17 | 0.395 |
| [2,2.5) | 11 | 0.256 |
| [2.5,3) | 6  | 0.140 |
| [3,3.5) | 1  | 0.023 |

The majority of the half-life values fall within the interval [1.5, 2.0), which accounts for 39.5% of the observations (17 patients). The smallest group is in the interval [3.0, 3.5), with 2.3% of the observations (1 patient). The distribution is concentrated primarily between [1.5, 2.5), suggesting that most patients have a half-life within this range.

## (6)

Frequency table for the dosage

| dosage  | Frequency | Relative Frequency |
|---------|-----------|--------------------|
| [0,2)   | 2         | 0.047              |
| [2,4)   | 19        | 0.442              |
| [4,6)   | 0         | 0.000              |
| [6,8)   | 4         | 0.093              |
| [8,10)  | 8         | 0.186              |
| [10,12) | 10        | 0.233              |

The majority of patients fall in the dosage interval [2.00, 4.00), accounting for 44.2% of the observations (19 patients). No patients have dosages in the interval [4.00, 6.00).

# R code

```
# 1. (a) create a data named "drug" from the data file drug.csv
drug <- read.csv(file = "drug.csv", stringsAsFactors = FALSE)
# 1. (b) look at a few first lines of data
head(drug)
# 1.(c) the names of the variables in the data
names(drug)
```

# 2. Describe the data set

```
# The purpose of the study is to analyze the half-life of aminoglycosides in trauma patients, which is affected by the drug type, dosage, and patient weight.
# This is part of a study by pharmacy researchers to understand drug behavior in a critical care setting.
# Variables in the dataset
```

```
# 1. patient: unique identifier for each patient
# 2. drug: type of drug administered to the patient
# 3. half.life: half-life of the drug in the patient's body
# 4. dosage: dosage of the drug administered to the patient
# sample size:
n <- length(drug$patient)
n # 43

# 3. Determine the Type of Each Variable
str(drug)
# patient: qualitative variable, nominal data
# drug: qualitative variable, nominal data
# half.life: quantitative variable, continuous data
# dosage: quantitative variable, continuous data

# 4. For the type of drugs, use R to construct a frequency table
table(drug$drug)
prop.table(table(drug$drug))
b <- prop.table(table(drug$drug))
round(b, digits = 3)

# 5. For the Half-life, use R to construct a frequency table
a <- cut(x = drug$half.life,
          breaks = c(0.50, 1.00, 1.50, 2.00, 2.50, 3.00, 3.50),
          right = FALSE)
a
b <- table(a)
b
d <- b / sum(b)
round(d, digits = 3)

# 6. For the dosage, use R to construct a frequency table
a <- cut(
  x = drug$dosage,
  breaks = c(0, 2.00, 4.00, 6.00, 8.00, 10.00, 12.00),
  right = FALSE)
b <- table(a)
b
d <- b / sum(b)
round(d, digits = 3)
```

Attach all of your R code at the end of your solution file or in a separate text file.

**Problem 1 (12 points, 1 point for each)**

The following multiple-choice questions are intended to provide practice in methods and reinforce some of the concepts presented in Chapter 1.

1.1 The scores of eight persons on the Stanford–Binet IQ test were:

95      87      96      110      150      104      112      110

The median is:

- (1) 107
- (2) 110
- (3) 112
- (4) 104
- (5) none of the above

1.2 The concentration of DDT, in milligrams per liter, is:

- (1) a nominal variable
- (2) an ordinal variable
- (3) a continuous variable
- (4) a discrete variable

1.3 If the interquartile range is zero, you can conclude that:

- (1) the range must also be zero (Note that the range of the data is the difference between the largest and smallest values)
- (2) the mean is also zero
- (3) at least 50% of the observations have the same value
- (4) all of the observations have the same value
- (5) none of the above is correct

1.4. The species of each insect found in a plot of cropland is:

- (1) a nominal variable
- (2) an ordinal variable
- (3) a continuous variable
- (4) a discrete variable

1.5 A sample of 100 IQ scores produced the following statistics:

$$\text{mean} = 95 \quad \text{lower quartile} = 70$$

$$\text{median} = 100 \quad \text{upper quartile} = 120$$

$$\text{standard deviation} = 30$$

Which statement(s) is (are) correct?

- (1) Half of the scores are less than 95.
- (2) The middle 50% of scores are between 100 and 120.
- (3) One-quarter of the scores are greater than 120.

1.6 A sample of 100 IQ scores produced the following statistics:

$$\text{mean} = 100 \quad \text{lower quartile} = 70$$

$$\text{median} = 95 \quad \text{upper quartile} = 120$$

$$\text{standard deviation} = 30$$

Which statement(s) is (are) correct?

- (1) Half of the scores are less than 100.
- (2) The middle 50% of scores are between 70 and 120.
- (3) One-quarter of the scores are greater than 100.

1.7 A sample of pounds lost in a given week by individual members of a weight reducing clinic produced the following statistics:

$$\text{mean} = 5 \text{ pounds} \quad \text{first quartile} = 2 \text{ pounds}$$

$$\text{median} = 7 \text{ pounds} \quad \text{third quartile} = 8.5 \text{ pounds}$$

$$\text{standard deviation} = 2 \text{ pounds}$$

Identify the correct statement:

- (1) One-fourth of the members lost less than 2 pounds.
- (2) The middle 50% of the members lost between 2 and 8.5 pounds.
- (3) Both (1) and (2) are correct.
- (4) Neither (1) and (2) is correct.

1.8 A measurable characteristic of a population is:

- (1) a parameter
- (2) a statistic
- (3) a sample
- (4) an experiment

1.9 What is the primary characteristic of a set of data for which the standard deviation is zero?

- (1) All values of the variable appear with equal frequency.
- (2) All values of the variable have the same value.
- (3) The mean of the values is also zero.
- (4) All of (1), (2), and (3) are correct.
- (5) Neither of (1), (2) or (3) is correct.

1.10 A subset of a population is:

- (1) a parameter
- (2) a population
- (3) a statistic
- (4) a sample
- (5) none of the above

1.11 The median is a better measure of central tendency than the mean if:

- (1) the variable is discrete
- (2) the distribution is skewed
- (3) the variable is continuous
- (4) the distribution is symmetric
- (5) none of the above is correct

1.12 A small sample of automobile owners at Texas A & M University produced the following number of parking tickets during a particular year: 4, 0, 3, 2, 5, 1, 2, 1, 0. The mean number of tickets (rounded to the nearest tenth) is:

- (1) 1.7
- (2) 2.0
- (3) 2.5

- (4) 3.0
- (5) none of the above

**Problem 2 (10 points)**

On ten days, a bank had 18, 15, 13, 12, 8, 3, 7, 14, 16, and 3 bad checks. Find the sample mean, sample median, sample variance, and sample standard deviation, of the number of bad checks. You need to show the details about how you calculate them.

**Problem 3 (24 points, 6 points for each part)**

Data file anl.csv contains the times in days from remission induction to relapse for 50 patients with acute nonlymphoblastic leukemia who were treated on a common protocol at university and private institutions in the Pacific Northwest. This is a portion of a larger study reported by Glucksberg et al. (1981). The data file only contains one variable “days”.

- (1) Draw a histogram using R function hist with default settings. Comments on the features of your plot. You must use appropriate main title and labels for x-axis and y-axis for the plot.
- (2) Draw a boxplot using R function boxplot. Comments on the features of your plot. You must use appropriate main title and labels for x-axis and y-axis for the plot.
- (3) Use R to calculate the following summary statistics: sample mean, sample median, sample variance, sample standard deviation, sample median, sample lower quartile, and sample upper quartile. You only need to present your final answer here. You do not need to present any details about your calculations.
- (4) Calculate the following values used in the boxplot: interquartile, step, lower inner fence, upper inner fence, lower outer fence, and upper outer fence. You need to show the details about how you calculate them based on summary the statistics obtained from (3).

1.1 The median is (1): 107 ✓

46/46

1.2 The concentration of DDT, in milligrams per liter, is: (3) a continuous variable ✓

1.3 If the interquartile range is zero, you can conclude that: (3) at least 50% of the observations have the same value ✓

1.4 The species of each insect found in a plot of cropland is: (1) a nominal variable ✓

1.5 Which statement(s) is (are) correct?: (3) One-quarter of the scores are greater than 120. ✓

1.6 Which statement(s) is (are) correct?: (2) The middle 50% of scores are between 70 and 120. ✓

1.7 Identify the correct statement: (3) Both (1) and (2) are correct. ✓

1.8 A measurable characteristic of a population is: (1) a parameter ✓

1.9 (2) All values of the variable have the same value. ✓

1.10 A subset of a population is: (4) a sample ✓

1.11 (2) the distribution is skewed ✓

1.12 The mean number of tickets (rounded to the nearest tenth) is: (2) 2.0 ✓

## Problem 2

Bad checks - 18, 15, 13, 12, 8, 3, 7, 14, 16, 3

Sample Mean:

Mean = Sum of all observations / Number of observations

Mean =  $(18 + 15 + 13 + 12 + 8 + 3 + 7 + 14 + 16 + 3) / 10 = 10.9$

Sample Median:

