STA501 E-Pack: Applied Statistical Methods

Cheng Peng

West Chester University

Contents

1 Introduction **7**

2 R, RStudio and RMarkdown **9**
2.1 R 9
2.2 RStudio 12
2.3 RMarkdown 14

3 Data Collection and Data Loading **17**
3.1 Sampling Plans 17
3.2 Study Designs 18
3.3 Loading Data to R from External Data Files 19
3.4 Data Frame and List 20

4 Descriptive Statistics **23**
4.1 Data Types 23
4.2 Tabular and Graphic Summary 24
4.3 Numerical Summary of Numerical Data 31
4.4 Assignment - Descriptive Statistics 36

5 Concepts of Probability Distributions **39**
5.1 Concepts of random variables 39
5.2 Types of random variables 40
5.3 Concepts of probability distributions 41
5.4 Probability distribution of random variables 43
5.5 Special Continuous Distributions 46
5.6 Summary 50
5.7 Assignment - Probability Distributions 53

6 Sampling Distributions **55**
6.1 Concepts of the sampling distribution 56
6.2 Sampling Distribution of Sampling Means 56
6.3 Sampling distribution of sample proportions 64
6.4 Summary 66
6.5 Assignment - Sampling Distributions 67

7 Confidence Intervals **69**
7.1 Some Terms . . . 69
7.2 How to find the Confidence Intervals for Means and Proportions? 71
7.3 Confidence Interval of Population Mean . . . 74
7.4 Confidence Interval of Proportion . . . 77
7.5 Two Sample Problems - Comparing Two Population Means . . . 78
7.6 Assignment - Confidence Intervals . . . 81

8 Testing Statistical Hypothesis **85**
8.1 Formulation of Statistical Hypothesis Testing . . . 86
8.2 Case Studies . . . 89
8.3 Unspecified population with an unknown variance - CLT . . . 93
8.4 Testing Population Proportion . . . 94
8.5 Testing the difference between two population means . . . 96
8.6 Practice Problems . . . 102

9 Analysis of Variance (ANOVA) **105**
9.1 The Question of One-way ANOVA . . . 105
9.2 Steps of ANOVA . . . 107
9.3 Welch's ANOVA . . . 109
9.4 Case Study: Mussel Length Example - Solution . . . 109
9.5 Practice Problems . . . 115

10 Correlation and Simple Linear Regression **117**
10.1 The question and the data . . . 117
10.2 Visual Inspection for Association . . . 118
10.3 Coefficient of Correlation . . . 119
10.4 Least square regression: structure, diagnostics, and applications . 120
10.5 Case Study: Amyotrophic lateral sclerosis analysis revisited . . . 126

11 Multiple Linear Regression **131**
11.1 The Practical Question . . . 131
11.2 The Process of Building A Multiple Linear Regression Model . . 133
11.3 Case Study 1 . . . 139
11.4 Case Study 2 . . . 145
11.5 Practice Problems . . . 147

12 Logistic Regression Models **149**
12.1 Motivational Example and Practical Question . . . 149
12.2 Logistic Regression Models and Applications . . . 152
12.3 Case Studies . . . 154
12.4 Practice Problems . . . 160

13 Contingency Table Analysis **163**
13.1 The Motivational Examples . . . 164
13.2 Two-way Contingency Tables and Analysis . . . 165
13.3 Measures of Association . . . 171

13.4 Case Studies 174
13.5 Practice Problems 177

14 Analysis of Counts and Rates **181**
14.1 Motivational Examples 181
14.2 Poisson Regression for Counts and Rates 183
14.3 Case Studies 186
14.4 Concluding Remarks 192
14.5 Practice Problems 192

15 Power Calculation and Sample Size Determination **195**
15.1 Two-sample Proportion Problems 195
15.2 Two-sample Mean Problems 200
15.3 Paired-sample Problems 206
15.4 Unequal Sample Sizes Problems 212
15.5 Conclusions 214

Chapter 1

Introduction

This *E-coursepack* is a self-contained homegrown eBook that contains all topics covered in current STA 501 at WCU.

The audience of this class is graduate students from life science. The objective is to equip students with applied statistical methods that are used to analyze data generated from their fields.

The coverage of this course is from descriptive statistics to ANOVA, contingency tables, and generalized linear models.

- Setting up computing tools - getting started with R, RStudio, and R Markdown
- Sampling and Experimental Design
- Data Visualization and Descriptive Statistics
- Standard scores, the Normal, t, Chi-Squared, and F Distributions
- Sampling distributions
- Confidence Intervals
- Tests of a single mean/proportion and two means/proportions
- ANOVA
- Correlation and Simple Linear Regression
- Multiple Linear Regression

- Binary Categorical Regression
- Frequency Count Regression
- Procedures Related to Nominal Data
- Power and Sample Size Determination

Most case studies are based on (field and laboratory) data taken from the fields of biology, ecology, clinical, health science, etc. A formal statistical programming language R is used for data analysis. Students are not assumed to have prior experience in R coding. In the meanwhile, RMarkdown is an R package that can be used to combine text, R code, and the output from the execution of that code. Therefore, we can use RMarkdown to do an analysis and report the analysis at the same time in the same document.

Chapter 2

R, RStudio and RMarkdown

This chapter introduces open-source free computation and technical writing tools for this course: R, RStudio, and RMarkdown.

2.1 R

R is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. – **Wikipedia**

The official R web page has download links: https://www.r-project.org/. You can download and install the most current version of R based on your machine's operating system.

2.1.1 Order of Operations

The order of basic operations that we will use in this class is given below.

PEMDAS: **P**arentheses => **E**xponential => **M**ultiplication => **D**ivision => **A**ddition => **S**ubtraction!

When calculating confidence intervals and test statistics based on given formulas, please keep PEMDAS in mind!

```
(2+7)/(3^2)*5-1 = 9/9*5-1 = 5 - 1 = 4
```

We check the answer directly by typing the following in the R Console:

```
(2+7)/(3^2)*5-1
```

```
## [1] 4
```

2.1.2 Basic R Objects

We introduce several basic R objects (or R data structures): vectors, matrices, lists, and data frames.

R is case-sensitive, so name and Name will refer to different objects!

```
Name <- 1
name <- 0
```

2.1.2.1 Vectors

- An R **vector** holds a set of numerical values or character strings. For example,

```
num.vec = c(1, 4, 2.1, log(5), pi, sin(2), exp(-0.5), 0) # numerical vector
num.vec
```

```
## [1] 1.0000000 4.0000000 2.1000000 1.6094379 3.1415927 0.9092974 0.6065307 0.0000000
```

```
char.vec = c("john", "david", "jones", "kate")           # character vector
char.vec
```

```
## [1] "john"  "david" "jones" "kate"
```

Note A single scalar is considered a vector (i.e., one-dimensional vector).

2.1.2.2 Matrices

An R **matrix** is a rectangular table that holds either numerical or character values, but not both types of values. The following are two examples.

```
num.mtx = matrix(num.vec, ncol=2, byrow = TRUE)
num.mtx
```

```
##           [,1]      [,2]
## [1,] 1.0000000 4.0000000
## [2,] 2.1000000 1.6094379
## [3,] 3.1415927 0.9092974
## [4,] 0.6065307 0.0000000
```

```
char.mtx = matrix(char.vec, ncol = 2, byrow = FALSE)
char.mtx
```

```
##      [,1]    [,2]
## [1,] "john"  "jones"
## [2,] "david" "kate"
```

Note that the values in vectors and matrices must be in the same data type. A scalar is also considered as a 1-by-1 matrix.

2.1.2.3 Lists

A list is an R structure that may contain objects of any other type, including other lists. Lots of the modeling functions produce lists as their return values. We define a list to hold vectors and matrices defined in the previous sub-sections.

```
my.list = list(numvec=num.vec, charvec=char.vec, nummtx =num.mtx, charmtx=char.mtx )
my.list
```

```
## $numvec
## [1] 1.0000000 4.0000000 2.1000000 1.6094379 3.1415927 0.9092974 0.6065307 0.0000000
##
## $charvec
## [1] "john"  "david" "jones" "kate"
##
## $nummtx
##           [,1]      [,2]
## [1,] 1.0000000 4.0000000
## [2,] 2.1000000 1.6094379
## [3,] 3.1415927 0.9092974
## [4,] 0.6065307 0.0000000
##
## $charmtx
##      [,1]    [,2]
## [1,] "john"  "jones"
## [2,] "david" "kate"
```

The following example shows how to access the objects in an R list.

```
my.list$nummtx
```

```
##           [,1]      [,2]
## [1,] 1.0000000 4.0000000
## [2,] 2.1000000 1.6094379
## [3,] 3.1415927 0.9092974
## [4,] 0.6065307 0.0000000
```

2.1.2.4 Data Frame

A data frame is a table or a two-dimensional array-like structure in which each column contains values of one variable and each row contains one set of values from each column.

The following are the characteristics of a data frame.

- The column names should be non-empty.

- The row names should be unique.
- The data stored in a data frame can be of numeric, factor, or character type.
- Each column should contain the same number of data items.

The following is an example of a data frame

```
# Create the data frame.
emp.data <- data.frame(
   emp.id = c (1:5),
   emp.name = c("Rick","Dan","Michelle","Ryan","Gary"),
   salary = c(623.3,515.2,611.0,729.0,843.25),

   start_date = as.Date(c("2012-01-01", "2013-09-23", "2014-11-15", "2014-05-11",
      "2015-03-27"))
)
# Print the data frame.
print(emp.data)
```

```
##   emp.id emp.name salary start_date
## 1      1     Rick 623.30 2012-01-01
## 2      2      Dan 515.20 2013-09-23
## 3      3 Michelle 611.00 2014-11-15
## 4      4     Ryan 729.00 2014-05-11
## 5      5     Gary 843.25 2015-03-27
```

2.2 RStudio

RStudio is a must-know tool for everyone who works with the R programming language. It's used in data analysis to import, access, transform, explore, plot, model data, and make predictions on data.

2.2.1 RStudio GUI

The RStudio interface consists of several windows. I insert an image of a regular RStudio GUI.

2.2.2 Console

We can type commands directly into the console, or write in a text file, and then send the command to the console. It is convenient to use the console if your task involves one line of code. Otherwise, we should always use an editor to write code and then run the code in the Console.

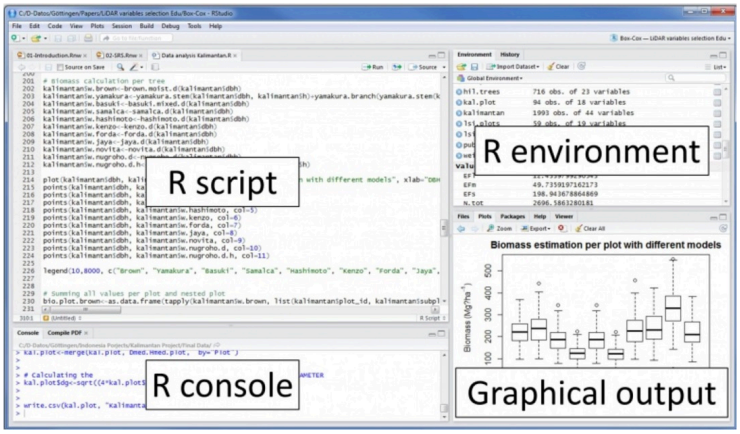


Figure 2.1: List of all variables and the description of each variable

2.2.3 Source Editor

Generally, we will want to write programs longer than a few lines. The Source Editor can help you open, edit, and execute these programs.

2.2.4 Environment Window

The Environment window shows the objects (i.e., data frames, arrays, values, and functions) in the environment (workspace). We can see the descriptive information such as the types and dimensions of the objects in your environment. We also choose data sources from the environment to view in the source window like a spreadsheet.

2.2.5 System and Graphic files

The Files tab has a navigable file manager, just like the file system on your operating system. The Plot tab is where the graphics you create will appear. The Packages tab shows you the packages that are installed and those that can be installed (more on this just now). The Help tab allows you to search the R documentation for help and is where the help appears when you ask for it from the Console.

2.2.6 RStudio offers numerous helpful features:

- A user-friendly interface
- The ability to write and save reusable scripts
- Easy access to all the imported data and created objects (like variables, functions, etc.)
- Exhaustive help on any object

- Code autocompletion
- The ability to create projects to organize and share your work with your collaborators more efficiently
- Plot previewing
- Easy switching between terminal and console

After you install R on your machine, you can go to https://posit.co/products/open-source/rstudio/ to download the free version of RStudio and install it. R will be automatically connected to RStudio. You can then open the Markdown through the GUI of RStudio.

2.3 RMarkdown

An R Markdown document is a text-based file format that allows you to include descriptive text, code blocks, and code output. It can be converted to other types of files such as PDF, HTML, and WORD that can include code, plots, and outputs generated from the code chunks.

2.3.1 Code Chunk

In R Markdown, we can embed R code in the code chunk defined by the symbol ```` ```{} ```` and closed by ```` ``` ````. The symbol ‘, also called **backquote** or **backtick**, can be found on the top left corner of the standard keyboard as shown in the following.

Figure 2.2: The location of backquote on the standard keyboard

There are two code chunks: executable and non-executable chunks. The following code chunk is non-executable since there is no argument specified in the `{}`.

````
```{}
This is a code chunk
```
````

Figure 2.3: Non-executable code chunk.

```
This is a code chunk
```

To write a code chunk that will be executed, we can simply put the letter `r` inside the curly bracket. If the code the code chunk is executable, you will the green arrow on the top-right corner of the chunk.

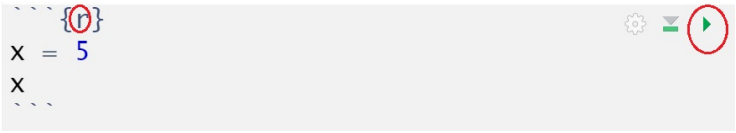


Figure 2.4: Executable code chunk.

We can define R objects with and without any outputs. In the above R code chunk, we define an R object under the name `x` and assign the value 5 to `x` (the first line of the code). We also request an output that prints the value of `x`. The above executable code chunk gives output `[1] 5` in the Markdown document. The same output in the knit output files is in a box with a transparent background in the form `## [1] 5`.

```
x = 5
x
```

```
## [1] 5
```

We can also use an argument in the code chunk to control the output. For example, the following code chunk will be evaluated when kitting to other formats of files. But we can still click the green arrow inside the code chunk to evaluate the code.

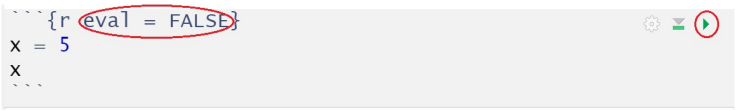


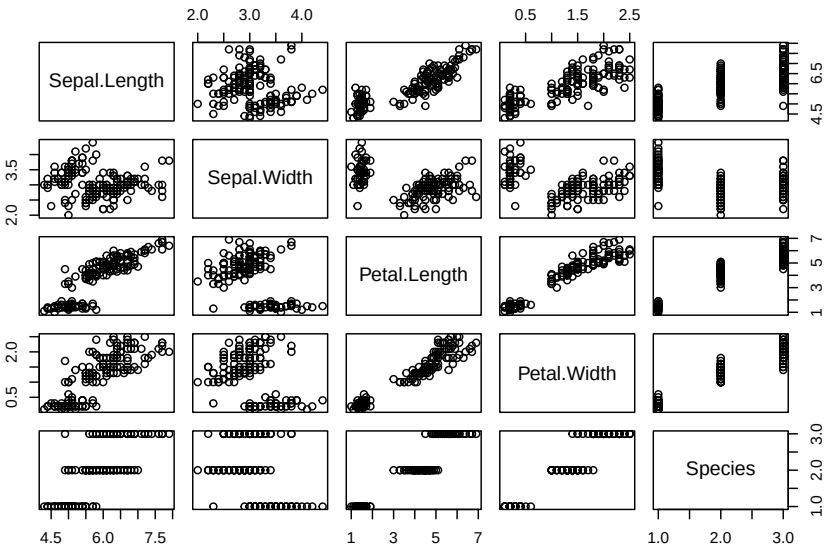
Figure 2.5: Executable code chunk with control options.

```
x = 5
x
```

2.3.2 Graphics Generated from R Code Chunks

In the previous sub-sections, we include images from external image files. In fact, can use the R function to generate graphics (other than interacting with plots, etc.) in the markdown file & knit. For instance, we can generate the following image from R and include it in the Markdown document and the knitter output files.

```
plot(iris)
```



Chapter 3

Data Collection and Data Loading

We briefly introduce the basic sampling plans and study design to collect valid data for statistical analysis and modeling. Since all analyses will be based on R, we will demonstrate to to load data sets in different formats to R.

3.1 Sampling Plans

In general, probability sampling plans generate samples suitable for inferential statistics (such as confidence intervals and testing hypotheses).

3.1.1 Simple Random Sampling (SRS)

In this sampling plan, each subject is randomly chosen in such a way that each subject in the population has an equal chance, or probability, of being selected. In mathematical statistics, there is a rigorous definition of SRS plan.

3.1.2 Systematic sampling

In this sampling plan, we (randomly) label all subjects in a population from 1, 2, ..., N (population size), randomly choose a label, say n_0, and then select every k^{th} subjects to include in the sample. That resulting sample is called a systematic sample. Because the

3.1.3 Stratified sampling

This sampling plan assumes that the target population of interest has several naturally defined sub-populations, then we take a sub-sample (SRS) from each

sub-population such that the sub-sample sizes are proportional to their corresponding sub-population sizes.

3.2 Study Designs

Depending on whether the variables in the study were modified or filtered by researchers, we can categorize the study designs into observational and experimental studies.

3.2.1 Observational Studies

Observational studies are ones where researchers observe the effect of a risk factor, diagnostic test, treatment, or other intervention **without trying to remove confounding factors**.

Cross-sectional studies collect information on a population by taking a snapshot or cross-section of the population. These studies usually involve one contact with the study population and are relatively cheap to undertake.

Pre-test/post-test Take two cross-sectional samples taken from randomly selected subjects before and after an intervention (such as a new treatment) applied to the randomly selected subjects in the study. The objective is to see whether a characteristic of the sample changed before and after the intervention.

Retrospective studies use historical information to study the characteristics of a population. Sometimes current information about the population is not available or difficult to obtain, we then use the historical information **under certain assumptions**.

Longitudinal studies follow study subjects over a period of time with repeated data collection throughout. Most are observational studies that seek to identify a correlation among various factors. Thus, longitudinal studies do not manipulate variables and are not often able to detect causal relationships.

3.2.2 Experimental Designs

Experimental studies are ones where researchers **carefully design experiments to remove potential confounding factors** introduce an intervention and then study the effects.

Prospective studies, also called **follow-up study**, in which you select subjects randomly and wait for a period of time to record the formation of interest to perform statistical analysis.

Randomized controlled trials (RCT) in clinical studies are always prospective studies and often involve following a “cohort” of individuals to determine the relationship between various variables.

Longitudinal studies can also be considered as experimental studies.

3.3 Loading Data to R from External Data Files

Three basic types of data files are common in practice: csv, txt, and xls(x). I have uploaded the well-known **iris* data set to the course web page in the aforementioned formats. You can practice the R functions to load these external data sets to R.

3.3.1 Text File (aka. Delimited Text File)

R command **read.table()** will load text files in R.

Caution: If the data file contains **missing values**, you need to handle the missing values before loading it to R. If you take a formal programming course, you will write several lines of code to clean the data. In this class, we only use basic R commands to do analysis.

The code in the following code chunk does not work!!!

```
placement.data = read.table("C:\\STA501\\w02\\placement.txt", header=TRUE)
```

3.3.1.1 Read files from a remote web server

```
### delimited text file
irisTXT="https://raw.githubusercontent.com/pengdsci/STA501/main/Data/w02-iris.txt"
iris.text = read.table(irisTXT, header = TRUE)
### csv file
irisCSV = "https://raw.githubusercontent.com/pengdsci/STA501/main/Data/w02-iris.csv"
iris.csv = read.csv(irisCSV, header = TRUE)
```

Note that there is no commands such as **read.table()** and **read.csv()** in the base R to read Excel file from URL. There are several R functions in different libraries, such as **read_xlsx()** and **read_xls()** in library **{readxl}**, can read Excel file from the local drive (see the example in the next sub-setion).

3.3.1.2 Read files from local folder

```
### delimited text file
iris.text.loc = read.table("C:\\STA501\\w02\\w02-iris.txt", header = TRUE)
### csv file
iris.csv.loc = read.table("C:\\STA501\\w02\\w02-iris.csv", header = TRUE)
## Excel file - no built-in command in the base R can read Excel file to R.
## need to load a command in a R library.
library(readxl)
iris.xlsx.loc = read_xlsx("C:\\STA501\\w02\\w02-iris.xlsx")
```

3.4 Data Frame and List

Data frame can hold different type of variables but requires variables to be equal in length. If you have variables in different types **and** having different length, you need to use list to hold these variables.

3.4.1 Data Frame

- We define following categrical and numeric vectors (data sets)

```
veca = 1:26           #  c() is unnecessary since colon (:) is a shortcut to
                      #  define a patterned sequence. 1, 2, 3, ..., 24, 25, 26.
vecb = 101:126
vecc = 201:226
vecd = letters        #  26 letters (lower case)
vece = LETTERS        #  upper-case letters
```

- **Example 1**. Define data frame with different types of variables

```
dataframe01 = data.frame(A=veca, B = vecb, D = vecd, E = vece)  # 26-by-3 data frame
```

- **Example 2**. Define a data frame using vectors with unequal width. The issue is that R will recycle the small vectors **whose lengths are factors of the length of the longest vector** to make equal lengths columns in the data frame.

```
## CAUTION: If the lengths of individual variables are different,
## R will recycle the values in the
##  mall data (short vector) to make the equal length across the columns
##
dframe2 = data.frame(A = veca,            # this is a vector with 100 values
                     X = 1:13,            # this vector has 13 values
                     c = sum(veca+vecb+vecc)
                                 # two comments: 1. c is reserved for defining
                                 # vectors. Should not be used a name
                                 # of any objects in R
                                 # 2. this is also a single value.
  )
```

- **Example 3**. The following code will produce an error because Y has 4 rows and A has 26 rows!

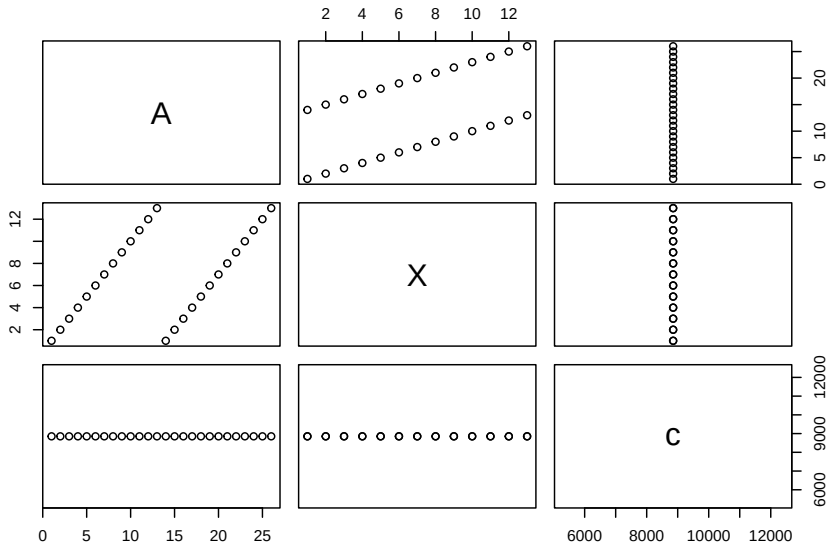
```
dframe2 = data.frame(A = veca,             # this is a vector with 100 values
                     Y = c(33,44,55,66),   #  4 values in vector b!
                     W = c(99,999)         #  2 values in W!
)
```

Error in data.frame(a = veca, b = c(33, 44, 55, 66), c = c(99, 999)) : arguments imply differing number of rows: 26, 4, 2

you can check whether the data frame is correctly defined. I added an option **eval = FALSE** to the following code chunk to avoid printing out the 100-by-3 data frame.

```
dframe2
```

```
plot(dframe2)    # no error. This may not be what you expected
```




```
                 # since the data frame has recycling issues.
```

The following code does not work since no specific column is specified to calculate the standard deviation!

```
sd(dframe2)    # sd() calculates the standard deviation of a SINGLE vector, but
               # There 3 in the data frame. The error message says that R cannot
               # cannot combine all columns to make a vector. In R, numerical
               # values are 'double'.
```

We can calculate the standard deviation of the variables in the data frame one by one. For example,

```
sd(dframe2$A)     # This will NOT generate an error since we calculate the
```

```
## [1] 7.648529
```

```
                  # standard deviation for a specific variable (column).
```

Chapter 4

Descriptive Statistics

The note outlines the basic descriptive statistics.

- Data Types
- Tabular and graphic summary of data
- Numerical summary of data

4.1 Data Types

There are different classifications of data types. We use the following simple one

- **Categorical Variables** - The values of these types of variables do not have numerical meaning in the sense that one can not perform arithmetic operations with the values of these types of variables.
  - *Ordinal categorical variables* - the values have a natural order. For example, course letter grades: A, B, C, D, and F.
  - *Nominal categorical variables* - the values do not have a natural order. For example, majors in a college: Mathematics, finance, music, biology, etc.
- **Numerical Variables** - As indicated in the name, the values of numerical variables are numbers.
  - *Discrete variables* - one can find two values of such variables such there are no meaningful values that fall between the two. For example, the number of children in a household: 1, 2, 3, 4, ... There is no such household that has 2.5 children.
  - *Continuous variables* - For any two distinct values of such variables, any value between the two is meaningful. For example, consider two arbitrarily selected human body temperatures (in Fahrenheit) 97.4

and 97.5, any number between 97.4 and 97.5 could be the temperature of someone in the population (although the person may not be part of the sample).

4.2 Tabular and Graphic Summary

Both tabular and graphic summaries are powerful and effective tools to **visualize** the (shape of the) distribution of the data.

4.2.1 Categorical Data

Example 1: [Status of Endangered Species]: The data was extracted from the U.S. Fish & Wildlife Service ECOS Environmental Conservation Online System. The data can be found at the following https://raw.githubusercontent.com/pengdsci/STA501/main/Data/EndangeredSpecies.csv

We want to summarize the status of endangered species (one of the columns in the data set).

4.2.1.1 Frequency Table

Since the values of categorical data sets are labels, constructing frequency tables of categorical data is straightforward.

```
speURL = "https://raw.githubusercontent.com/pengdsci/STA501/main/Data/EndangeredSpecies
Species = read.csv(speURL, header = TRUE)   # read the csv data from the URL
kable(t(head(Species[1:4,])))     # list first 4 rows of the data
```

	1	2	3
scientificName	Acanthorutilus handlirschi	Accipiter fasciatus natalis	Accipiter francesii pusillus
commonName	Cicek (minnow)	Christmas Island goshawk	Anjouan Island sparrowha
criticalHabitat	N/A	N/A	N/A
speciesGroup	Fishes	Birds	Birds
Status	Endangered	Endangered	Endangered
specialRules	N/A	N/A	N/A
whereListed	Wherever found	Wherever found	Wherever found

Next, we create a frequency table to include all four types of frequencies.

```
speciesGroup = Species$speciesGroup   # extract the column of endangered species
freq = table(speciesGroup)            # frequency count
rel.freq = freq/sum(freq)             # relative frequency
cum.freq = cumsum(freq)               # cumulative frequency
cum.rel.freq = cum.freq/sum(freq)     # cumulative relative frequency
freq.table = cbind(freq =freq,
                   rel.freq = rel.freq,
                   cum.freq = cum.freq,
```

```
                    cum.rel.freq = cum.rel.freq)
kable(freq.table)     # kable() makes a nice-looking table
```

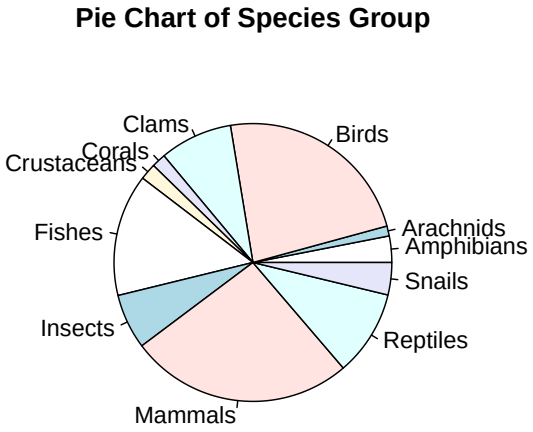
	freq	rel.freq	cum.freq	cum.rel.freq
Amphibians	45	0.0307377	45	0.0307377
Arachnids	17	0.0116120	62	0.0423497
Birds	342	0.2336066	404	0.2759563
Clams	124	0.0846995	528	0.3606557
Corals	24	0.0163934	552	0.3770492
Crustaceans	28	0.0191257	580	0.3961749
Fishes	208	0.1420765	788	0.5382514
Insects	94	0.0642077	882	0.6024590
Mammals	381	0.2602459	1263	0.8627049
Reptiles	146	0.0997268	1409	0.9624317
Snails	55	0.0375683	1464	1.0000000

4.2.1.2 Bar Chart and Pie Chart

We use R to create both charts based on the frequency table created in the previous sub-section in the following.

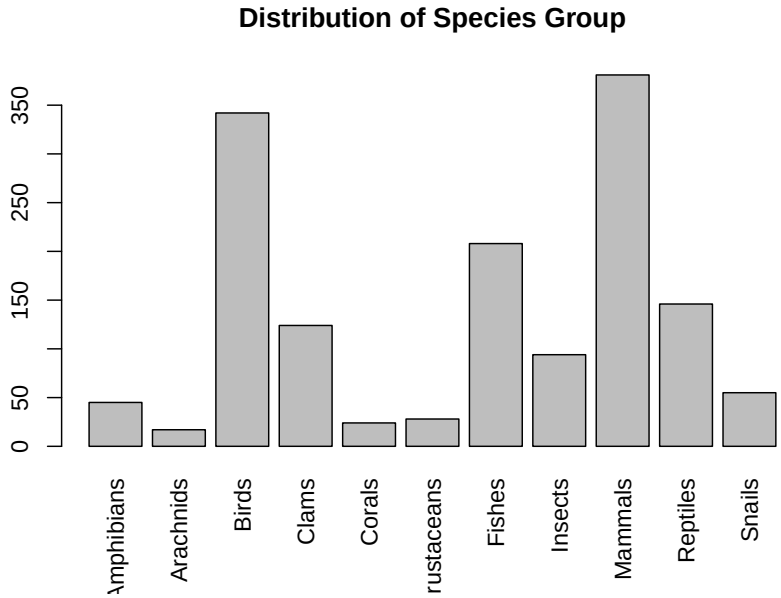
We first draw a simple pie chart. You can add different colors and additional information to the chart. You can visit https://www.statmethods.net/graphs/pie.html for more examples.

```
freq = table(speciesGroup)
group = names(freq)
pie(freq, labels = group, main="Pie Chart of Species Group")
```



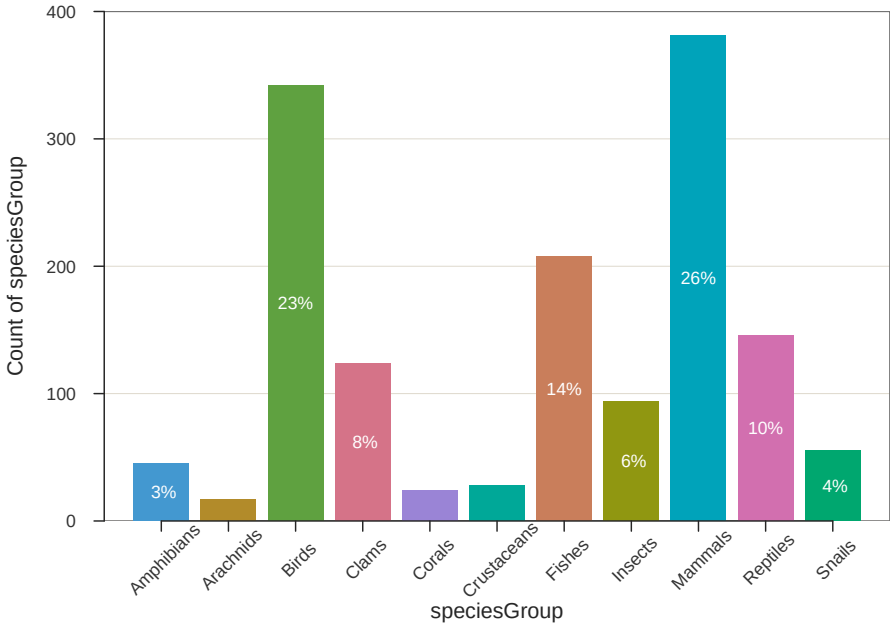
Since there are too many slices in the pie chart, it is not easy to add frequencies to the chart. This is not a good visualization. Next, we create a bar chart to represent the distribution of the same data set.

```
freq = table(speciesGroup)
group = names(freq)          # categories
barplot(freq,                # frequency table
        names.arg=group,     # tick marks
        las=3,               #
        main="Distribution of Species Group" )
```



The above bar plot was created using the function in Base R. We can also use relevant functions in different R packages to make a bar chart that may contain additional information. For example, the R function **BarChart()** in library **{lessR}** generates bar charts with more information based on the original data values. This is different from **barplot()** which uses the frequency tables.

```
# library(lessR)          # placed at the beginning of the document
BarChart(speciesGroup,  rotate_x=45)
```



There are more examples to use **BarChart()** in a nice blog https://cran.r-project.org/web/packages/lessR/vignettes/BarChart.html.

4.2.2 Numerical Data

To summarize numerical data sets, we use frequency tables and histograms to visualize the underlying distributions.

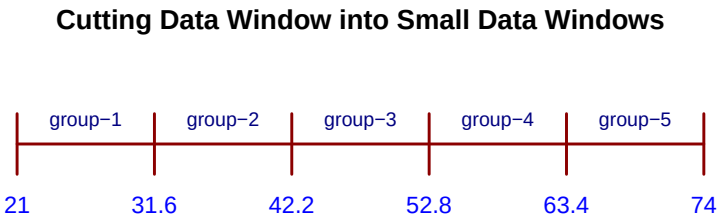
We will use the following data set to illustrate the steps to construct frequency tables and histograms using R. The data set https://raw.githubusercontent.com/pengdsci/STA501/main/Data/diet.csv was used to study the effect of three different diets on weight loss.

4.2.2.1 Frequency Tables

Unlike categorical data in which the data values are category labels, in numerical data, we need to group data values to create data groups and then construct the corresponding frequency table.

Think about creating a **data window** by the maximum and minimum data values in the data set then cut the data window into several small data windows with equal width. The data values in each small data window form a data group. In the figure, we assume there is a data set with a minimum value of 21 and a maximum value of 74. We plan to split the data window [21, 74] into five small data windows **with equal width**. The cut-off points are [21.0, 31.6, 42.2, 52.8, 63.4, 74.0] (including minimum and maximum values). We can use

the R command to find these cut-offs if we provide the minimum, maximum, and number of small windows to be used for creating the frequency tables and histogram.