Arrange the bad checks in arranging order

Bad checks in ascending order: 3, 3, 7, 8, 12, 13, 14, 15, 16, 18 ✓

Median =  $(12 + 13) / 2 = 12.5$  ✓

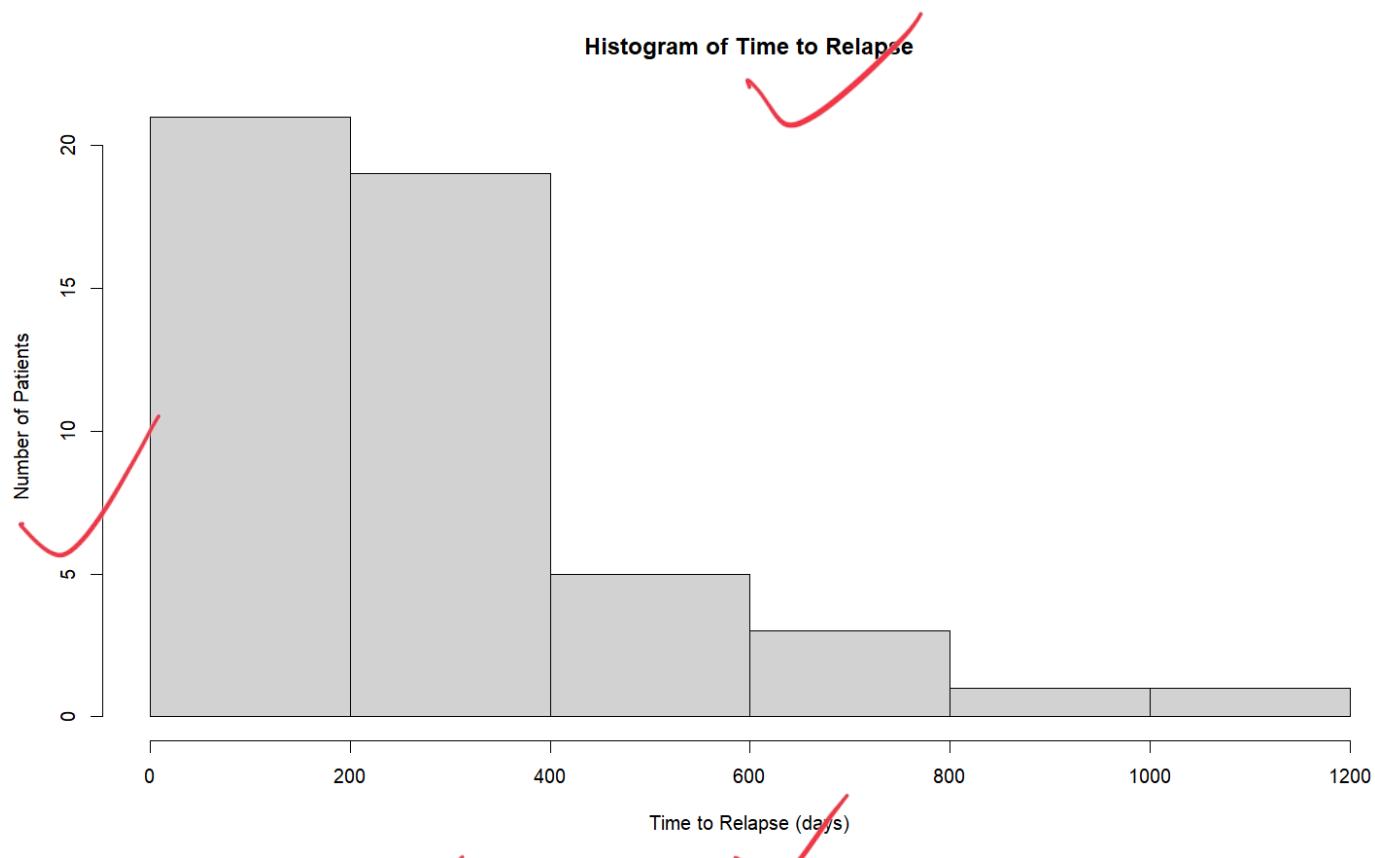
Sample Variance =  $\Sigma(x_i - \mu)^2 / (n - 1)$  where  $\mu$  is mean

Sample Variance =  $(18 - 10.9)^2 + (12 - 10.9)^2 + \dots + (3 - 10.9)^2 / (10 - 1) = 28.54444$

Sample Standard deviation (s) =  $\sqrt{\text{sample variance}} = \sqrt{28.54444} = 5.3427$  ✓

Problem 3)

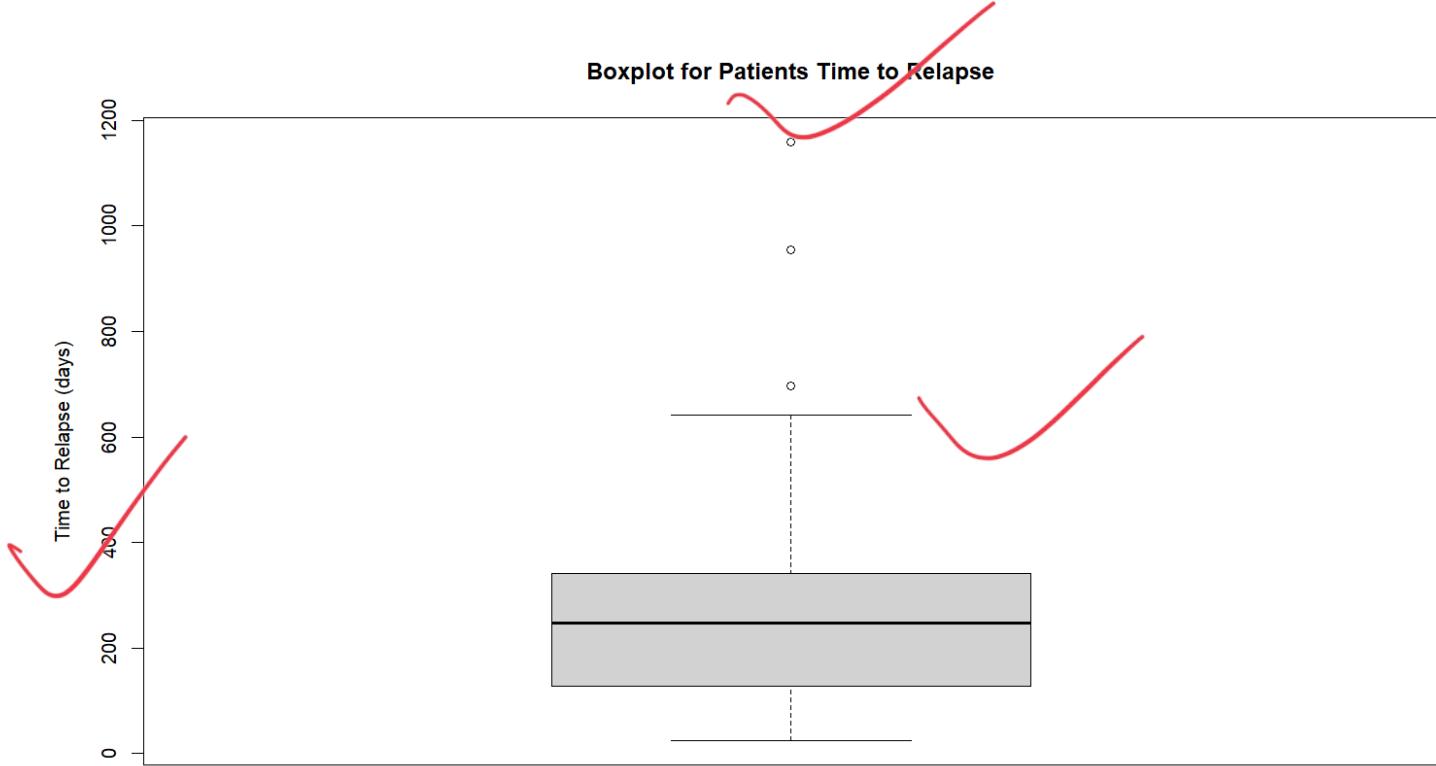
(1) Histogram using R with function hist with default settings for patients Time to Relapse in days



The distribution is right-skewed, as most relapse times occur at lower values (0 - 400) with tail extending at higher values (1200). Most patients' relapse occurs in the 0 - 200 days range. The data shows a wide range, with relapse times extending from close to 0 days up to 1200 days. Very few relapse in higher values.

(2) Boxplot using R function boxplot for patients time to relapse (days)

Boxplot for Patients Time to Relapse



From the above boxplot we can say that the distribution is skewed. There appears to be three outliers from the boxplot. The 1st Quartile and 3rd Quartile lie between above 100 and under 400. The median lies between 400 and 200.

(3)

Sample Mean: 287.88

Sample Median: 248

Sample Variance: 53065.33

Sample Standard Deviation: 230.3591

Lower Quartile: 131.75

Upper Quartile: 338.75

(4)

Interquartile Range (IQR):  $Q_3 - Q_1$

Step:  $Step = 1.5 * IQR$

Lower Inner Face:  $Q_1 - step$

Upper Inner Face:  $Q_2 + step$

Lower outer fence:  $Q_1 - 2 * IQR$

Upper outer fence:  $Q_3 + 2 * IQR$

Interquartile Range (IQR): 207

Step: 310.5

Lower Inner Fence: -178.75

Upper Inner Fence: 649.25

Lower Outer Fence: -282.25

Upper Outer Fence: 752.75

```

# R code

# Problem 1
# 1.1 data: 95 87 96 110 150 104 112 110
# sort the data in ascending order
scores <- c(95, 87, 96, 110, 150, 104, 112, 110)
median_value <- median(scores)
print(median_value)

# problem 2
bad_check <- c(3,3,7,8,12,13,14,15,16,18 )

mean(bad_check)
median(bad_check)
sample_var <- var(bad_check)
sample_var
std_dev <- sqrt(sample_var)
std_dev

# problem 3
anl_data <- read.csv("anl.csv", stringsAsFactors = FALSE)
anl_data
str(anl_data)

# histogram
hist(
  anl_data$days,
  main = "Histogram of Time to Relapse",
  xlab = "Time to Relapse (days)",
  ylab = "Number of Patients",
)

# box plot
boxplot(
  anl_data$days,
  main = "Boxplot for Patients Time to Relapse",
  ylab = "Time to Relapse (days)",
  xlab = ""
)

# Calculate summary statistics
mean_value <- mean(anl_data$days)
median_value <- median(anl_data$days)
variance_value <- var(anl_data$days)

```

```
sd_value <- sd(anl_data$days)
quartiles <- quantile(anl_data$days)

# Print the results
cat("Sample Mean:", mean_value, "\n")
cat("Sample Median:", median_value, "\n")
cat("Sample Variance:", variance_value, "\n")
cat("Sample Standard Deviation:", sd_value, "\n")
cat("Lower Quartile:", quartiles["25%"], "\n")
cat("Upper Quartile:", quartiles["75%"], "\n")

# Calculate boxplot values
IQR_value <- IQR(anl_data$days)
step <- 1.5 * IQR_value
lower_inner <- quartiles["25%"] - step
upper_inner <- quartiles["75%"] + step
lower_outer <- quartiles["25%"] - 2 * IQR_value
upper_outer <- quartiles["75%"] + 2 * IQR_value

# Print the results
cat("Interquartile Range (IQR):", IQR_value, "\n")
cat("Step:", step, "\n")
cat("Lower Inner Fence:", lower_inner, "\n")
cat("Upper Inner Fence:", upper_inner, "\n")
cat("Lower Outer Fence:", lower_outer, "\n")
cat("Upper Outer Fence:", upper_outer, "\n")
```

Attach all of your R code at the end of your solution file or in a separate text file.