We can use the R function **seq(min, max, length = number-of-windows + 1)** to find cut-off points. For example, in the above figure, the following code yields the cut-off.

```
# round off the cut-offs to 1 decimal point
cutoff = round(seq(21, 74, length =5+1),1)
# kable() produces a nice looking table in PDF
kable(data.frame(cutoff), align = 'l')
```

cutoff
21.0
31.6
42.2
52.8
63.4
74.0

Example 2: [**Effectiveness of Diets Data**] We are interested in creating a histogram of the weights of all participants in the study before starting the three diets. We want to create a frequency table with 6 rows. That is, We will create 6 small data windows to define 6 groups. R function **cut(x = data-set, breaks=cutoff-points)** .

```
dietURL = "https://raw.githubusercontent.com/pengdsci/STA501/main/Data/diet.csv"
diet = read.csv(dietURL, header=TRUE)
pre.weight=diet$initial.weight   # extract pre.weight from the data set.
### calculate the cut-offs that yield 6 small data windows with equal widths
cutoff.pt = seq(min(pre.weight), max(pre.weight), length = 6+1)
cutoff.pt = round(cutoff.pt, 1)       # rounding off to keep 1 decimal place
### use R function **cut()** to split the data window into 6 small data windows
```

```
data.group = cut(x = pre.weight, breaks=cutoff.pt, include.lowest = TRUE)
## use R function **table** to get the frequency table
freq.count=table(data.group)    # regular frequency counts
kable(freq.count, align = 'l')
```

data.group	Freq
[58,63]	12
(63,68]	14
(68,73]	17
(73,78]	15
(78,83]	11
(83,88]	7

We can also use the same steps to find relative and cumulative frequencies.

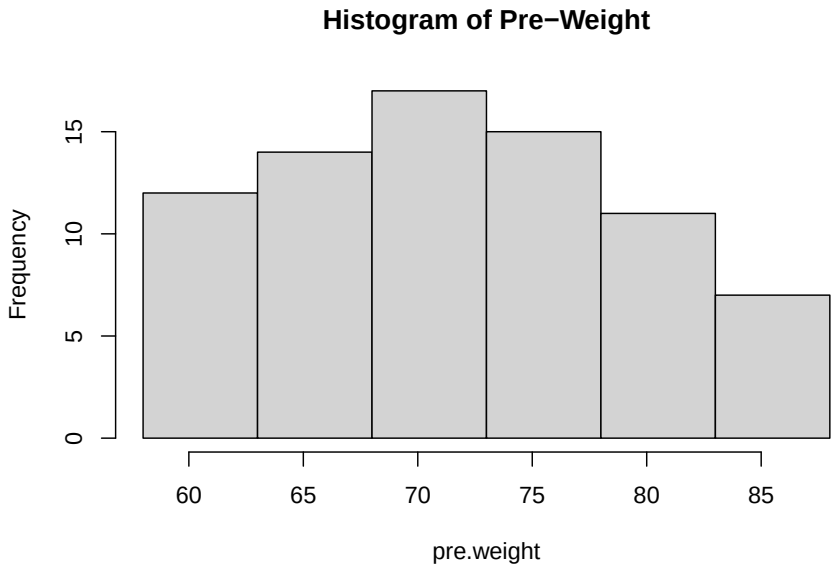
```
freq = table(data.group)              # frequency count
rel.freq = freq/sum(freq)             # relative frequency
cum.freq = cumsum(freq)               # cumulative frequency
cum.rel.freq = cum.freq/sum(freq)     # cumulative relative frequency
freq.table = cbind(freq =freq,
                   rel.freq = round(rel.freq,3),   # keep 3 decimal places
                   cum.freq = cum.freq,
                   cum.rel.freq = round(cum.rel.freq ,3)) # keep 3 decimal places
kable(freq.table, align = 'l')
```

	freq	rel.freq	cum.freq	cum.rel.freq
[58,63]	12	0.158	12	0.158
(63,68]	14	0.184	26	0.342
(68,73]	17	0.224	43	0.566
(73,78]	15	0.197	58	0.763
(78,83]	11	0.145	69	0.908
(83,88]	7	0.092	76	1.000

4.2.2.2 Graphic Summary - Histogram

As mentioned earlier, we use a histogram to visualize the distribution of the numerical data set. R function **hist(x=data-set, breaks = cutoff)**. We still use the same **pre.weight** and the same cut-off obtained in the previous subsections to construct the histogram.

```
hist(pre.weight, breaks = cutoff.pt,
     main = "Histogram of Pre-Weight")
```



We can see that the distribution of pre-weights is **skewed to the right** since the above histogram has a long right tail.

4.3 Numerical Summary of Numerical Data

Three family measures are outlined in this section: central tendency, variation, and location. We will still use **pre-weight** as an example to show how to basic R functions to calculate these numerical measures.

4.3.1 Central Tendency

We will not list all relevant measures of centers. Three three R functions **mean()** and **median()** are used to calculate the mean and median of a given data set.

Mean - the average of the values in the data set.

```
avg.pre.weight = mean(pre.weight)
kable(data.frame(avg.pre.weight), align = 'l')
```

avg.pre.weight
72.28947

Median - a cut-off value that splits the data values into two parts (the cut-off in both parts) such that at least 50% of data values are **greater than or equal to** and at least 50% of data values are **less than or equal to the** cut-off value.

```
middle.numer = median(pre.weight)
kable(data.frame(middle.numer), align = 'l')
```

middle.numer
72

The more general quantile function **quantile()** can also be used to find the median. In fact, **quantile()** can be any percentile. The 50th percentile is the median.

```
quantile.mid.num = quantile(pre.weight, # data set name
                            0.5,        # percentile, 0.5 = 50%
                            type=2      # there are different interpolations.
                                        # We use type 2.
                            )
fifty.percentile= data.frame(quantile.mid.num)
kable(fifty.percentile, align = 'l')
```

	quantile.mid.num
50%	72

4.3.2 Variations

We use R functions and variable **pre-weight** to calculate variance, standard deviation, and inter-quartile range (IQR).

- **Variance** - measure the spread of the data. R function **var()** calculates the sample variance.

```
sample.var = var(pre.weight)
kable(data.frame(sample.var), align = 'l')
```

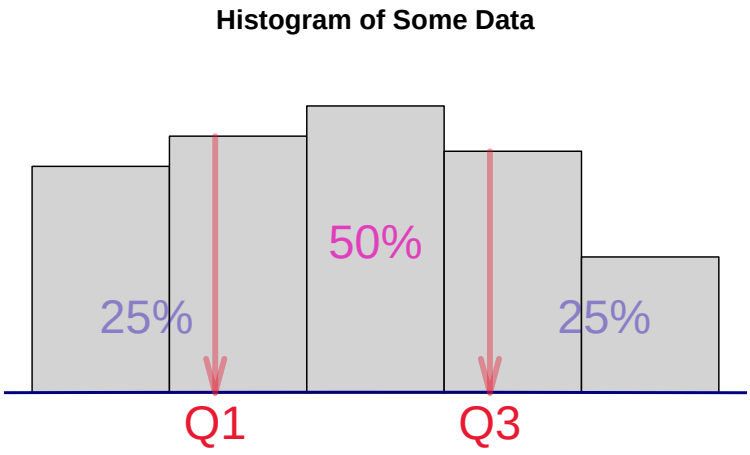
sample.var
63.59509

- **Standard Deviation** - measures the spread of the data and is equal to the square root of the variance.

```
stdev = sd(pre.weight)
kable(data.frame(stdev), align = 'l')
```

stdev
7.974653

- **Inter-quartile Range (IQR)** - the range of the middle 50% data values. That is, we throw out the bottom and upper 25% of data values and use the difference between the maximum and the minimum values to define IQR. The idea is illustrated in the following figure [I exclude the code in the output file. You can find the RMD document].



where **Q1** and **Q3** are the first and third quartiles which can be found using **quantile()**. The inter-quartile range is defined to be **IQR = Q3 - Q3**. We still use the **pre.weight** to illustrate how to find the IQR with the following code.

```
IQR = quantile(pre.weight, 0.75, type = 2) - quantile(pre.weight, 0.25, type = 2)
IQR = as.vector(IQR)
kable(data.frame(IQR), align = 'l')
```

IQR
12

4.3.3 Location

4.3.3.1 Z-score Transforamtion

The z-score transformation converts any given numerical data set to a new standardized data set such the new data set has zero mean and unit standard deviation. Let's denote the original data set to be $X = \{x_1, x_2, \cdots, x_n\}$. let $Z = \{z_1, z_2, \cdots, z_n\}$ be the standardized data set. The formula that transforms X to Z is given by

$$z_i = \frac{x_i - \bar{x}}{s}$$

where $\bar{x}$ is the sample mean and s is the standard deviation of X.

I use the toy data $X = \{1, 3, 5, 7, 9\}$ as an example to perform the z-score transformation.

```
X= c(1, 3, 5, 7, 9)      # type in data values
xbar = mean(X)           # sample mean
s = sd(X)                # sample standard deviation
Z=(X-xbar)/s             # z-score transformation
kable(data.frame(Z), align = 'l', format = "pipe")   # make a nice looking
```

Z
-1.2649111
-0.6324555
0.0000000
0.6324555
1.2649111

4.3.3.2 Quantile

A k-th quantile **of a data set** (also called sample k-th quantile) is defined as a cut-off value that splits the data into two parts such that **at least** $100k\%$ of data values are **bigger than or equal to** the cut-off and **at least** $100(1-k)\%$ data values are **less than or equal to** the cut-off value, where $0 < k < 100$. Special quantiles are the quartile (quarter) and percentiles (hundredth).

Please keep in mind that the calculation of quantile is based on the sorted data and involves interpolations. Several interpolations were implemented in R. There is a minor difference between these different interpolations. The simple interpolation that is commonly used is the *so-called* type 2 interpolation. The type 1 interpolation is the default type in **quantile(dataset, k/100, type=2)**.

Example [**Pre-weight data**] - We want to find 25% and 68% percentiles of pre-weights.

```
q.25 = quantile(pre.weight, 0.25, type = 2)
kable(data.frame(q.25), align = 'l')
```

	q.25
25%	66

```
q.68 = quantile(pre.weight, 0.68, type = 2)
kable(data.frame(q.68), align = 'l')
```

	q.68
68%	77

We can call **quantile()** to find the two quantiles simultaneously.

```
q.25.68 = quantile(pre.weight, c(0.25, 0.68), type = 2)
kable(data.frame(q.25.68), align = 'l')
```

	q.25.68
25%	66
68%	77

4.3.3.3 Five-number Summary and Box-plot

The five-number summary consists of 5 numbers: minimum (0%), 1st quartile (25%), 2nd quartile(50%, median), 3rd quartile (75%), and maximum (100%). R function **fivenum()** is dedicated to finding the five-number summary.

```
fivenum(pre.weight)
```

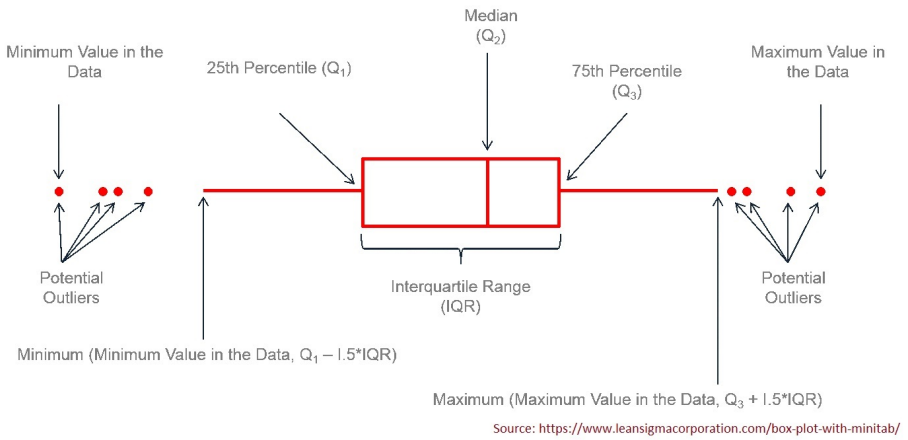
```
## [1] 58 66 72 78 88
```

We can also use **quantile()** to find the five-number summary in the following.

```
quantile(pre.weight, c(0, 0.25, 0.5, 0.75, 1), type = 2)
```

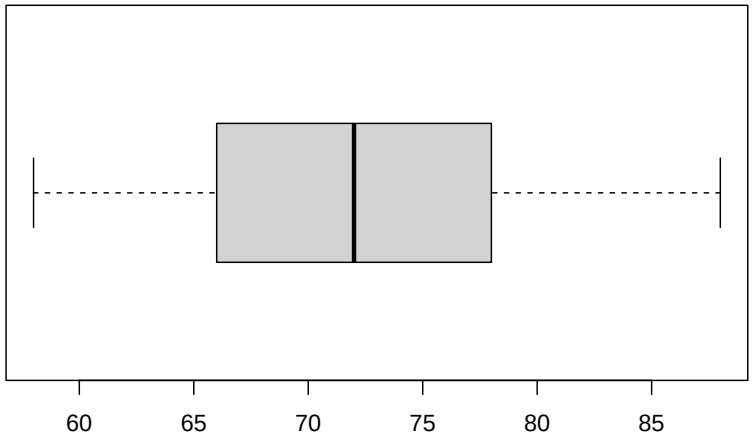
```
##   0%  25%  50%  75% 100%
##   58   66   72   78   88
```

A box plot is the graphic representation of the five-number summary.



R function **boxplot()** will make the box-plot. We only present a simple box plot in the following.

```
boxplot(pre.weight, horizontal = TRUE)
```



4.4 Assignment - Descriptive Statistics

The **Diabetes** data set to be used in this assignment is taken from Vanderbilt's Biostatistics Datasets.

The following is the description from the web page:

These data are courtesy of Dr. John Schorling, Department of Medicine, University of Virginia School of Medicine. The data consists of 19 variables on 403 subjects from 1046 subjects who were interviewed in a study to understand the prevalence of obesity, diabetes, and other cardiovascular risk factors in central Virginia for African Americans. According to Dr. John Hong, Diabetes Mellitus Type II (adult-onset diabetes) is associated most strongly with obesity. The waist/hip ratio may be a predictor of diabetes and heart disease. DM II is also associated with hypertension - they may both be part of "Syndrome X". The 403 subjects were the ones who were actually screened for diabetes. Glycosolated hemoglobin > 7.0 is usually taken as a positive diagnosis of diabetes. For more information about this study see

Willems JP, Saunders JT, DE Hunt, JB Schorling: Prevalence of coronary heart disease risk factors among rural blacks: A community-based study. *Southern Medical Journal* 90:814-820; 1997

Schorling JB, Roach J, Siegel M, Baturka N, Hunt DE, Guterbock TM, Stewart

HL: A trial of church-based smoking cessation interventions for rural African Americans. *Preventive Medicine* 26:92-101; 1997.

```
diaURL = "https://raw.githubusercontent.com/pengdsci/STA501/main/Data/diabetes.csv"
diabetes = read.csv(diaURL, header = TRUE)
kable(t(head(diabetes)))
```

	1	2	3	4	5	6
id	1000	1001	1002	1003	1005	1008
chol	203	165	228	78	249	248
stab.glu	82	97	92	93	90	94
hdl	56	24	37	12	28	69
ratio	3.6	6.9	6.2	6.5	8.9	3.6
glyhb	4.31	4.44	4.64	4.63	7.72	4.81
location	Buckingham	Buckingham	Buckingham	Buckingham	Buckingham	Buckingham
age	46	29	58	67	64	34
gender	female	female	female	male	male	male
height	62	64	61	67	68	71
weight	121	218	256	119	183	190
frame	medium	large	large	large	medium	large
bp.1s	118	112	190	110	138	132
bp.1d	59	68	92	50	80	86
bp.2s	NA	NA	185	NA	NA	NA
bp.2d	NA	NA	92	NA	NA	NA
waist	29	46	49	33	44	36
hip	38	48	57	38	41	42
time.ppn	720	360	180	480	300	195

We can see from the first 6 observations that there are 15 numerical variables and 3 categorical variables. Variable **bp.2s** and **bp.2d** have missing values. To complete this week's assignment, you need to choose one numerical variable and one categorical variable with NO missing values.

The following code shows how to extract variables from the data frame. I will use the two variables with missing values as an example. You can modify the code to extract your variables for the assignment.

```
bp.2s <- diabetes$bp.2s
bp.2d <- diabetes$bp.2d
```

4.4.1 Summarizing Categorical Data

- Use the categorical variable you selected to perform the following analysis
  - **Construct a relative frequency table**. Write a few sentences to describe the distribution of the variable. Note that you are encouraged to construct a frequency table with all four types of frequencies as I did in the class note.

  - **Construct a pie-chart** to represent the distribution of the categorical variable.

- Using the numerical variable you chose from the diabetes data to answer the following questions.
  - **Construct a relative frequency table of the numerical variable** with 10 categories. In other words, the frequency table should have 10 rows. You are encouraged to include all 4 frequencies in the table. Please provide a brief description of the relative frequencies.
  - **Construct a histogram of the numerical variable** with 10 vertical bars. In other words, the histogram is a geometric representation of the frequency table. Explain the distribution of the variable. Is it skewed to the left or the right?
  - **Construct a box-plot** and explain it. That is, can you tell whether the distribution is skewed to the right or the left?

Chapter 5

Concepts of Probability Distributions

In this chapter, we provide a non-technical description of random variables and their corresponding probability distributions.

5.1 Concepts of random variables

A random variable is a variable and its value is dependent on chance. What is the difference between a "regular" variable we learned from middle school and a **random variable**?

- **Example 1**: Let x be a variable in the equation $3x+4=8$. x is unknown before you solve for it from the equation. Most importantly, it is a **fixed** value although it is unknown.

- **Example 2**: Let Y be the height of the WCU student population. Y is unknown before you measure the height of a student from this population. However, which student is selected to measure his/her height is dependent on **the chance**!

We can see from examples 1 and 2 that x and Y are variables. x is a **"regular" variable** and Y is a **random variable**!

Because the value of a random variable is **dependent on the chance**, we need additional mathematical tools to characterize the **chance** - probability distribution. This will be described in a non-technical manner in the next section.

5.2 Types of random variables

There are basic types of random variables: discrete and continuous random variables.

- **Discrete random variables**: A random variable is said to be discrete if its value is obtained by **counting**. A discrete random variable may have either finite or infinite distinct values.
  - **Example 3**: *Coin Flipping Experiment* - Consider flipping an unfair coin (like the ones in the left panel of Figure 1) 10 times. Let $X =$ the number of **heads** observed. X is a discrete random variable since it can take finite values (11 distinct values): 0, 1, 2, ..., 10. Further, it can only take more than 11 distinct values since the unfair coin was flipped 10 times! Moreover, X can never be 2.7! It is discrete!
  - **Example 4**: *Quadrat Sampling (right panel of Figure 1)* - Consider estimating the total number of dandelions in a field. We know or we can measure the area of the field. The area of the quadrat is fixed. We can throw the quadrat randomly to the different regions in the field multiple times and count the number of dandelions in the squared plots sampled. Since the ratio of the sampled area and the total area of the field is equal to the ratio of the total number of dandelions in the sampled area and the total number of dandelions in the field. We then can solve the equation for the estimated total number of dandelions in the field. Now, let Y be the number of dandelions inside the quadrat in each sampled region. Clearly, Y is discrete. Is Y finite? Technically speaking, the number of dandelions in the quadrat cannot be infinite no matter where it is placed. However, unlike X in Example 3 which is naturally capped by 11, no cap can be placed on Y. Theoretically speaking, Y is infinite!

(1). Unfair Coin (2). Estimating total number of dandelions in a field.

Figure 5.1: Left: unfair coins. Right: quadrat for ecology sampling - estimating the number of dandelions in a field.

- **Continuous random variables** A random variable is continuous if its value is obtained by **measuring**.
  - **Example 5**: Let Y be the pH of arterial plasma (i.e., the acidity

of the blood) of people of a population. Y is a typical continuous random variable. It has uncountably many values **between 0 and 14**. It is continuous since any value between any selected pHs could be the pH of a person in the population.

5.3 Concepts of probability distributions

We first briefly describe the concept of probability and then outline the probability distributions of random variables.

5.3.1 Concepts of probability

Before introducing the definition of probability, we list the following concepts.

- **(Statistical) Experiment** - a process that produces well-defined outcomes. For example, consider an experiment of flipping a fair coin, the possible outcomes of this experiment are {heads, tails}.
- **Sample Space** - The set of all possible outcomes is called sample space. The **sample space** of the above coin-toss example is **S = {heads, tails}**.
  - **Example 6**: Consider an experiment of flipping a fair coin *sequentially* three times. We use **T** to denote **tails** and **H** for **heads**. The sample space of this experiment is given by **S = {TTT, TTH, THT, HTT, HHH, HHT, HTH, THH}** which can be explained by the following figure.

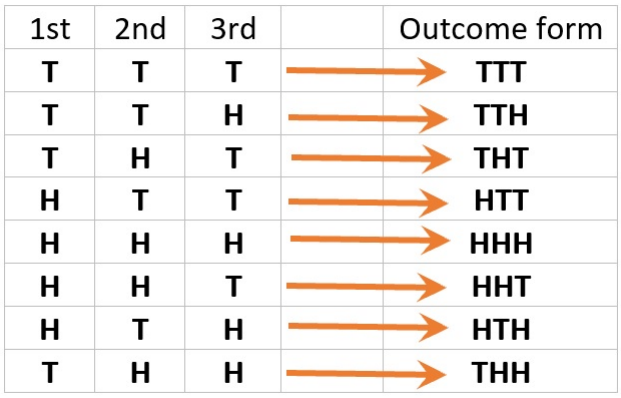


Figure 5.2: Sample space of the experiment of flipping a coin three times.

- **Event** - a subset of sample space. Two extreme events are the impossible event (i.e., the subset is empty) and the sure event (i.e., the subset is equal to the sample space).

  - **Example 7**: We will define a few events based on the experiment in **Example 6** in the following.
    * E1 = {observing at least 2 heads} = {HHH, HTH, THH, HHT}.
    * E2 = {observing exactly one heads} = {HTT, THT, TTH}.
    * E3 = {observing 5 heads} = {} = empty set = impossible event.
    * E4 = {observing at one heads **OR** one tails} = **S** = sure event.
  - **Example 8**: We still use the experiment in **Example 6** with sample space **S = {TTT, TTH, THT, HTT, HHH, HHT, HTH, THH}**. Define Y = the number of **heads** observed in the experiment. We now define **Events** based on the value of random variable Y.
    * E.0 = {Y=0} = {TTT}.
    * E.1 = {Y=1} = {TTH, THT, HTT}.
    * E.2 = {Y=2} = {THH, HTH, HHT}.
    * E.3 = {Y=1} = {HHH}.

- **Definitions of probability** - Two technical definitions of probability measure the chance of occurrence of an event.

  - *Classical probability (based on equally likely outcome)*: P(E) = (# outcomes in E)/(# outcomes in S).

    * **Example 9**: The experiment in **Example 6** is an equally likely outcome experiment. Based on the above definition, we can calculate the probability of the event in **Example 7**:
      · P(E1) = #E1/#S = 4/8 = 1/2.
      · P(E2) = #E2/#S = 3/8.
      · P(E3) = #E3/#S = 0/8 = 0. That is, an impossible event has a probability of 0.
      · P(E4) = #E4/#S = #S/#S = 1. That is, a sure event has probability 1.

  - *Relative frequency approximation* - If an event is defined based on an unequally likely experiment, we need to repeat the experiment multiple times to observe the number of occurrences and then use the relative frequency to **approximate** the probability of the event. This definition is used in most practical applications.

    * **Example 10**: Chronic arsenic toxicity, which is due to low-concentration exposure over a long period of time, impairs the same organs and tissues and is a threat to public health. The following map shows the distribution of people with arsenic levels $>$ 10 $\mu g/L$ by US counties. As an example, Maine is one of the few states with a high level of arsenic. Let's consider a remote northern Maine community where no public water system is available, the drinking water is from private wells. What is the probability that a long-term resident of the community has an arsenic level of more than 10 $\mu g/L$? Apparently, that probability is NOT 0.5

(i.e., this is not an equally likely outcome experiment). We survey many residents to measure the arsenic level for each selected resident and record whether the arsenic level is higher than 10 $\mu g/L$. The desired probability is approximated by the relative frequency of residents with an arsenic level higher than 10 $\mu g/L$.

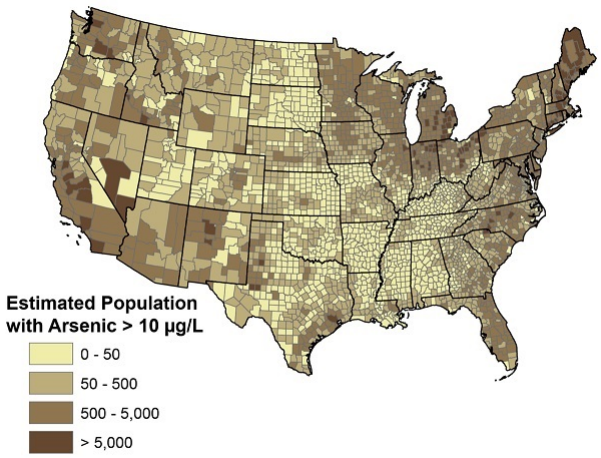


Figure 5.3: Distribution of people with arsenic level > 10 ug/L by US counties

5.4 Probability distribution of random variables

The probability distribution of a population (or random variable) contains **all information** in the population (random variable). The primary questions we need to answer frequently about the distributions.

- **Finding probabilities**: For any given two values, say x_1 and x_2, (including one or both of the extreme values), we can find the probability $P(x_1 < X < x_2)$.
- **Finding quantiles (specific values of the random variable)**: For any value x (including one of the extreme values) and a probability, say p_0, of a well-defined event, we can find the specific value of Y, say x_0 that was used to define the valid event, from the equation $P(x < X < x_0) = p_0$ or $P(x_0 < X < x) = p_0$.

There are types of random variables: discrete random variables and continuous random variables. The probability distribution of a random variable provides a way to find the probability of an event defined by a value or a set of values of the random variable.

Table 5.1: Probability Distribution Table

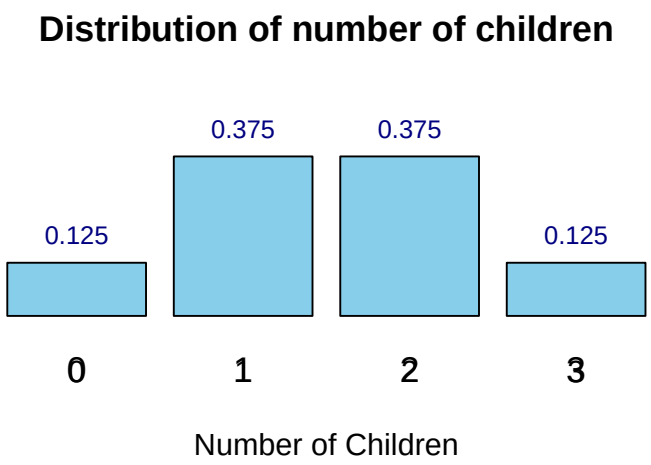
Y	Prob
0	0.125
1	0.375
2	0.375
3	0.125

5.4.1 Discrete probability distribution

The probability distribution of a discrete random variable is a description of the relative frequencies of the corresponding district values.

- **Example 11**: Refer to **Example 9**, Let Y be the number of children. Y has 4 possible values: 0, 1, 2, 3. Then the probability of each distinct value of Y is summarized in the following table.

With the above table, we can find the probability of all events defined based on the values of random variable Y. The above table is called the **probability distribution table**. The following graphic representation of the probability distribution table is called the **probability distribution histogram**.



For example, with the above table, we can two types of questions.

- **Finding probabilities**
  - $P(Y < 2) = P(0 < X < 2) = P(Y = 0) + P(Y = 1) = 0.125 + 0.375 = 0.5$.
  - $P(Y > 2) = P(2 < X \leq 3) = P(Y = 3) = 0.125$.

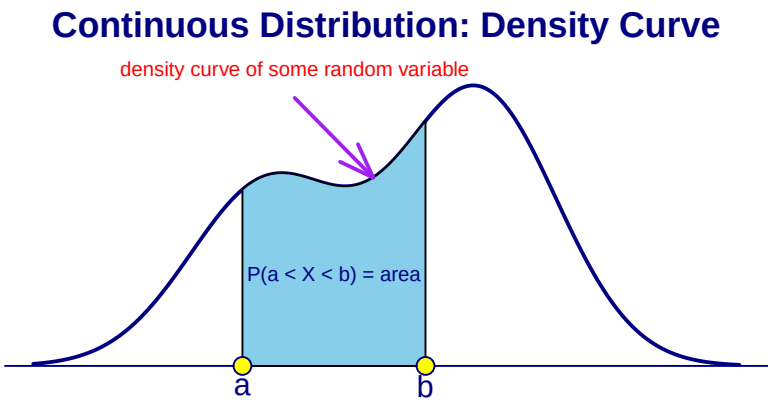
- **Finding quantiles**
  - $P(0 < X < x_0) = 0.5 \rightarrow x_0 = 1$.
  - $P(1 < X < x_0) = 0.375 \rightarrow x_0 = 0.375$.

Caution: The above two types of questions can be complicated in discrete distribution. This blog post explains this complexity in some detail. We will NOT use the discrete distribution directly in this class, instead, we use will the normal distribution (in the next section).

Remark: We can see from the above examples that the definition of an event associated with a discrete random variable is a value or set of values.

5.4.2 Continuous probability distribution

A continuous random variable has uncountably many distinct values. That is, for any two distinct values of the continuous random variable, no matter how close they are, there are still uncountably many values in between. because this property, **an event associated with a continuous random variable is defined to be an interval or the union of some intervals of the values of the random variable**.



Unlike in the case of the discrete random variable in which the height of the vertical bar in the probability histogram is defined to be the probability of observing that corresponding value, **for any continuous random variable**, we define the probability of an event that is defined based on an interval $[a, b]$ to be

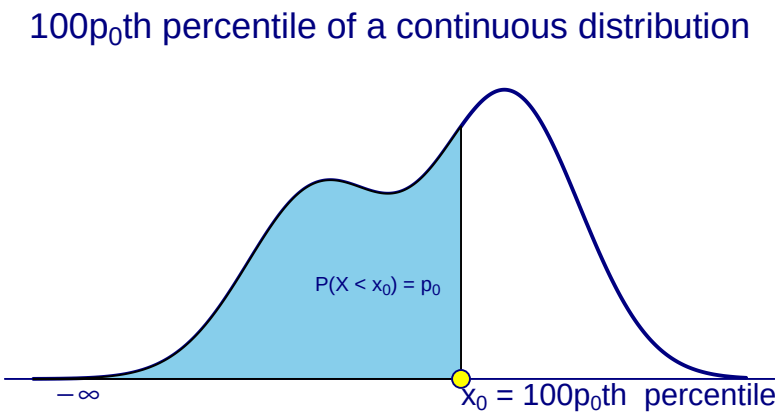
- \$P(a < X < b) = \$ the area of the region defined by a, b and the density curve (see the above figure).

- As a special case, $P(X = c) = 0$. Moreover, $P(X = c_1, c_2, c_3, \cdots) = 0$. That is, **the probability of observing countably many values of a continuous random variable is ALWAYS ZERO!**.

- **Two Basic Types of Questions**: finding probabilities and quantiles.

  - *Finding probabilities*: for any given two values (including possibly one of or both $-\infty$ and ∞), say a, b, then $P(a < X < b) =$ area of the shaded region as shown in the above Figure. As a special case, if $a = b$, then $P(X = a) = 0$.

  - *Finding quantiles*: for a given value, say x, of X (possibly including ∞ or $-\infty$) and a probability, say p_0, we can find the other value x_0 from $P(x < X < x_0) = p_0$ if $(x \leq x_0)$. As a special case, if the given value is ∞ and $-\infty$ (both are not valid values of X!), then x_0 that satisfies $P(X < x_0) = P(-\infty < X < x_0) = p_0$ is called the $100p_0^{th}$ quantile (see the following figure).



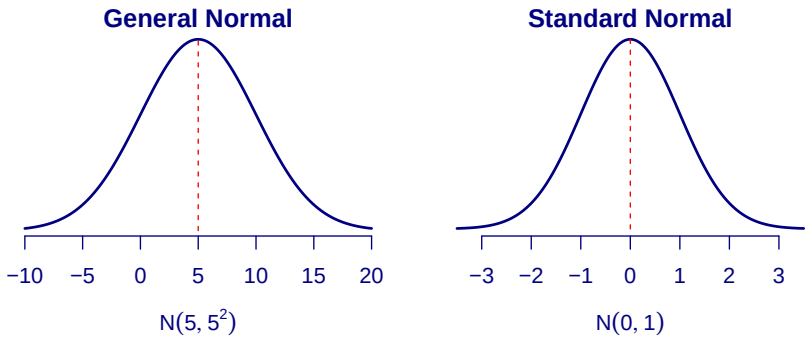
In the next section, we will introduce several distributions of special continuous random variables.

5.5 Special Continuous Distributions

Four distributions will be used in this course. We introduce the first two of them: normal distribution and t-distribution.

5.5.1 Normal Distribution

The general normal distribution has a bell-shaped distribution as shown in the following figure. A normal distribution is uniquely determined by its mean and variance. We usually use notation $N(\mu, \sigma^2)$, where μ and σ^2 are the mean and variance of the normal distribution. When $\mu = 0$ and $\sigma^2 = 1$, the normal distribution $N(0, 1)$ is called the **standard normal distribution**. The



- **Two Types of Questions in Normal Distribution** are related to the left-tail area and quantile. R has two functions for finding **left-tail area** and **quantile** for any given normal distribution.

```
pnorm(quantile, mean, sd)
# The above function finds the left-tail area for a given quantile.
# mu = mean, sd = standard deviation.
qnorm(left.tail.prob, mean, sd)
# The above function finds the quantile for a given left-tail area.
# mu = mean, sd = standard deviation.
```

- **Example 12**: We find the left-tail area of general normal distributions $X \to N(16, 4^2)$. $P(X < 13) = ?$

```
normal.left.tail <- round(pnorm(13, mean = 16, sd = 4),4)
normal.left.tail
```

```
## [1] 0.2266
```

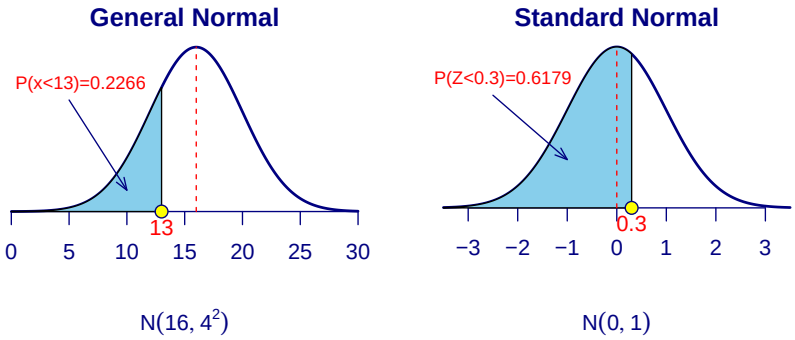
*__Example 13__: We find the left-tail area of the standard normal distributions $X \to N(0, 1)$. $P(X < 0.3) = ?$

```
sd.norm.left.tail <- round(pnorm(0.3, mean = 0, sd = 1),4)
sd.norm.left.tail
```

```
## [1] 0.6179
```

We next give a graphical representation of the tail area in **Example 12** and

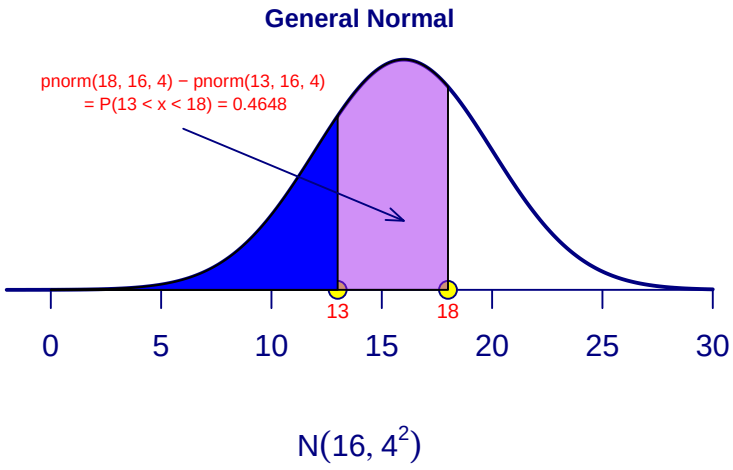
Example 13.