### Problem 1 (16 points, 4 points for each part)

The data file used for this problem is `texas-house.csv`. Here we will consider two qualitative variables: the exterior type (“exter”) and the last digit of the zip code (“zip”).

- (1) Construct a two-dimensional frequency data with the values of “exter” as the row and the values of zip code as the column.
- (2) Construct a percentage table based on the table from (1). The percentage is calculated based on each column.
- (3) Construct a bar plot based on the table from (2). Use appropriate main title and labels for the x-axis and y-axis.
- (4) Comment on features of plot from (3).

### Problem 2 (10 points)

Someone wants to know whether the direction of price movements of the general stock market, as measured by the New York Stock Exchange (NYSE) Composite Index, can be predicted by directional price movements of the New York Futures Contract for the next month. Data on these variables have been collected for a 46-day period. The data file is: `nyse.csv`. The variables in this data set are:

index: the percentage change in the NYSE composite index for a one-day period.

future: the percentage change in the NYSE futures contract for a one-day period.

Construct a scatterplot relating these variables and comments on your findings. You need to decide which variable should be used for the x-axis and which variable should be used for y-axis. Use appropriate main title and labels for the x-axis and y-axis.

### Problem 3 (15 points)

The use of placement exams in elementary statistics courses has been a controversial topic in recent times. Some researchers think that the use of a placement exam can help determine whether a student will successfully complete a course (or program). A recent study in a large university resulted in the data listed in the data file `exam.csv`. The placement test administered was an in-house written general mathematics test. The course was Elementary Statistics. The students were

told that the test would not affect their course grade. After the semester was over, students were classified according to their status. The variables are:

score: the students' scores on the placement test (from 0 to 100)

status: the status of the student (coded as 0 = passed the course, 1 = failed the course, and 2 = dropped out before the semester was over)

Construct a boxplot stratified by the status of the students. First discuss features of each boxplot then comment on your findings by comparing those boxplots. Three boxplots should be in one single plot. Use appropriate main title and labels for the x-axis and y-axis.

Problem 1

10/41

1)

The last digit of the zip code

| Exterior | 1 | 2  | 3 | 4  |
|----------|---|----|---|----|
| Brick    | 4 | 10 | 4 | 30 |
| Frame    | 1 | 1  | 5 | 1  |
| other    | 1 | 2  | 7 | 3  |

**Interpretation:** There are a lot of brick exterior type houses when compared to others. Zip code 1 area have fewer houses.

2)

The last digit of the zip code

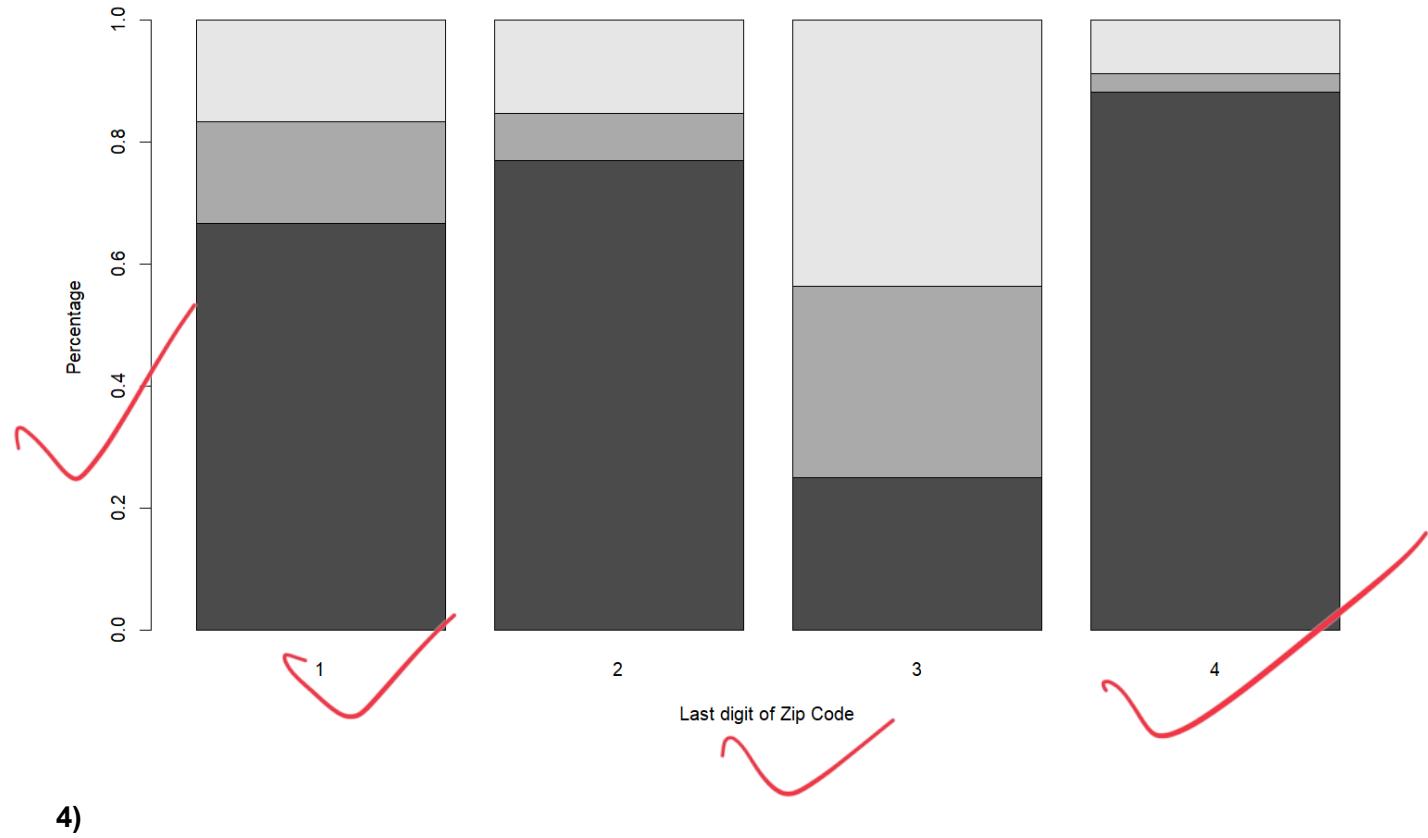
| Exterior | 1          | 2          | 3          | 4          |
|----------|------------|------------|------------|------------|
| Brick    | 0.66666667 | 0.76923077 | 0.25000000 | 0.88235294 |
| Frame    | 0.16666667 | 0.07692308 | 0.31250000 | 0.02941176 |
| other    | 0.16666667 | 0.15384615 | 0.43750000 | 0.08823529 |

3)

-1

use 2 or 3 decimal digits

Bar plot for percentage distribution of Exterior Type by Zip Code Last Digit

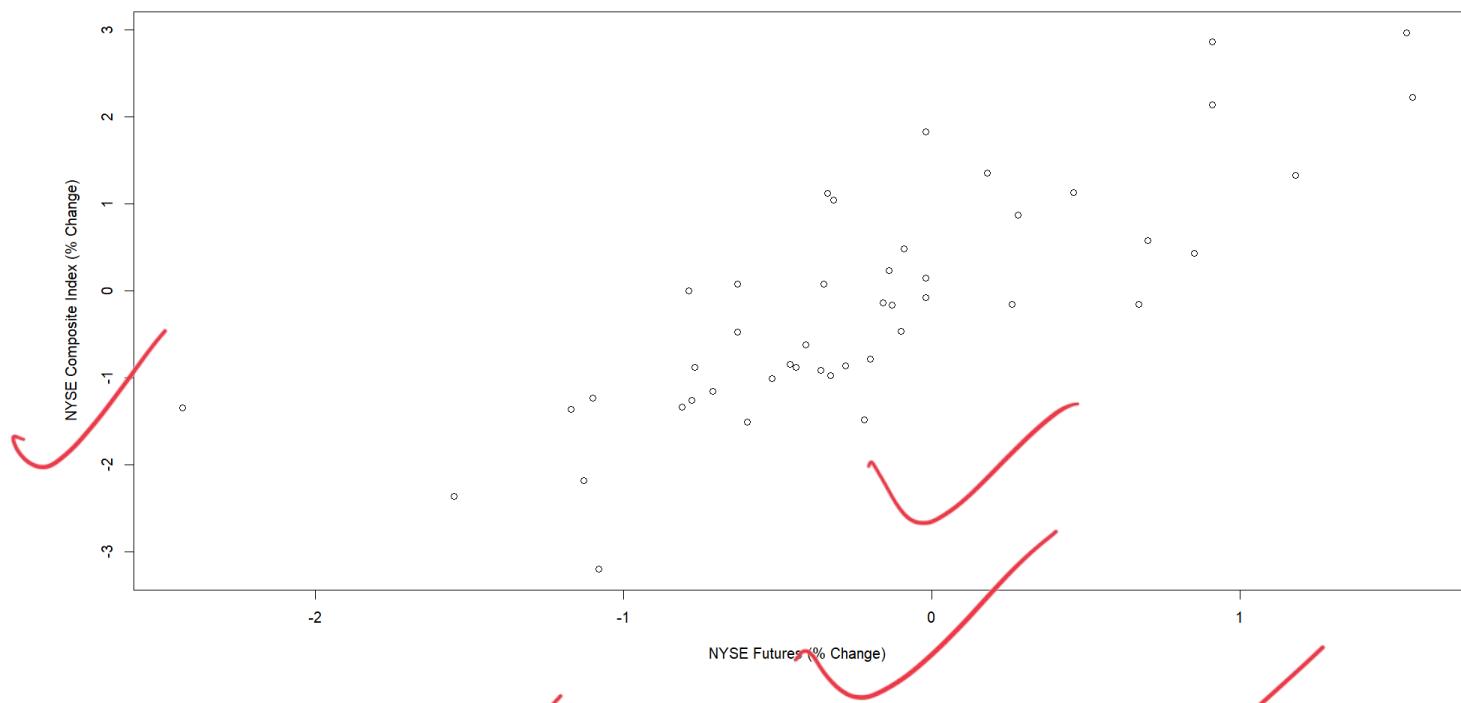


4)

**Features:** In all area zip codes except 3 there are more brick exterior type houses than others. In zip code 1 exterior brick type is high and frame and other are the same percentage. In zip code 2 the exterior brick type is more followed by other and then frame type. In zip code 3 other types are more when compared to brick and frame. In zip 4 exterior brick type is most of it and the other are very less.

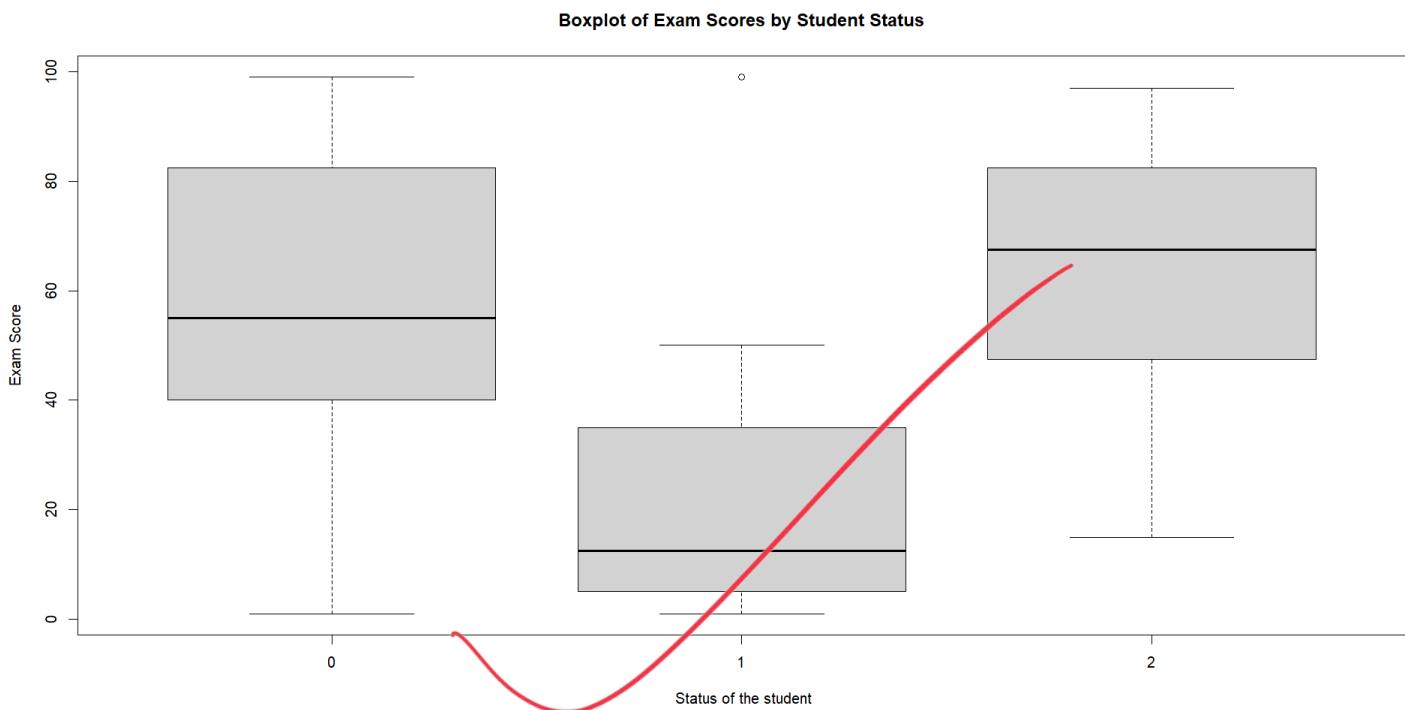
Problem 2

Scatter Plot of NYSE Composite Index vs. NYSE Futures



The scatter plot shows a positive linear trend between NYSE Futures(% Changes) and NYSE Composite Index( % Change). This indicates that percentage change in NYSE Futures increases, the percentage change in the NYSE Composite Index also tends to increase. Most data points lie in the range between -1 and 0.

3)



the status of the student:  
coded as 0 = passed the course  
1 = failed the course

2 = dropped out before the semester was over

## Features:

### Status 0

Median is around close to 60, IQR roughly 40 to 80

students who passed have the highest scores, with their median and upper quartiles being well above the other group 1. High variance when compared to others.

### Status 1

Median around 20, IQR roughly above zero to below 40. One outlier is present.

Students who failed performed poorly, with scores concentrated at the lower end.

### Status 2

Median around above 60, IQR 40 to 80

Students who dropped out have similar scores compared to students who passed. There might be several reasons students drop out. If they continued, there may be a chance that they would have passed given the scores in the test.

Given the test scores if the student has a high score they will pass or might drop out but given the student score is less they might fail.

## R Code:

```
# problem 1
```

```
# 1.1
```

```
texas_house_data <- read.csv("texas-house.csv")
```

```
a <- table(texas_house_data$exter, texas_house_data$zip)
```

```
a
```

```
# 1.2
```

```
b <- prop.table(a, margin = 2)
```

```
b
```

```
# 1.3
```

```
barplot(b, main = "Bar plot for percentage distribution of Exterior Type by Last Digit of Zip Code", xlab = "Last digit of Zip Code", ylab = "Percentage")
```

```
# Problem 2
```

```
nyse_data <- read.csv("nyse.csv")
```

```
plot(nyse_data$future, nyse_data)index,
```

```
main = "Scatter Plot of NYSE Composite Index vs. NYSE Futures",
```

```
xlab = "NYSE Futures (% Change)",
```

```
ylab = "NYSE Composite Index (% Change)")
```

```
# Problem 3
```

```
placement_test_data <- read.csv("exam.csv")
```

```
boxplot(
```

```
placement_test_data$score ~ placement_test_data$status,
```

```
main = "Boxplot of Exam Scores by Student Status",
```

```
xlab = " Status of the student",  
ylab = "Exam Score"
```

```
)
```

**Note:** Show sufficient details when you do the calculation for Problems from 2 and 3, otherwise no credit will be given. Partial credit will be given for the correct parts of your work.

### Problem 1 (6 points, 1.5 points for each part)

This section consists of some true/false questions regarding concepts of statistical inference. Indicate if a statement is true or false and, **if false, indicate what is required to make the statement true.** One and half points will be deducted if you indicate a true statement as false. One and half points will be deducted if you state a false statement as true. Half point will be deducted if you indicate a false statement as false but fail to provide what is required to make the statement to be true.

- (1) If two events are mutually exclusive, then  $\Pr(A \text{ or } B) = \Pr(A) + \Pr(B)$ .
- (2) If  $A$  and  $B$  are two events, then  $\Pr(A \text{ and } B) = \Pr(A) * \Pr(B)$ , no matter what the relation between  $A$  and  $B$ .
- (3) The probability distribution function of a discrete random variable cannot have a value greater than 1.
- (4) The probability distribution function of a continuous random variable can take on any value, even negative ones.

### Problem 2 (15 points, 5 points for each part)

From the weather channel, we find that the chance that there will be a snowstorm at Houghton this Wednesday is 70% while the chance that there will be a snowstorm at Yellowstone Park this Wednesday is 40%. Since the distance between these two cities, Houghton and Yellowstone, is long so we can assume the event that there will be a snowstorm at Houghton and the event that there will be a snowstorm at Yellowstone are independent. Find the probability of the following events.

- (1) There will be a snowstorm at Houghton this Wednesday and there will be a snowstorm at Yellowstone this Wednesday.
- (2) There will be a snowstorm at Houghton this Wednesday or there will be a snowstorm at Yellowstone this Wednesday.
- (3) There will be a snowstorm at Houghton this Wednesday but there will not be a snowstorm at Yellowstone this Wednesday.

**Problem 3 (23 points)**

Toss two defective dices D1 and D2. When you toss a die, you can get an integer number from 1 to 6. Suppose the probability distributions of the number from tossing these dices are given in the following table. We further assume that the number obtained from tossing D1 is independent of the number obtained from tossing D2. The following events are defined:

- A - an odd number obtained from tossing D1
- B – an even number obtained from tossing D2
- C – either 1 or 2 obtained from tossing D2
- D – the sum of two numbers obtained from tossing D1 and D2 is 6.

|                | <b>1</b> | <b>2</b> | <b>3</b> | <b>4</b> | <b>5</b> | <b>6</b> |
|----------------|----------|----------|----------|----------|----------|----------|
| <b>Dice D1</b> | 0.2      | 0.1      | 0.1      | 0.2      | 0.2      | 0.2      |
| <b>Dice D2</b> | 0.1      | 0.1      | 0.2      | 0.3      | 0.2      | 0.1      |

- (1) **(8 points)** Calculate  $\Pr(A)$ ,  $\Pr(B)$ ,  $\Pr(C)$ , and  $\Pr(D)$ .
- (2) **(5 points)** Calculate  $\Pr(A \text{ and } C)$  and  $\Pr(A \text{ or } C)$ .
- (3) **(5 points)** Calculate  $\Pr(A \text{ and } D)$  and  $\Pr(A \text{ or } D)$
- (4) **(5 points)** Let  $Y$  be the number obtained by tossing D1. Find its mean and variance.