- **Example 14**: Let $X \to N(16, 4^2)$. Find $P(13 < X < 18) = ?$ The following R code calculates the probability. Note that `P(X<18) = pnorm(18, 16, 4)` and `P(X<13) = pnorm(13, 16, 4)`. Therefore `P(13 < X < 18) = pnorm(18, 16, 4) - pnorm(13, 16, 4)`. The probability is the area of the purple region. See the following Figure.

```
p.18 = pnorm(18, 16, 4)
p.13 = pnorm(13, 16, 4)
p.13.to.18 = round(p.18 - p.13, 4)
p.13.to.18
```

```
## [1] 0.4648
```



- **Example 15**: Consider the standard normal distribution $N(0, 1)$. Find z_0 if $P(Z > z_0) = 0.2345$. The answer is given in the following R code: `qnorm(0.2345, 0, 1)` or simply `qnorm(0.2345)`. The latter form does not specify `mean=` and `sd=` since the default `mean = 0` and `sd = 1`.

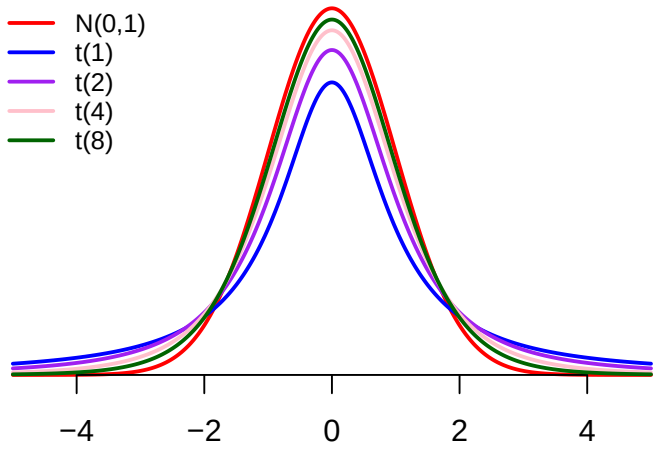
```
qnorm(0.2345, 0, 1)
```

```
## [1] -0.7241071
```

That is, 23.45% of the values in the standard normal population are less than or equal to −0.7241 and 23.45% are greater than or equal to −0.7241.

5.5.2 t distribution

The t distribution is symmetric with respect to the vertical axis with a mean of 0. The shape is **ALWAYS** flatter than the **standard normal distribution**. The shape of a t distribution is uniquely determined by the **degrees of freedom**. The following chart describes the relationship between the standard normal and t distributions.



R has two functions for finding left-tail area (also called tail probability) and quantile:

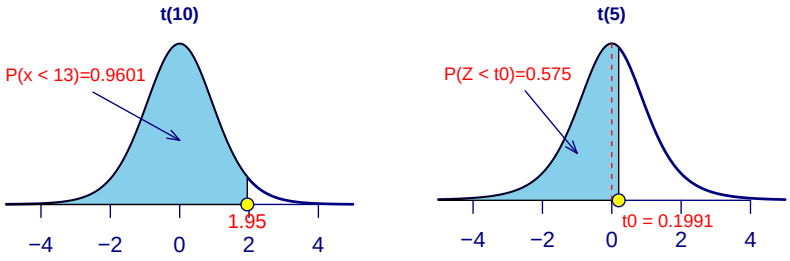
```
pt(quantile, df)          # left-tail probability (left-tail area)
qt(left-tail-area, df)    # quantile
```

- **Example 16**: Answer the following questions about t-distributions.
  - Consider $t(10)$, t-distribution with 10 degrees of freedom. $P(T < 1.95) = ?$
  - Consider $t(5)$, t-distribution with 10 degrees of freedom. What is t_0 if $P(T < t_0) = 0.575$?

```
# problem 1
problem01 = pt(1.95, df=10)
# Problem 2
problem02 = qt(0.575, df = 5)
## display the two result
cbind(problem01 = problem01, problem02 = problem02)
```

```
##      problem01 problem02
## [1,] 0.9601258 0.1991374
```

The above results are also reflected in the following figures.



5.6 Summary

In this module, we introduced some basic concepts of probability, random variables, and distributions of random variables. Left-tail probability and quantile of normal and t distributions are the most important quantities that will be used in this course. We also introduced R functions to calculate left-tail the probability and the quantile of normal and t distributions. These R functions are summarized in the following.

```
# normal distribution
pnorm(quantile, mean, sd) # for the left-tail probability of normal distribution
qnorm(left.tail.prob, mean, sd) # for the quantile of normal distribution
```

```
# t-distribution
pt(quantile, df)        # for the left-tail probability of normal distribution
qt(left.tail.prob, df)  # for the quantile of t distribution
```

5.6.1 Numerical Examples Based on Normal and t Distributions

- **Example 17**: In the United States, males between the ages of 40 and 49 eat on average 103.1 g of fat every day with a standard deviation of 4.32 g ("What we eat," 2012). Assume that the amount of fat a person eats is normally distributed.

  1. State the random variable.
  2. Find the probability that a man in the age group of 40-49 in the U.S. eats more than 110 g of fat every day.
  3. Find the probability that a man in the age group of 40-49 in the U.S. eats less than 93 g of fat every day.
  4. Find the probability that a man in the age group of 40-49 in the U.S. eats less than 65 g of fat every day.
  5. If you found a man in the age group of 40-49 in the U.S. who says he eats less than 65 g of fat every day, would you believe him? Why or why not?
  6. What daily fat level do 5% of all men in the age group of 40-49 in the U.S. eat more than?

Solution: We use R to find the answers to the above questions.

1. Y = amount of fat a person eats every day. Calculations of problems 2 - 6 are given in the following R code chunk.

```
p2 = 1- pnorm(110, mean =103.1, sd = 4.32)  # 1-left-tail-prob = right-tail-prob
p3 = pnorm(93, mean =103.1, sd = 4.32)      # left-tail probability
p4 = pnorm(65, mean =103.1, sd = 4.32)      # left-tail probability
p5 = pnorm(65, mean =103.1, sd = 4.32)      # left-tail probability
p6 = qnorm(1-0.05, mean =103.1, sd = 4.32)  # quantile
ans = cbind(problem2 = round(p2,4), problem3 =  round(p3,4),
            problem4 = round(p4,4), problem5 =  round(p5,4),
            problem6= p6)
row.names(ans) ="prob or quantile"
kable(t(ans))
```

	prob or quantile
problem2	0.0551
problem3	0.0097
problem4	0.0000
problem5	0.0000
problem6	110.2058

2. The probability that a man in the age group of 40-49 in the U.S. eats more than 110 g of fat every day is 5.51%.
3. The probability that a man in the age group of 40-49 in the U.S. eats less than 93 g of fat every day is 0.997%.
4. The probability that a man in the age group of 40-49 in the U.S. eats less than 65 g of fat every day is close to 0.00%.
5. I will not believe a man in the age group of 40-49 in the U.S. who says he eats less than 65 g of fat every day because the chance of eating less than 65 g in that age group is almost 0.00%.
6. 5% of all men in the age group of 40-49 in the U.S. eat more than 110.2058 g.

- **Example 18**: The mean cholesterol levels of women aged 45-59 in Ghana, Nigeria, and Seychelles is 5.1 mmol/l and the standard deviation is 1.0 mmol/l (Lawes, Hoorn, Law & Rodgers, 2004). Assume that cholesterol levels are normally distributed.

    1. State the random variable.
    2. Find the probability that a woman aged 45-59 in Ghana, Nigeria, or Seychelles has a cholesterol level above 6.2 mmol/l (considered a high level).
    3. Find the probability that a woman aged 45-59 in Ghana, Nigeria, or Seychelles has a cholesterol level below 5.2 mmol/l (considered a normal level).
    4. Find the probability that a woman aged 45-59 in Ghana, Nigeria, or Seychelles has a cholesterol level between 5.2 and 6.2 mmol/l (considered borderline high).
    5. If you found a woman aged 45-59 in Ghana, Nigeria, or Seychelles having a cholesterol level above 6.2 mmol/l, what could you conclude?
    6. What value do 5% of all women ages 45-59 in Ghana, Nigeria, or Seychelles have a cholesterol level less than?

Solution

1. Random variable Y = mean cholesterol levels of women aged 45-59 in Ghana, Nigeria, and Seychelles.

The calculation of problems 2 - 6 is given in the following R code chunk.

```
p2.2 = 1- pnorm(6.2, mean = 5.1, sd = 1)      # 1-left-tail-prob = right-tail-prob
p2.3 = pnorm(5.2, mean = 5.1, sd = 1)         # left-tail probability
p2.4 = pnorm(6.2, mean = 5.1, sd = 1) -  pnorm(5.2, mean = 5.1, sd = 1)
p2.5 = 1 -pnorm(6.2, mean = 5.1, sd = 1)      # left-tail probability
p2.6 = qnorm(0.05, mean = 5.1, sd = 1)      # quantile
ans.p2 = cbind(problem2 = round(p2.2,4), problem3 =  round(p2.3,4),
               problem4 = round(p2.4,4), problem5 =  round(p2.5,4),
               problem6 = p2.6)
row.names(ans.p2) ="prob or quantile"
```

```
kable(t(ans.p2))
```

	prob or quantile
problem2	0.135700
problem3	0.539800
problem4	0.324500
problem5	0.135700
problem6	3.455146

2. The probability that a woman aged 45-59 in Ghana, Nigeria, or Seychelles has a cholesterol level above 6.2 mmol/l is 13.57%.
3. Find the probability that a woman aged 45-59 in Ghana, Nigeria, or Seychelles has a cholesterol level below 5.2 mmol/l is 53.98%.
4. Find the probability that a woman aged 45-59 in Ghana, Nigeria, or Seychelles has a cholesterol level between 5.2 and 6.2 mmol/l 32.45%.
5. If a woman aged 45-59 in Ghana, Nigeria, or Seychelles has a cholesterol level above 6.2 mmol/l, he has a high cholesterol level.
6. 5% of all women ages 45-59 in Ghana, Nigeria, or Seychelles have a cholesterol level of less than 3.455 mmol/l.

5.7 Assignment - Probability Distributions

We focus on the concepts of random variables and their corresponding distributions. Two basic types of questions are will face very often in this course are to find (1) left-tail probability (area under the density curve) and (2) quantile under given conditions for all continuous distributions. We present numerical examples of normal and t distributions in the last part of the class note. In this assignment, you will do problems similar to the last two examples. You can use the codes provided in the class note [HTML or PDF]. I also include RMD on the course web page but not required. In case some of you want to learn more advanced R coding, the source RMD is useful.

kable() function is in library **{knitr}**. If you have not installed this package on your computer, you will receive an error `could not find function "kable" calls:`. The following screenshot shows how to install a package on your computer.

This assignment focuses on the two types of questions using normal distributions. Please prepare an R Markdown document to complete the assignment.

- **Problem 1**

The size of fish is very important to commercial fishing. A study conducted in 2012 found the length of Atlantic cod caught in nets in Karlskrona to have a mean of 49.9 cm and a standard deviation of 3.74 cm (Ovegard, Berndt & Lunneryd, 2012). Assume the length of fish is normally distributed.

1. State the random variable.

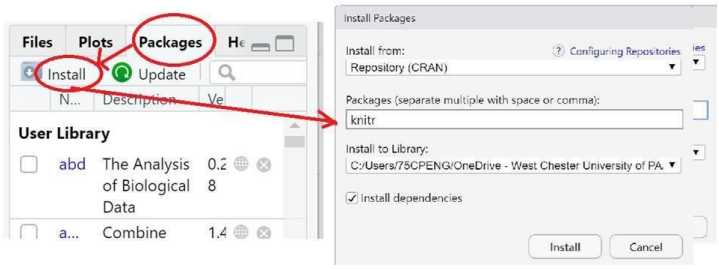


Figure 5.4: Install Packages on computer.

2. Find the probability that an Atlantic cod has a length of less than 52 cm.
3. Find the probability that an Atlantic cod has a length of more than 74 cm.
4. Find the probability that an Atlantic cod has a length between 40.5 and 57.5 cm.
5. If you found an Atlantic cod to have a length of more than 74 cm, what could you conclude?
6. What length are 15% of all Atlantic cod longer than?

- **Problem 2**

The mean yearly rainfall in Sydney, Australia, is about 137 mm and the standard deviation is about 69 mm ("Annual maximums of," 2013). Assume rainfall is normally distributed.

1. State the random variable.
2. Find the probability that the yearly rainfall is less than 100 mm.
3. Find the probability that the yearly rainfall is more than 240 mm.
4. Find the probability that the yearly rainfall is between 140 and 250 mm.
5. If a year has a rainfall of less than 100mm, does that mean it is an unusually dry year? Why or why not?
6. What rainfall amounts are 90% of all yearly rainfalls more than?

Chapter 6

Sampling Distributions

In this section, we study the random behavior of quantities obtained from random samples using the probability distributions introduced in the previous weeks.

Some technical terms:

- **Population** - set of all subjects of interest. For example, if we want the average height of WCU students, then the set of all WCU students is the **population**.
- **Sample** - a subset of subjects of the population. For example, all students in the Department of Biology form a subset of all WCU students. In other words, the set of all students in Biology is a **sample**. **However**, this **sample** does not represent the **population** since WCU has many other majors.
- **Random Sample** - a subset of subjects that represent the population. For example, we can use one of the methods introduced in week #2 to collect a subset from the WCU student population - to obtain a random sample.
- **Parameter** - numerical characteristic of the population. For example, the average height of the WCU student population is denoted by μ. Apparently, μ is unknown but fixed.
- **Statistic** - numerical characteristic of the population calculated from the random sample. For example, we select a random sample of 150 students from WCU and find the average height, denoted by $\bar{X}$. Apparently, $\bar{X}$ is random since its value is dependent on the random sample.

Since $\bar{X}$ is a random variable, we use probability to characterize the random behavior of $\bar{X}$.

6.1 Concepts of the sampling distribution

We see from the examples in the previous section that the sample mean is a random variable. In fact, all population parameters evaluated at a random sample taken from the population are random variables. To characterize the behavior, we need to use probability distributions.

- **A sampling distribution** is the distribution of a **sample statistic** such as sample mean, sample variance, sampling coefficient of variation, sample correlation coefficient, etc.

A population has many different numerical characteristics that require different probability distributions to characterize them. In this course, we focus on the mean and proportion and some coefficients of regression that affect the mean and proportion. In the next few modules, we introduce procedures for constructing confidence intervals and testing hypothesis inferences for population means and proportions.

In the next sections, we introduce sampling distributions of sample means and proportions under different assumptions.

6.2 Sampling Distribution of Sampling Means

Inferential statistics is all about making inferences about population parameters by using the information of individual subjects. The estimated population parameters could be used to predict the individual level.

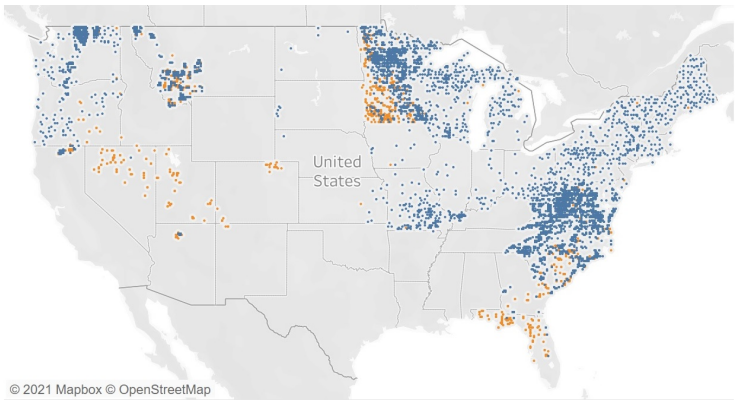
The information in the estimate is dependent on the amount of information in the sample and the population as well. In the next few sections, we introduce the sampling distribution of sample means under different assumptions.

6.2.1 Working Data Set: Plant Diversity

This data set includes the geographic location (lat/lon) for 15,136 plots, as well as the herbaceous species richness, climate, soil pH, and other variables related to the plots.

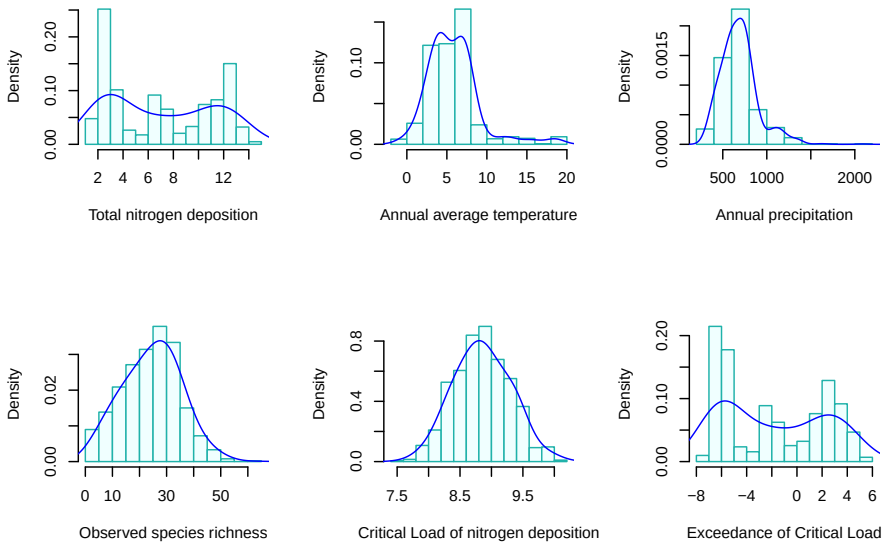
This data set is associated with the following publications: Simkin, S., C. Clark , W. Bowman, E. Allen, J. Belnap, and L. Pardo. The conditional vulnerability of plant diversity to atmospheric nitrogen deposition across the United States. PNAS (PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES). National Academy of Sciences, WASHINGTON, DC, USA, 113(15): 4086-4091, (2016).

```
knitr::include_graphics("img05/w05-site-map.jpg")
```



We choose a special subset that contains data associated with **Perennial graminoid vegetation**. That has 1152 records. The distributions of 6 numerical variables are given in the following histograms.

```
plant="https://raw.githubusercontent.com/pengdsci/STA501/main/Data/w05-plant-diversity.csv"
plant.diversity = read.csv(plant, header = TRUE)
```



We can see from the above histograms that the **Critical Load of nitrogen deposition** is close to the normal distribution. Other variables are **not** normally distributed. In the rest of this note, I will assume the above data sets to be "populations" and take random samples from appropriate populations listed above in different examples. Some of the following examples will be based on the above populations. You may want to use the distributional information (the

shapes of histograms) to choose appropriate sampling distributions of sample statistics based on the sample taken from one of the above populations.

6.2.2 Normal Population with given mean (μ) and Variance (σ^2)

Result #1: Assume that X is a normal random variable with an unknown mean μ and the known variance σ^2. That is,

$$X \to N(\mu, \sigma^2)$$

Let random sample $\{x_1, x_2, \cdots, x_n\} \to N(\mu, \sigma_0^2)$, where population mean μ is unknown and variance σ_0^2. Then the sample mean

$$\bar{X} = \frac{\sum_{i=1}^{n}}{n}$$

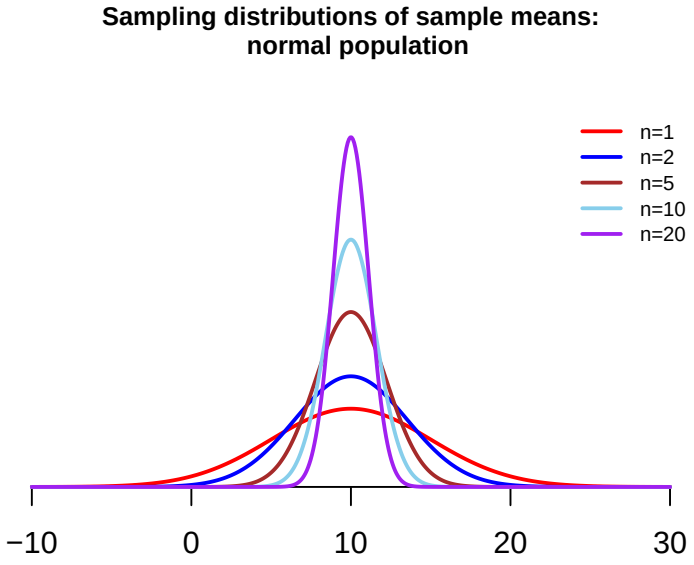
is normally distributed with mean μ and variance σ^2/n. In other words,

$$\bar{X} \to N(\mu, \sigma_0^2/n)$$

Remarks: The simple comparison of X and $\bar{X}$.

- The means of the distributions of X and $\bar{X}$ are equal.
- The variances of the distributions of X and $\bar{X}$ are **not** equal. In fact, the variance of $\bar{X}$ is less than the variance of X.
- As the sample size increases, the variance of $\bar{X}$ decreases since the denominator of the variance of $\bar{X}$ contains the sample size n.
- Since the R functions **pnorm()** and **qnorm()** require the standard deviation as an argument.

The following figure shows the variance of the sample means with different sample sizes.



The above figure shows that, as sample size increases, the variance of $\bar{X}$ decreases (the density curve becomes skinnier as sample size increases).

Example 1. We use **Critical Load (CL) of nitrogen deposition** in the **plant diversity** data set. Assume that **Critical Load (CL) of nitrogen deposition** follows a normal distribution with a mean of $\mu = 10$ and known variance $\sigma^2 = 0.2$. Use this information to answer the following questions.

1. If a random of 20 Critical Load (CL) of nitrogen deposition sample values are taken from the population, what is the probability that the mean critical load nitrogen deposition ($\bar{X}$) is greater than 10.1 kg N/ha/yr?

2. What is the level of critical load (CL) of nitrogen deposition that is higher than the 95% mean level of critical load (CL) of nitrogen deposition with the same sample size 20?

Solution: Since the critical load (CL) of nitrogen deposition is a normal population, therefore, the sampling distribution of $\bar{X} \rightarrow N(10, .2/20)$. The solutions to problems are based on this sampling distribution.

1. $P(\bar{X} > 10.05)$ is equal to the right tail area of $N(10, (\sqrt{0.2/20})^2$. Note that, by default, the R function **pnorm()** only gives the left-tail area. The desired probability is equal to $1 - left.tail.area$. The R code is given in the following.

```
1-pnorm(10.1, mean = 10, sd = sqrt(0.2/20))
```

```
## [1] 0.1586553
```

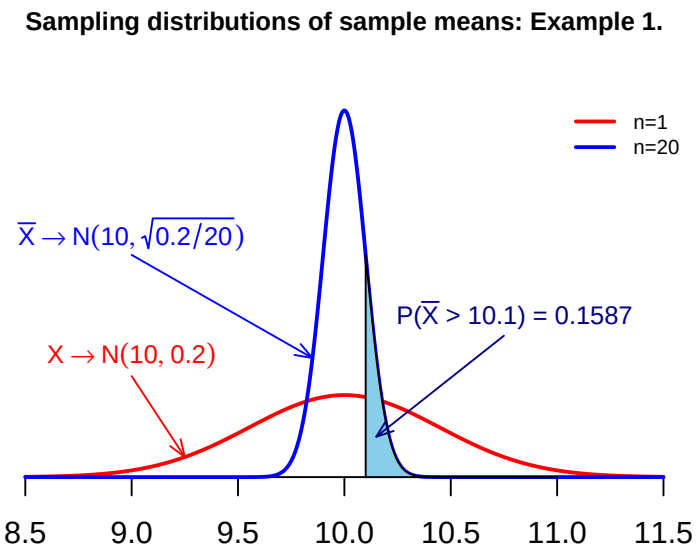
Therefore, $P(\bar{X} > 11.5) = 0.1587$. The tail probability is labeled in the following figure.

2. the desired cut-off if actually the 95% quantile of the distribution of $\bar{X}$ which can be found by **qnorm()**.

```
qnorm(0.95, mean = 10, sd = sqrt(0.2/20))
```

```
## [1] 10.16449
```

Therefore the 95% quantile is 10.16449.



6.2.3 Normal Population with a given mean (μ) and an Unknown Variance (σ^2)

If the population variance is unknown, then we have to use sample variance to characterize the distribution $\bar{X}$. Unlike in the case of the normal population with known variance in which we can specify the sampling distribution $\bar{X}$ directly, we cannot but have the following result based on the standardized statistic.

Result #2: Let random sample $\{X_1, X_2, \cdots, X_n\} \to N(\mu, \sigma^2)$ and $\bar{X}$ and s^2 be the sample mean and sample variance, respectively. then we have

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \to t_{n-1}$$

Where t_{n-1} is a t-distribution with $n-1$ degrees of freedom. The two basic types of questions associated with $t-$distribution were discussed in the previous module.

Example 2. Assume that the sample of 11 levels of critical load (CL) of nitrogen deposition $\{8.67, 9.38, 9.27, 8.56, 8.72, 8.83, 9.72, 8.36, 8.76, 8.91, 9.49\}$ is randomly selected from all sites in the study. Its **sample standard deviation** is 0.430. What is the percent of the sample means based on the same sample size 11 will be bigger than 10.2? Note that the population mean is $\mu = 10$.

Solution The question to answer is about the sample mean $\bar{X}$ with size $n = 11$. Since the population variance is unknown, $\bar{X}$ is not a normal distribution.

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \to t_{n-1}$$

Therefore,

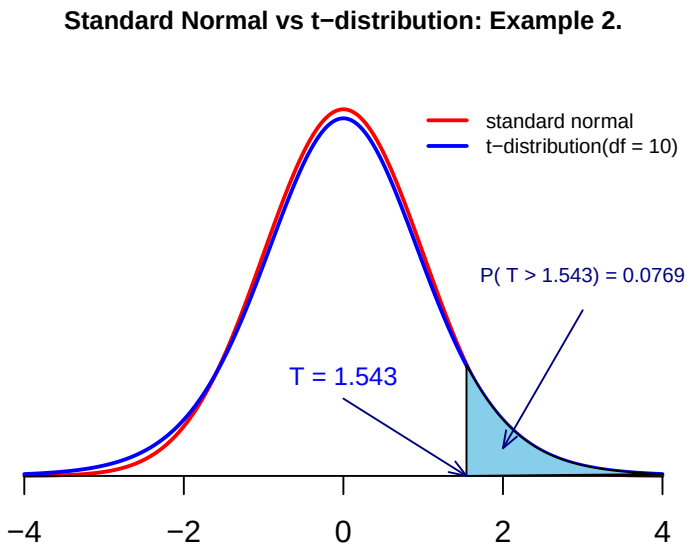
$$P(\bar{X} > 10.2) = P\left(\frac{\bar{X} - \mu}{s/\sqrt{n}} > \frac{10.2 - 10}{0.43/\sqrt{11}}\right) = P(T > 1.543)$$

As we know, $P(T > 1.543)$ is the right-tail area of $t_{11-1} = t_{10}$. The desired probability is $1-$ left-tailed area. The following R code finds the above probability.

```
1- pt(1.543, df = 11-1)
```

```
## [1] 0.07693013
```

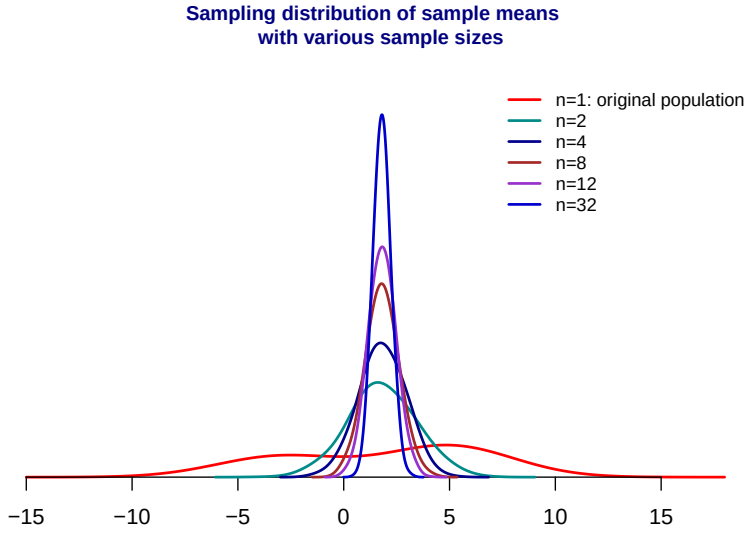
Therefore, $P(T > 1.543) = 0.0769$.



6.2.4 Unspecified Population with a given mean (μ)and an unspecifed variance (σ^2)

In the previous two cases, we assume the population to be normal. Quite frequently we need to deal with a population without a specified distribution of real-world applications. This means that the population distribution is unknown. As usual, we assume the population mean is known. The population variance is either given or unknown. If we still want to use the sampling distribution of sample means to make inferences about the population mean, what result(s) we can use to characterize the sampling distribution of the sample mean?

In this section, we discuss this type of sampling distribution of sample means under certain conditions. Before we present the result, we perform a simulation to show the patterns of the sampling distributions of sample means using various sample sizes. We simulate a non-normal distribution with two peaks and then take samples from that population with different sample sizes, and use these sample means to estimate the corresponding density curves.



The figure shows that, as the sample size increases, the sampling distribution of the sampling mean approaches a normal distribution regardless of the distribution of the original population. The following theorem explicitly specifies the sampling distribution.

Results #3: **Central Limit Theorem:** Let $X \to (\mu, \sigma^2)$ (**caution**: the population is not necessarily to be normal.). Let $\bar{X}$ be the sample mean with size n. If n is large, the sampling $\bar{X}$ is **approximately** normally distributed with a mean $\mu_{\bar{X}} = \mu$ and variance $\sigma^2_{\bar{X}} = \sigma^2/n$. That is,

$$\bar{X} \to N\left(\mu, \frac{\sigma^2}{n}\right).$$

Comments: The following are comments related to the normal approximations.

- The baseline population is not specified in the theorem. The distribution could be discrete or continuous.
- When the sample size is large, the sample mean is approximately normally distributed. The question is how large is called "large"? As a convention, we call the sample size large if $n > 30$.

Example 3. The blood cholesterol levels of a population of workers have a mean of 202 and a standard deviation of 14. If a sample of 36 workers is selected, approximate the probability that the sample mean of their blood cholesterol levels will lie between 198 and 206.

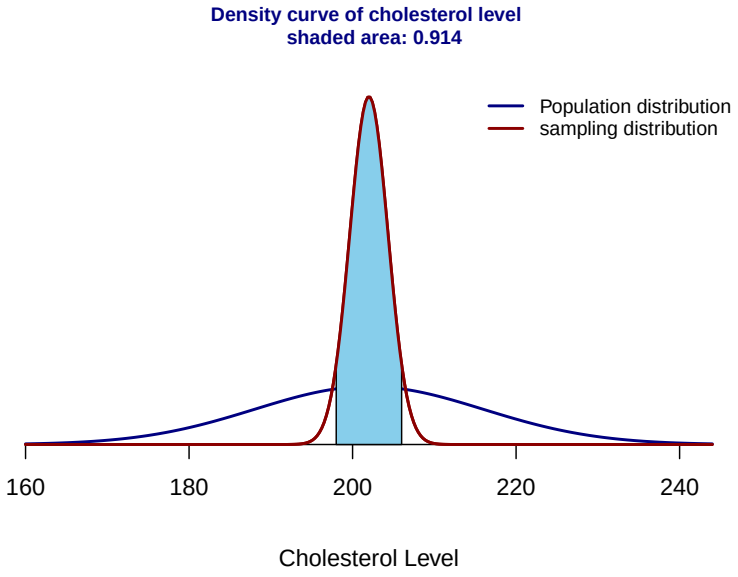
Solution Let X be the blood cholesterol levels of a population of workers. Then $X \to (\mu = 202, \sigma = 14)$. Since $n = 36 > 30$, using the C.L.T, we have $\bar{X} \to N(202, 14/\sqrt{36})$. Therefore,

$$P(198 < \bar{X} < 206) = P(\bar{X} < 206) - P(\bar{X} < 198) = 0.9135237$$

```
pnorm(206, mean = 202, sd = 14/sqrt(36)) - pnorm(198, mean = 202, sd = 14/sqrt(36))
```

```
## [1] 0.9135237
```

The above probability is the area of the shaded area in the following figure.



6.3 Sampling distribution of sample proportions

For a population of binary data that only takes on exactly two possible values such as "success" vs "failure", "diseased" vs "disease-free", etc., its distribution is uniquely determined by the proportion of one of the categories.

Let $X = \{x_1, x_2, \cdots, x_n\}$ be a random sample taken from a population of binary data taking on only two possible values "success" or "failure". That is, $x_i =$ "success" or "failure", for $i = 1, 2, \cdots, n$. If we are interested in the proportion of "success" of the population, we can do the following numerical coding on the sample: 1 = "success" and 0 = "failure". With this numerical coding, we calculate the sample mean

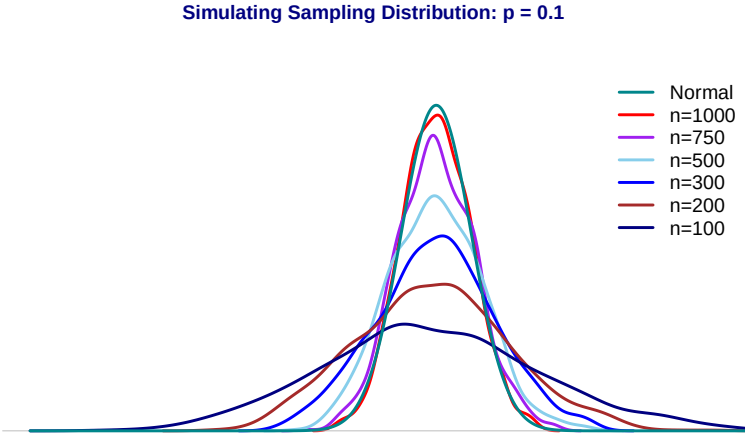
$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\#1s}{n} = \frac{\#succeses}{n} = proportion.of.successes.$$

Therefore, the sample proportion is actually a **sample mean**. Therefore, according to the central limit theorem, $\hat{p}$ is approximately normally distributed under certain conditions. This idea is formalized in the following result.

Result #4. Let p be the proportion of one of the categories, say "success" and $\hat{p}$ the sample proportion based on a random sample with size n. If $np > 5$ and $n(1-p) > 5$, then

$$\hat{p} \to N\left(p, \sqrt{\frac{p(1-p)}{n}}\right).$$

In some applications, the standard deviation $\sqrt{p(1-p)/n}$ is estimated by $\sqrt{\hat{p}(1-\hat{p})/n}$.



The above simulated sampling distributions with samples taken from a binary population with true $p = 10\%$ use different sample sizes. We see that as the sample size increases, the sampling distribution approaches the normal distribution.

Example 4 The frequency of color blindness (dyschromatopsia) in the Caucasian American male population is estimated to be about 8%. We take a

random sample of size 125 from this population. What is the probability that more than 9% of the 125 subjects have color blindness?

Solution Since $np = 125 \times 0.08 = 10 > 5$, $n(1-0.08) = 115 > 5$, the C.L.T can be used to approximate the sampling distribution of the sample proportion to the normal distribution as outlined in the **result #3**. That is,

$$\hat{p} \to N\left(0.08, \sqrt{\frac{0.08(1-0.08)}{125}}\right) = N\left(0.08, 0.0243\right).$$

The probability we want to find is $P(\hat{p} > 0.09) = 1 - P(\hat{p} < 0.09) = 0.34$

```
1 - pnorm(0.09, mean = 0.08, sd = 0.0243)
```

```
## [1] 0.3403447
```

Example 5 Suppose 60% of seniors who get flu shots remain healthy, and independent from one person to the next. If we selected a random sample from the complex of 100, what is the probability that the sample proportion will be greater than 50%?