### Problem 1

44/44 great!

- (1) True ✓
- (2) False. To make  $\Pr(A \text{ and } B) = \Pr(A) * \Pr(B)$  holds True only if A and B independent events. To make that statement true we need to explicitly mention that A and B are independent. ✓
- (3) True . ✓
- (4) False. The probability distribution function (PDF) of a continuous random variable must be non-negative. To make the statement true, it should state that the PDF can take any non-negative value. ✓

### Problem 2

Probability of a snowstorm at Houghton this Wednesday is 70% ,  $\Pr(A) = 0.7$

Probability of a snowstorm at Yellowstone Park this Wednesday is 40%,  $\Pr(B) = 0.4$

The distance between these two cities, Houghton and Yellowstone, is long so we can assume the event that there will be a snowstorm at Houghton and the event that there will be a snowstorm at Yellowstone are independent meaning A and B are independent events.

- (1) Probability of there will be a snowstorm at Houghton this Wednesday and there will be a snowstorm at Yellowstone this Wednesday

$$\Pr(A \cap B) = \Pr(A) * \Pr(B) \quad A, B \text{ independent events}$$

$$\Pr(A \cap B) = 0.7 * 0.4$$

$$\Pr(A \cap B) = 0.28$$

- (2) Probability of there will be a snowstorm at Houghton this Wednesday or there will be a snowstorm at Yellowstone this Wednesday

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$

$$\Pr(A \cup B) = 0.7 + 0.4 - (0.7 * 0.4)$$

$$\Pr(A \cup B) = 0.82$$

- (3) Probability of there will be a snowstorm at Houghton this Wednesday but there will not be a snowstorm at Yellowstone this Wednesday

$$\Pr(A \cap B^c) = \Pr(A) * (1 - \Pr(B))$$

$$\Pr(A \cap B^c) = 0.7 * (1 - 0.4)$$

$$\Pr(A \cap B^c) = 0.42$$

### Problem 3

- (1)

$\Pr(A)$  - Probability of getting an odd number from tossing D1

$$\Pr(A) = \Pr(D1 = 1) + \Pr(D1 = 3) + \Pr(D1 = 5)$$

$$\Pr(A) = 0.2 + 0.1 + 0.2$$

$$\Pr(A) = 0.5$$

$\Pr(B)$  - Probability of getting an even number from tossing D2

$$\Pr(B) = \Pr(D2 = 2) + \Pr(D2 = 4) + \Pr(D2 = 6)$$

$$\Pr(B) = 0.1 + 0.3 + 0.1$$

$$\Pr(B) = 0.5$$

$\Pr(C)$  - Probability of getting 1 or 2 from tossing D2

$$\Pr(C) = \Pr(D2 = 1) + \Pr(D2 = 2)$$

$$\Pr(C) = 0.1 + 0.1$$

$$\Pr(C) = 0.2$$

$\Pr(D)$  - Probability of getting the sum of two numbers from tossing D1 and D2 equal to 6

Pairs that equal to 6

$$(1, 5) : \Pr(D1 = 1) * \Pr(D2 = 5) = 0.2 * 0.2 = 0.04$$

$$(2, 4) : \Pr(D1 = 2) * \Pr(D2 = 4) = 0.1 * 0.3 = 0.03$$

$$(3, 3) : \Pr(D1 = 3) * \Pr(D2 = 3) = 0.1 * 0.2 = 0.02$$

$$(4, 2) : \Pr(D1 = 4) * \Pr(D2 = 2) = 0.2 * 0.1 = 0.02$$

$$(5, 1) : \Pr(D1 = 5) * \Pr(D2 = 1) = 0.2 * 0.1 = 0.02$$

$$\Pr(D) = 0.04 + 0.03 + 0.02 + 0.02 + 0.02$$

$$\Pr(D) = 0.13$$

(2)

$\Pr(A \cap C)$ , since A and C are independent events

$$\Pr(A \cap C) = \Pr(A) * \Pr(C)$$

$$\Pr(A \cap C) = 0.5 * 0.2$$

$$\Pr(A \cap C) = 0.1$$

$$\Pr(A \cup C) = \Pr(A) + \Pr(C) - \Pr(A \cap C)$$

$$\Pr(A \cup C) = 0.5 + 0.2 - 0.1$$

$$\Pr(A \cup C) = 0.6$$

(3)

$\Pr(A \cap D)$ , Probability of D1 being odd and sum is 6. Valid Pairs (1, 5), (3, 3), (5, 1)

$$\Pr(A \cap D) = 0.2 * 0.2 + 0.1 * 0.2 + 0.2 * 0.1$$

$$\Pr(A \cap D) = 0.04 + 0.02 + 0.02$$

$$\Pr(A \cap D) = 0.08$$

$$\Pr(A \cup D) = \Pr(A) + \Pr(D) - \Pr(A \cap D)$$

$$\Pr(A \cup D) = 0.5 + 0.13 - 0.08$$

$$\Pr(A \cup D) = 0.55$$

(4)

Y be the number obtained by tossing D1.

Mean - Expected Value of Y

$$E[Y] = \sum_{i=1}^6 x_i * p(Y = x_i)$$

$$E[Y] = (1)(0.2) + (2)(0.1) + (3)(0.1) + (4)(0.2) + (5)(0.2) + (6)(0.2)$$

$$E[Y] = 0.2 + 0.2 + 0.3 + 0.8 + 1 + 1.2$$

$$E[Y] = 3.7$$

Variance of Y

$$\text{Var}(Y) = \sum_{i=1}^6 (x_i - u)^2 * p(Y = x_i) \text{ i.e } E[Y^2] - (E[Y])^2$$

$$\text{Var}(Y) = (1 - 3.7)^2 * (0.2) + (2 - 3.7)^2 * (0.1) + (3 - 3.7)^2 * (0.1) + (4 - 3.7)^2 * (0.2) + (5 - 3.7)^2 * (0.2) + (6 - 3.7)^2 * (0.2)$$

$$\text{Var}(Y) = 3.21$$

**Problem 1 (15 points, 5 points for each part)**

Let  $Y$  be a discrete random variable that represents the number of defected products from a factory and has the following probability mass function:

|        |      |      |      |      |
|--------|------|------|------|------|
| $y$    | 0    | 1    | 2    | 3    |
| $f(y)$ | 0.94 | 0.03 | 0.02 | 0.01 |

Calculate (1)  $\Pr(Y > 1)$ ; (2) the mean of  $Y$ ; and (3) the variance of  $Y$ .

**Problem 2 (12 points, 6 points for each part)**

Suppose that the probability to get the head from tossing a coin is  $p$ . If we toss the coin two times and assume the outcome from the first toss is independent of the outcome from the second toss.

- (1) Define the random  $X$  as the number of **tails**. Find the probability mass function, the expected value, and the variance of  $X$ .
- (2) Define the random  $Y$  as the following:  $Y = 1$  if the outcome from the second toss is same as the outcome from the first toss and  $Y = 0$  otherwise. Find the probability mass function, the expected value, and the variance of  $Y$ .

**Problem 3 (10 points)**

Based on data from the 2007 National Health Interview Survey, it is estimated that “10% of adults experienced feelings of sadness for all, most, or some of the time” during the 30 days prior to the interview. You interviewed a random sample of 68 people who have recently filed for unemployment benefits in your county and asked this same question in your survey.

- (1) **(2 points)** Identify the implied target population for your study.
- (2) **(2 points)** Let  $Y$  be the number of people with these feelings among your sample. We can assume that the proportion of your population with these feelings (feelings of sadness for all, most, or some of the time) is the same as the 10% nationally. What is the distribution of  $Y$ ?
- (3) **(3 points)** Using R find the probability that your sample will have 12 or more people with these feelings.
- (4) **(3 points)** Using R find the probability that your sample will have at most 8 people with these feelings.

## Problem 1

(1)  $\Pr(Y > 1)$

$Y > 1$  meaning  $Y$  can take 2 and 3 values.

$$\Pr(Y > 1) = \Pr(Y = 2) + \Pr(Y = 3)$$

$$\Pr(Y > 1) = 0.02 + 0.01 = 0.03$$

37/37

great!

(2) The mean of  $Y$

Mean of  $Y = E[Y]$

$$E[Y] = \sum_{i=0}^3 x_i * p(Y = x_i)$$

$$E[Y] = 0 * 0.94 + 1 * 0.03 + 2 * 0.02 + 3 * 0.01$$

$$E[Y] = 0.10$$

(3) The variance of  $Y$

$$\text{Var}(Y) = \sum_{i=0}^3 (x_i - u)^2 * p(Y = x_i) \text{ i.e. } E[Y^2] - (E[Y])^2$$

$$\text{Var}(Y) = (0 - 0.1)^2 * 0.94 + (1 - 0.1)^2 * 0.03 + (2 - 0.1)^2 * 0.02 + (3 - 0.1)^2 * 0.01$$

$$\text{Var}(Y) = 0.19$$

## Problem 2

(1) Define the random variable  $X$  as the number of tails.

We toss a coin two times. Let the probability of getting a head be  $p$  (so the probability of a tail is  $1-p$ ), and assume the tosses are independent.

The sample space for two tosses is  $\{\text{HH}, \text{HT}, \text{TH}, \text{TT}\}$ . We count tails as follows:

- For HH: 0 tails      Probability =  $p \times p = p^2$ .
- For HT: 1 tail      Probability =  $p \times (1-p)$ .
- For TH: 1 tail      Probability =  $(1-p) \times p$ .
- For TT: 2 tails      Probability =  $(1-p) \times (1-p) = (1-p)^2$ .