Solution: We are given that $p = 0.6$. Since $np = 200 \times 0.6 = 120 > 5$ and $n(1-p) = 200 \times 0.4 = 80 > 5$. Using **result #3**, we have

$$\hat{p} \to N\left(0.6, \sqrt{\frac{0.6(1-0.6)}{200}}\right) = N(0.6, 0.0346).$$

Therefore, $P(\hat{p} > 0.5) = 1 - P(\hat{p} < 0.5) = 0.2859$

```
1-pnorm(0.6, 0.0346)
```

```
## [1] 0.2859009
```

6.4 Summary

In this module, we introduced the sampling distribution of sample means and proportions under various assumptions. Let random sample $\{x_1, x_2, \cdots, x_n\}$ be taken from a population with sample mean

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

The sampling distribution of $\bar{X}(\hat{p})$ under various conditions is summarized in the following.

- **Sampling distribution of sampling means $\bar{X}$**

– Population is normal with known variance (σ_0^2).

$$\bar{X} \to N\left(\mu, \frac{\sigma_0^2}{n}\right) \Rightarrow \frac{\bar{X}-\mu}{\sigma_0/\sqrt{n}} \to N(0,1).$$

– Population is normal with unknown variance (σ^2).

$$\bar{X} \to N\left(\mu, \frac{\sigma^2}{n}\right) \Rightarrow \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \to N(0,1).$$

However, if the sample variance s^2 is used,

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \to t_{n-1}.$$

– Population is unspecified: **if the sample size is large**,

$$\bar{X} \to N\left(\mu, \frac{\sigma^2}{n}\right) \Rightarrow \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \to N(0,1) \quad \text{and} \quad \frac{\bar{X}-\mu}{s/\sqrt{n}} \to N(0,1).$$

where s is the sample standard deviation.

- **Sampling distribution of proportions ($\hat{p}$)**: if $np > 5$ and $n(1-p) > 5$, then

$$\hat{p} \to N\left(p, \frac{p(1-p)}{n}\right) \Rightarrow \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \to N(0,1).$$

Note also that

$$\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \to N(0,1).$$

6.5 Assignment - Sampling Distributions

This assignment focuses on the applications of sampling distributions. Please check the assumptions of each of the four results and compare them with the information given in each of the problems in the following to determine the correct sampling distribution.

- **Problem 1**

Suppose it is known that in a certain large human population, cranial length is approximately **normally distributed** with a mean of 185.6mm and a standard deviation of 12.7 mm.

(1). Find the mean and standard deviation of the sampling distribution respectively.

(2). What is the probability that a random sample of size 10 from this population will have a mean greater than 190?

- **Problem 2**

The National Health and Nutrition Examination Survey of 1988–1994 (NHANES III) estimated the mean serum cholesterol level for U.S. females aged 20–74 years to be 204 mg/dl. The estimate of the standard deviation was approximately 44. Using these estimates as the mean μ and standard deviation σ for the U.S. population, consider the sampling distribution of the sample mean based on samples of size 49 drawn from women in this age group.

(1). Find the mean and the standard deviation of the sampling distribution respectively

(2). Find the probability that the sample mean serum cholesterol level will be between 170 and 195.

(3). Find the probability that the sample mean serum cholesterol level will be less than 210.

(4). Find the probability that the sample mean serum cholesterol level will be bigger than 195.

- **Problem 3**

Smith et al. performed a retrospective analysis of data on 782 eligible patients admitted with myocardial infarction to a 46-bed cardiac service facility. Of these patients, 248 (32 percent) reported a past myocardial infarction. Use .32 as the population proportion. Suppose 50 subjects are chosen at random from the population.

(1). Find the mean and the standard deviation of the sampling distribution respectively.

(2). What is the probability that over 40 percent would report previous myocardial infarctions?

Chapter 7

Confidence Intervals

Objectives: We want to estimate the population parameters such as mean, standard deviation, and proportion from a random sample and build a confidence interval (also called interval estimate) to show how good the estimate is, finally, we should be able to interpret the estimate.

7.1 Some Terms

- The sample is random. That is, the values in the sample are representative of the whole population.
- **An estimate** is a specific value or range of values obtained from a random sample that is used to approximate a population parameter.
- **A point estimate** is a single value (obtained from the random sample) that is used to approximate a population parameter.

  Example 1: The sample mean $\bar{X}$ is the best point estimate of the population mean μ.

  Example 2: The sample variance s^2 is the best point estimate of the population variance σ^2.
- **The bias of the point estimate** is equal to the difference between the estimate and the true parameter. For example, the bias of the sample mean is $\bar{x} - \mu$. The bias of an estimate measures the accuracy of the estimate.

Since the point estimate of a parameter is obtained from an underlying random sample, it is a random variable. The variance of the point estimate measures the precision of the point estimate of the corresponding population parameter.

The goodness of an estimate – no bias (accurate) and small variance (precise).

7.1.1 What are the issues of a point estimate?

With a point estimate, we can only say that it is close to the true population parameter. The question is how close is called "close"?

Example 1: The distribution of heights of WCU students is approximately normal with mean μ inches and standard deviation σ inches. To estimate μ, we select a random sample $\{x_1, \cdots, x_n\}$

Then

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

is close to the unknown true average height of WCU students. For example, assuming the true mean is 69 inches, a sample of 50 heights yields a sample average of 68.5 inches.

Is 68.5 close to 69?

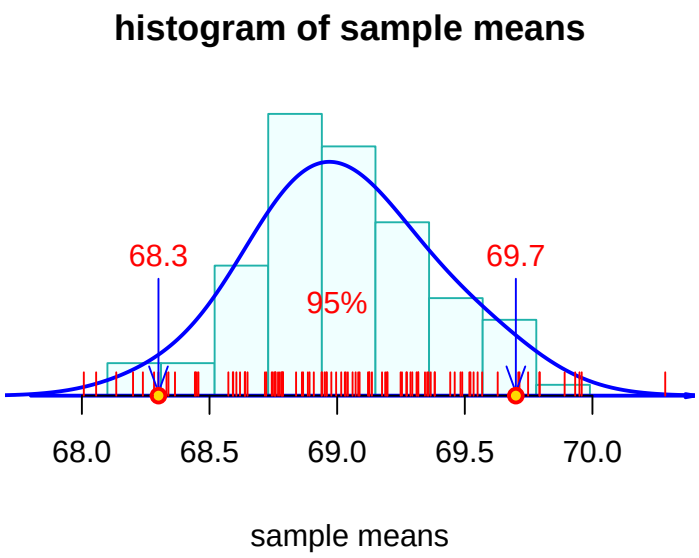
Apparently, we cannot answer the above question with a single sample mean without any additional information.

7.1.2 How About Multiple Samples?

If I invite 100 students to help collect 100 random samples of the same size 50 from the WCU student population. Then we will have the following 100 sample means, say

```
68.2 68.3 68.3 68.4 68.4 68.5 68.6 68.6 68.6 68.6 68.6 68.7 68.7 68.7 68.7 68.7
68.7 68.7 68.8 68.8 68.8 68.8 68.8 68.8 68.8 68.8 68.8 68.8 68.8 68.8 68.8 68.8
68.9 68.9 68.9 68.9 68.9 68.9 68.9 68.9 68.9 68.9 68.9 69.0 69.0 69.0 69.0
69.0 69.0 69.0 69.0 69.0 69.0 69.0 69.1 69.1 69.1 69.1 69.1 69.1 69.1 69.1 69.1
69.1 69.1 69.1 69.2 69.2 69.2 69.2 69.2 69.2 69.2 69.2 69.2 69.2 69.3 69.3 69.3
69.3 69.3 69.3 69.4 69.4 69.4 69.5 69.5 69.5 69.5 69.5 69.5 69.6 69.6 69.6 69.6
69.7 69.7 69.7 69.9
```

Each of the above means is supposed to be close to 69. Can we now answer the question about how close is called "close"? Let's make a histogram of all sample means in the following.



We can see from the above histogram that most of the means are around ("close" to) 69. In fact, 95% of the sample means are within the interval [68.3, 69.7]. If we consider sample means within the interval to be "close" to the true mean, the above interval [68.3, 69.7] reveals the following information.

- The interval has a 95% chance to include the true mean - accuracy and confidence;
- The width of the confidence interval reflects the precision.

Therefore, this interval contains all desired information, but What is more important is that the interval was obtained from the distribution of sample means. In other words, only the sampling distribution of sample mean can provide such intervals that provide both accuracy and precision of the interval for a given confidence level!

7.2 How to find the Confidence Intervals for Means and Proportions?

The above section explained the concept of the confidence interval of the population mean intuitively with an example. This section provides a slightly formal approach to confidence intervals for population means and proportions.

7.2.1 Framework Of Confidence Interval

The ways of constructing confidence intervals for different parameters may be slightly different from a procedural perspective, but they follow the same framework of using the distribution of a random variable that contains the parameter of interest and its point estimate.

Definition: a pivotal quantity is a random quantity of both parameters and sample statistics and its distribution is independent of the parameters.

The following are a few examples related to the distribution we learned in the previous weeks.

Example 1. Recall in **result #1** of the previous note, that if a random sample is taken from a normal population with a known variance σ_0^2, the sampling distribution of the sample mean $barx$ is given by

$$\bar{x} \to N(\mu, \sigma_0^2/n).$$

$\bar{x}$ is not a pivotal quantity since its distribution $N(\mu, \sigma_0^2/n)$ is dependent on the unknown parameter μ. However,

$$Z_1 = \frac{\bar{x} - \mu}{\sigma_0/\sqrt{n}} \to N(0, 1).$$

Since Z_1 has a standard normal distribution that is independent of parameters and statistics, Z_1 is a pivotal quantity.

Example 2. If a random sample is taken from a normal population with unknown variance. Let s be the sample standard deviation. From **Result #2**, we have

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}} \to t_{n-1}$$

T is also a pivotal quantity since its distribution is independent of the population parameters.

Example 3. If a random sample is taken from an unspecified population with an unknown variance. Let s be the sample standard deviation. If the sample size is large, by the C.L.T,

$$\bar{x} \to N(\mu, \sigma^2/n)$$

The standardized form is independent of unknown parameters

$$Z_2 = \frac{\bar{X} - \mu}{s/\sqrt{n}} \to N(0, 1).$$

Therefore, Z_2 is a pivotal quantity.

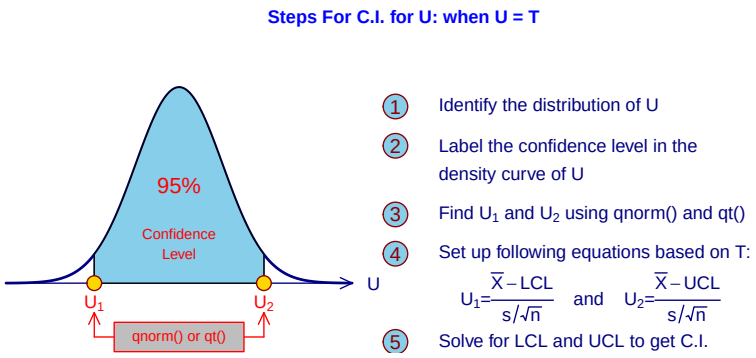
Example 4 Let $\{x_1, x_2, \cdots, x_n\}$ be a random sample taken from a binary population Y with $P(X = 1) = p$. If $np > 5$ and $n(1-p) > 5$, according to **result #4**, $\hat{p} \to N(p, \sqrt{p(1-p)/n})$. The following standardized Z_3 has a distribution

$$Z_3 = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \to N(0, 1).$$

Z_3 is also a pivotal quantity.

From the above four examples, we can see that the pivotal quantity associated with a sample mean or sample proportion is either the standard normal distribution or a t-distribution. Both distributions are symmetric with respect to the vertical axis.

For ease of illustration, we use U to denote either one of the above four quantities (Z_1, Z_2, Z_3, and T). Now, for a given confidence level of 95% (or other confidence levels), we can find two cut-off U values, denoted by U_1 and U_2 such that $P(U_1 < U < U_2) = 95\%$.



The above figure gives the steps for constructing the confidence interval for population means and proportions.

- **Confidence Level** - The area of the shaded region in the above figure is called the confidence level. Note that 95% of the U values will be in $[U_1, U_2]$ is equivalent to sat the 95% of sample means based on the same size will be in [LCL, UCL].
- **Critical Value** - U_2 on the right-hand side of the density curve of the pivotal quantity is called the **critical value**. R functions **qnorm()** and **qt()** can be used to find the normal critical and t-critical values respectively.

7.3 Confidence Interval of Population Mean

This section introduces confidence intervals based on two sampling distributions: normal and t-distributions under different assumptions. We will use examples to illustrate how to construct confidence intervals with given confidence intervals.

From the histogram of the mean heights of WCU students, we can see that a 95% confidence interval of the population mean under normal sampling distribution is 2.5% and 97.5% quantiles. Therefore, R function **qnorm()** will be used to construct the confidence intervals of means and proportions.

7.3.1 Normal Population with Known Variance.

According to result #1 in the previous module, we have

$$\bar{X} \to N(\mu, \sigma_0^2/n) \Longleftrightarrow \frac{\bar{X}-\mu}{\sigma_0/\sqrt{n}} \to N(0,1).$$

If the confidence interval $1-\alpha = 95\%$, then $U_1 = U_{0.025} = qnorm(0.025)$ and $U_2 = U_{0.975} = qnorm(0.975)$.

$$P\left(U_1 < \frac{\bar{X}-\mu}{\sigma_0/\sqrt{n}} < U_2\right) = 0.95 \Longleftrightarrow P\left(\bar{x} + U_1\frac{\sigma_0}{\sqrt{n}} < \mu < \bar{x} + U_2\frac{\sigma_0}{\sqrt{n}}\right) = 0.95$$

The 95% confidence interval of μ is given by

$$\left(\bar{x} + U_1\frac{\sigma_0}{\sqrt{n}}, \bar{x} + U_2\frac{\sigma_0}{\sqrt{n}}\right) = \left[qnorm(0.025, mean = \bar{x}, sd = \sigma_0), qnorm(0.975, mean = \bar{x}, sd = \sigma_0)\right].$$

The left-hand side equation is the confidence interval and the right-hand side of the equation is the confidence interval expressed in R functions **qnorm()**.

Example 1. Suppose a researcher, interested in obtaining an estimate of the average level of some enzyme in a certain human population, takes a sample of 10 individuals, determines the level of the enzyme in each, and computes a sample mean of $\bar{x} = 22$. Suppose further it is known that the variable of interest is approximately normally distributed with a variance $\sigma_0^2 = 45$ We wish to construct a confidence interval of the population mean μ at a confidence level of 95%.

Solution Let $\bar{x}$ be the sample mean calculated based on a random sample with size $n = 10$. Since the population is approximately distributed with a known variance $\sigma_0^2 = 45$, according to result #1 in the last note, we have the sampling distribution of the sample mean in the following form

$$\bar{x} \to N(\mu, \frac{45}{10}).$$

When we use **pnorm()** to find the quantiles, we replace the unknown mean μ with the $\bar{x}$. The following code is used to find the confidence interval.

```
qnorm(c(0.025, 0.975), mean = 22, sd = sqrt(45/10))
```

```
## [1] 17.84229 26.15771
```

Conclusion. We are 95% confident that the interval [17.84, 26.16] includes the true mean level of some enzyme in a certain human population.

Example 2. Some studies of Alzheimer's disease (AD) have shown an increase in CO2 production in patients with the disease. In one such study, the following CO2 values were obtained from 16 neocortical biopsy samples from AD patients.

```
1009 1280 1180 1255 1547 2352 1956 1080 1776 1767 1680 2050 1452 2857 3100 1621
```

Assume that the population of such values is normally distributed with a standard deviation of 350.

Solution Based on the condition, we have $\bar{x} \to N(\mu, 350/\sqrt{16})$. The unknown population will be estimated by the sample mean $\bar{x}$ in the following R code.

```
normal.sample = c(1009, 1280, 1180, 1255, 1547, 2352, 1956, 1080, 1776, 1767,
                  1680, 2050, 1452, 2857, 3100, 1621)
sample.avg = mean(normal.sample)    # sample mean
## 95% confidence intervals
qnorm(c(0.025, 0.975), mean = sample.avg, sd = 350/sqrt(16))
```

```
## [1] 1576.128 1919.122
```

Conclusion. We are 95% confident that the interval [1576.1, 1919.1] includes the true mean level of CO2 in neocortical biopsy samples from AD patients with size 16.

7.3.2 Normal Population with Unknown Variance

Recall in **Results #2**, the standardized pivotal quantity $U = T_{n-1}$, t-distribution with $n-1$ degrees of freedom.

$$\frac{\bar{x}-\mu}{s/\sqrt{n}} \to t_{n-1}.$$

Using the same logic, we have the following 95 confidence intervals based on the t-distribution.

$$\left(\bar{x} + qt(0.025, df = n-1)\frac{s}{\sqrt{n}}, \bar{x} + qt(0.975, df = n-1)\frac{s}{\sqrt{n}}\right)$$

Unfortunately, there is no direct sampling distribution of $\bar{x}$, so we have to use the above formula to find the t-confidence interval if we are given the sample mean

($\bar{x}$) and sample standard deviation (s). However, if we were given a data set, the R function **t.test()** can be used to generate the above confidence interval.

Example 3 In a study of the effects of early Alzheimer's disease on non-declarative memory, the Category Fluency Test was used to establish baseline persistence semantic memory, and language abilities. The eight subjects in the sample had Category Fluency Test scores of 11, 10, 6, 3, 11, 10, 9, and 11. Assume that the eight subjects constitute a simple random sample from a normally distributed population of similar subjects with early Alzheimer's disease.

(a) What is the point estimate of the population mean?

(b) What is the standard deviation of the sample?

(c) What is the estimated standard error of the sample mean?

(d) Construct a 95 percent confidence interval for the population mean category fluency test score.

Solution The following R code gives the answers to the above questions.

```
test.score= c(11, 10, 6, 3, 11, 10, 9, 11)
n = length(test.score)                  # sample size
Q.a = mean(test.score)
Q.b = round(sd(test.score),3)
Q.c = round(sd(test.score)/sqrt(n),3)
Q.d = round(t.test(test.score, conf.level = 0.95)$conf.int,3)
kable(as.data.frame(cbind(Q.a = Q.a,
      Q.b = Q.b,
      Q.c = Q.c,
      Q.d = paste("[",Q.d[1], ", ", Q.d[2], "]"))))
```

Q.a	Q.b	Q.c	Q.d
8.875	2.9	1.025	[6.45 , 11.3]

The answers to questions a, b, and c are given in the above table. The 95% confidence interval for the fluency test score is [6.45, 11.3] meaning that interval [6.45, 11.3] has a 95% chance to include the true mean test score.

Example 4: A sample of 16 ten-year-old girls had a mean weight of 71.5 and a standard deviation of 12 pounds, respectively. Assuming normality, find a 99 percent confidence interval for μ.

Solution Based on the given assumption, the pivotal quantity has a t-distribution with 15 degrees of freedom. We cannot use **t.test()** to find the confidence interval since it requires the raw data set. We can then use the given formula to find the confidence interval.

```
xbar = 71.5
s = 12
LCL = xbar + qt(0.005, df = 15)*s/sqrt(16)
```

```
UCL = xbar + qt(0.995, df = 15)*s/sqrt(16)
CI = cbind(LCL = LCL, UCL = UCL)
CI = round(CI,2)
kable(CI)
```

LCL	UCL
62.66	80.34

The 95% confidence interval of the mean weight is [62.66, 80.34]. Therefore, we are 99% confident that the mean weight is in [62.66, 80.34].

7.3.3 Unspecified Population with Unknown Variance but with Large Sample Sizes

When the sample size is large, we then use the central limit theorem (CLT). The pivotal quantity is approximately normally distributed. The confidence interval can be similarly found using the sample code as used in **Example 1**.

Example 5 A physical therapist wished to estimate, with 99 percent confidence, the mean maximal strength of a particular muscle in a certain group of individuals. A sample of 64 subjects who participated in the experiment yielded a mean of 84.3 and a sample variance of 144.

Solution Since sample size $n = 64$, according to the CLT, we have

$$\bar{x} \to N(84.3, 144/64)$$

We can use the above sampling distribution to find the 99% confidence interval given in the following code.

```
LCL = qnorm(0.005, mean = 84.3, sd = sqrt(144/64))
UCL = qnorm(0.995, mean = 84.3, sd = sqrt(144/64))
CI = cbind(LCL, UCL)
CI = round(CI,2)
kable(CI)
```

LCL	UCL
80.44	88.16

The 99% confidence interval is [80.44, 88.16]. Therefore, we are 99% confident that the mean maximal strength of a particular muscle in a certain group of individuals is between 80.44 and 88.16.

7.4 Confidence Interval of Proportion

According to **Result 4** in the previous section, we can find a 95% confidence interval of the population proportion based on the following sampling distribution of $\hat{p}$

$$\hat{p} \to N(p, \sqrt{\hat{p}(1-\hat{p})/n})$$

Example 6 To study patients who were mechanically ventilated in the intensive care unit of six hospitals in Buenos Aires, Argentina. The researchers found that of 472 mechanically ventilated patients, 63 had clinical evidence of ventilator-associated pneumonia (VAP). Construct a 95 percent confidence interval for the proportion of all mechanically ventilated patients at these hospitals who may be expected to develop VAP and interpret the confidence interval.

Solution Since $n\hat{p} = 472 \times (63/472) = 63 > 5$ and $n(1-\hat{p}) = 409 > 5$, according to **Result 4**, $\hat{p} \to N(p, \sqrt{\hat{p}(1-\hat{p})/472})$. The following R code generates the 95% confidence interval of the proportion.

```
n = 472
phat = 63/472
s.phat = sqrt((63/472)*(1-63/472)/472)
LCL = qnorm(0.025, mean =phat, sd = s.phat)
UCL = qnorm(0.975, mean =phat, sd = s.phat)
CI=cbind(LCL, UCL)
CI = round(CI,2)
kable(CI)
```

LCL	UCL
0.1	0.16

The 95% confidence interval of the proportion of all mechanically ventilated patients at these hospitals who may be expected to develop VAP is [0.1, 0.16]. Therefore, [0.10, 0.16] has a 95% chance to include the true proportion.

7.5 Two Sample Problems - Comparing Two Population Means

We only introduce two sample problems based on independent samples with large sample sizes. One important fact is that the `variances` of two **independent** variables are additive. To be more specific, let X and Y be independent random variables, then $Var(X \pm Y) = Var(X) + Var(Y)$. **However**, standard deviations are **NOT** additive, $sd(X \pm Y) = sd(X) + sd(Y)$.

In real-world applications, one of the practical questions is to compare the difference between two population means $(\mu_1 - \mu_2)$. One way to address this issue is to construct a confidence interval for $\mu_1 - \mu_2$.

7.5.1 Confidence Interval of the Difference of Two Unspecified Populations Means

Let $\bar{x}_1$ and $\bar{x}_2$ be sample means of two independent populations. The corresponding sample standard deviations are s_1 and s_2, and $n_1 > 30$ and $n_2 > 30$ are sample sizes. The sampling distribution of $\bar{x}_1 - \bar{x}_2$ is given by

$$\bar{x}_1 - \bar{x}_2 \to N\left(\mu_1 - \mu_2, \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right)$$

Using the above sampling distribution, we can construct a 95% confidence interval of the difference between two population means.

Example 7 Despite common knowledge of the adverse effects of doing so, many women continue to smoke while pregnant. To examine the effectiveness of a smoking cessation program for pregnant women. The mean number of cigarettes smoked daily at the close of the program by the 328 women who completed the program was 4.3 with a standard deviation of 5.22. Among 64 women who did not complete the program, the mean number of cigarettes smoked per day at the close of the program was 13 with a standard deviation of 8.97. We wish to construct a 99 percent confidence interval for the difference between the means of the populations from which the samples may be presumed to have been selected.

Solution Since $n_1 = 328 > 30$ and $n_2 = 64 > 30$, then

$$\bar{x}_1 - \bar{x}_2 \to N\left(\mu_1 - \mu_2, \sqrt{\frac{5.22^2}{328} + \frac{8.97^2}{64}}\right)$$

The 95% confidence interval is given by

```
n1 = 328
n2 = 64
x1bar = 4.3
x2bar = 13
s1 = 5.22
s2 = 8.97
LCL = qnorm(0.005, mean = x1bar -x2bar, sd = sqrt(s1^2/n1 + s2^2/n2))
UCL = qnorm(0.995, mean = x1bar -x2bar, sd = sqrt(s1^2/n1 + s2^2/n2))
CI=cbind(LCL, UCL)
CI=round(CI,2)
kable(CI)
```

LCL	UCL
-11.68	-5.72

Therefore, we are 99% confident that the difference between the two population means $\mu_1 - \mu_2$ is in [-11.68, -5.72]. Since both confidence limits are negative, we

can claim that $\mu_1 - \mu_2 < 0$, that is, $\mu_2 > \mu_1$.

7.5.2 Confidence Interval of the Difference of Two Normal Populations Means

We only discuss a special case in which

- both populations are normal.
- population variances are unknown but equal.

Because the two population variances are equal, we combine two samples to estimate the common variance using the following formula.

$$s_{pool} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2}}$$

With the above pooled standard deviation, we have

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{pool}\sqrt{1/n_1 + 1/n_2}} \to t_{n_1+n_2-2}$$

Let $U_1 = qt(0.025, df = n_1 + n_2 - 2)$ and $U_2 = qt(0.975, df = n_1 + n_2 - 2)$, then the 95% confidence interval of $\mu_1 - \mu_2$ is given by

$$\left[(\bar{x}_1 - \bar{x}_2) + U_1 \frac{s_{pool}}{\sqrt{1/n_1 + 1/n_2}}, (\bar{x}_1 - \bar{x}_2) + U_2 \frac{s_{pool}}{\sqrt{1/n_1 + 1/n_2}}\right]$$

Example 8 To determine the effectiveness of an integrated outpatient dual-diagnosis treatment program for mentally ill subjects. The authors were addressing the problem of substance abuse issues among people with severe mental disorders. A retrospective chart review was performed on 50 consecutive patient referrals to the Substance Abuse/Mental Illness program at the VA San Diego Healthcare System. One of the outcome variables examined was the number of inpatient treatment days for a psychiatric disorder during the year following the end of the program. Among 18 subjects with schizophrenia, the mean number of treatment days was 4.7 with a standard deviation of 9.3. For 10 subjects with bipolar disorder, the mean number of psychiatric disorder treatment days was 8.8 with a standard deviation of 11.5. We wish to construct a 95 percent confidence interval for the difference between the means of the populations represented by these two samples. We assume that both populations are normal and have equal variances.

Solution Based on the given information, we pivotal quantity is given by

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \to t_{n_1+n_1-2}$$

Therefore, the t-confidence interval of the difference of the two population means is given by the following code.

```
n1 = 18
n2=10
x1bar = 4.7
x2bar = 8.8
s1 = 9.3
s2=11.5
s.pool = sqrt(((n1-1)*s1^2 + (n2-1)*s2^2)/(n1+n2-2))
LCL = (x1bar-x2bar)+qt(0.025, df = n1+n2-2)*s.pool*sqrt(1/n1 + 1/n2)
UCL = (x1bar-x2bar)+qt(0.975, df = n1+n2-2)*s.pool*sqrt(1/n1 + 1/n2)
CI=cbind(LCL, UCL)
CI = round(CI, 2)
kable(CI)
```

LCL	UCL
-12.3	4.1

Therefore, the 95% confidence interval of $\mu_1 - \mu_2$ is [-12.3, 4.1]. Since the 95% confidence interval does contain 0. We are 95% confident that there is no significant difference between the two population means.

Conclusion Remarks

- We can discuss the difference between two independent proportions using the same logic in the previous sections.
- In the next module, we will discuss testing hypotheses - another type of inference.

7.6 Assignment - Confidence Intervals

Please study the examples in the class note and complete the following problems. Please note that you are expected to interpret each of the confidence intervals you obtained. You can modify my code to calculate confidence intervals.

Problem 1

We wish to estimate the mean serum indirect bilirubin level of 4-day-old infants. The mean for a sample of 16 infants was found to be 5.98 mg/100 cc. Assume that bilirubin levels in 4-day-old infants are approximately normally distributed with a standard deviation of 3.5 mg/100 cc. Construct a 95% confidence interval of the mean serum indirect bilirubin level of 4-day-old infants (μ) and interpret the interval.

Problem 2

10 obstetrics and gynecology interns participated in a study conducted by researchers at the University of Colorado Health Sciences Center. The researchers wanted to assess competence in performing clinical breast examinations. One of the baseline measurements was the number of such examinations performed. The following data give the number of breast examinations performed for this sample of 10 interns.

30, 40, 8, 20, 26, 35, 35, 20, 25, 20

Construct a 95% confidence interval for the number of breast examinations and give an interpretation of the confidence interval.

Problem 3

The punctuality of patients in keeping appointments is of interest to a research team. In a study of patient flow through the offices of general practitioners, it was found that a sample of 35 patients was 17.2 minutes late for appointments, on average. Previous research had shown the standard deviation to be about 8 minutes. The population distribution was felt to be non-normal. Find the 90 percent confidence interval for μ, the true mean amount of time late for appointments and interpret the confidence interval.

Problem 4

The following are the activity values (micromoles per minute per gram of tissue) of a certain enzyme measured in normal gastric tissue of 35 patients with gastric carcinoma.

```
0.360, 1.189, 0.614, 0.788, 0.273, 2.464, 0.571, 1.827, 0.537, 0.374, 0.449, 0.262,
0.448, 0.971, 0.372, 0.898, 0.411, 0.348, 1.925, 0.550, 0.622, 0.610, 0.319, 0.406,
0.413, 0.767, 0.385, 0.674, 0.521, 0.603, 0.533, 0.662, 1.177, 0.307, 1.499
```

We wish to construct a 95 percent confidence interval for the population mean and interpret the confidence interval. It is not necessary to assume that the sampled population of values is normally distributed.

Problem 5

In a study, 136 subjects with syncope or near syncope were studied. Syncope is the temporary loss of consciousness due to a sudden decline in blood flow to the brain. Of these subjects, 75 also reported having cardiovascular disease. Construct a 99 percent confidence interval for the population proportion of subjects with syncope or near syncope who also have cardiovascular disease and interpret the interval.

Problem 6

In a study of factors thought to be responsible for the adverse effects of smoking on human reproduction, cadmium level determinations (nanograms per gram) were made on the placenta tissue of a sample of 14 mothers who were smokers

and an independent random sample of 18 nonsmoking mothers. The results were as follows:

Nonsmokers: 10.0, 8.4, 12.8, 25.0, 11.8, 9.8, 12.5, 15.4, 23.5, 9.4, 25.1, 19.5, 25.5, 9.8, 7.5, 11.8, 12.2, 15.0

Smokers: 30.0, 30.1, 15.0, 24.1, 30.5, 17.8, 16.8, 14.8, 13.4, 28.5, 17.5, 14.4, 12.5, 20.4

Assume that both smokers and non-smoker populations are normally distributed and have equal variance.

Does it appear likely that the mean cadmium level is higher among smokers than nonsmokers? Why Do you reach this conclusion? [Hint: answer the above questions by constructing a 95% confidence interval of the difference between population means.]

Chapter 8

Testing Statistical Hypothesis

There are two basic statistical inferences: confidence interval and testing hypothesis. In confidence interval inference, we estimate the population parameter(s) by constructing an interval that reveals the information about the precision and accuracy of the estimate and the level of confidence of the estimate to be correct.

The other type of inference is to justify a statement about a population parameter. For example, if someone **claims** that the average height of students at a university is higher than 70 inches, how to justify the claim? To reach the binary decision of either supporting the claim or rejecting the claim, we need to gather information from the population and then use the sample evidence to make the statistical decision.

The logic of statistical hypothesis testing is similar to the medical diagnostic decision process and jury trial - both are evidence-based decision processes. Analogies between the statistical testing hypothesis and the aforementioned two processes are summarized in the following table.

Analogies of Statistical Hypothesis Testing		
Medical Diagnostics	**Test of Statistical Significance**	**Court Trial**
Every incoming patient has a disease until proved disease-free	Null Hypothesis	Every defendant is innocent until proved guilty
Order labs and gathering health info	data (information) collection	Investigation and evidence gathering
summarize clinical evidence	Information Aggregation: test Statistics	Summary of evidence exhibitions
Clinical standards	Statistical Decision Rules	Jury's instruction: cross-examination
Diagnostic decision	Statistical Decision	Verdict
Claim a diseased patient to be disease-free	**Type I Error**	Conviction of innocent defendant
Claim a disease-free patient to be diseased	**Type II Error**	Acquittal of a criminal

Figure 8.1: Analogies of statistical hypothesis testing to jury trial and medical diagnostics

Note that no matter whichever statistical decision is made, there will be two

possible errors: Type I and type II errors. We can see from the above analogies that **the type I error is more serious than the type II error**. Because of this relationship between the two types of errors, in practice, we control the type I error and then minimize the type II error - this is the basis of Neymann-Pearson's Lemma for statistical hypothesis testing.

With the above conceptual understanding of the testing hypothesis, we next formally formulate the statistical testing hypothesis so that we can implement it in real-world applications.

8.1 Formulation of Statistical Hypothesis Testing

The formal statistical testing hypothesis is summarized in the following steps.

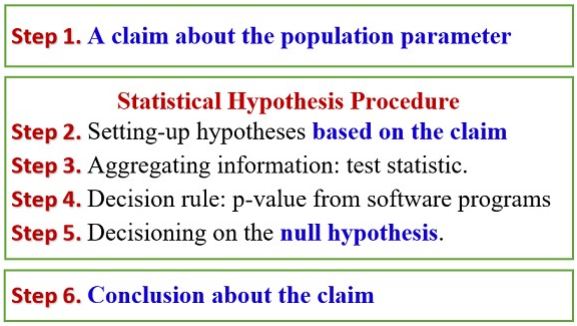


Figure 8.2: Formulation of statistical hypothesis testing

Unlike the formulation used in the textbook, we separate the **statistical hypothesis testing procedure** from the general testing problem. Steps 2-5 are actual statistical procedures. The general workflow of conducting a testing hypothesis is summarized in the following few sections.

8.1.1 Null and Alternative Hypotheses

We start with a practical question that involves data and a statement claiming population parameters.

Example The yield of alfalfa from a random sample of six test plots is 1.4, 1.6, 0.9, 1.9, 2.2, and 1.2 tons per acre. Assume that the random sample comes from a normal population. Test at the 0.05 level of significance whether this supports the contention that the average yield for this kind of alfalfa is 1.5 tons per acre.

- **Data Set**: `{1.4, 1.6, 0.9, 1.9, 2.2, 1.2}`
- **Claim**: `The average yield for this kind of alfalfa is 1.5 tons per acre.`

If we use the Greek letter μ to denote the population mean, then the claim is $\mu \neq 1.5$ of this particular example. In general, there are six possible claims

$$\mu = 1.5, \mu \neq 1.5, \mu > 1.5, \mu \leq 1.5, \mu < 1.5, \mu > 1.5$$

Three of the six potential claims **have an equal sign** and the other three potential claims **do not have an equal sign**.

8.1.2 Relationship between the claim and the statistical hypotheses

The statistical **null hypothesis** must contain an **equal sign** since we assume that **the null hypothesis is true**. The **alternative hypothesis** is the opposite of the **null hypothesis**!

In other words, if the claim has an **equal sign**, the corresponding **null hypothesis** is the same as the claim. Otherwise, the **opposite of the claim will be the null hypothesis**.

The above relationship is summarized in the following table.

The guideline for setting-up Ha is given below:

a). If **the original claim** contains an "=" (i.e., $\leq$, =, $\geq$), we choose the opposite of the original claim (i.e., >, $\neq$, <) as **the alternative hypothesis** H_a;

b). If **the original claim** doesn't contain an "="(i.e., >, $\neq$, <), we choose simply the original claim (i.e., >, $\neq$, <) as **the alternative hypothesis** H_a.

Figure 8.3: Formulation of statistical hypothesis testing

Example 1 (Revisited) Based on the above description, the **claim** is $\mu = 1.5$. It contains an **equal sign**. Therefore, the **null hypothesis (Ho)** and the **alternative hypothesis (Ha)** are given by

Ho: $\mu = 1.5$ v.s. Ha: $\mu \neq 1.5$.

8.1.3 Test Statistics - Statistical Evidence

In the construction of confidence intervals of population mean and proportion, we used the distribution of **pivotal quantity** since it contains all the information needed for constructing a confidence interval. **The same amount of information is needed for testing the hypothesis**. We introduced four

major types of **pivotal quantities** in the previous module for constructing the confidence intervals of the single population mean μ and proportion p.

In testing hypotheses, we assume the **null hypothesis** is true. Therefore, we replace the unknown population parameter with the claimed value of the parameter in the **null hypothesis**. Therefore, the four test statistics are under various assumptions.

Normal Population with known variance $TS = \frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}} \sim N(0,1)$	Unspecified population with a large sample size $TS = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim N(0,1)$
Normal Population with unknown variance $TS = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$	Proportion problem satisfying np>5 and n(1-p)>5 $TS = \frac{\hat{p} - p_0}{\sqrt{\hat{p}(1-\hat{p})/n}} \sim N(0,1)$

Figure 8.4: Test statistics under different assumptions

The above four statistics are either standard normal or t distribution. **If the claimed value in the null hypothesis is close to the true population parameter, we would expect the value of the test statistic to be around zero**. However, if the value of the test statistic is **far away from zero**, we intend to reject the **null hypothesis** and accept the **alternative hypothesis**.

A similar question that was asked when we constructed confidence intervals needs to be answered in the testing hypothesis. The **null hypothesis** is rejected if the test statistic is far ways from zero. **How Far Is Far?**

8.1.4 Types of Test and Rejection Region

The rejection of a test is dependent on its specific type (right-, left-, and two-tailed) and the type I error (also called significance level α, the same α used in the confidence level $1 - \alpha$). Next, we assume the significance to be α. The reject region of each of the three types of tests is summarized in the following.

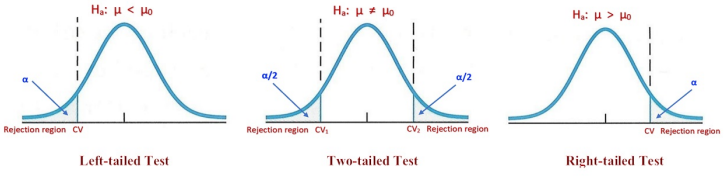


Figure 8.5: Three different types of hypothesis tests

From the above figure, we can see the location of the rejection region from the form of the alternative test. The rejection region of the left-tailed test is on the left tail of **the distribution of the test statistic**. The area of the rejection

regions is equal to the **significance level** α. For a two-tailed test, there are two rejection regions located on both sides of the tails of the sampling distribution of the test statistic. The tail regions on both tails are equal to $\alpha/2$. We can also similarly interpret the right-tailed test.

8.1.5 Statistical Decision Rule: p-value

In this subsection, we define the **p-value** based on the statistic and the significance level to determine whether the test statistic is in the rejection region.

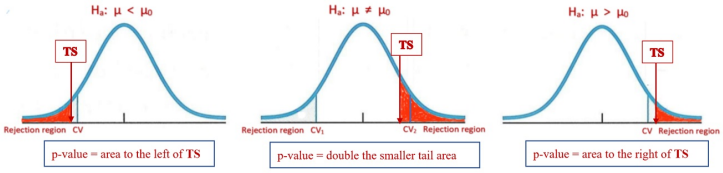


Figure 8.6: Definitions of p-values for different types of tests

The definition of the p-value for different types of tests is given in the above figures. The statistical decision rule is summarized in the following.

- If the p-value is less than the significance level α, the **null hypothesis is rejected** and the alternative is accepted.
- If the p-value is greater than the significance level α, the **null hypothesis is accepted** and the alternative is rejected.

8.1.6 Summary of the Test Hypothesis Procedure

The above subsection describes the steps for performing a formal statistical hypothesis. When we implement the testing procedure, we need to follow the above-mentioned steps. Particularly, write the original clearly and then based on the description in section 2.1.1 set up the **null hypothesis (Ho)** and the **alternative hypothesis (Ha)** correctly.

In the next section, we will present numerical examples with different claims and assumptions about the populations. For the two-sample comparison problems, we only focus on testing the difference between two population means.

Some of the examples will be based on the raw data set(s). The analysis related to your research will be based on raw data you generated from lab experiments or fieldwork.

8.2 Case Studies

One important piece of advice is to draw a density curve of the distribution of the underlying distribution of the test statistics and label all information

about the population, samples, and the null hypothesis (rejection regions) on the density curve so you can choose a correct R function from **pnorm()** or **pt()** to find the p-value. As a convention, if you are given the significance level α, you are expected to use the default significance level $\alpha = 0.05$ meaning that your resulting statistical decision only allows less than 5% of chance to be wrong.

8.2.1 Case I: normal population with a known variance.

This situation is not common in real-life applications unless you have some prior information about the population variance and you believe that variance remains unchanged.

Example 2: Researchers are interested in the mean age of a certain population. The data available to the researchers are the ages of a simple random **sample of 10 individuals** drawn from the population of interest with mean $\bar{x} = 27$. It is assumed that the sample comes from a population whose ages are approximately normally distributed. Let us also assume that the population has a known variance of $\sigma_0^2 = 20$. The question that researchers want to ask is: Can we conclude that the mean age of this population is different from 30 years? Assuming that the mean age of the population is equal to 30.

Solution: First of all, the **claim** is the mean age of the research population remains unchanged meaning that $\mu_0 = 20$. Since the original claim has an equal sign in it, the null hypothesis is identical to the claim. Therefore, the null and alternative hypotheses are given by

$$H_o : \mu = 30 \leftrightarrow H_a : \mu \neq 30.$$

Since the study population is normal with known variance $\sigma_0^2 = 20$. Therefore, the test statistic

$$TS = \frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}} \rightarrow N(0, 1)$$

The calculation of the p-value is given in the following code.

```
# given conditions
xbar = 27
sig.sq.0 = 20
mu0 = 30
n = 10
alpha = 0.05   # not given, use the default
# two-tailed normal test
TS = (xbar - mu0)/(sqrt(sig.sq.0/n))    # test statistics
right.tail.area = 1 - pnorm(TS)      # TS is standard normal
left.tail.area = pnorm(TS)
p.value = 2*min(right.tail.area, left.tail.area)   # double the smaller tail area
p.value
```

```
## [1] 0.03389485
```

p-value = 0.03389 is less than the default significance level of 0.05. We reject the NULL HYPOTHESIS. We conclude the current mean age of the study population is significantly different from 30 years ago.

Remarks

1. The statistical decision is always made on the **Null Hypothesis**!
2. If we are given a raw data set, we need to find the sample mean and sample size and then use the above code to find the p-value.

Example 3: Refer to **Example 2**. Suppose, instead of asking if they could conclude that $\mu_0 = 30$, the researchers had asked: Can we conclude that $\mu < 30$?

Solution: To this question, the claim is $\mu < 30$ that does NOT have an equal sign in it, we need to the opposite the claim as the null hypothesis. The alternative hypothesis will be identical to the claim. That is, we have the following null and alternative hypotheses.

$$H_o : \mu \geq 30 \leftrightarrow H_a : \mu < 30.$$

This is a left-tailed test. The rejection region is on the left tail of the density curve of the test statistic. The p-value is calculated by

```
# given conditions
xbar = 27
sig.sq.0 = 20
mu0 = 30
n = 10
alpha = 0.05   # not given, use the default
# two-tailed normal test
TS = (xbar - mu0)/(sqrt(sig.sq.0/n))    # test statistics
right.tail.area = pnorm(TS)          # TS is standard normal
p.value = right.tail.area
p.value
```

```
## [1] 0.01694743
```

Since the p-value is less than 0.05. We reject the null hypothesis that $\mu \geq 30$, and we conclude the actual mean age of the population less than 30.

8.2.2 Case II: Normal population with an unknown variance

From theory, if the sample was taken from a normal population with unknown variance, the test statistic is always a random variable that follows t-distribution with degrees of freedom $df = n - 1$.

$$TS = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \rightarrow t_{n-1}$$

If the sample is large, the t-distribution is close to the standard normal distribution. In this case, you can use either normal or t-distribution to find the p-value. **However**, if the sample size small (n $<$ 30), we **MUST** use the t -distribution to find the p-value.

Example 4: Nakamura et al. studied subjects with medial collateral ligament (MCL) and anterior cruciate ligament (ACL) tears. Between February 1995 and December 1997, 17 consecutive patients with combined acute ACL and grade III MCL injuries were treated by the same physician at the research center. One of the variables of interest was the length of time in days between the occurrence of the injury and the first magnetic resonance imaging (MRI). The data are shown in the following.

14, 0, 28, 14, 9, 10, 24, 9, 18, 4, 24, 26, 8, 2, 12, 21, 3

We wish to know if we can conclude that the mean number of days between injury and initial MRI **is not 15 days** in a population presumed to be represented by these sample data. The original problem did not mention the normal distribution of the population, but we assumed the normality of the population in order to perform a t-test about the population mean.

Solution: The claim about the population mean is $\mu \neq 15$. This is a two-tailed test.

$$H_o : \mu = 15 \leftrightarrow H_a : \mu \neq 15.$$

Since we have a small sample (n = 17) from an implicitly assumed population with an unknown variance. The test statistic

$$TS = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \rightarrow t_{n-1}$$

Since this is a t-test with a given set of raw data values, we can use a convenient R function **t.test()** was built based on the above formulas and the definition of p-values. You type `?t.test` in the R console to find the help document of this function and accompanying examples as well.

```
# defining a data set
days = c(14, 0, 28, 14, 9, 10, 24, 9, 18, 4, 24, 26, 8, 2, 12, 21, 3)
t.test(days,                          # data set
       mu = 15,                     # claimed value in the null hypothesis
       conf.level = 0.95,      # confidence level 0.95 => significant level 1-0.95.
       alternative ="two.sided")  # alternative hypothesis
```

```
##
##  One Sample t-test
##
## data:  days
## t = -0.79148, df = 16, p-value = 0.4402
## alternative hypothesis: true mean is not equal to 15
## 95 percent confidence interval:
##   8.725081 17.863155
## sample estimates:
## mean of x
##  13.29412
```

The p-value is 0.4403 which is greater than 0.05. We **fail to reject** the null hypothesis. Therefore, the sample DOES NOT have evidence to support the claim that $\mu \neq 15$.

Remark We use the formula to find the p-value using **pt()** to find the p-values as shown in the following.

```
days = c(14, 0, 28, 14, 9, 10, 24, 9, 18, 4, 24, 26, 8, 2, 12, 21, 3)
xbar = mean(days)
s = sd(days)
mu0 = 15
n = length(days)
alpha = 0.05    # not given, use the default
# two-tailed normal test
TS = (xbar - mu0)/(s/sqrt(n))    # test statistics
right.tail.area = 1 - pt(TS, df = n-1)         # TS is standard normal
leftt.tail.area = pt(TS, df = n-1)
p.value = 2*min(right.tail.area,leftt.tail.area)
p.value
```

```
## [1] 0.4402394
```

We can see the p-values generated from the two methods are identical.

8.3 Unspecified population with an unknown variance - CLT

This is a direct application of the CLT. The key is that the sample size MUST be large!

Example 5: The goal of a study by Klingler et al. was to determine how symptom recognition and perception influence clinical presentation as a function of race. They characterized symptoms and care-seeking behavior in African-American patients with chest pain seen in the emergency department. One of the presenting vital signs was systolic blood pressure. Among 157 African-

American men, the mean systolic blood pressure was 146mm Hg with a standard deviation of 27. We wish to know if, on the basis of these data, we may conclude that the mean systolic blood pressure for a population of African-American men **is greater than** 140.

Solution: The claim is that the mean systolic blood pressure for a population of African-American men **is greater than** 140. That is $\mu > 140$. Since the claim does not have an equal sign in it, its opposite will be the null hypothesis.

$$H_o : \mu \leq 140 \leftrightarrow H_a : \mu > 140.$$

We are also given the sample size $n = 157 > 30$, by the CLT, the test statistic

$$TS = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \rightarrow N(0, 1)$$

The following code calculates the p-value

```
# sample information
n = 157
xbar = 146
s = 27
mu0 = 140
###
TS = (xbar - mu0)/(s/sqrt(157))
# This is a right-tailed test. We need the right tail area using 1- pnorm(TS)
p.value = 1-pnorm(TS)
p.value
```

```
## [1] 0.002681041
```

8.4 Testing Population Proportion

The most important step is to check whether conditions $np > 5$ and $n(1-p) > 5$ before claiming that

$$TS = \frac{\hat{p} - p_0}{\sqrt{\hat{p}(1-\hat{p})/n}} \rightarrow N(0, 1)$$

R has a function **prop.test()** for testing the proportion and equality of two proportions. Of course, we can also translate the above formula and appropriate distribution of the test statistics to calculate the p-value.

Example 6: Wagenknecht et al. collected data on a sample of 301 Hispanic women living in San Antonio, Texas. One variable of interest was the percentage of subjects with impaired fasting glucose (IFG). IFG refers to a metabolic stage intermediate between normal glucose homeostasis and diabetes. In the study, 24

women were classified in the IFG stage. The article cites population estimates for IFG among Hispanic women in Texas as 6.3 percent. Is there sufficient evidence to indicate that the population of Hispanic women in San Antonio has a prevalence of IFG higher than 6.3 percent?

Solution: Note that the claim is that the population of Hispanic women in San Antonio has a prevalence of IFG higher than 6.3 percent. That is, $p > 0.063$. The null and alternative hypotheses are given by

$$H_o : p \leq 0.063 \leftrightarrow H_a : p > 0.063.$$

Since $\hat{p} = 24/301$ $n\hat{p} = 301 \times \hat{p} = 24 > 5$, $n(1-\hat{p}) = 301 \times (1 - 24/301) > 5$. We can claim that the test statistic is approximately normally distributed.

$$TS_0 = \frac{\hat{p} - p_0}{\sqrt{\hat{p}(1-\hat{p})/n}} \rightarrow N(0,1)$$

R also has a built-in function **prop.test()** that can be used for testing a single proportion or the equality of two proportions. Here we use the **prop.test()** to conduct the above right-tailed test. The test used in the R function has the following form

$$TS_1 = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \rightarrow N(0,1)$$

The p-value reported from **prop.test()** is slightly different from the formula in which the denominator uses $\hat{p}$.

```
prop.test(x=24,                      # number of successes
          n = 301,                   # sample size
          p = 0.063,                 # claimed probability in the null hypothesis
          alternative = "greater",   # right-tailed test
          conf.level = 0.95)         # significance level = 1 - confidence level
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  24 out of 301, null probability 0.063
## X-squared = 1.1585, df = 1, p-value = 0.1409
## alternative hypothesis: true p is greater than 0.063
## 95 percent confidence interval:
##  0.05623225 1.00000000
## sample estimates:
##          p
## 0.07973422
```

The p-value is 0.1409 which is greater than 0.05. We fail to reject the null hypothesis that $p \leq 0.063$. Therefore, the sample does not support the claim that $p > 0.063$.

Remark Both TS_0 and TS_1 are valid statistics. They are derived using different methods. The following code shows that the two statistics result in similar p-values.

```
TS1 = (24/301-0.063)/(sqrt((24/301)*(1-24/301)/301))
TS2 = (24/301-0.063)/(sqrt((0.063)*(1-0.063)/301))
p.value1 = 1 - pnorm(TS1)
p.value2 = 1 - pnorm(TS2)
cbind(p.value1, p.value2)
```

```
##       p.value1  p.value2
## [1,] 0.1419072 0.1160539
```

8.5 Testing the difference between two population means

Comparing two population means is common in practice. In this subsection, we introduce two special procedures for testing two population means under different conditions.

8.5.1 Both populations are unspecified and sample sizes are large.

Assume that $\{x_1, x_2, \cdots, x_{n_1}\}$ and $\{y_1, y_2, \cdots, y_{n_2}\}$ are taken from two independent populations with means μ_1 and μ_2, respectively. Assume $n_1 > 30$ and $n_2 > 30$. By CLT, we have

$$\bar{x} - \bar{y} \to N\left(\mu_1 - \mu_2, \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right).$$

Then the test statistic for testing $\mu_1 - \mu_2$ is defined to be

$$TS = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \to N(0, 1)$$

Example 7: To identify the role of various disease states and additional risk factors in the development of thrombosis. One focus of the study was to determine if there were differing levels of the anticardiolipin antibody IgG in subjects with and without thrombosis. The following table summarizes the researchers' findings in a study

```
Sample statistics:

                     mean   size   stdev
Thrombosis (x)      59.01    53    44.89
No thrombosis (y)   46.61    54    34.85
```

We wish to know if we may conclude, on the basis of these results, that, in general, persons with thrombosis have, on average, higher IgG levels than persons without thrombosis.

Solution: The **claim** is that persons with thrombosis have, on average, higher IgG levels than persons without thrombosis, that is, $\mu_1 - \mu_2 > 0$. The opposite of the claim will be the null hypothesis.

$$H_o : \mu_1 - \mu_2 \leq 0 \leftrightarrow H_a : \mu_1 - \mu_2 > 0.$$

This is a right-tailed test. The test statistic is

$$TS = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(59.01 - 46.61) - (0)}{\sqrt{\frac{44.89^2}{53} + \frac{34.85^2}{54}}}$$

We use the following code to find the p-value.

```
#                      mean   size   stdev
# Thrombosis (x)      59.01    53    44.89
# No thrombosis (y)   46.61    54    34.85
xbar = 59.01
ybar = 46.61
s1 = 44.89
s2 = 34.85
n1 = 53
n2 = 54
##
TS = ((xbar - ybar) - 0)/sqrt(s1^2/n1 + s2^2/n2)
## right-tailed test
p.value = 1 - pnorm(TS)
p.value
```

```
## [1] 0.05546304
```

Since p-value $= 0.0555 > 0.05$, we fail to reject the **null hypothesis** at level 0.05. We conclude that the $\mu_1 - \mu_2 > 0$ meaning that persons with thrombosis have, on average, higher IgG levels than persons without thrombosis.

Remark: If we are given two raw data sets, we use **mean()** and **var()** and **length()** to calculate the means, variances, and sample sizes.

8.5.2 Both populations are normal with unknown but equal variances

Assume that $\{x_1, x_2, \cdots, x_{n_1}\}$ and $\{y_1, y_2, \cdots, y_{n_2}\}$ are taken from two independent **normal** populations with means μ_1 and μ_2, respectively. Let s_1^2 and s_2^2 be the corresponding sample variances. Since we assume that the two population variances are equal.

$$s_{pool} = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

Then the test statistic for testing $\mu_1 - \mu_2$ is defined to be

$$TS = \frac{(\bar{x}-\bar{y}) - (\mu_1-\mu_2)}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}} \to N(0,1)$$

We can use the above formulas to find the p-value based on the type of test. In R, **t.test()** can also be used to generate the result directly if we are given the raw data.

Example 8: To investigate wheelchair maneuvering in individuals with lower-level spinal cord injury (SCI) and healthy controls (C). Subjects used a modified wheelchair to incorporate a rigid seat surface to facilitate the specified experimental measurements. Interface pressure measurement was recorded by using a high-resolution pressure-sensitive mat with a spatial resolution of four sensors per square centimeter taped on the rigid seat support. During static sitting conditions, average pressures were recorded under the ischial tuberosities (the bottom part of the pelvic bones). The data for measurements of the left ischial tuberosity (in mm Hg) for the SCI and control groups are shown in the following

```
Control: 131 115 124 131 122 117 88 114 150 169
SCI:      60 150 130 180 163 130 121 119 130 148
```

We wish to know if we may conclude, on the basis of these data, that, in general, healthy subjects exhibit lower pressure than SCI subjects.

Solution: The **claim** is healthy subjects exhibit lower pressure than SCI subjects, $\mu_c < \mu_s$. Since the claim $\mu_c - \mu_s < 0$ has no equal sign in it,

$$H_o : \mu_c - \mu_s \geq 0 \leftrightarrow H_a : \mu_1 - \mu_2 < 0.$$

This is a left-tailed test. We use the following code to test the above hypothesis.

```r
## R is case-sensitive, we encourage to use of all lowercase letters to name variables
control = c(131, 115, 124, 131, 122, 117, 88, 114, 150, 169)
sci= c(60, 150, 130, 180, 163, 130, 121, 119, 130, 148)
```

```
##
t.test(control, sci,
       conf.level = 0.95,
       alternative = "less",
       var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  control and sci
## t = -0.56936, df = 18, p-value = 0.2881
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 14.31935
## sample estimates:
## mean of x mean of y
##     126.1     133.1
```

We can also use the above formulas to find the p-value.

```
control = c(131, 115, 124, 131, 122, 117, 88, 114, 150, 169)
sci= c(60, 150, 130, 180, 163, 130, 121, 119, 130, 148)
xbar.c = mean(control)
s.sq.c = var(control)
ybar.s = mean(sci)
s.sq.s = var(sci)
n.c = length(control)
n.s = length(sci)
##
s.pool = sqrt(((n.c-1)*s.sq.c+(n.s-1)*s.sq.s)/(n.c+n.s-2))
##
TS = (xbar.c - ybar.s)/(s.pool*sqrt(1/n.c + 1/n.s))
## left-tailed test
p.value = pt(TS, df = n.c+n.s-2)
p.value
```

```
## [1] 0.2880734
```

With p-value $= 0.288 > 0.05$, we fail to reject the **null hypothesis** and reject the alternative hypothesis. We don't have the sample evidence to support the claim that healthy subjects exhibit lower pressure than SCI subjects.

Remark: **t.test()** can conduct two-sample test with unequal variances.

8.5.3 Paired t-test

The paired t-test is widely used in clinical studies. For example, an investigator wants to assess the effect of an intervention in reducing systolic blood pressure

(SBP) in a pre-post design. Here, for each patient, there would be two observations of SBP, that is, before and after. Here instead of individual observations, the difference between pairs of observations would be of interest and the problem reduces to the one-sample situation where the null hypothesis would be to test the mean difference in SBP equal to zero against the alternate hypothesis of mean SBP being not equal to zero. The underlying assumption for using paired t-test is that under the null hypothesis **the population of difference is normally distributed** and this can be judged using the sample values.

t.test() can do the paired test. We can also convert the two-sample problem to a single-sample problem. Let $\{x_1, x_2, \cdots, x_n\}$ be the before sample and $\{y_1, y_2, \cdots, y_n\}$ be the after sample. We can take the difference of the corresponding before-after sample values to create a single sample $\{y_1 - x_1, y_2 - x_2, \cdots, y_n - x_n\}$. We can use the one-sample procedure to test the difference between before and after means.

Example 9: To study the effects of reminiscence therapy for older women with depression. She studied 15 women 60 years or older residing for 3 months or longer in an assisted living long-term care facility. For this study, depression was measured by the Geriatric Depression Scale (GDS). Higher scores indicate more severe depression symptoms. The participants received reminiscence therapy for long-term care, which uses family photographs, scrapbooks, and personal memorabilia to stimulate memory and conversation among group members. Pretreatment and post-treatment depression scores are given in the following table. Can we conclude, based on these data, that subjects who participate in reminiscence therapy experience, on average, a decline in GDS depression scores? Let $\alpha = 01$.

```
Pre-GDS:  12 10 16 2 12 18 11 16 16 10 14 21 9 19 20
Post-GDS: 11 10 11 3  9 13  8 14 16 10 12 22 9 16 18
```

Solution: The **claim** is that subjects who participate in reminiscence therapy experience, on average, a decline in GDS depression scores. $\mu_{after} - \mu_{before} < 0$. This is a left-tailed test.

$$H_o : \mu_{after} - \mu_{before} \geq 0 \leftrightarrow H_a : \mu_{after} - \mu_{before} < 0.$$

```
pre.GDS = c(12, 10, 16, 2, 12, 18, 11, 16, 16, 10, 14, 21, 9, 19, 20)
post.GDS = c(11, 10, 11, 3,  9, 13,  8, 14, 16, 10, 12, 22, 9, 16, 18)
##
t.test(post.GDS, pre.GDS,
       conf.level = 0.95,
       alternative = "less",
       paired = TRUE)
```

```
##
##  Paired t-test
```

```
##
## data:  post.GDS and pre.GDS
## t = -3.167, df = 14, p-value = 0.003428
## alternative hypothesis: true mean difference is less than 0
## 95 percent confidence interval:
##        -Inf -0.7101668
## sample estimates:
## mean difference
##            -1.6
```

```
pre.GDS = c(12, 10, 16, 2, 12, 18, 11, 16, 16, 10, 14, 21, 9, 19, 20)
post.GDS = c(11, 10, 11, 3,  9, 13,  8, 14, 16, 10, 12, 22, 9, 16, 18)
dif.GDS = post.GDS - pre.GDS
t.test(dif.GDS,
       mu = 0,
       conf.level = 0.95,
       alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  dif.GDS
## t = -3.167, df = 14, p-value = 0.003428
## alternative hypothesis: true mean is less than 0
## 95 percent confidence interval:
##        -Inf -0.7101668
## sample estimates:
## mean of x
##      -1.6
```

8.5.4 Concluding Remarks

The R built-in function **t.test(x,y)** is a black-box method that is convenient to perform the t-tests. For a two-sample test, the difference in the hypothesis testing is defined to be $\mu_x - \mu_y$. **Reversing the order will result in a wrong answer**.

prop.test() can also be used to test the equality of two proportions. We also need to pay attention to the order of the two proportions.

Using built-in functions **t.test()** and **prop.test()** is convenient, but we have to know how these functions were set up to avoid unnecessary mistakes. Using formulas to perform test hypotheses can help enhance the understanding of the concept of these testing procedures.

8.6 Practice Problems

This is an open book and open note exam. The level of detail in your solution should be similar to that in the examples in the class notes. Please keep in mind that interpretation of results is as important as generation of the results. You can use either the built-in R functions or the formulas given in the class notes to complete the exams.

For confidence interval and testing hypothesis problems, you need to **justify** the sampling distributions and interpret the results. The default confidence level is 0.95 and the default significance level is 0.05.

Problem 1

In a study of the physical endurance levels of male college freshmen, the following composite endurance scores based on several exercise routines were collected.

```
254, 281, 192, 260, 212, 179, 225, 179, 181, 149,
182, 210, 235, 239, 258, 166, 159, 223, 186, 190,
180, 188, 135, 233, 220, 204, 219, 211, 245, 151,
198, 190, 151, 157, 204, 238, 205, 229, 191, 200,
222, 187, 134, 193, 264, 312, 214, 227, 190, 212,
165, 194, 206, 193, 218, 198, 241, 149, 164, 225,
265, 222, 264, 249, 175, 205, 252, 210, 178, 159,
220, 201, 203, 172, 234, 198, 173, 187, 189, 237,
272, 195, 227, 230, 168, 232, 217, 249, 196, 223,
232, 191, 175, 236, 152, 258, 155, 215, 197, 210,
214, 278, 252, 283, 205, 184, 172, 228, 193, 130,
218, 213, 172, 159, 203, 212, 117, 197, 206, 198,
169, 187, 204, 180, 261, 236, 217, 205, 212, 218,
191, 124, 199, 235, 139, 231, 116, 182, 243, 217,
251, 206, 173, 236, 215, 228, 183, 204, 186, 134,
188, 195, 240, 163, 208
```

Use the above data to construct a frequency table and a histogram. Describe your findings from the histogram.

Problem 2

Iron deficiency anemia is an important nutritional health problem in the United States. A dietary assessment was performed on 51 boys 9 to 11 years of age whose families were below the poverty level. The mean daily iron intake among these boys was found to be 12.50 mg with a standard deviation of 4.75 mg. Suppose the mean daily iron intake among a large population of 9- to 11-year-old boys from all income strata is 14.44 mg. We want to test whether the mean iron intake among the low-income group is different from that of the general population. Carry out the hypothesis test using the critical-value method with an α level of .05. State the hypotheses that we can use to consider this question and summarize your findings.

Problem 3

A topic of recent clinical interest is the possibility of using drugs to reduce infarct size in patients who have had a myocardial infarction within the past 24 hours. Suppose we know that in untreated patients the mean infarct size is 25 (ck-g-EQ/m2). Furthermore, in 8 patients treated with a drug, the mean infarct size is 16 with a standard deviation of 10. Is the drug effective in **reducing** infarct size? Assuming that the infarct size is normally distributed.

Problem 4

Drug A was prescribed for a random sample of 12 patients complaining of insomnia. An independent random sample of 16 patients with the same complaint received drug B. The number of hours of sleep experienced during the second night after treatment began was as follows.

```
A: 3.5, 5.7, 3.4, 6.9, 17.8, 3.8, 3.0, 6.4, 6.8, 3.6, 6.9, 5.7
B: 4.5, 11.7, 10.8, 4.5, 6.3, 3.8, 6.2, 6.6, 7.1, 6.4, 4.5, 5.1, 3.2, 4.7, 4.5, 3.0
```

Construct a 95 percent confidence interval for the difference between the population means. Assume that the populations are normal and variances are unknown but equal.

Problem 5

Can we conclude that, on average, lymphocytes and tumor cells differ in size? The following are the cell diameters (μm) of 40 lymphocytes and 50 tumor cells obtained from biopsies of tissue from patients with melanoma.

```
## Lymphocytes
9.0,  9.4,  4.7,   4.8,  8.9,  4.9,  8.4,  5.9,  6.3,  5.7,
5.0,  3.5,  7.8,  10.4,  8.0,  8.0,  8.6,  7.0,  6.8,  7.1,
5.7,  7.6,  6.2,   7.1,  7.4,  8.7,  4.9,  7.4,  6.4,  7.1,
6.3,  8.8,  8.8,  5.2,  7.1,  5.3,   4.7,  8.4,  6.4,  8.3

## Tumor Cells
12.6, 14.6, 16.2, 23.9, 23.3, 17.1, 20.0, 21.0, 19.1, 19.4,
16.7, 15.9, 15.8, 16.0, 17.9, 13.4, 19.1, 16.6, 18.9, 18.7,
20.0, 17.8, 13.9, 22.1, 13.9, 18.3, 22.8, 13.0, 17.9, 15.2,
17.7, 15.1, 16.9, 16.4, 22.8, 19.4, 19.6, 18.4, 18.2, 20.7,
16.3, 17.7, 18.1, 24.3, 11.2, 19.5, 18.6, 16.4, 16.1, 21.5
```

What statistical method you are going to use to conduct the analysis? What are the assumptions for the method? Interpret your result appropriately.

Problem 6

To evaluate the analgesic effectiveness of a daily dose of oral methadone in patients with chronic neuropathic pain syndromes. The researchers used a visual analog scale (0–100 mm, a higher number indicates higher pain) ratings for maximum pain intensity over the course of the day. Each subject took either 20

mg of methadone or a placebo each day for 5 days. Subjects did not know which treatment they were taking. The following table gives the mean maximum pain intensity scores for the 5 days on methadone and the 5 days on placebo.

subject ID:	1	2	3	4	5	6	7	8	9	10	11
methadone:	29.8	73.0	98.6	58.8	60.6	57.2	57.2	89.2	97.0	49.8	37.0
placebo:	57.2	69.8	98.2	62.4	67.2	70.6	67.8	95.6	98.4	63.2	63.6

Do these data provide sufficient evidence, at the .05 level of significance, to indicate that, in general, the maximum pain intensity is lower on days when methadone is taken? Perform a formal inferential procedure and interpret the result.

Problem 7

A study was conducted on genetic and environmental influences on cholesterol levels. The data set used for the study was obtained from a twin registry in Sweden. Specifically, four populations of adult twins were studied: (1) monozygotic (MZ) twins reared apart, (2) MZ twins reared together, (3) dizygotic (DZ) twins reared apart, and (4) DZ twins reared together. One issue is whether it is necessary to correct **potential differences** in sex before performing more complex genetic analyses. The data in the following table were presented for total cholesterol levels for MZ twins reared apart, by sex.

	Men	Women
Mean:	253.3	271.0
sd:	44.1	44.1
size(n):	44	48

If we assume (a) serum cholesterol is normally distributed, (b) the samples are independent, and (c) the standard deviations for men and women are the same.

Using a two-sided test. State the hypotheses being tested, and implement the method. Report a p-value and interpret the result.

Chapter 9

Analysis of Variance (ANOVA)

In the previous module, we introduced the test on equality of two population variances. A natural question is how to compare three, four, and more population means.

9.1 The Question of One-way ANOVA

Figure 9.1: Mussule Mytilus trossulus

Consider here are some data on a shell measurement in the mussel Mytilus trossulus from five locations: Tillamook, Oregon; Newport, Oregon; Petersburg, Alaska; Magadan, Russia; and Tvarminne, Finland, taken from a much larger data set.

Tillamook	Newport	Petersburg	Magadan	Tvarminne
(μ_1)	(μ_2)	(μ_3)	(μ_4)	(μ_5)
0.0571	0.0873	0.0974	0.1033	0.0703
0.0813	0.0662	0.1352	0.0915	0.1026
0.0831	0.0672	0.0817	0.0781	0.0956
0.0976	0.0819	0.1016	0.0685	0.0973
0.0817	0.0749	0.0968	0.0677	0.1039
0.0859	0.0649	0.1064	0.0697	0.1045
0.0735	0.0835	0.105	0.0764	
0.0659	0.0725		0.0689	
0.0923				
0.0836				

The statistical **null hypothesis** is that the mean lengths of mussel shell *are the same across the five locations*. That is, the null hypothesis has the following form.

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu 5.$

The **alternative hypothesis** is that the mean lengths of the mussel shells of the five locations are *not all equal*.

H_a : at least one of the means is different from the other.

9.1.1 What is ANOVA

ANOVA is a statistical technique that assesses potential differences in a continuous dependent variable by a categorical variable having 2 or more categories. For example, an ANOVA can examine potential differences in the starting salaries of undergraduate students majoring in Biology, Mathematics, and Psychology. We have learned the two-sample t-test to compare the difference between the two population means. To compare three or more population means, we need a new procedure - analysis of variance (ANOVA).

The use of ANOVA depends on the research designs (see the class note of week #2). One-way and k-way ANOVAs are commonly used in practice.

- **One-way ANOVA** involves a single factor variable that has several categories. The one-way ANOVA addresses whether the means across categories are equal. If the means are not equal, the ANOVA does not tell which specific difference. Separate procedures are needed to perform pairwise comparisons to detect the specific difference.
- **Two-way ANOVA** involves two-factor variables. Each factor variable has several categories. The null hypotheses could be different from situation to situation. This is not the main to cover in this module. We will use multiple linear regression approaches to this problem.

9.1.2 Assumptions of ANOVA

There are two basic assumptions for the **classical ANOVA**.

- The response variable has a normal distribution with potentially different means at different factor levels.
- The variance of the response variable has constant variance (the variances of different categories are equal).

The ANOVA procedures are sensitive to the normality assumption of the response variable. In practice, we need to perform diagnostic analysis and make sure the assumptions are satisfied. If the assumptions are violated, the resulting p-values might not be correct.

If the constant variance assumption is not satisfied, the distribution of the variance ratio in the ANOVA table is NOT the specified F distribution. One way to handle multiple comparisons with this unequal variance data is the well-known Welch ANOVA. We will use it in the data analysis of the case study.

9.2 Steps of ANOVA

In this section, we outline the steps for the analysis of variance.