The probability mass function (pmf) for  $X$  is:

$$\Pr(X = 0) = p^2,$$

$$\Pr(X = 1) = p(1-p) + (1-p)p = 2p(1-p),$$

$$\Pr(X = 2) = (1-p)^2.$$

Since  $X$  is the count of tails in two independent tosses, it follows a binomial distribution with parameters  $n = 2$  and success probability  $1-p$ .

The expected value is  $E(X) = 2(1-p)$ ,

The variance is  $\text{Var}(X) = 2(1-p) \cdot p$ .

(2)

Define the random variable  $Y$  such that

$Y = 1$  if the outcome of the second toss is the same as the outcome of the first toss, and

$Y = 0$  otherwise.

The outcomes:

- HH: both tosses are the same, so  $Y = 1$ ; probability =  $p \times p = p^2$ .
- TT: both tosses are the same, so  $Y = 1$ ; probability =  $(1-p) \times (1-p) = (1-p)^2$ .
- HT or TH: tosses differ, so  $Y = 0$ ; combined probability =  $p(1-p) + (1-p)p = 2p(1-p)$ .

Thus, the pmf of  $Y$  is:

$$\Pr(Y = 1) = p^2 + (1-p)^2,$$
$$\Pr(Y = 0) = 2p(1-p).$$

The expected value of  $Y$  is

$$E(Y) = 1 * [p^2 + (1-p)^2] + 0 * [2p(1-p)] = p^2 + (1-p)^2.$$

Since  $Y$  is a binary random variable (taking values 0 and 1), its variance is given by

$$\text{Var}(Y) = \sum (x_i - u)^2 * p(Y = x_i) \text{ i.e } E[Y^2] - (E[Y])^2$$

$$\text{Var}(Y) = [p^2 + (1-p)^2] \times [1 - (p^2 + (1-p)^2)].$$

### Problem 3

(1)

Target population: All adults who recently filed for unemployment benefits in the county

(2)

Let  $Y$  be the number of people in the sample who experience these feelings of sadness. Assuming that the proportion of adults with these feelings in the target population is the same as the national rate (10%), and that each person's response is independent,  $Y$  follows a binomial distribution with parameters  $n = 68$  and  $p = 0.10$ .

$$Y \sim \text{Binomial}(68, 0.10).$$

(3)

$$\Pr(Y \geq 12) = 1 - \text{pbinom}(11, \text{size} = 68, \text{prob} = 0.10)$$

$$\Pr(Y \geq 12) = 0.0362$$

(4)

$$\Pr(Y \leq 8) = \text{pbinom}(8, \text{size} = 68, \text{prob} = 0.10)$$

$$\Pr(Y \leq 8) = 0.763$$

# R Code

```
1 - pbinom(11, size = 68, prob = 0.10)
pbinom(8, size = 68, prob = 0.10)
```

**Problem 1 (10 points, 1 point for each part)**

Simply indicate it as true if the statement is always true and state it as false otherwise. You do not need to explain why the statement is true or false.

- (1) The probability that a continuous random variable lies in the interval 4 to 7, inclusively, is the sum of  $\Pr(4) + \Pr(5) + \Pr(6) + \Pr(7)$ . Here  $\Pr(4)$  is the probability that the random variable equals to 4.
- (2) The variance of the number of successes in a binomial experiment of  $n$  trials is  $\sigma^2 = np(1 - p)$ .
- (3) A normal distribution is characterized by its mean and its degrees of freedom.
- (4) The standard normal distribution has mean zero and variance  $\sigma^2$ .
- (5) The standard deviation of the sample mean increases as the sample size increases.
- (6) As  $\alpha$  increases, the value of  $z_\alpha$  will decrease.

**Problem 2 (4 points, 1 point for each part)**

- (1) Use a normal distribution table to find  $\Pr(Z > 0.374)$  where  $Z \sim N(0,1)$ .
- (2) Use R to find  $\Pr(Y > 0.374)$  where  $Z \sim N(0,1)$ .
- (3) Use a normal distribution table to find  $z_{0.12}$ .
- (4) Use R to find  $z_{0.12}$ .

**Problem 3 (25 points, 5 points for each part)**

Suppose that  $Y$  is normally distributed random variable with  $\mu = 10$  and  $\sigma = 2$  and  $X$  is also normally distributed with  $\mu = 5$  and  $\sigma = 5$ .  $X$  and  $Y$  are independent. Calculate the following probabilities according to a **normal distribution table**. You need to show sufficient details. For example, if you want to calculate  $\Pr(Z < 0.10)$ , you cannot directly write out  $\Pr(Z < 0.10) = 0.5398$ . You should use  $\Pr(Z < 0.10) = 1 - \Pr(Z > 0.10) = 1 - 0.4602 = 0.5398$ . These steps show that the normal table is indeed used.

- (1)  $\Pr(Y > 12)$  and  $\Pr(3 < X < 6)$ .
- (2)  $\Pr(Y > 12 \text{ and } 3 < X < 6)$ . [Hint:  $X$  and  $Y$  are independent]
- (3)  $\Pr(Y > 12 \text{ or } 3 < X < 6)$  (You can use the results from (1) and (2)).
- (4) The value of  $C$  such that  $\Pr(Y < C) = 0.94$ .
- (5) The value of  $D$  such that  $\Pr(X > D) = 0.40$ .

**Problem 4 (10 points)**

The average modulus of rupture (MOR) for a particular grade of pencil lead is known to be 6500 psi with a standard deviation of 250 psi. Again, you need to use calculation should be based on a normal table.

- (1) **(8 points)** Find the probability that a random sample 16 pencil leads will have an average MOR between 6400 and 6550 psi.
- (2) **(2 points)** What did you assume to in order to find this probability?

**Problem 5 (15 points, 5 points for each part)**

The Kaufman Assessment Battery for Children is designed to measure achievement and intelligence with a special emphasis on nonverbal intelligence. Its global measures, such as its Sequential Processing score, are scaled to have a mean of 100 and a standard deviation of 15. Assume that the Sequential Processing score has a normal distribution. You can use a normal table or R for your calculations.

- (1) Find a value that divides the children with the highest 10% of the scores from those with the lower 90%.
- (2) What proportion of children will have Sequential Processing scores between 90 and 110?
- (3) In a sample of 20 children, what is the probability the sample mean will differ from the population mean by more than 3 points (either positive or negative)?

### Problem 1

- (1) False ✓
- (2) True ✓
- (3) False ✓
- (4) False ✓
- (5) False ✓
- (6) True ✓

59/60

### Problem 2

(1)  $\Pr(Z > 0.374) = 1 - \Pr(Z \leq 0.374) = 0.354$

need 0.37  
or 0.38

(2)  $\text{pnorm}(0.374, \text{lower.tail} = \text{FALSE}) = 0.354$

(3)  $z_{0.12} = 1.175$

→ how did you find it from table?

(4)  $\text{qnorm}(1 - 0.12) = 1.175$

### Problem 3

(1)

$$\Pr(Y > 12) = \Pr(Z > (12 - 10) / 2) = \Pr(Z > 1) = 0.1587$$

$$Z_1 = (3 - 5) / 5 = -0.4, Z_2 = (6 - 5) / 5 = 0.2$$

$$\Pr(3 < X < 6) = \Pr(Z < 0.2) - \Pr(Z < -0.4) = 0.5793 - 0.3446 = 0.2347$$

(2)

Since X and Y are independent

$$\Pr(Y > 12 \text{ and } 3 < X < 6) = 0.1587 \times 0.2347 = 0.0372.$$

(3)

$$\Pr(Y > 12 \text{ or } 3 < X < 6) = \Pr(Y > 12) + \Pr(3 < X < 6) - \Pr(Y > 12 \text{ and } 3 < X < 6)$$

$$\Pr(Y > 12 \text{ or } 3 < X < 6) = 0.1587 + 0.2347 - 0.0372$$

$$\Pr(Y > 12 \text{ or } 3 < X < 6) = 0.3562$$

(4)

$$\Pr(Y < C) = 0.94$$

The z-score for 0.94 is approximately 1.5548

$$C = 10 + 1.5548 \times 2$$

$$C = 13.11$$



(5)

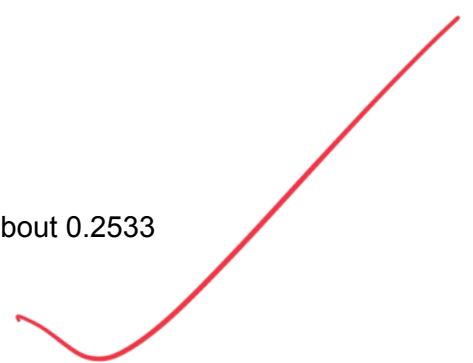
$$\Pr(X > D) = 0.40$$

$$\Pr(X \leq D) = 0.60$$

The z-score for 0.60 is about 0.2533

$$D = 5 + 0.2533 \times 5$$

$$D = 6.27$$



#### Problem 4

(1)

Probability that the sample mean is between 6400 and 6550 psi:

The sampling distribution of the mean has

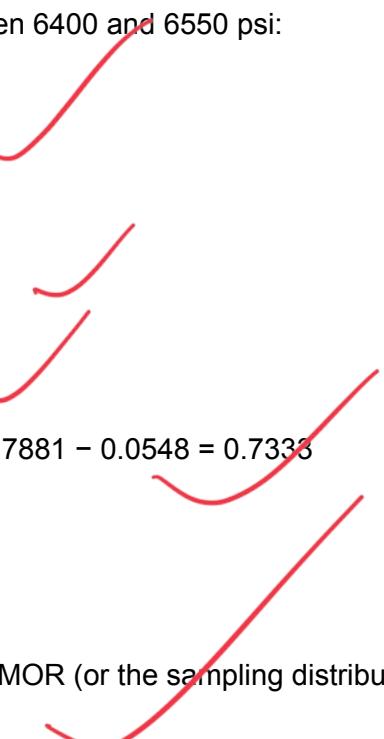
$$SE = 250/\sqrt{16} = 62.5 \text{ psi}$$

Standardize the endpoints:

$$\text{For } 6400: Z = (6400 - 6500) / 62.5 = -1.6$$

$$\text{For } 6550: Z = (6550 - 6500) / 62.5 = 0.8$$

$$\Pr(6400 < X_{\text{mean}} < 6550) = z_{0.8} - z_{-1.6} = 0.7881 - 0.0548 = 0.7333$$



(2)

We assume that the underlying distribution of the MOR (or the sampling distribution of the mean) is normal (or that the Central Limit Theorem applies).