9.2.1 ANOVA Tables

All computer programs that perform ANOVA generate the following ANOVA table.

Source	SS	DF	MS	F
Between	SS_B	$K-1$	$MS_B = SS_B/(K-1)$	$F = MS_B/MSE_W$
Within	SS_W	$N-K$	$MS_W = SS_W/(N-K)$	
Total	SS_T	$N-1$		

I will use the mussel length data as an example to calculate the quantities in the above table.

Tillamook	Newport	Petersburg	Magadan	Tvarminne
(μ_1)	(μ_2)	(μ_3)	(μ_4)	(μ_5)
0.0571	0.0873	0.0974	0.1033	0.0703
0.0813	0.0662	0.1352	0.0915	0.1026
0.0831	0.0672	0.0817	0.0781	0.0956
0.0976	0.0819	0.1016	0.0685	0.0973
0.0817	0.0749	0.0968	0.0677	0.1039
0.0859	0.0649	0.1064	0.0697	0.1045
0.0735	0.0835	0.105	0.0764	

Tillamook	Newport	Petersburg	Magadan	Tvarminne
0.0659	0.0725		0.0689	
0.0923				
0.0836				
$n_1, \bar{x}_1, s_1$	$n_2, \bar{x}_2, s_2$	$n_3, \bar{x}_3, s_3$	$n_4, \bar{x}_4, s_4$	$n_5, \bar{x}_5, s_5$

In the last row of the table, we calculated the sample mean ($\bar{x}_i$), sample standard deviation (s_i), and sample size (n_i) of each of the $K = 5$ locations. Let $\bar{x}$ is the mean of the combined sample mean and $N = n_1 + n_2 + n_3 + n_4 + n_5$.

- The sum of squared deviations between groups.

$$SS_B = \sum n_i(\bar{x}_i - \bar{x})^2.$$

- The sum of squared deviation within groups

$$SS_W = \sum (n_i - 1)s_i^2.$$

- The sum of the total error

$$SS_T = SS_B + SS_W.$$

Important Result: If the ANOVA assumptions are satisfied, $F \approx F_{K-1,N-K}$.

If the p-value of the **F test** is less than the given significance level, we reject the null hypothesis.

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu 5$.

This implies that at least one of the means is different from the other.

9.2.2 Post hoc Pairwise Comparison

If the null hypothesis of the ANOVA is rejected, we need to perform a pairwise comparison the identify the significant difference between the means. There are different types of comparisons: pairwise comparison and various simultaneous comparison procedures are available in R.

9.2.3 How to Summarize ANOVA Results

- **Describe the test type you used and the purpose of the test** - An ANOVA is appropriate for multiple test subjects. In the above example, the measurements of the same shell were taken from 5 locations.
- **Write whether or not a significant difference existed between the means of each test group**. Write "There was" or "There was not a significant effect of the factor."

- **Write the results of the F test**. Write the F test, followed by a parenthesis, then the two sets of degrees of freedom values separated by a comma, followed by an equal sign and the F value. The statement is something like: "F (two sets of degrees of freedom) = F value, p = p-value."
- **Write "Post hoc test comparison" if a significant result is present**. Write the post hoc test you used.
- **Recap the results in an easy-to-understand sentence or two**. We could write "A significant difference existed between locations.

9.3 Welch's ANOVA

The classical ANOVA assumes that the population is normal and variance is constant (i.e., group variances are equal). This is the base to define the F test. If the assumptions are violated, the variance ratio (F test statistic) does not follow an F distribution.

One method commonly used in practice is the Welch ANOVA in which the degrees of freedom are modified to approximate the F distribution. The "Games-Howell" test is commonly used for the post-hoc multiple comparisons for the Welch ANOVA. The Games-Howell post-hoc test is another nonparametric approach to compare combinations of groups or treatments.

In R, `oneway.test()` in **stats** performs Welch ANOVA. It is routine to use both classical ANOVA and Welch's ANOVA. If there are no violations of the model, both ANOVAs yield similar results. If the two results are different, Welch ANOVA is more appropriate. The classical ANOVA is easy to interpret and understandable. If there are no violations, we always report the classical ANOVA.

9.4 Case Study: Mussel Length Example - Solution

We will use R to conduct ANOVA and post hoc comparison on the mussel length example.

9.4.1 Creating An ANOVA Table

Next, we create an R data set (data frame) based on the given data table using the following R code to perform the ANOVA procedure.

```
x1 = c(0.0571,0.0813, 0.0831, 0.0976, 0.0817, 0.0859, 0.0735, 0.0659, 0.0923, 0.0836)
x2 = c(0.0873,0.0662, 0.0672, 0.0819, 0.0749, 0.0649, 0.0835, 0.0725)
x3 = c(0.0974,0.1352, 0.0817, 0.1016, 0.0968, 0.1064, 0.1050)
```

Table 9.4: Analysis of variance table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
location	4	0.0045197	0.0011299	7.12102	0.0002812
Residuals	34	0.0053949	0.0001587	NA	NA

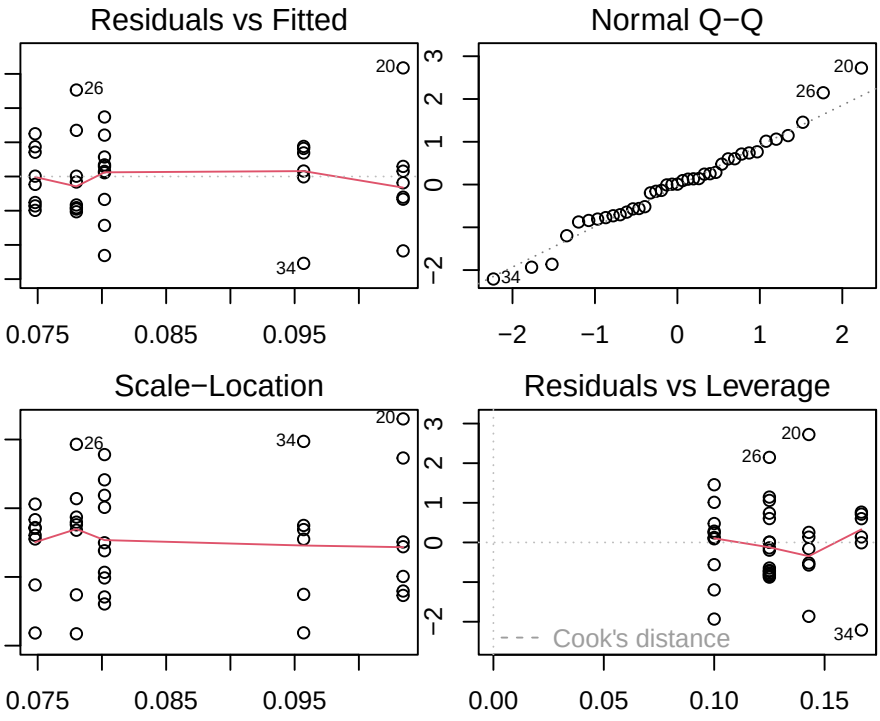
```
x4 = c(0.1033,0.0915, 0.0781, 0.0685, 0.0677, 0.0697, 0.0764, 0.0689)
x5 = c(0.0703,0.1026, 0.0956, 0.0973, 0.1039, 0.1045)
mussel.len = c(x1, x2, x3, x4, x5)      # pool all sub-samples of lengths
location = c(rep("Tillamook", length(x1)),
             rep("Newport", length(x2)),
             rep("Petersburg", length(x3)),
             rep("Magadan", length(x4)),
             rep("Tvarminne", length(x5)))  # location vector matches the lengths
data.matrix = cbind(len = mussel.len, location = location)   # data a data table
musseldata = as.data.frame(data.matrix)        # data frame
model01 = lm(len ~ location, data = musseldata)  # a model for extracting ANOVA
anova.model = anova(model01)   # creating the ANOVA table
kable(anova.model, caption = "Analysis of variance table")
```

The ANOVA test indicates that not all means are equal ($F(4, 34) = 7.12, p = 0.00028$). The follow-up pair-wise comparisons of the group means are given in the next subsection.

9.4.2 Diagnostic Analysis

Since the ANOVA is sensitive to the normality assumption of the population and the constant variance. We need to check whether there are violations of the assumption. The following diagnostic plots are used to check the appropriateness of the ANOVA test.

```
aov.model = aov(len~location, data=musseldata)
par(mfrow=c(2,2), mar=c(2,1,2,1))
plot(aov.model)
```



Constant variance: If all points on the top-left graph are evenly spread around the horizontal axis (the red curve overlaps with the horizontal axis), then the constant variance assumption is satisfied.

Normal population: If all points on the top right graph are all close to the off-diagonal line, then the normality assumption is satisfied.

The above plot showed that there are no significant violations of the assumptions of the ANOVA test. Next, we perform pair-wise comparisons.

```
welch = oneway.test(mussel.len ~ location)
F.stats = as.vector(welch$statistic)
num.df = as.vector(welch$parameter[1])
denom.df = as.vector(welch$parameter[2])
p.value = as.vector(welch$p.value)
cap.text = welch$method
kable(cbind(F.stats = F.stats, num.df = num.df,
            denom.df=denom.df, p.value = p.value), caption = cap.text)
```

9.4.3 Welch ANOVA

Although no significant violations were observed in the above residual plots, we still perform the Welch ANOVA for illustrative purposes.

Table 9.5: One-way analysis of means (not assuming equal variances)

F.stats	num.df	denom.df	p.value
5.664479	4	15.69546	0.0050796

Table 9.6: Descriptive statistics of sub-populations

	n	means	variances
Magadan	8	0.0780125	0.0001676
Newport	8	0.0748000	0.0000739
Petersburg	7	0.1034429	0.0002627
Tillamook	10	0.0802000	0.0001431
Tvarminne	6	0.0957000	0.0001680

```
source("https://raw.githubusercontent.com/pengdsci/STA501/main/ref/anova-posthoc-test.t
# posthoc.tgh <- function(y, x, method=c("games-howell", "tukey"), digits=2)
posthoc = posthoc.tgh(mussel.len, location, method="tukey",digits=2 )
descriptives.stats = posthoc$intermediate$descriptives
kable(descriptives.stats, caption = "Descriptive statistics of sub-populations")
```

We can see that both classical and Welch ANOVA yield the same result (both p-values are < 0.01). Therefore, we report the results of classical ANOVA and Tukey's HSD procedure.

9.4.4 Post-hoc Multiple Comparisons

Since the null hypothesis was rejected, it is necessary to quantify the differences between groups in order to determine which groups significantly differ from each other.

There are several multiple comparison tests with implementations in various R libraries. We will use Tukey's Honest Significant Differences (HSD). The Tukey HSD procedure will run a pairwise comparison of all possible combinations of groups and test these pairs for significant differences between their means, all while adjusting the p-value. If the assumption of constant variance is violated, the Welch ANOVA will be carried out. The Games-Howell test will be used if the null hypothesis in the Welch ANOVA is rejected.

The following R code performs the pair-wise comparisons between the group means.

```
aov.model = aov(len~location, data=musseldata)
kable(TukeyHSD(aov.model)$location, caption = "Tukey multiple comparisons of means")
```

The multiple comparisons show that four differences are significant at level 0.05.

Table 9.7: Tukey multiple comparisons of means

	diff	lwr	upr	p adj
Newport-Magadan	-0.0032125	-0.0213487	0.0149237	0.9857956
Petersburg-Magadan	0.0254304	0.0066576	0.0442031	0.0036924
Tillamook-Magadan	0.0021875	-0.0150180	0.0193930	0.9959794
Tvarminne-Magadan	0.0176875	-0.0019019	0.0372769	0.0928839
Petersburg-Newport	0.0286429	0.0098701	0.0474156	0.0009253
Tillamook-Newport	0.0054000	-0.0118055	0.0226055	0.8934665
Tvarminne-Newport	0.0209000	0.0013106	0.0404894	0.0317354
Tillamook-Petersburg	-0.0232429	-0.0411181	-0.0053676	0.0056547
Tvarminne-Petersburg	-0.0077429	-0.0279230	0.0124373	0.8028001
Tvarminne-Tillamook	0.0155000	-0.0032310	0.0342310	0.1446987

Table 9.8: Games-Howell multiple comparisons

	t	df	p
Magadan:Newport	0.5847249	12.169510	0.9748911
Magadan:Petersburg	3.3254180	11.496217	0.0412241
Magadan:Tillamook	0.3684022	14.550533	0.9956174
Magadan:Tvarminne	2.5281745	10.915427	0.1539440
Newport:Petersburg	4.1880670	8.857240	0.0156494
Newport:Tillamook	1.1127299	15.868239	0.7976041
Newport:Tvarminne	3.4248678	8.205773	0.0506123
Petersburg:Tillamook	3.2279514	10.436333	0.0530007
Petersburg:Tvarminne	0.9564490	10.967062	0.8686901
Tillamook:Tvarminne	2.3828489	9.970454	0.1972836

These differences will be reported in the summary.

For illustrative purposes, we carry the Games-Howell test for multiple comparisons in the following.

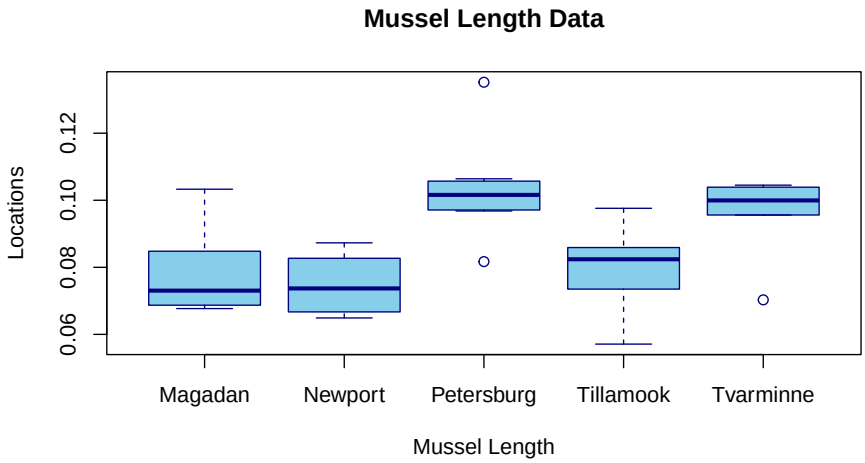
```
Games.howell = posthoc$output$games.howell
kable(Games.howell, caption="Games-Howell multiple comparisons")
```

The results are slightly different from Tukey's HSD test. Next, we present a visual presentation before summarizing the ANOVA result and post-hoc comparisons.

9.4.5 Visual Comparison - Boxplots

```
# Box-plot of mussel length by locations
boxplot(mussel.len ~ location,
```

```
        main="Mussel Length Data", # plot title
        xlab="Mussel Length",      # label of x-axis
        ylab="Locations",          # label y-axis
        border = "navy",           # border color box-plot
        col="skyblue")             # box color
```



The box plot also confirms that the group means are not identical.

9.4.6 ANOVA Reporting

We conducted a one-way ANOVA test on the equality of the mean lengths of mussel shells in the five locations and found a statistical significance with F(4, 34) = 7.12, p-value = 0.00028. This implies that the means at different locations are not identical.

After we conducted pairwise comparisons of the means of the lengths of mussel shells across five locations using the HSD test, we found that, among the 10 pairwise comparisons, four of them are significant with an adjusted p-value less than 0.05: Petersburg - Magadan = 0.0254304 (p = 0.0037), Petersburg - Newport = 0.0286429 (p = 0.0009), Tvarminne - Newport = 0.0209 (p = 0.032), and Tillamook - Petersburg = -0.0232429 (p = 0.0057). The group box plot also showed the difference between the locations.

9.5 Practice Problems

The effects of thermal pollution on Asiatic clams at three different geographical locations were analyzed by researchers. Sample data on clamshell length, width, and height are displayed in the following table.

Location 1			Location 2			Location 3		
Length	Width	Height	Length	Width	Height	Length	Width	Height
7.20	6.10	4.45	7.25	6.25	4.65	5.95	4.75	3.20
7.50	5.90	4.65	7.23	5.99	4.20	7.60	6.45	4.56
6.89	5.45	4.00	6.85	5.61	4.01	6.15	5.05	3.50
6.95	5.76	4.02	7.07	5.91	4.31	7.00	5.80	4.30
6.73	5.36	3.90	6.55	5.30	3.95	6.81	5.61	4.22
7.25	5.84	4.40	7.43	6.10	4.60	7.10	5.75	4.10
7.20	5.83	4.19	7.30	5.95	4.29	6.85	5.55	3.89
6.85	5.75	3.95	6.90	5.80	4.33	6.68	5.50	3.90
7.52	6.27	4.60	7.10	5.81	4.26	5.51	4.52	2.70
7.01	5.65	4.20	6.95	5.65	4.31	6.85	5.53	4.00
6.65	5.55	4.10	7.39	6.04	4.50	7.10	5.80	4.45
7.55	6.25	4.72	6.54	5.89	3.65	6.81	5.45	3.51
7.14	5.65	4.26	6.39	5.00	3.72	7.30	6.00	4.31
7.45	6.05	4.85	6.08	4.80	3.51	7.05	6.25	4.71
7.24	5.73	4.29	6.30	5.05	3.69	6.75	5.65	4.00
7.75	6.35	4.85	6.35	5.10	3.73	6.75	5.57	4.06
6.85	6.05	4.50	7.34	6.45	4.55	7.35	6.21	4.29
6.50	5.30	3.73	6.70	5.51	3.89	6.22	5.11	3.35
6.64	5.36	3.99	7.08	5.81	4.34	6.80	5.81	4.50
7.19	5.85	4.05	7.09	5.95	4.39	6.29	4.95	3.69
7.15	6.30	4.55	7.40	6.25	4.85	7.55	5.93	4.55
7.21	6.12	4.37	6.00	4.75	3.37	7.45	6.19	4.70
7.15	6.20	4.36	6.94	5.63	4.09	6.70	5.55	4.00
7.30	6.15	4.65				7.51	6.20	4.74
6.35	5.25	3.75				6.95	5.69	4.29
						7.50	6.20	4.65

Source: Data provided courtesy of John Brooker, M.S. and the Wright State University Statistical Consulting Center.

Figure 9.2: Clam shell data table

The objective is to determine if there is a significant difference in mean length, height, or width (measured in mm) of the clamshell at the three different locations by performing three analyses.

To save you some time, I created three data sets for length, width, and height respectively in the following code chunk. **You only choose ONE of the three sets** to complete this week's assignment.

```
# length
Len.loc.1 = c(7.20, 7.50, 6.89, 6.95, 6.73, 7.25, 7.20, 6.85, 7.52, 7.01, 6.65,
              7.55, 7.14, 7.45, 7.24, 7.75, 6.85, 6.50, 6.64, 7.19, 7.15, 7.21,
              7.15, 7.30, 6.35)
Len.loc.2 = c(7.25, 7.23, 6.85, 7.07, 6.55, 7.43, 7.30, 6.90, 7.10, 6.95, 7.39,
              6.54, 6.39, 6.08, 6.30, 6.35, 7.34, 6.70, 7.08, 7.09, 7.40, 6.00,
              6.94)
```

```
Len.loc.3 = c(5.95, 7.60, 6.15, 7.00, 6.81, 7.10, 6.85, 6.68, 5.51, 6.85, 7.10,
              6.81, 7.30, 7.05, 6.75, 6.75, 7.35, 6.22, 6.80, 6.29, 7.55, 7.45,
              6.70, 7.51, 6.95, 7.50)

# Width
Width.loc.1 = c(6.10, 5.90, 5.45, 5.76, 5.36, 5.84, 5.83, 5.75, 6.27, 5.65, 5.55,
                6.25, 5.65, 6.05, 5.73, 6.35, 6.05, 5.30, 5.36, 5.85, 6.30, 6.12,
                6.20, 6.15,5.25)
Width.loc.2 = c(6.25, 5.99, 5.61, 5.91, 5.30, 6.10, 5.95, 5.80, 5.81, 5.65, 6.04,
                5.89, 5.00, 4.80, 5.05, 5.10, 6.45, 5.51, 5.81, 5.95, 6.25, 4.75,
                5.63)
Width.loc.3 = c(4.75, 6.45, 5.05, 5.80, 5.61, 5.75, 5.55, 5.50, 4.52, 5.53, 5.80,
                5.45, 6.00, 6.25, 5.65, 5.57, 6.21, 5.11, 5.81, 4.95, 5.93, 6.19,
                5.55, 6.20, 5.69, 6.20)

## Height
Height.loc.1 = c(4.45, 4.65, 4.00, 4.02, 3.90, 4.40, 4.19, 3.95, 4.60, 4.20, 4.10,
                 4.72, 4.26, 4.85, 4.29, 4.85, 4.50, 3.73, 3.99, 4.05, 4.55, 4.37,
                 4.36, 4.65, 3.75)
Height.loc.2 = c(4.65, 4.20, 4.01, 4.31, 3.95, 4.60, 4.29, 4.33, 4.26, 4.31, 4.50,
                 3.65, 3.72, 3.51, 3.69, 3.73, 4.55, 3.89, 4.34, 4.39, 4.85, 3.37,
                 4.09)
Height.loc.3 = c(3.20, 4.56, 3.50, 4.30, 4.22, 4.10, 3.89, 3.90, 2.70, 4.00, 4.45,
                 3.51, 4.31, 4.71, 4.00, 4.06, 4.29, 3.35, 4.50, 3.69, 4.55, 4.70,
                 4.00, 4.74, 4.29, 4.65)
```

The following are the steps for completing the assignment:

1. Pick **ONE** of the three data sets (length, width, and height) from the above code chunk.
2. Modify the code in 4.1 to create an R data set and perform the classical ANOVA analysis. Interpret the ANOVA results as I did in the class note.
3. Perform a diagnostic analysis using the residual plot as shown in Section 4.2. Explain whether you observe any violations of the model assumptions.
4. If the null hypothesis is rejected, carry out multiple comparisons between the three locations using either Tukey's HSD or Games-Howell procedures.
5. Report the ANOVA analysis results similar to what was reported in the note.

Chapter 10

Correlation and Simple Linear Regression

In the previous module, we discussed the relationship between a continuous variable (the length of mussel shells) and a categorical variable (location, also called factor variable). If there is no association between the two variables, the means of all populations are equal. If there is an association between the continuous variable and the factor variable, then the means of the populations are not identical.

The continuous variable is assumed to normal distribution. The continuous variable is also called the response variable (dependent variable) and the factor variable is called the predictor variable (or explanatory variable).

A natural question is how to characterize the relationship between two continuous variables.

10.1 The question and the data

Example: Amyotrophic lateral sclerosis (ALS) is characterized by a progressive decline of motor function. The degenerative process affects the respiratory system. To investigate the longitudinal impact of nocturnal noninvasive positive-pressure ventilation on patients with ALS. Prior to treatment, they measured the partial pressure of arterial oxygen (Pao2) and partial pressure of arterial carbon dioxide (Paco2) in patients with the disease. The results were as follows:

Source: M. Butz, K. H.Wollinsky, U.Widemuth-Catrinescu, A. Sperfeld, S. Winter, H. H. Mehrkens, A. C. Ludolph, and H. Schreiber, "Longitudinal Effects of Noninvasive Positive-Pressure Ventilation in Patients with Amyotrophic Lateral Sclerosis," *American Journal of Medical Rehabilitation*, 82 (2003) 597–604.

PTID	Paco2	Pao2	PTID	Paco2	Pao2
1	40.0	40.0	16	41.8	41.8
2	47.0	47.0	17	33.0	33.0
3	34.0	34.0	18	43.1	43.1
4	42.0	42.0	19	52.4	52.4
5	54.0	54.0	20	37.9	37.9
6	48.0	48.0	21	34.5	34.5
7	53.6	53.6	22	40.1	40.1
8	56.9	56.9	23	33.0	33.0
9	58.0	58.0	24	59.9	59.9
10	45.0	45.0	25	62.6	62.6
11	54.5	54.5	26	54.1	54.1
12	54.0	54.0	27	45.7	45.7
13	43.0	43.0	28	40.6	40.6
14	44.3	44.3	29	56.6	56.6
15	53.9	53.9	30	59.0	59.0

Figure 10.1: Length of clamshells

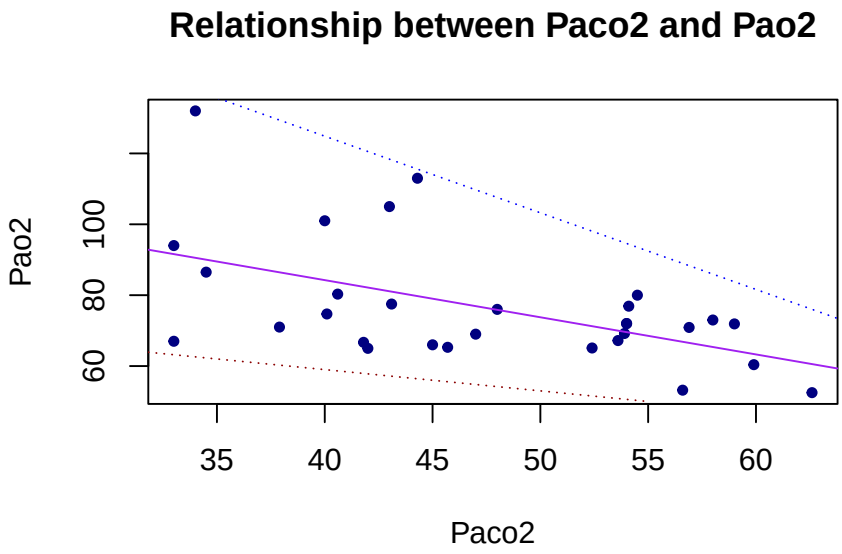
The layout of the data table is shown below. Each subject (patient ID) has one **record** with two pieces of information Paco2 and Pao2.

10.2 Visual Inspection for Association

We define two vectors to store the sample values of the partial pressure of arterial oxygen (Pao2) and the partial pressure of arterial carbon dioxide (Paco2). Before conducting the analysis, we make a scatter plot to visualize the relationship between the two continuous variables.

```
# define the data sets based on the given data table.
Paco2 =c(40.0, 47.0, 34.0, 42.0, 54.0, 48.0, 53.6, 56.9, 58.0, 45.0, 54.5, 54.0,
         43.0, 44.3, 53.9, 41.8, 33.0, 43.1, 52.4, 37.9, 34.5, 40.1, 33.0, 59.9,
         62.6, 54.1, 45.7, 40.6, 56.6, 59.0)
Pao2 = c(101.0, 69.0, 132.0, 65.0, 72.0, 76.0, 67.2, 70.9, 73.0, 66.0, 80.0,
         72.0, 105.0, 113.0, 69.2, 66.7, 67.0, 77.5, 65.1, 71.0, 86.5, 74.7,
         94.0, 60.4, 52.5, 76.9, 65.3, 80.3, 53.2, 71.9)
## scatter plot
plot(Paco2, Pao2,
     pch = 20,
     col = "navy",
     main = "Relationship between Paco2 and Pao2",
     xlab = "Paco2",
     ylab = "Pao2"
```

```
)
## The following two lines are not required when you make this scatter plot
segments(30, 65, 55, 50, lty = 3, col = "darkred")
segments(33, 140, 70, 60, lty = 3, col = "blue")
abline(lm(Pao2 ~ Paco2), col = "purple")
```



We can see from the above scatter plot that there is a negative association between Paco2 and Pao2 since Pao2 decreases as Paco2 increases. We can also see that **the variance** of Pao2 is also decreasing as Paco2 increases.

How to quantify the above association?

10.3 Coefficient of Correlation

The strength of linear correlation can be measured by the linear correlation coefficient. We will introduce one such measure - the Pearson correlation coefficient.

10.3.1 Definition of Pearson correlation coefficient

One well-known quantity for measuring the **linear association** between two numerical variables is the Pearson correlation coefficient. The sample Pearson correlation coefficient is defined as

Table 10.1: Pearson correlation coefficient

r
-0.5307874

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

10.3.2 Interpretation of correlation coefficient

The interpretation of the Pearson correlation coefficient is customarily given in the following

- if $r > 0$, then x and y are positively correlated; if $r < 0$, then x and y are negatively correlated;
- if $r = 0$, there is **no** linear correlation between x and y.
- if $|r| < 0.3$, there is a **weak** linear correlation between x and y.
- if $0.3 < |r| < 0.7$, there is a **moderate** linear correlation between x and y.
- if $0.7 < |r| < 1.0$, there is a **strong** linear correlation between x and y.
- if $r = 1$, there is a **perfect** linear correlation between x and y.

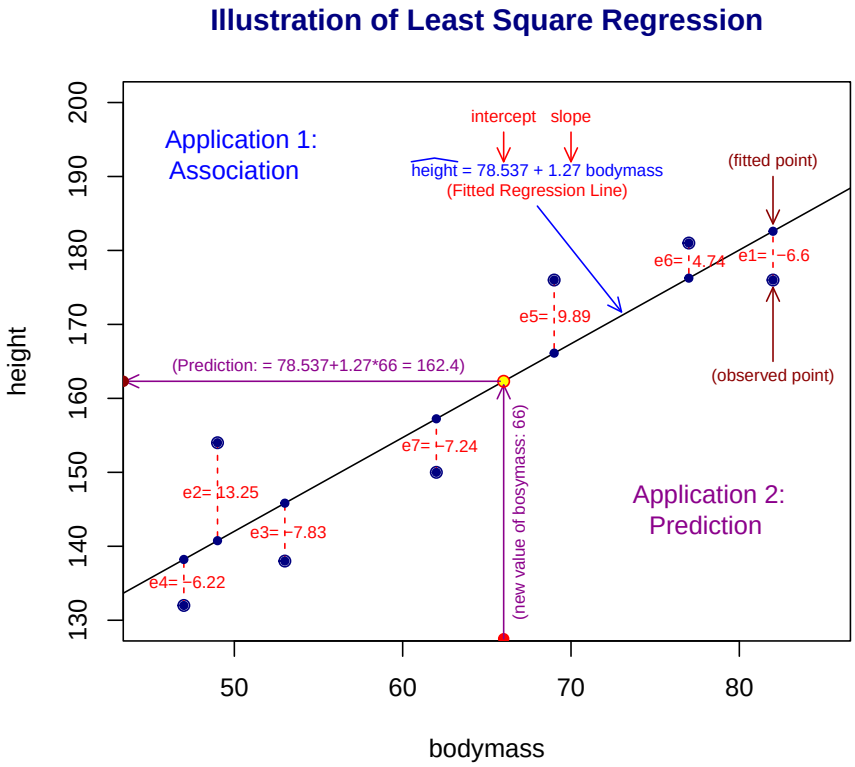
In R, we use the command **cor()** to calculate the above Pearson correlation coefficient.

```
Pearson.correlation = cbind(r=cor(Paco2, Pao2))
kable(Pearson.correlation, caption = "Pearson correlation coefficient",
      align="c")
```

Therefore, there is a weak negative linear correlation between the partial pressure of arterial oxygen (Pao2) and the partial pressure of arterial carbon dioxide (Paco2).

10.4 Least square regression: structure, diagnostics, and applications

For illustration purposes, we make the following plot based on an artificial data set to introduce several important concepts of the least square regression model.



10.4.1 Definitions

The following concepts are annotated in the above figure.

- The variable associated with the vertical variable is called the **response** variable. The *response** variable is always placed on the left-hand side of the model formula.
- The variable that impacts the value of the response variable is called the explanatory variable (also called predictor, independent variables).
- The points in the figure plotted based on the data set are called observed data points.
- The line in the figure is called the **fitted regression line**.
- The points on the fitted regression line are called **fitted data points**.
- The difference between coordinates observed and fitted points is called ** the residual** of the observed data point. The residual is, in fact, called

the fitted error. It reflects the goodness of the fitted regression line. e_i $(i = 1, 2, \cdots, n)$ are the estimated residual errors.

- The **best** regression line is obtained by minimizing the sum of the squared errors, $e_1^2 + e_2^2 + \cdots + e_2^2$. This is the reason why we call the **best** regression line the **least square regression line** .
- The intercept and slope completely determine a straight line. The intercept and slope of the **least square** regression line are obtained by minimizing the sum of the squared residual errors.
- The statistical term *linear model* is the equation of the fitted regression line.
- There are two possible applications of a linear regression model.
  - **Association analysis** - basic describes how the change of explanatory variable impacts the value of the response variable.
  - **Prediction analysis** - predict the values of the response variable based on the corresponding **new values** of the explanatory variable.

In this module, we only discuss the least square regression with ONE explanatory variable. In the next module, we will generalize this model to multiple explanatory variables.

10.4.2 Assumptions of Least Square Regression

The assumptions of the least square linear regression are identical to the ones of the ANOVA.

- The response variable is a **normal random variable**. Its mean is dependent on the **non-random** explanatory variable.
- The variance of the response variable is constant (i.e., its variance is NOT dependent on the **non-random** explanatory variable).
- The relationship between the response and explanatory variables is assumed to be correctly specified.

10.4.3 Model Building and Diagnostics

We will use **lm()** and the artificial data used in the above plot to find the least square estimate of **parameters**: intercept and slope.

```
height <- c(176, 154, 138, 132, 176, 181, 150)
bodymass <- c(82, 49, 53, 47, 69, 77, 62)
ls.reg <- lm(height ~ bodymass)
parameter.estimates <- ls.reg$coef
kable(parameter.estimates,
```

Table 10.2: Least square estimate of the intercept and slope

	x
(Intercept)	78.627588
bodymass	1.267897

```
      caption = "Least square estimate of the intercept and slope",
      align='c')
```

The least-square estimated intercept and slope are approximately equal to 78.63 and 1.27. Therefore, the fitted least square regression line is $\widehat{height}_i = 78.63 + 1.27 \times bodymass_i$. The residual error $e_i = height_i - \widehat{height}_i$, for $i = 1, 2, \cdots, n$.

Since the explanatory variable is implicitly assumed to be a non-random variable, we can see the relationship between the response variable and the residual in the following general representation.

$$response.variable = \alpha + \beta \times predictor.variable + \epsilon$$

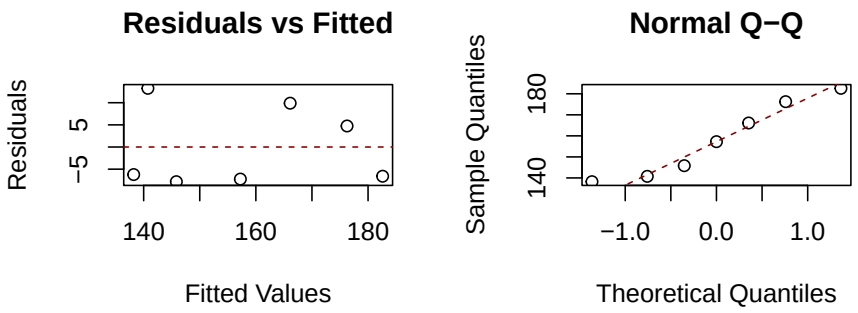
The assumption that the response variable is normal with a constant variance is equivalent to that ϵ is a normal random variable with mean 0 and constant variance σ_0^2.

The residual ϵ is estimated by the errors e_i. Therefore, we can look at the distribution of $\{e_1, e_2, \cdots, e_n\}$ to see potential violations of the model assumptions.