## Problem 5

(1)

We need the 90th percentile since 90% are below this value:

$$\text{Cutoff} = 100 + 15 \times \text{qnorm}(0.90)$$

$$\text{Cutoff} = 100 + 15 \times 1.2816$$

$$\text{Cutoff} = 119.22$$



(2)

$$\Pr(90 < X < 110) = \Pr(z < 0.6667) - \Pr(z < -0.6667) = 0.495$$

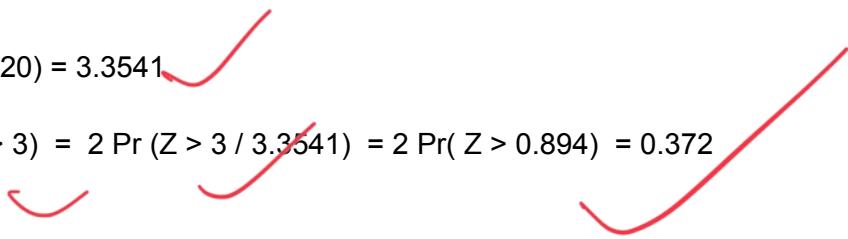
(3)

Probability that in a sample of 20 children the sample mean differs from 100 by more than 3 points:

The sampling distribution of the mean has standard error:

$$SE = 15 / \sqrt{20} = 3.3541$$

$$\Pr(|X_{\text{mean}} - 100| > 3) = 2 \Pr(Z > 3 / 3.3541) = 2 \Pr(Z > 0.894) = 0.372$$



**Problem 1 (20 points, 4 points for each part)**

Consider the filling operation for 20-oz bottles of a popular soft drink. Assume the volume of drink filled in the bottle follows a normal distribution with the mean of 20.2 oz and the standard deviation of 0.40. A recent random sample of 12 bottles yielded these volumes:

|      |      |      |      |      |      |
|------|------|------|------|------|------|
| 20.1 | 20.1 | 20.0 | 19.9 | 20.5 | 20.9 |
| 20.1 | 20.4 | 20.2 | 19.1 | 20.1 | 20.0 |

- (1) Use R to find the sample mean, sample variance, and sample standard deviation of this data. You can use the following R code to create the data:

```
drink <- c(20.1, 20.1, 20.0, 19.9, 20.5, 20.9, 20.1, 20.4, 20.2, 19.1, 20.1, 20.0)
```

- (2) Draw the Q-Q plot for this data and comment on its normality.
- (3) Find the probability that the volume of drink filled in a random selected bottle is less than 20.1 oz. Please provide sufficient details and use the normal table for your calculation.
- (4) For a sample of 12 bottles, find the probability that the sample mean is less than 20.1 oz. Please provide sufficient details and use the normal table for your calculation.
- (5) Use R function pnorm() to calculate the probabilities in (3) and (4).

### Problem 1

(1)

sample mean - 20.1167

sample variance - 0.1779

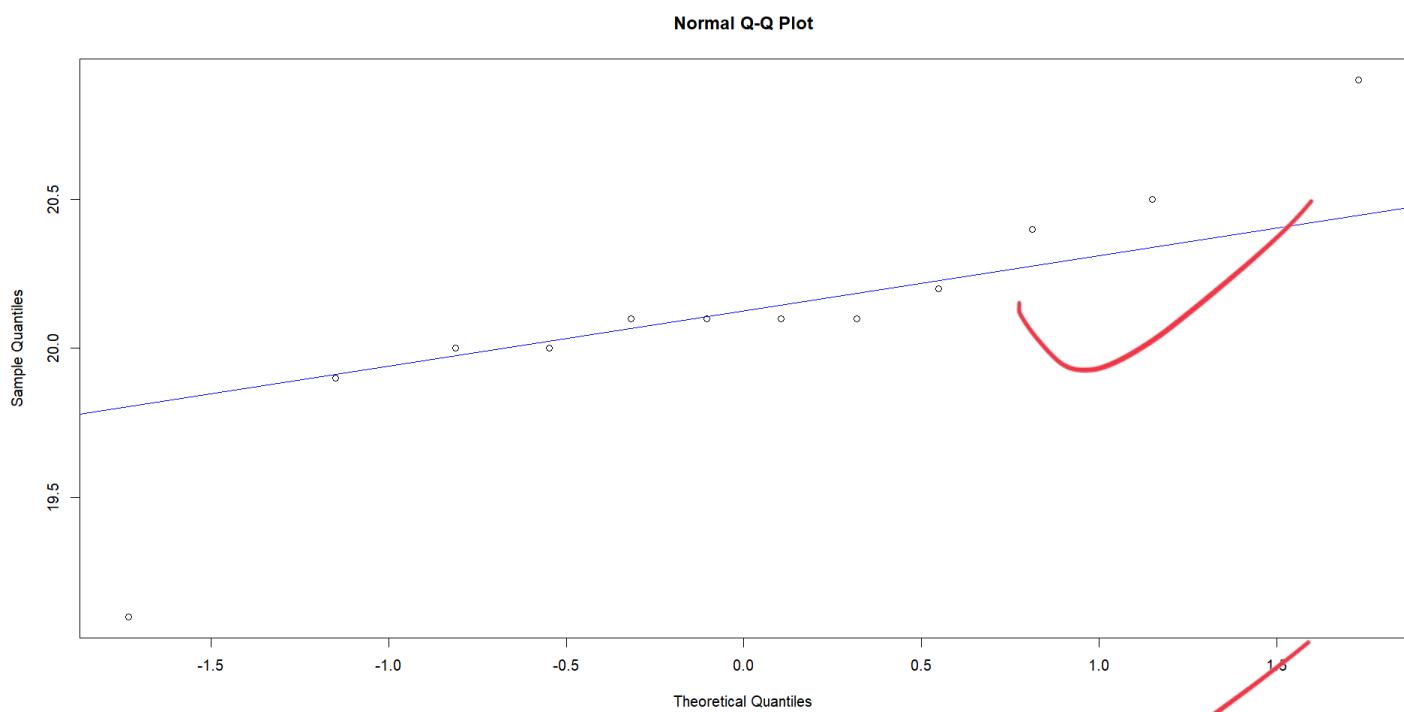
sample standard deviation - 0.4218

use 20.12

19/20

(2)

Q-Q plot



The majority of the points align closely with the reference line, suggesting that the central portion of the data follows a normal distribution. However, there is a significant deviation at the lower end.

(3)

The population of bottle fills is normally distributed with mean  $\mu=20.2$  oz and standard deviation  $\sigma=0.40$  oz. We need to calculate the  $\Pr(X < 20.1)$  where  $X \sim N(20.2, 0.4^2)$

$$Z\text{- score} = Z = (20.1 - 20.2) / 0.4 = (-0.1) / 0.4 = -0.25$$

$$\Pr(Z < -0.25) = 0.4013$$

$$\Pr(Z < -0.25)$$

$$= \Pr(Z > 0.25) = \dots$$

We need to calculate the  $\Pr(\bar{X} < 20.1)$

$$Z\text{- score} = (20.1 - 20.2) / (0.4 / \sqrt{12}) = -0.1 / 0.11547 = -0.866$$

From the standard normal table,

$$\Pr(Z < -0.866) = 0.1932$$

(5)

Probability for a single bottle volume  
pnorm(q = 20.1, mean = 20.2, sd = 0.4)  
0.4013

Probability for the sample mean of 12 bottles

# Standard error = 0.4 / sqrt(12)  
pnorm(q = 20.1, mean = 20.2, sd = 0.4 / sqrt(12))  
0.1932

use either

$$\Pr(Z < -0.86)$$

$$\text{or } \Pr(Z < -0.87)$$

### # R code

```
# (1)
```

```
drink <- c(20.1, 20.1, 20.0, 19.9, 20.5, 20.9, 20.1, 20.4, 20.2, 19.1, 20.1, 20.0)
mean(drink)
var(drink)
std(drink)
```

```
#(2)
```

```
qqnorm(drink, main = "Q-Q Plot for Drink Volume")
qqline(drink, col = "blue")
```

**Problem 1 (8 points, 1 point for each part)**

This problem consists of some true/false questions regarding concepts of statistical inference. Indicate if a statement is true or false. You do not need to explain why it is true or false.

- 1.1 In a hypothesis test, the  $p$  value is 0.043. This means that the null hypothesis would be rejected at  $\alpha = 0.05$ .
- 1.2 If the null hypothesis is rejected by a one-tailed hypothesis test at  $\alpha$ , then it will also be rejected by a two-tailed test at  $\alpha$ .
- 1.3 If a null hypothesis is rejected at the 0.01 level of significance, it will also be rejected at the 0.05 level of significance.
- 1.4 The probability of a type II error is controlled in a hypothesis test by establishing a specific significance level.
- 1.5 If we decrease the confidence coefficient for a fixed  $n$ , we decrease the width of the confidence interval.
- 1.6 If a 95% confidence interval on  $\mu$  was from 50.5 to 60.6, we would reject the null hypothesis that  $\mu = 60$  at the 0.05 level of significance.
- 1.7 If the sample size is increased and the level of confidence is decreased, the width of the confidence interval will increase.
- 1.8 A research article reports that a 95% confidence interval for mean reaction time is from 0.25 to 0.29 seconds. Therefore, about 95% of individuals will have reaction times in this interval.