10.4.4 Residual Diagnostics and Remedies

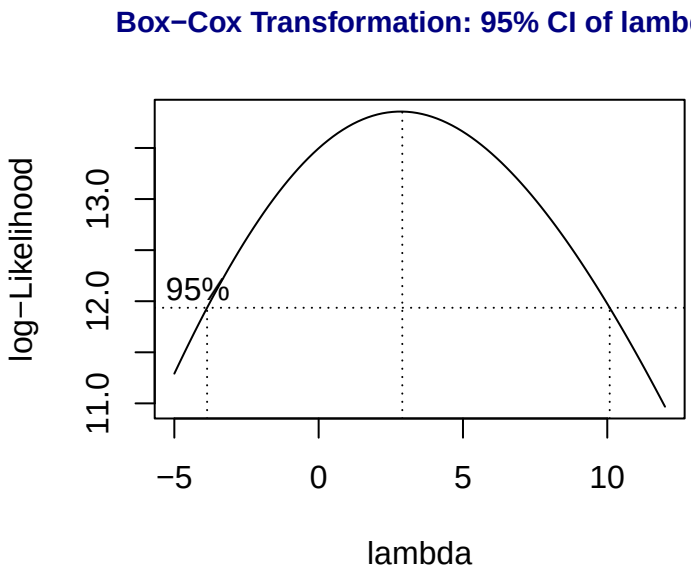
We make four default residual plots from R function **lm()** in R.

```
par(mfrow = c(1,2))  # par => graphic parameter
                     # mfrow => splits the graphic page into panels
#
plot(ls.reg$fitted.values, ls.reg$residuals,
     main="Residuals vs Fitted",       # title of the plot
     xlab = "Fitted Values",           # label of X-axis
     ylab = "Residuals"                # label of y-axis
     )
abline(h=0, lty=2, col = "darkred")
##
qqnorm(ls.reg$fitted.values, main = "Normal Q-Q")
qqline(ls.reg$fitted.values, lty = 2, col = "darkred")
```



We can see that there seems to be a minor violation of the assumption of constant variance from the residual plot in the top left figure. In statistics, we have a transformation to stabilize the variable. We will introduce the well-known Box-Cox transformation in this module. It is a generic transformation and was designed to identify the optimal power transformation to stabilize the variance and maintain the normality. Sometimes it works really well but not always. It is always worth a try in case we observe the pattern of non-constant variance.

```
library(MASS)
boxcox(height ~ bodymass,
       lambda = seq(-5, 12, length = 100),
       xlab="lambda")
##
title(main = "Box-Cox Transformation: 95% CI of lambda",
      col.main = "navy", cex.main = 0.9)
```



The above plot shows the 95% confidence interval of the power (λ) on the potential power transformation of the response variable *height*.

In practice, we choose the **most convenient** number in the interval as the power to transform the response variable. Since 1 is in the interval, it is necessary to perform a power transformation. We can also choose the logarithmic transformation of height since $\lambda = 0$ is also in the interval.

Note: A special power transformation is the logarithmic transformation of the response if we choose $\lambda = 0$.

10.4.5 Final Model with Applications

Once the final model is identified, we can use it in two different ways: association and prediction.

In the **association analysis**, we interpret the regression coefficient associated with the significant explanatory variable.

In this toy example, the summarized statistics are extracted and summarized in the following

```
kable(summary(ls.reg)$coef, caption = "Summary of regression model")
```

The p-value of testing the null hypothesis that the slope parameter is zero is 0.0075. We reject the null hypothesis H_0 : *slope* $= 0$. This implies that

Table 10.3: Summary of regression model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	78.627588	18.7505570	4.193347	0.0085442
bodymass	1.267897	0.2929519	4.328005	0.0075135

Table 10.4: The predicted height with body mass: 75

fit	lwr	upr
173.7199	172.9821	174.4577

bodymass impacts **height**. To be more specific, as the value of **bodymass** increases by one unit, the response variable **height** will increase by 1.27 cm.

In the **prediction analysis**, we can predict the **height** with any given new **bodymass** that is not in the data set. For example, assume **bodymass = 75**, we can then use the R function **predict()** to predict the **height**.

```
pred.height = predict(ls.reg, newdata = data.frame(bodymass=75),
                      interval = "prediction", level=0.05)
kable(pred.height, caption="The predicted height with body mass: 75")
```

For a given person with a body mass of 75 units, the predicted height of that person is 173.7cm with a 95% predictive interval [173.0, 174.5].

In summary: It is dependent on the objectives of your data analysis,

- if the objective is association analysis, the summary will focus on the interpretation of the regression coefficient with a p-value < 0.05.
- if the objective is prediction, the summary will focus on the predicted value and its predictive interval.
- if the objectives are both association analysis and prediction, then summarize both results as shown above.

10.5 Case Study: Amyotrophic lateral sclerosis analysis revisited

We only perform the least square regression analysis about the relationship between the partial pressure of arterial oxygen (Pao2) and partial pressure of arterial carbon dioxide (Paco2) in patients with the disease.

The scatter plot of Paco2 versus Pao2 in section 1 indicates a negative association between them. Next, we assume the linear relationship between Paco2 and Pao2 and build the least square regression in the following steps.

10.5.1 Objectives

The objective of this analysis is to build a linear regression model and then use this model to

- assess how Pao2 impacts Paco2.
- predict the value of Paco2 for given Pao2

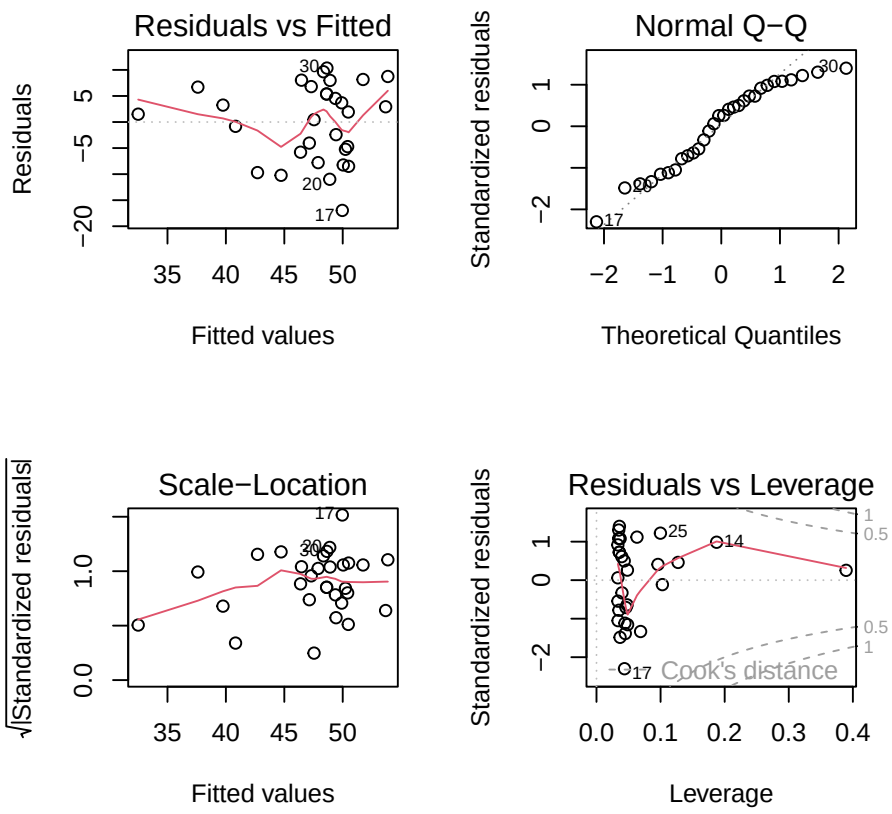
10.5.2 Model Fitting

We fit a least square regression to the data first. Based on the objective of the analysis, Paco2 will be the response variable and Pao2 will be the explanatory variable. We first fit the following model

$$Paco2 = \alpha + \beta \times Pao2 + \epsilon$$

α, β, and ϵ are called intercept, slope, and residuals, respectively. Then carry out the diagnostics of the above model and find the potential remedy.

```
Paco2 =c(40.0, 47.0, 34.0, 42.0, 54.0, 48.0, 53.6, 56.9, 58.0, 45.0, 54.5, 54.0,
         43.0, 44.3, 53.9, 41.8, 33.0, 43.1, 52.4, 37.9, 34.5, 40.1, 33.0, 59.9,
         62.6, 54.1, 45.7, 40.6, 56.6, 59.0)
Pao2 = c(101.0, 69.0, 132.0, 65.0, 72.0, 76.0, 67.2, 70.9, 73.0, 66.0, 80.0, 72.0,
         105.0, 113.0, 69.2, 66.7, 67.0, 77.5, 65.1, 71.0, 86.5, 74.7, 94.0, 60.4,
         52.5, 76.9, 65.3, 80.3, 53.2, 71.9)
##
ls.reg0 <- lm(Paco2 ~ Pao2)   # fitting a least square regression
par(mfrow = c(2,2))   # split the graphic page into 4 panels
plot(ls.reg0)
```



The residual diagnostic plots show that there are violations of the model assumptions. The Q-Q plot does not support the normality assumption of the residuals. The top-left residual plot does not support the constant variance assumption since the variance of the residual increases as the fitted value increases.

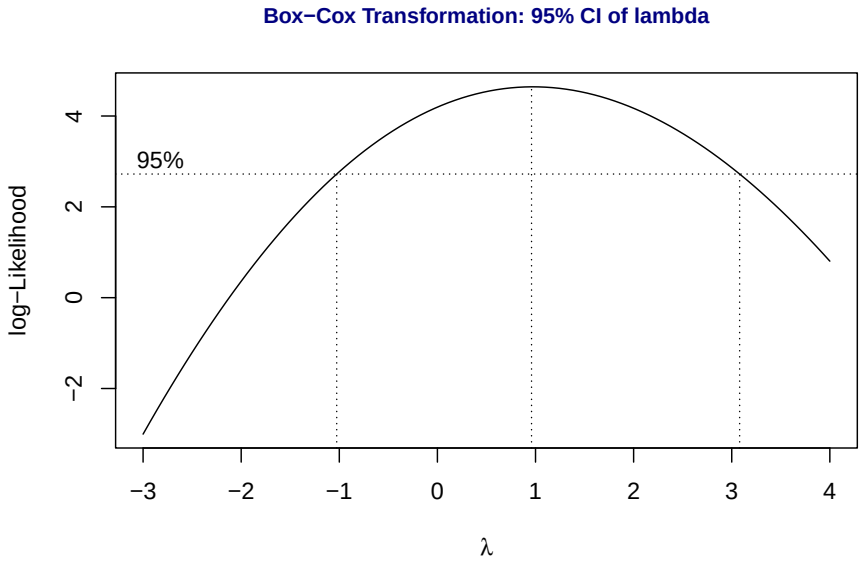
Next, we perform the Box-Cox transformation.

10.5.3 Box-Cox Transformation

In the Box-cox transformation, the range of potential λ is selected by trial and error so that the figure should contain the 95% confidence interval. We will use a function in the library **{MASS}**. If you don't have the library on your computer, you need to install it and then load it to the workspace.

```
library(MASS)
boxcox(Paco2 ~ Pao2, lambda = seq(-3, 4, length = 10),
       xlab=expression(paste(lambda)))
```

```
title(main = "Box-Cox Transformation: 95% CI of lambda",
      col.main = "navy", cex.main = 0.9)
```



The above Box-Cox procedure indicates that the power transformation will not improve the residual plots. I will not try other transformations in this course and simply use the above model as the final working model for prediction and perform association analysis.

10.5.4 Model Applications

Recall that we will use the final working model to assess the association between the Paco2 and Pao2 and predict the value of Paco2 with the new Pao2 as well.

We first present summary statistics of the least square regression model in the following table.

```
ls.reg.final <-lm(Paco2 ~ Pao2)
kable(summary(ls.reg.final)$coef,
      caption ="Summary of the final least square regression model")
```

We can see that Pao2 significantly impacts Paco2 with a p-value = 0.0025. To be more specific, as Pao2 increases by a unit, the Paco2 **decreases** by about 0.27. The negative sign of the estimated slope indicates the negative linear association between Paco2 and Pao2.

Now, let's assume that there are two new patients who Pao2 levels 63 and 75, respectively. Note that these two Pao2 are **within the range of Pao2**.

Table 10.5: Summary of the final least square regression model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	67.9716010	6.3535426	10.698221	0.0000000
Pao2	-0.2687739	0.0811017	-3.314037	0.0025473

```
library(pander)
## put the new observations in the form of the data frame.
new.pao2 = data.frame(Pao2 = c(63,75))
##
pred.new = predict(ls.reg.final, newdata = new.pao2,
                   interval = "prediction",
                   level = 0.05)
pred.new.cbind = cbind(Pao2.new=c(63,75), pred.new)
pander(pred.new.cbind, caption = "95% predictive intervals of Paco2")
```

Table 10.6: 95% predictive intervals of Paco2

Pao2.new	fit	lwr	upr
63	51.04	50.55	51.53
75	47.81	47.33	48.3

The above predictive table indicates that the predicted value of Paco2 with Pao2 = 63 is about 51.04 with a 95% predictive interval [50.55, 51.23]. The predicted Paco2 is 47.81 with a 95% predictive interval [47.23, 48.30] for Pao2 = 73.

Chapter 11

Multiple Linear Regression

We discussed the relationship between variables in the previous two modules. The continuous variable with a normal distribution is called the response (dependent) variable and the other variable is called the explanatory (predictor, independent, or risk) variable. If the predictor variable is a factor variable, the model is called the ANOVA model which focuses on comparing the means across all factor levels. If the predictor variable is **continuous**, the model is called simple linear regression (SLR). Note that all predictor variables are assumed to be non-random.

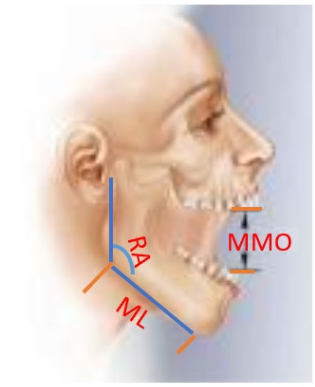
11.1 The Practical Question

Maximum mouth opening (MMO) is also an important diagnostic reference for dental clinicians as a preliminary evaluation. Establishing a normal range for MMO could allow dental clinicians to objectively evaluate the treatment effects and set therapeutic goals for patients performing mandibular functional exercises.

To study the relationship between maximum mouth opening and measurements of the lower jaw (mandible). A researcher randomly selected a sample of 35 subjects and measured the dependent variable, maximum mouth opening (MMO, measured in mm), as well as predictor variables, mandibular length (ML, measured in mm), and angle of rotation of the mandible (RA, measured in degrees) of each of the 35 subjects.

The question is whether the maximum mouth opening (MMO) is determined by **two variables simultaneously**. We want to assess how these two variables (ML and RA) impact MMO **simultaneously**.

If we pick one predictor variable at a time, ML, to build a simple linear regression model and ignore the other predictor variable (RA), you only get the



MMO = maximum mouth opening

ML = mandible length

RA = angle of rotation of the mandible

Figure 11.1: MMO, ML, and RA

MMO (Y)	ML (X_1)	RA (X_2)	MMO (Y)	ML (X_1)	RA (X_2)
52.34	100.85	32.08	50.82	90.65	38.33
51.90	93.08	39.21	40.48	92.99	25.93
52.80	98.43	33.74	59.68	108.97	36.78
50.29	102.95	34.19	54.35	91.85	42.02
57.79	108.24	35.13	47.00	104.30	27.20
49.41	98.34	30.92	47.23	93.16	31.37
53.28	95.57	37.71	41.19	94.18	27.87
59.71	98.85	44.71	42.76	89.56	28.69
53.32	98.32	33.17	51.88	105.85	31.04
48.53	92.70	31.74	42.77	89.29	32.78
51.59	88.89	37.07	52.34	92.58	37.82
58.52	104.06	38.71	50.45	98.64	33.36
62.93	98.18	43.89	43.18	83.70	31.93
57.62	91.01	41.06	41.99	88.46	28.32
65.64	96.98	41.92	39.45	94.93	24.82
52.85	97.85	35.25	38.91	96.81	23.88
64.43	96.89	45.11	49.10	93.13	36.17
57.25	98.35	39.44			

Figure 11.2: Dental Data for the multiple linear regression model (MLR)

marginal relationship between MMO and ML since you implicitly assume that the relationship between MMO and ML will not be impacted by RA. This implicit assumption is, in general, incorrect. We need to consider all predictor variables at the same time. This is the motivation for studying multiple linear regression (MLR).

11.2 The Process of Building A Multiple Linear Regression Model

The previous motivation example involves two continuous predictor variables. In real-world applications, it is common to have many predictor variables. Predictor variables are also assumed to be non-random. They could be categorical, continuous, or discrete. In a specific application, you may have a set of categorical, continuous, and discrete predictor variables in one data set.

11.2.1 Assumptions of MLR

There are several assumptions of multiple linear regression models.

- The response variable is a normal random variable and its mean is influenced by explanatory variables but not the variance.
- The explanatory variables are assumed to be non-random.
- The explanatory variables are assumed to be uncorrelated to each other.
- The functional form of the explanatory variables in the regression model is correctly specified.
- The data is a random sample taken independently from the study population with a specified distribution.

Some of these assumptions will be used directly to define model diagnostic measures. The idea is to assume all conditions are met (at least temporarily) and then fit the model to the data set.

11.2.2 The Structure of MLR

Assume that there are p predictor variables $\{x_1, x_2, \cdots, x_p\}$, the first-order linear regression is defined in the following form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

β_0 is the intercept, $\beta_1, \beta_2, \cdots, \beta_p$ are called slope parameters. if $\beta_i = 0$, the associated predictor variable x_i is uncorrelated with response vararible y. If $\beta_i > 0$, then y and x_i are positively correlated. In fact, β_1 is the increment of y as x_i increases one unit and other predictors remain unchanged.

The response variable is assumed to be a normal random variable with constant variance. If the first-order linear regression function is correct, then

$$y \to N(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p, \sigma^2).$$

This also implies that $\epsilon \to N(0, 1)$. The residual of each data point can be estimated from the data with an assumed linear regression model.

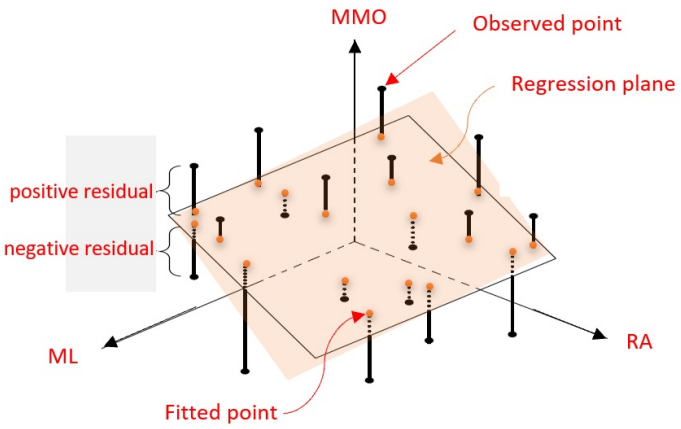


Figure 11.3: Illustrative regression plane: MMO vs ML and RA

For ease of illustration, let's consider the case of the MLR with two predictor variables in the motivation example.

$$MMO = \beta_0 + \beta_1 ML + \beta_2 RA + \epsilon$$

is the first-order linear regression model. The following figure gives the graphical annotations of the fundamental concepts in linear regression models. This is a generalization of the regression line (see the analogous figure in the previous module for the simple linear regression model).

Since MMO is a normal random variable with constant variance, $MMO \to N(\beta_0 + \beta_1 ML + \beta_2 RA, \sigma^2)$, or equivalently, $\epsilon \to N(0, \sigma^2)$. The residuals are defined to be the directional vertical distances between the observed points and the regression plane.

In some practical applications, we may need **the second-order** linear regression model to reflect the actual relationship between predictor variables and the response variable. For example,

$$MMO = \alpha_0 + \alpha_1 ML + \alpha_2 RA + \alpha_3 ML^2 + \alpha_4 RA^2 + \alpha_5 ML \times RA + \epsilon$$

is called (the second-order) linear regression model. With the second-order terms in the regression function, we obtain the regression surface as shown in Figure.

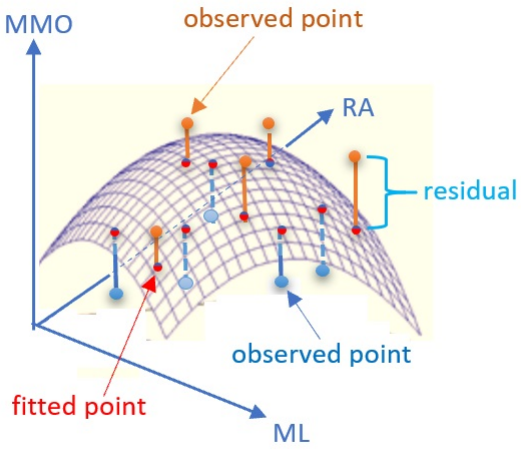


Figure 11.4: Illustrative regression surface: MMO vs ML and RA

If the second-order linear regression is appropriate, then $\epsilon \rightarrow N(0, \sigma^2)$ and $E[MMO] = \alpha_0 + \alpha_1 ML + \alpha_2 RA + \alpha_3 ML^2 + \alpha_4 RA^2 + \alpha_5 ML \times RA$. The residuals of the second-order linear regression model are defined to be the directional distance between the observed points and the regression surface.

11.2.3 More on Model Specifications

In the above section, we introduced both first- and second-order polynomial regression models. In general, it is not common to use high-order polynomial regression models in real-world applications.

- **Interaction effect** - It is common to include interaction terms (i.e., the cross product of two or more predictor variables) in the multiple linear regression models when the effect of one variable on the response variable is dependent on the other predictor variable. In other words, the interaction terms capture the **joint effect** of predictor variables. **It is rare to have third-order or higher-order interaction terms in a regression model**.
- **Dummy variables** - All categorical predictor variables are automatically converted into dummy variables (binary indicator variables). If categorical variables in the data are numerically coded, we have to turn these numerically coded variables into factor variables in the regression model.

- **Discretization and Regrouping** - Discretizing numerical predictor variables and regrouping categorical or discrete predictor variables are two basic pre-process procedures that are actually very common in many practical applications.
  - Sometimes these two procedures are required to satisfy certain model assumptions. For example, if a categorical variable has a few categories that have less than 5 observations, the resulting p-values based on certain hypothesis tests will be invalid. In this case, We have to regroup some of the categories in **meaningful ways** to resolve the **sparsity** issues in order to obtain valid results.
  - In many other applications, we want the model to be easy to interpret. Discretizing numerical variables is common. For example, we can see grouped ages and salary ranges in different applications.

11.2.4 Estimation of Regression Coefficients

A simple and straightforward method for estimating the coefficients of linear regression models is to minimize the sum of the squared residuals - least square estimation (LSE). To find the LSE of the regression coefficients, we need to

- choose the (first-order, second-order, or even high-order) regression function (see 3D hyper-plane or hyper-surface in the above two figures as examples).
- find the distances between the observed points and the hyper-plane (or hyper-surface). These distances are the residuals of the regression - which is dependent on the regression coefficients.
- calculate the sum of squared residuals. This sum of the residuals is still dependent on the regression coefficients.
- find the values for the regression coefficients that minimize the sum of the squared residuals. These values are called the least square estimates (LSEs) of the corresponding regression coefficients.

R function **lm()** implements the above the LSE algorithm to find the regression coefficients. We have used this function in ANOVA and simple linear regression models.

11.2.5 Model Diagnostics

Unlike simple linear regression models, the primary assumptions of the regression model focus on the normal distribution of the response variable and the correct regression function. For multiple linear regression models, we need to impose a couple of assumptions in addition to those in the simple linear regression models

- **Residual Diagnostics**

One of the fundamental assumptions of linear regression modeling is that the response variable is normally distributed with a constant variance. This implies $\epsilon \to N(0, \sigma^2)$.

After obtaining LSE of the regression coefficients, we can estimate the residuals and use these estimated residuals to detect the potential violations of the normality assumption of the response variable. To be more specific, we consider the first-order polynomial regression, the estimated residual of i-th observation is defined to be $e_i = MMO - \hat{\beta}_0 + \hat{\beta}_1 ML + \hat{\beta}_2 RA$

If there is no violation of the normality assumption, we would expect the following residual plot and Q-Q plot.

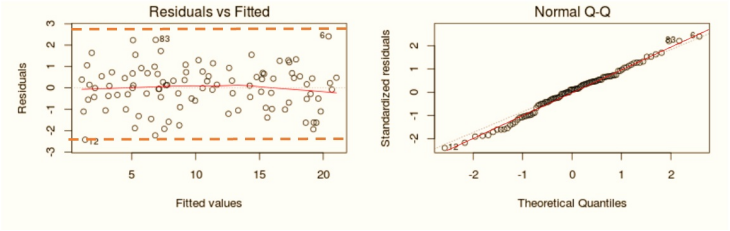


Figure 11.5: Good residual plot and normal Q-Q plot

Some of the commonly seen poor residual plots represent different violations of various assumptions. We can try to use various transformations (such as Box-Cox power transformation) of the response variable to correct the issue.

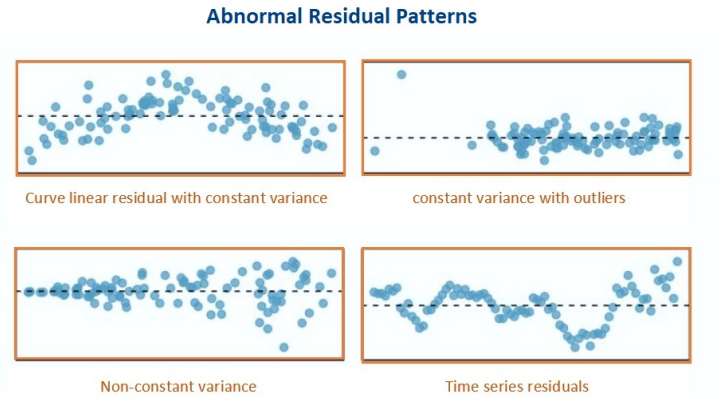


Figure 11.6: Poor residual plots representing various violations of the model assumptions

- **Multicollinearity**

Some of the predictor variables are linearly correlated. The consequence of multi-collinearity causes to unstable LSE of the regression coefficients (i.e., the LSEs of the regression coefficients are sensitive to a small change in the model). It also reduces the precision of the estimate coefficients and, hence, the p-values are not reliable.

Multicollinearity affects the coefficients and p-values, but it does not influence the predictions, precision of the predictions, and the goodness-of-fit statistics. If our primary goal is to make predictions, we don't need to understand the role of each independent variable and we don't need to reduce severe multicollinearity.

If the primary goal is to perform association analysis, we need to reduce collinearity since both LSE and p-values are the keys to association analysis.

To detect multicollinearity, we can use the variance inflation factor (VIF) to inspect the multicollinearity of the individual predictor variable. There are some different methods to reduce multicollinearity. Centering predictor variables is one of them and works well sometimes. Some other advanced modeling-based methods are covered in more advanced courses.

11.2.6 Goodness-of-fit and Variable Selection

Several different goodness-of-fit measures are available for the linear regression model due to the assumption of the normality assumption of the response variable.

- **Coefficient of Determination**

We only introduce **the coefficient of determination** R^2 which measures the percentage of variability within the -values that can be explained by the regression model. In simple linear regression models, **the coefficient of determination** R^2 is simply the square of the sample Pearson correlation coefficient.

- **Statistical Significance and Practical Importance**

A small p-value of the significant test for a predictor variable indicates the variable is statistically significant but may not be practically important. On the other hand, some practically important predictor variables may not achieve statistical significance due to the limited sample size. In the practical applications, **we may want to include some of the practically important predictor variables in the final model regardless of their statistical significance**.

- **Model Selection**

One of the criteria for assessing the goodness-of-fit is the parsimony of the model. A parsimonious model is a model that accomplishes the desired level of explanation or prediction with as few predictor variables as possible. There are generally two ways of evaluating a model: Based on predictions and based on goodness of fit on the current data such as R^2 and some likelihood-based measures.

R has an automatic variable selection procedure, **step()**, which uses the goodness-of-fit measure AIC (Akaike Information Criterion) which is not formally introduced in this class due to the level of mathematics needed in the definition, but we can still use it to perform the automatic variable selection. This tutorial gives detailed examples on how to use **step()** (link).

11.3 Case Study 1

We use the dental data in the motivation example for the case study.

```
MMO=c(52.34, 51.90, 52.80, 50.29, 57.79, 49.41, 53.28, 59.71, 53.32, 48.53,
      51.59, 58.52, 62.93, 57.62, 65.64, 52.85, 64.43, 57.25, 50.82, 40.48,
      59.68, 54.35, 47.00, 47.23, 41.19, 42.76, 51.88, 42.77, 52.34, 50.45,
      43.18, 41.99, 39.45, 38.91, 49.10)
##
ML=c(100.85, 93.08, 98.43, 102.95, 108.24, 98.34, 95.57, 98.85,98.32, 92.70,
     88.89, 104.06, 98.18, 91.01, 96.98, 97.85, 96.89, 98.35, 90.65, 92.99,
     108.97, 91.85, 104.30, 93.16, 94.18, 89.56, 105.85, 89.29, 92.58, 98.64,
     83.70, 88.46, 94.93, 96.81, 93.13)
##
RA = c(32.08, 39.21, 33.74, 34.19, 35.13, 30.92, 37.71, 44.71, 33.17, 31.74,
       37.07, 38.71, 43.89, 41.06, 41.92, 35.25, 45.11, 39.44, 38.33, 25.93,
       36.78, 42.02, 27.20, 31.37, 27.87, 28.69, 31.04, 32.78, 37.82, 33.36,
       31.93, 28.32, 24.82, 23.88, 36.17)
DentalData = as.data.frame(cbind(MMO = MMO, ML = ML, RA = RA))
```

- **Pair-wise Scatter Plot**

This pairwise scatter plot tells whether there are significant correlations between **numerical predictor variables**.

We can see the following patterns from the above pair-wise scatter plot.

(1). Both ML and RA are linearly correlated with the response variable MMO. This is what we expected.

(2). ML and RA are not linearly correlated. This indicates that there is no collinearity issue.

(3). We also don't see any special patterns such as outliers and extremely skewed distribution. There is no need to perform discretization and regrouping procedures on the predictor variables.

(4). In this data set, there is no categorical variables or categorical variable with a numerical coding system in this data set. There is no need to create dummy variables.

- **Initial model**

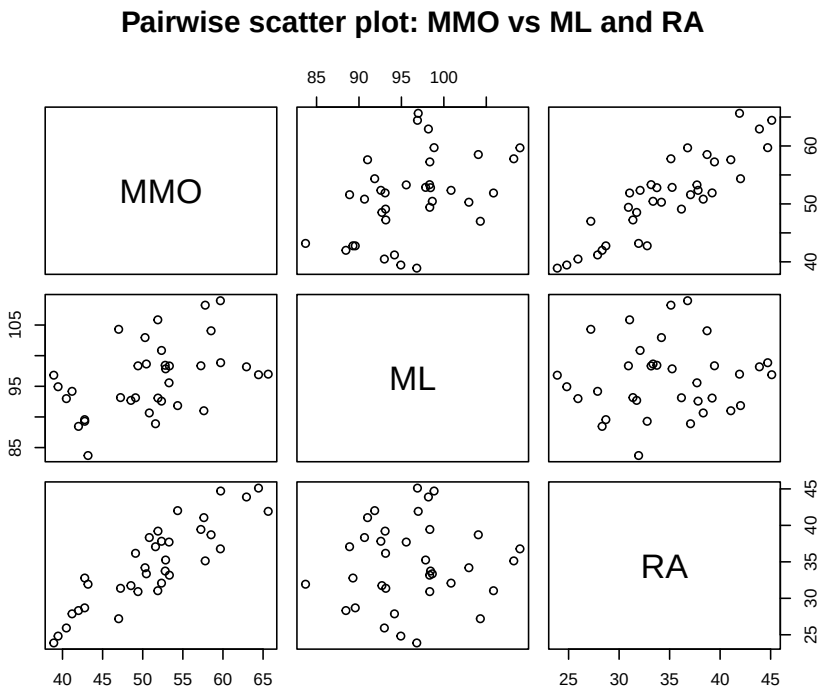


Figure 11.7: Pair-wise scatter plot

The following initial model includes all predictor variables and then performs the residual diagnostics immediately afterward.

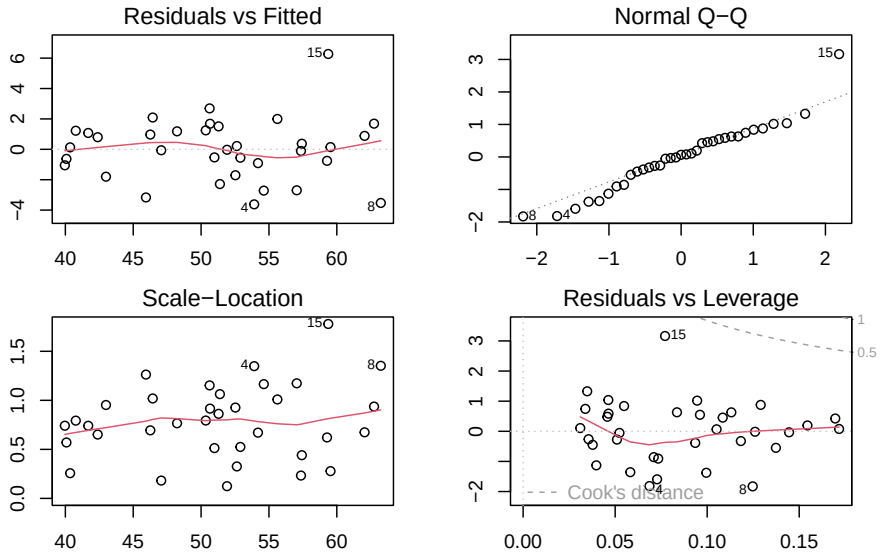
The residual plots indicate that

(1). One of the observations seems to be an outlier (observation 15);

(2). There is a minor violation of the assumption of constant variance.

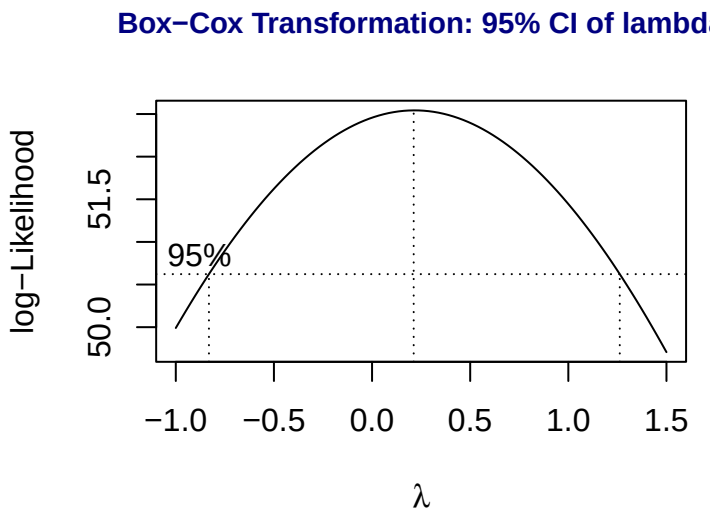
(3). There is also a minor violation of the assumption of normality of the distribution of the residuals.

```
ini.model = lm(MMO ~ ML + RA, data = DentalData)    #  interaction effect
par(mfrow=c(2,2), mar=c(2,3,2,2))
plot(ini.model)
```



Next, we will carry the Box-Cox transformation to identify a potential power transformation of the response variable MMO.

```
library(MASS)
boxcox(MMO ~ ML + RA,
       data = DentalData,
       lambda = seq(-1, 1.5, length = 10),
       xlab=expression(paste(lambda)))
title(main = "Box-Cox Transformation: 95% CI of lambda",
      col.main = "navy", cex.main = 0.9)
```

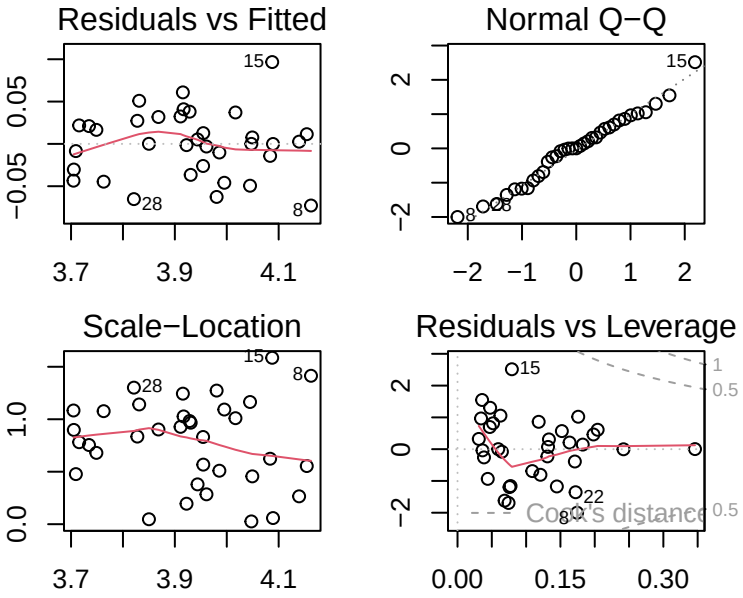


Since both 0 and 1 are in the 95% confidence interval of λ, technically speaking, there is no need to perform the power transformation. By the optimal λ is closer to 0, we try to perform the log transformation (corresponding to $\lambda = 0$) to see whether there will be some improvement of the initial model

```
transform.model = lm(log(MMO) ~ ML * RA, data  = DentalData)
par(mfrow=c(2,2), mar = c(2,2,2,2))
plot(transform.model)
```

Table 11.1: Summarized statistics of the regression coefficients of the model with log-transformed response

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.9957108	1.0023242	1.9910831	0.0553479
ML	0.0124400	0.0104503	1.1903999	0.2429256
RA	0.0296960	0.0293716	1.0110477	0.3198204
ML:RA	-0.0000884	0.0003059	-0.2890324	0.7744807



The above residual plots indicate an improvement in model fit. We will use the transformed response to build the final model.