**Problem 2 (9 points, 1 point for each part)**

Multiple choice questions. Only one of the statements is correct.

- 2.1 In a hypothesis test the  $p$  value is 0.043. This means that we can find statistical significance at:
  - a) both the 0.05 and 0.01 levels
  - b) the 0.05 but not at the 0.01 level
  - c) the 0.01 but not at the 0.05 level
  - d) neither the 0.05 or 0.01 levels
  - e) none of the above

2.2 A research report states: The differences between public and private school seventh graders' attitudes toward minority groups were statistically significant at the  $\alpha = 0.05$  level. This means that:

- a) It has been proven that the two groups are different.
- b) There is a probability of 0.05 that the attitudes of the two groups are different.
- c) There is a probability of 0.95 that the attitudes of the two groups are different.
- d) If there is no difference between the groups, the difference observed in the sample would occur by chance with a probability of no more than 0.05.
- e) None of the above is correct.

2.3 If the null hypothesis is really false, which of these statements characterizes a situation where the value of the test statistic falls in the rejection region?

- a) The decision is correct.
- b) A type I error has been committed.
- c) A type II error has been committed.
- d) Insufficient information has been given to make a decision.
- e) None of the above is correct.

2.4 If the null hypothesis is really false, which of these statements characterizes a situation where the value of the test statistic does not fall in the rejection region?

- a) The decision is correct.
- b) A type I error has been committed.
- c) A type II error has been committed.
- d) Insufficient information has been given to make a decision.
- e) None of the above is correct.

2.5 If the value of any test statistic does not fall in the rejection region, the decision is:

- a) Reject the null hypothesis.
- b) Reject the alternative hypothesis.
- c) Fail to reject the null hypothesis.
- d) Fail to reject the alternative hypothesis.
- e) There is insufficient information to make a decision.

2.6 For a particular sample, the 95% two-sided confidence interval for the population mean is from 11 to 17. You are asked to test the hypothesis that the population mean is 18 against a two-sided alternative. Your decision is:

- a) Fail to reject the null hypothesis,  $\alpha = 0.05$ .
- b) Reject the null hypothesis,  $\alpha = 0.05$ .
- c) There is insufficient information to decide.

2.7 If we decrease the confidence level, the width of the confidence interval will:

- a) Increase
- b) remain unchanged
- c) decrease
- d) double
- e) none of the above

2.8 If the value of the test statistic falls in the rejection region, then:

- a) We cannot commit a type I error.
- b) We cannot commit a type II error.
- c) We have proven that the null hypothesis is true.
- d) We have proven that the null hypothesis is false.
- e) None of the above is correct.

2.9 You are reading a research article that states that there is no significant evidence that the median income in the two groups differs, at  $\alpha = 0.05$ . You are interested in this conclusion, but prefer to use  $\alpha = 0.01$ .

- a) You would also say there is no significant evidence that the medians differ.
- b) You would say there is significant evidence that the medians differ.
- c) You do not know whether there is significant evidence or not, until you know the  $p$  value.

### Problem 3 (10 points, 5 points for each part)

Suppose that for a given population with  $\sigma = 7.2$  we want to test  $H_0: \mu = 80$  against  $H_1: \mu < 80$  based on a sample of  $n = 100$ . For this problem, you can use R to find the probabilities associated with the standard normal and  $z_\alpha$  but you need to present sufficient details about your calculations.

- (1) If the null hypothesis is rejected when  $\bar{y} < 76$ , what is the probability of a type I error?
- (2) What would be the rejection region if we wanted to have a level of significance of exactly 0.05?

**Problem 4 (25 points)**

The family incomes in a certain city in 1970 had a mean of \$14,200 with a standard deviation of \$2600. A random sample of 75 families taken in 1975 produced  $\bar{y} = \$15,300$  (adjusted for inflation). For this problem, you can use R to find the probabilities associated with the standard normal and  $z_\alpha$  but you still need to present sufficient details about your calculations.

- (1) **(11 points)** Assume the standard deviation has remained unchanged. Use the rejection region method to test if the mean family income has changed at a 0.05 level of significance. Please clearly specify 5 steps used in the test.
- (2) **(6 points)** Calculate the power of the test from (1) for the mean income of \$15300 and \$15600.
- (3) **(3 points)** Find the  $p$ -value associated with this test.
- (4) **(5 points)** Construct a 99% two-sided confidence interval on the mean family income in 1975. Based on this confidence interval, state if you would like to reject the null hypothesis from part (1) with a significance level of 0.01.

### Problem 1

- 1.1 True ✓
- 1.2 False ✓
- 1.3 True ✓
- 1.4 False ✓
- 1.5 True ✓
- 1.6 False ✓
- 1.7 False ✓
- 1.8 False ✓

51/52

### Problem 2

- 2.1 b ✓
- 2.2 d ✓
- 2.3 a ✓
- 2.4 c ✓
- 2.5 c
- 2.6 b ✓
- 2.7 c ✓
- 2.8 e ✗
- 2.9 a ✓



### Problem 3

1

we have  $H_0: \mu = 80$  vs  $H_1: \mu < 80$ ,  $n = 100$ ,  $\sigma = 7.2$

A type I error is rejecting the null hypothesis ( $H_0$ ) when the null hypothesis ( $H_0$ ) is true.  
Under  $H_0$ ,  $\bar{Y}$  has mean 80 and standard deviation

$$SE = \sigma / \sqrt{n} = 7.2 / 10 = 0.72$$

$$P(\text{Type I error}) = P(Y_{\bar{}} < 76 | \mu = 80)$$
$$P(Y_{\bar{}} < 76) = P(Z < (76 - 80) / 0.72)$$

$$z = (76 - 80) / 0.72 = -5.556$$

$$\text{pnorm}(-5.5556) = 1.383299e-08$$

The probability of Type I error is  $1.383299e-08$

2

Rejection region for  $\alpha = 0.05$

For one-tailed test at level  $\alpha = 0.05$ , we want the cutoff  $c$  such that

$$P(Y_{\bar{}} < c | \mu = 80) = 0.05$$

$$P(z < (c - 80) / 0.72) = 0.05$$

$$z_{\alpha} = \text{qnorm}(0.05) = -1.645$$

$$(c - 80) / 0.72 = -1.645$$

$$c = -1.645 * 0.72 + 80$$

$$c = 78.8156$$

So the rejection region (for a 5% left-tailed test) is  $y_{\bar{}} < 78.8156$

#### Problem 4

1

$$\mu_0 = \$14,200, \sigma = \$2,600,$$

A random sample of  $n = 75$ ,  $y_{\bar{}} = \$15,300$ , assume still  $\sigma = \$2,600$

We want to test  $H_0 : \mu = 14200$  vs  $H_1 : \mu \neq 14200$  at  $\alpha = 0.05$

**Step 1:** Hypothesis testing

$$H_0 : \mu = 14200 \text{ vs } H_1 : \mu \neq 14200$$

**Step 2:** Significance level:

$$\alpha = 0.05$$

this is step 1

two-sided test,  $\alpha / 2 = 0.025$

### Step 3: test statistic

Under  $H_0$ , the test statistic is

$$z = (\bar{y} - \mu_0) / (\sigma / \sqrt{n})$$

$$z = (15300 - 14200) / (2600/\sqrt{75})$$

$$z = 3.6639$$

### Step 4: Rejection Region

It is two-sided at  $\alpha = 0.05$ , we reject if  $|z| > z_{0.025}$

$$z_{0.025} = 1.96$$

The rejection region is  $z < -1.96$  or  $z > 1.96$

### Step 5: Conclusion

$z = 3.6639$  greater than 1.96. Therefore we reject the null hypothesis ( $H_0$ ).

There is significant evidence at the 5% level that the mean family income in 1975 differs from \$14,200.

2

Power of the test  $\mu = 15300$  and  $\mu = 15600$

$$\text{power} = 1 - \beta(\mu) = P(\text{Reject } H_0 \mid \mu \text{ is the true mean})$$

Reject  $H_0$  if

$$z = (\bar{y} - 14200) / (\sigma / \sqrt{n}) > 1.96 \text{ or } z < -1.96$$

the rejection region in terms of  $\bar{y}$

$$\bar{y} < 14200 - 1.96 * (\sigma / \sqrt{n}) \text{ or } \bar{y} > 14200 + 1.96 * (\sigma / \sqrt{n})$$

$$\text{SE} = (\sigma / \sqrt{n}) = 2600 / \sqrt{75} = 300.2221$$

Cutoffs are

$$L = 14200 - 1.96 * 300.2221 = 13611.56$$

$$R = 14200 + 1.96 * 300.2221 = 14788.44$$

We reject the null hypothesis if  $y_{\bar{}} < 13611.56$  or  $y_{\bar{}} > 14788.44$

Power for  $\mu = 15300$

Under true mean  $\mu = 15300$ ,

$$P(\text{reject}) = P(y_{\bar{}} < 13611.56 \text{ or } y_{\bar{}} > 14788.44 | \mu = 15300)$$

$$P(y_{\bar{}} < 13611.56) = 0$$

$$P(y_{\bar{}} > 14788.44) = P(Z > (14788.44 - 15300) / 300)$$

$$P(Z > -1.7052) = 0.9560$$

The power is about 0.9560

Power for  $\mu = 15600$

$$P(y_{\bar{}} > 14788 | \mu = 15600)$$

$$P(z > (14788.44 - 15600) / 300)$$

$$P(z > -2.7052) = 0.9967$$

power is about = 0.9966

**3**

p - value

Test statistic  $z = 3.6639$ . Two sided test

$$p = 2 * P(Z > 3.6639)$$

$$p = 0.00024$$

**4**

99% confidence Interval

$$CI = 15300 \pm (2.576 * 300.2221)$$

$$CI = 15300 + 773.3721, 15300 - 773.3721$$

$$CI = [14526.63, 16073.37]$$

Since 14200 is outside the 99% interval, we reject the null hypothesis  $H_0 : \mu = 14200$  at  $\alpha = 0.01$