- **Final Model**

The model based on the log-transformed response is summarized in the following.

```
kable(summary(transform.model)$coef,
caption = "Summarized statistics of the regression
      coefficients of the model with log-transformed response")
```

we can see that the interaction effect is insignificant in the model. We drop the highest term in the regression model either manually or automatically. In the next code chunk, we use the automatic variable selection method to find the final model.

Table 11.2: Summarized statistics of the regression coefficients of the final model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.2833535	0.1176375	19.410076	0
ML	0.0094391	0.0011705	8.064482	0
RA	0.0212140	0.0012017	17.653748	0

Table 11.3: Coefficients of correlation of the three candidate models

ini.model	transfd.model	final.model
0.9204481	0.9257218	0.9255216

```
transform.model = lm(log(MMO)~ML*RA, data = DentalData)
## I will use automatic variable selection function to search the final model
final.model =  step(transform.model, direction = "backward", trace = 0)
kable(summary(final.model)$coef,
caption = "Summarized statistics of the regression
      coefficients of the final model")
```

Now we have three candidate models to select from. We extract the coefficient of determination (R^2) of each of the three candidate models.

```
r.ini.model = summary(ini.model)$r.squared
r.transfd.model = summary(transform.model)$r.squared
r.final.model = summary(final.model)$r.squared
##
Rsquare = cbind(ini.model = r.ini.model, transfd.model = r.transfd.model,
                final.model = r.final.model)
kable(Rsquare, caption="Coefficients of correlation of the three candidate models")
```

The second and the third models have almost the same R^2, 92.56% and 92.57%. Both models are based on the log-transformed MMO. The interpretations of these two models are not straightforward. The initial model has a slightly lower 92.0%. Since the initial model has a simple structure and is easy to interpret, we chose the initial model as the final model to report. The summarized statistic is given in the following table.

```
summary.ini.model = summary(ini.model)$coef
kable(summary.ini.model, caption = "Summary of the final working model")
```

In summary, both ML and RA are statistically significant (p-value $\approx$ 0) and both are positively correlated to MMO. Further, for a given angle of rotation of the mandible (RA), when mandibular length (ML) increases by 1mm, the maximum mouth opening (MMO) increases by 0.473 mm. However, for holding ML, a 1-degree increase in RA will result in a 1.071 mm increase in MMO.

Table 11.4: Summary of the final working model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-31.4247984	6.1474668	-5.111829	0.0000144
ML	0.4731743	0.0611653	7.735992	0.0000000
RA	1.0711725	0.0627967	17.057792	0.0000000

11.4 Case Study 2

We discussed the ANOVA model in module 8. In fact, the ANOVA model is a special linear regression model. The location is a factor variable. We now build a linear regression using mussel shell length as the response and the location as the predictor variable in the following (code is copied from module 8).

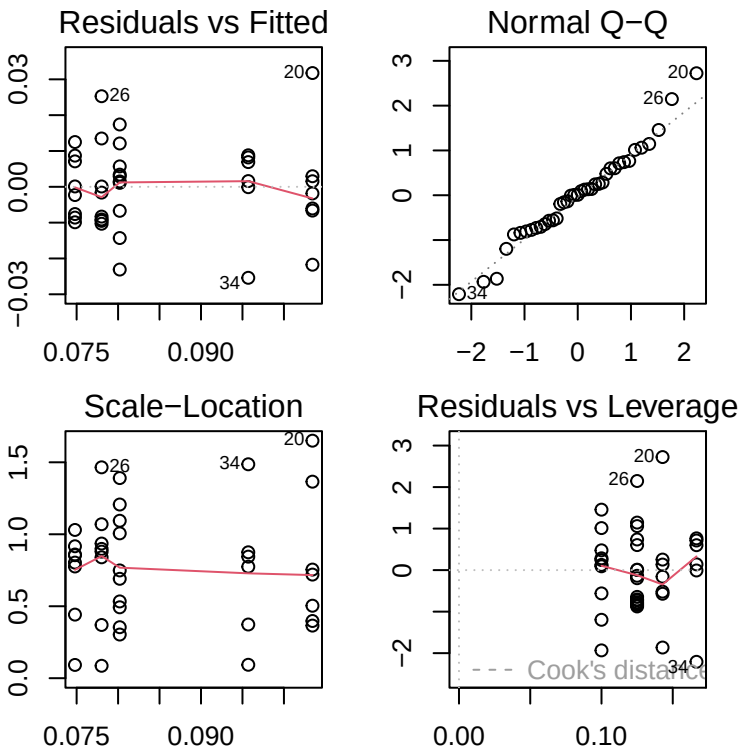
Since predictor variable location is a categorical factor variable, R function **lm()** will automatically define four dummy variables for each category except for the baseline category is, by default, the smallest character values (alphabetical order). In our example, the value **Magadan** is the smallest. Other categories will be compared with the baseline category through the corresponding dummy variable.

To be more specific, the four dummy variables associated with the four categories will be defined by

1. locationNewport = 1 if the location is Newport, 0 otherwise;
2. locationPetersburg = 1 if the location is Petersburg, 0 otherwise;
3. locationTillamook = 1 if the location is Tillamook, 0 otherwise;
4. locationTvarminne = 1 if the location is Tvarminne, 0 otherwise.

```
x1 = c(0.0571,0.0813, 0.0831, 0.0976, 0.0817, 0.0859, 0.0735, 0.0659,
       0.0923, 0.0836)
x2 = c(0.0873,0.0662, 0.0672, 0.0819, 0.0749, 0.0649, 0.0835, 0.0725)
x3 = c(0.0974,0.1352, 0.0817, 0.1016, 0.0968, 0.1064, 0.1050)
x4 = c(0.1033,0.0915, 0.0781, 0.0685, 0.0677, 0.0697, 0.0764, 0.0689)
x5 = c(0.0703,0.1026, 0.0956, 0.0973, 0.1039, 0.1045)
len  = c(x1, x2, x3, x4, x5)      # pool all sub-samples of lengths
location = c(rep("Tillamook", length(x1)),
             rep("Newport", length(x2)),
             rep("Petersburg", length(x3)),
             rep("Magadan", length(x4)),
             rep("Tvarminne", length(x5))) # location vector matches the lengths
data.matrix = cbind(len =  len, location = location)   # data a data table
musseldata = as.data.frame(data.matrix)        # data frame
## end of data set creation
##
```

```
## starting building ANOVA model
anova.model.01 = lm(len ~ location, data = musseldata)  # define a model for generatin
##
par(mfrow=c(2,2), mar = c(2,2,2,2))
plot(anova.model.01)
```



The above residual plots indicate no serious violation of the model assumption. The model that generates the above residual plot will be used as the final working model. The inference of the regression coefficients is summarized in the following table.

```
sum.stats = summary(anova.model.01)$coef
kable(sum.stats, caption = "Summary of the ANOVA model")
```

From the above summary tale, we can see that P-values associated with location dummy variables locationNewport, locationTillamook are bigger than 0.05 meaning the means associated with **Newport**, **Tillamook**, and the baseline **Magadan** (not appearing in the summary table). The p-values associated with **Petersburg** and **Tvarminn** are less the 0.05 which implies that the mean length of these two locations is significantly different from that of the baseline location **Magadan**. Further, the coefficient associated with dummy

Table 11.5: Summary of the ANOVA model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0780125	0.0044536	17.5168782	0.0000000
locationNewport	-0.0032125	0.0062983	-0.5100593	0.6133053
locationPetersburg	0.0254304	0.0065193	3.9007522	0.0004300
locationTillamook	0.0021875	0.0059751	0.3661039	0.7165558
locationTvarminne	0.0176875	0.0068029	2.5999834	0.0136962

variable **locationPetersburg** indicates that the mean length of mussel shell in **Petersburg** is 0.0543 units longer than that in the baseline location **Magadan**. We can also interpret the coefficients associated with **locationTvarminne**.

11.5 Practice Problems

Family caregiving of older adults is more common in Korea than in the United States. A research team studied 100 caregivers of older adults with dementia in Seoul, South Korea. The dependent variable was caregiver burden as measured by the Korean Burden Inventory (KBI). Scores ranged from 28 to 140, with higher scores indicating a higher burden. Explanatory variables were indexes that measured the following:

ADL: total activities of daily living (low scores indicate that the elderly perform activities independently).

MEM: memory and behavioral problems (higher scores indicate more problems).

COG: cognitive impairment (lower scores indicate a greater degree of cognitive impairment).

The reported data are given in the following code chunk.

```
Y = c(28, 68, 59, 91, 70, 38, 46, 57, 89, 48, 74, 78, 43, 76, 72, 61, 63, 77,
      85, 31, 79, 92, 76, 91, 78, 103, 99, 73, 88, 64, 52, 71, 41, 85, 52, 68,
      57, 84, 91, 83, 73, 57, 69, 81, 71, 91, 48, 94, 57, 49, 88, 54, 73,
      87, 47, 60, 65, 57, 85, 28, 40, 87, 80, 49, 57, 32, 52, 42, 49, 63, 89,
      67, 43, 47, 70, 99, 53, 78, 112, 52, 68, 63, 49, 42,
      56, 46, 72, 95, 57, 88, 81, 104, 88, 115, 66, 92, 97, 69, 112, 88)

X1 =c(39, 52, 89, 57, 28, 34, 42, 52, 88, 90, 38, 83, 30, 45, 47, 90, 63, 34,
      76, 26,68,  85, 22, 82, 80, 80, 81, 30, 27, 72, 46, 63, 45, 77, 42, 60,
      33, 49, 89, 72, 45, 73, 58, 33, 34, 90, 48, 47, 32, 63, 76, 79, 48, 90,
```

```
      55, 83, 50, 44, 79, 24, 40, 35, 55, 45, 46, 37, 47, 28, 61, 35, 68, 80,
      43, 53, 60, 63, 28, 35, 37, 82, 88, 52, 30, 69, 52, 59, 53, 65, 90, 88,
      66, 60, 48, 82, 88, 63, 79, 71, 66, 81)

X2 =c(4, 33, 17, 31, 35, 3, 16, 6, 41, 24, 22, 41, 9, 33, 36, 17, 14, 35, 33,
      13, 34, 28,12, 57, 51, 20, 20, 7, 27, 9, 15, 52, 26, 57, 10, 34, 14, 30,
      64, 31, 24, 13, 16, 17, 13, 42, 7, 17, 13, 32, 50, 44, 57, 33, 11, 24,
      21, 31, 30, 5, 20, 15, 9, 28, 19, 4, 29, 23, 8, 31, 65, 29, 8, 14, 30,
      22, 9, 18, 33, 25, 16, 15, 16, 49, 17, 38, 22, 56, 12, 42, 12, 21, 14,
      41, 24, 49, 34, 38, 48, 66)

X3 =c(18, 9, 3, 7, 19, 25, 17, 26, 13, 3, 13, 11, 24, 14, 18, 0, 16, 22, 23,
      18, 26, 10, 16, 3, 3, 18, 1, 17, 27, 0, 22, 13, 18, 0, 19, 11, 14, 15,
      0, 3, 19, 3, 15, 21, 18, 6, 23, 18, 15, 15, 5, 11, 9, 6, 20, 11, 25, 18,
      20, 22, 17, 27, 21, 17, 17, 21, 3, 21, 7, 26, 6, 10, 13, 18, 16, 18, 27,
      14, 17, 13, 0, 0, 18, 12, 20, 17, 21, 2, 0, 6, 23, 7, 13, 13, 14, 5, 3,
      17, 13, 1)
```

In this assignment, you are expected to replicate the analysis in case study #1 in the weekly note. To be more specific, you can proceed with the analysis with the following steps.

1. Perform exploratory data analysis: pair-wise scatter plot. Describe the patterns you observed from the pair-wise plot.
2. Build an initial model that includes all variables and perform the residual analysis. Inspect the residual plot and describe potential abnormal patterns you observed from the residual plots.
3. Explore potential power transformation using Box-cox transformation. Use trial and error to identify an appropriate range so you can view the 95% confidence interval for λ. Please keep in mind that $\lambda = 0$ implies log transformation.
4. Use the automatic variable selection to identify the **best** model.
5. Use the coefficient of determination to select the final model from the candidate models (initial, the **best** model from step 3.)
6. Interpret the final working model.

Chapter 12

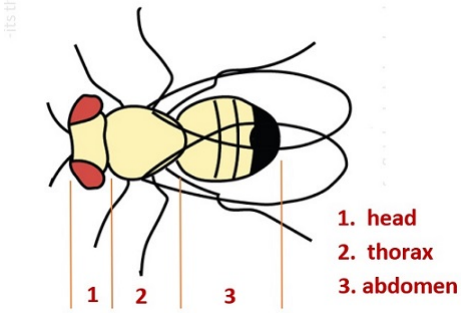
Logistic Regression Models

Linear regression models are used to assess the association between the continuous response variable and other predictor variables. If the response variable is a binary categorical variable, the linear regression model is not appropriate. We need a new model, the logistic regression model, to assess the association between the binary response variable and other predictor variables.

This module focuses on the regression model with a binary response.

12.1 Motivational Example and Practical Question

Example : Longevity in male fruit flies is positively associated with adult size. However, reproduction has a high physiological cost that could impact longevity.



The original study looks at the association between longevity and adult size in male fruit flies kept under one of two conditions. One group is kept with sexually active females over the male's life span. The other group is cared for in the same way but kept with females who are not sexually active.

Longevity	ThxLength	IndReprod	Longevity	ThxLength	IndReprod
34	0.78	0	46	0.84	0
42	0.76	0	56	0.76	1
30	0.8	0	76	0.92	1
46	0.88	0	65	0.8	1
40	0.82	0	42	0.76	0
49	0.68	1	19	0.64	0
56	0.8	1	19	0.68	0
70	0.88	1	70	0.88	1
64	0.76	1	26	0.8	0
54	0.88	0	64	0.72	1
85	0.84	1	76	0.92	1
76	0.84	1	33	0.72	0
54	0.88	0	81	0.84	1
61	0.88	0	34	0.72	0
56	0.88	0	54	0.82	0
46	0.76	1	65	0.84	1
44	0.92	0	37	0.68	1
76	0.94	1	39	0.76	1
70	0.84	1	35	0.84	0
64	0.84	1	35	0.64	1
70	0.8	1	30	0.76	0
35	0.84	0	46	0.84	0
65	0.76	1	16	0.64	0
34	0.74	0	46	0.72	1

Figure 12.1: Fruit Flies Data Table

The above table gives the longevity in days for the male fruit flies allowed to reproduce (IndReprod = 0) and for those deprived of the opportunity (IndReprod = 1).

The data was collected from a case-control study design. The original study association analysis using the multiple linear regression model in which Longevity was a response variable and Thorax and IndReprod were used as predictor variables. In this example, we build a logistic regression to assess the association between longevity and reduction. Due to the case-control study design, the resulting logistic regression cannot be used as a predictive model.

Since the response variable is binary (i.e., it can only take on two distinct values, 0 and 1 in this example), the linear regression line is a bad choice since (1) the response variable is not continuous, (2) the fitted regression line can take on any values between negative infinity and positive infinity. The response variable takes on only either 0 or 1 (see the dark red straight line).

A meaningful approach to assess the association between the binary response variable and the predictor variable by looking at how the predictor variables impact the probability of observing the **bigger** value (i.e., the one that has a higher alphabetical order) of the response variable. The above **S** curve describes the relationship between the values of the predictor variable(s) and the probability of observing the **bigger** value of the response variable.

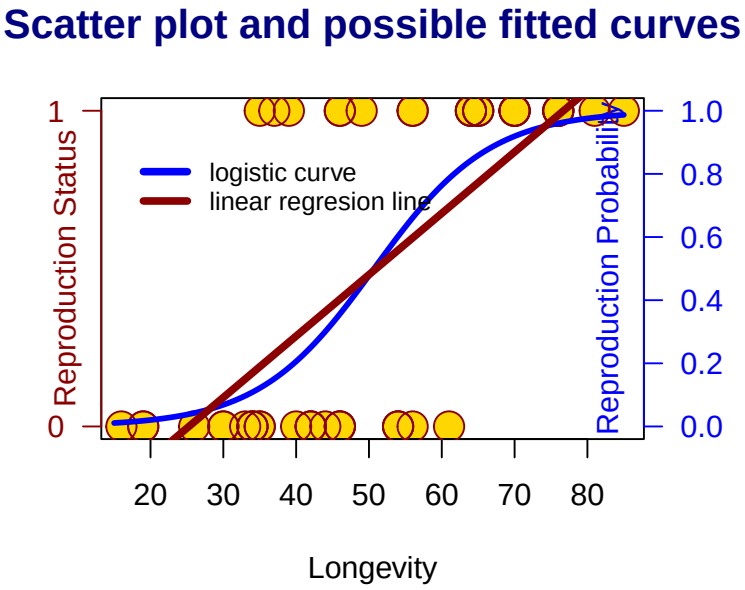


Figure 12.2: The scatter plots of a binary response v.s. a continuous predictor variable

12.2 Logistic Regression Models and Applications

Logistic regression is a member of the generalized linear regression (GLM) family which includes linear regression models. The modeling-building process is the same as that in linear regression modeling.

12.2.1 The Structure of Logistic Regression Model

The actual probability function of observing the **bigger** value of the response variable for giving the predictor variable is given by

$$P(Y = 1) = \frac{\exp(\beta_0 + \beta_1 \text{Longevity})}{1 + \exp(\beta_0 + \beta_1 \text{Longevity})}$$

If β_1 (also called slope parameter) is equal to zero, the longevity and the reproduction status are NOT associated. Otherwise, there is an association between the response and the predictor variables. The sign of the slope parameter reflects the positive or negative association.

12.2.2 Assumptions and Diagnostics

There are assumptions for the binary logistic regression model.

- The response variable must be binary (taking on only two distinct values).
- Predictor variables are assumed to be uncorrelated.
- The functional form of the predictor variables is correctly specified.

The model diagnostics for logistic regression are much more complicated than in the linear regression models. We will not discuss this topic in this course.

12.2.3 Coefficient Estimation and Interpretation

The estimation of the logistic regression coefficients is not as intuitive as we saw in the linear regression model (regression lines and regression plane or surface). An advanced mathematical method needs to be used to estimate the regression coefficient. The R function **glm()** can be used to find the estimate of regression coefficients.

The interpretation of the coefficients of the logistic regression model is also not straightforward. In the motivational example, the value of β_1 is the change of log odds of observing reproduction status to be 1. As usual, we will not make an inference of the intercept. In case-control data, the intercept is inestimable.

The output of **glm()** contains information similar to what has been seen in the output of **lm()** in the linear regression model.

12.2.4 Use of glm() and Annotations

We use the motivational example to illustrate the setup of **glm()** and the interpretation of the output.

```
## Fit the logistic regression
mymodel =glm(IndReprod ~ Longevity,    # model formula (response on the left).
          family=binomial,             # must be binomial for the binary response!
          data=fruitflies)             # data frame name
```

```
glm(formula = IndReprod ~ Longevity, family = binomial, data = fruitflies)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7548  -0.6794  -0.1583   0.5525   2.0535

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.39071    1.72165  -3.712 0.000206 ***
Longevity    0.12605    0.03358   3.753 0.000175 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 67.908  on 48  degrees of freedom
Residual deviance: 39.975  on 47  degrees of freedom
AIC: 43.975

Number of Fisher Scoring iterations: 5
```

The only inforamtion we need for this class.

Figure 12.3: The complete output of glm()

The output has four major pieces of information: the model formula, the five-number-summary of the deviance residuals, significant test results of the predictor variables, and goodness-of-fit measures. In this course, we only focus on the significance tests.

12.2.5 Applications of Logistic Regression Models

Like other regression models, the logistic regression models have two basic applications: association analysis and prediction analysis.

The association analysis focuses on the interpretation of the regression coefficients that have information about whether predictor variables are associated with the response variable through the probability of the **bigger** value of the response variable.

Since the logistic regression builds the relationship between the probability of observing the **bigger** value of the response and the predictor variable, predicting the **value of the response variable** requires a cut-off probability to assign a value of 1 or 0 to the response variable. The prediction process of a logistic regression model is depicted in the following figure.

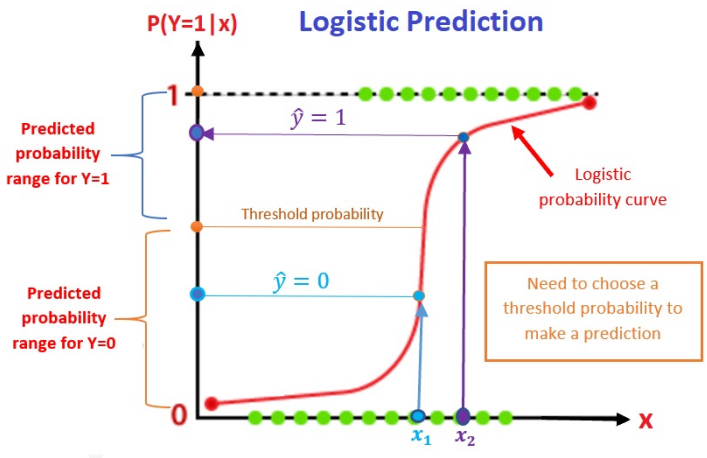


Figure 12.4: Prediction process of the logistic regression models

12.3 Case Studies

We present two examples in this section.

12.3.1 The simple logistic regression model

Suzuki et al. (2006) measured sand grain size on 28 beaches in Japan and observed the presence or absence of the burrowing wolf spider Lycosa ishikariana on each beach. Sand grain size is a measurement variable, and spider presence or absence is a nominal variable. Spider presence or absence is the dependent variable; if there is a relationship between the two variables, it would be sand grain size affecting spiders, not the presence of spiders affecting the sand.

```
grainsize=c(0.245, 0.247, 0.285, 0.299, 0.327, 0.347, 0.356, 0.360, 0.363, 0.364,
            0.398, 0.400, 0.409, 0.421, 0.432, 0.473, 0.509, 0.529, 0.561, 0.569,
            0.594, 0.638, 0.656, 0.816, 0.853, 0.938, 1.036, 1.045)
status=c(0, 0, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1,
         1, 1, 1, 1)
spider = as.data.frame(cbind(grainsize = grainsize, status = status))
```

Fitting a Simple Logistic Regression Model

Since there is only one predictor variable in this study, simply choose the simple linear regression model to this data set.

```
spider.model = glm(status ~ grainsize,
                   family = binomial,
                   data = spider)
significant.tests = summary(spider.model)$coef
```

Grain Size (mm)	Status	Numerical Status	Grain Size (mm)	status	Numerical Status
0.245	absent	0	0.432	absent	0
0.247	absent	0	0.473	present	1
0.285	present	1	0.509	present	1
0.299	present	1	0.529	present	1
0.327	present	1	0.561	absent	0
0.347	present	1	0.569	absent	0
0.356	absent	0	0.594	present	1
0.36	present	1	0.638	present	1
0.363	absent	0	0.656	present	1
0.364	present	1	0.816	present	1
0.398	absent	0	0.853	present	1
0.4	present	1	0.938	present	1
0.409	absent	0	1.036	present	1
0.421	present	1	1.045	present	1

Figure 12.5: Spider Data Table

```
pander(significant.tests, caption = "Summary of the significant tests of
      the logistic regression model")
```

Table 12.1: Summary of the significant tests of the logistic regression model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.648	1.354	-1.217	0.2237
grainsize	5.122	3.006	1.704	0.08844

Association Analysis

The above significant tests indicate that the grain size does not achieve significance (p-value = 0.08844) at level 0.05. Note that the p-value is calculated based on the sample, it is also a random variable. Moreover, the sample size in this study is relatively small. We will claim the association between the two variables. As the grain size increases by one unit, the log odds of observing the wolf spider burrowing increase by 5.121553. In other words, the grain size and the presence of spiders are positively associated.

Prediction Analysis

As an example, we choose two new grain sizes 0.33 and 0.57, and want to predict the presence of the wide spiders on the beaches with the given grain sizes. We used the R function **predict()** in linear regression, we used the same function to predict the logistic regression model.

```
spider.model = glm(status ~ grainsize,
                   family = binomial,
```

```
                    data = spider)
##
mynewdata = data.frame(grainsize=c(0.275, 0.57))
pred.prob = predict(spider.model, newdata = mynewdata,
        type = "response")
## threshold probability
cut.off.prob = 0.5
pred.response = ifelse(pred.prob > cut.off.prob, 1, 0)  # predicts the response
pred.table = cbind(new.grain.size = c(0.275, 0.57),
                    pred.response = pred.response)
pander(pred.table, caption = "Predicted Value of response variable
      with the given cut-off probability")
```

Table 12.2: Predicted Value of response variable with the given cut-off probability

new.grain.size	pred.response
0.275	0
0.57	1

12.3.2 Multiple Logistic Regression Model

In this case study, we used a published study on bird introduction in New Zealand. The objective is to predict the success of avian introduction to New Zealand. Detailed information about the study can be found in the following article. https://stat501.s3.amazonaws.com/w11-Correlates-of-introduction-success-in-exotic.pdf. The data is included in the article. A text format data file was created and can be downloaded or read directly from the following URL: https://stat501.s3.amazonaws.com/w11-birds-data.txt.

The response variable: Status - status of success Predictor variables:

- length - female body length (mm)
- mass = female body mass (g)

- range = geographic range (% area of Australia)
- migr = migration score: 1= sedentary, 2 = sedentary and migratory, 3 = migratory
- insect = the number of months in a year with insects in the diet
- diet = diet score: 1 = herbivorous; 2 = omnivorous, 3 = carnivorous.
- clutch = clutch size
- broods = number of broods per season
- wood = use as woodland scored as frequent(1) or infrequent(2)
- upland = use of the upland as frequent(1) or infrequent(2)
- water = use of water scored as frequent(1) or infrequent(2)
- release = minimum number of release events
- indiv = minimum of the number of individuals introduced.

We next read the data from the given URL directly to R. Since there are some records with missing values. We drop those records with at least one missing value.

There are several categorical variables with numerical codings. Among them, **migr** and **diet** have three categories and the rest of the categorical variables have two categories. In practice, a **categorical variable with more than two categories must be specified as factor variables** so R can define dummy variables to capture the difference across the difference.

We conducted an exploratory analysis on **nigr** and **diet** and found a flat discrepancy across the effect.We simply treat them as a discrete numerical variable using the numerical coding as the values of the variables.

```
birds = "https://raw.githubusercontent.com/pengdsci/STA501/main/Data/w11-birds-data.txt"
NZbirds=read.table(birds, header=TRUE)
birds = na.omit(NZbirds)
```

- **Build an Initial Model**

We first build a logistic regression model that contains all predictor variables in the data set. This model is usually called the full model. Note that the response variable is the success status (1 = success, 0 = failure). Species is a kind of ID, it should not be included in the model.

```
initial.model = glm(status ~ length + mass + range + migr + insect + diet +
                      clutch + broods + wood + upland + water + release + indiv,
                    family = binomial, data = birds)
coefficient.table = summary(initial.model)$coef
pander(coefficient.table, caption = "Significance tests of logistic regression model")
```

Table 12.3: Significance tests of logistic regression model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-6.338	5.717	-1.109	0.2676

	Estimate	Std. Error	z value	Pr(>|z|)
length	-0.002815	0.005317	-0.5294	0.5965
mass	0.002668	0.001674	1.594	0.111
range	-0.1316	0.3502	-0.3758	0.7071
migr	-2.044	1.116	-1.831	0.06704
insect	0.148	0.2124	0.6969	0.4859
diet	2.029	1.883	1.077	0.2814
clutch	0.07938	0.2683	0.2959	0.7673
broods	0.02177	0.9283	0.02345	0.9813
wood	2.49	1.642	1.517	0.1293
upland	-4.713	2.865	-1.645	0.09991
water	0.2349	2.67	0.08799	0.9299
release	-0.01292	0.1932	-0.06685	0.9467
indiv	0.01593	0.008324	1.913	0.05571

The p-values in the above significance test table are all bigger than 0.5. We next search for the best model by dropping some of the insignificant predictor variables. Since there are so many different ways to drop variables, next we use an automatic variable procedure to search the final model.

- **Automatic Variable Selection**

R has an automatic variable selection function **step()** for searching the final model. We will start from the initial model and drop insignificant variables using AIC as an inclusion/exclusion criterion.

In practice, sometimes, there may be some practically important predictor variables. Practitioners want to include these practically important variables in the model regardless of their statistical significance. Therefore, we can fit the smallest model that includes only those practically important variables. The final model should be **between** the smallest model, which we will call a **reduced model**, and the initial model, which we will call a **full model**. For illustration, we assume **insect** and **range** are practically important, we want to include these two variables in the final model regardless of their statistical significance.

In summary, we define two models: the full model and the reduced model. The final best model will be the model between the full and reduced models. The summary table of significant tests is given below.

```
full.model = initial.model  # the *biggest model* that includes all predictor variable
reduced.model = glm(status ~ range + insect , family = binomial, data = birds)
final.model =  step(full.model,
                    scope=list(lower=formula(reduced.model),upper=formula(full.model))
                    data = birds,
                    direction = "backward",
                    trace = 0)   # trace = 0: suppress the detailed selection process
```

```
final.model.coef = summary(final.model)$coef
pander(final.model.coef , caption = "Summary table of significant tests")
```

Table 12.4: Summary table of significant tests

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.438	2.138	-1.608	0.1079
mass	0.001939	0.0007326	2.647	0.008119
range	0.00001412	0.3098	0.00004558	1
migr	-2.024	0.9603	-2.108	0.03506
insect	0.2705	0.1426	1.897	0.05785
wood	1.949	1.317	1.479	0.139
upland	-4.731	2.098	-2.255	0.02415
indiv	0.01381	0.004058	3.404	0.0006648

- **Interpretation - Association Analysis**

The summary table contains the two practically important variables **range** and **insect**. **range** does not achieve statistical significance (p-value ≈ 1) and **insect** is slightly higher than the significance level 0.005. Both variables are seemingly positively associated with the response variable.

The following interpretation of the individual predictor variable assumes that other life-history variables and introduction effort variables.

migr and **upland** are negatively associated with the response variable. The odds of success of introducing migratory birds are lower than the sedentary birds. Similarly, birds using upland infrequently have lower odds of being successfully introduced than those using upland frequently.

insect is significant (p-value =0.058). The **odds of success** increase as the number of months of having insects in diet increases.

mass and **indiv** are positively associated with the response variable. The odds of success increase and the body mass increases Similarly, the odds of success increase as the number of minimum birds of the species increases.

wood does not achieve statistical significance but seems to be positively associated with the response variable.

- **Predictive Analysis**

As an illustration, we use the final model to predict the status of successful introduction based on the new values of the predictor variables associated with two species. See the numerical feature given in the code chunk.

```
mynewdata = data.frame(mass=c(560, 921),
                       range = c(0.75, 1.2),
                       migr = c(2,1),
```

```
                        insect = c(6, 12),
                        wood = c(1,1),
                        upland = c(0,1),
                        indiv = c(123, 571))
pred.cuccess.prob = predict(final.model, newdata = mynewdata, type="response")
##
## threshold probability
cut.off.prob = 0.5
pred.response = ifelse(pred.cuccess.prob > cut.off.prob, 1, 0)# predicts response
## add the new predicted response to Mynewdata
mynewdata$Pred.Response = pred.response
##
pander(mynewdata, caption = "Predicted Value of response variable
      with the given cut-off probability")
```

Table 12.5: Predicted Value of response variable with the given cut-off probability

mass	range	migr	insect	wood	upland	indiv	Pred.Response
560	0.75	2	6	1	0	123	0
921	1.2	1	12	1	1	571	1

The predicted status of the successful introduction of the two species is attached to the two new data records.

12.4 Practice Problems

Framingham Heart Study (FHS), a long-term research project developed to identify risk factors of cardiovascular disease, the findings of which had far-reaching impacts on medicine. Indeed, much common knowledge about heart disease—including the effects of smoking, diet, and exercise—can be traced to the Framingham study. The study's findings further emphasized the need for preventing, detecting, and treating risk factors of cardiovascular disease in their earliest stages

The dataset is a rather small subset of possible FHS datasets, having 4240 observations and 16 variables. The variables are as follows:

- sex : the gender of the observations. The variable is a binary named "male" in the dataset.
- age : Age at the time of medical examination in years.
- education : A categorical variable of the participants' education, with the levels: Some high school (1), high school/GED (2), some college/vocational school (3), college (4) - caution: This is a numerically

coded categorical variable, we need to use the form of factor(education) in the model formulas.

- currentSmoker: Current cigarette smoking at the time of examinations
- cigsPerDay: Number of cigarettes smoked each day
- BPmeds: Use of Anti-hypertensive medication at exam
- prevalentStroke: Prevalent Stroke (0 = free of disease)
- prevalentHyp: Prevalent Hypertensive. The subject was defined as hypertensive if treated
- diabetes: Diabetic according to criteria of the first exam treated
- totChol: Total cholesterol (mg/dL)
- sysBP: Systolic Blood Pressure (mmHg)
- diaBP: Diastolic blood pressure (mmHg)
- BMI: Body Mass Index, weight (kg)/height (m)^2
- heartRate: Heart rate (beats/minute)
- glucose: Blood glucose level (mg/dL)

And finally the response variable:

- TenYearCHD: The 10-year risk of coronary heart disease(CHD).

3658 of these 4240 records are complete cases, and the rest have some missing values. The following code chunk will clean the data and only keep the complete records in the final data set.

```
fhs = "https://raw.githubusercontent.com/pengdsci/STA501/main/Data/w11-FraminghamCHD.csv"
FHS.data = na.omit(read.csv(fhs))
```

The above **FHS.data** has 3658 complete records and 16 variables listed above. The goals of this assignment are both **association analysis** and **predictive analysis**. To be more specific, you are expected to follow the analysis in Section 4.2 in the class note.

- Use the above FHS.data to fit an initial model that includes all predictor variables.
- BMI and glucose are two clinically important variables. Both variables will be included in the final model. That is, you can define the smallest model using the two important clinical variables.
- Using an automatic variable selection procedure to identify the final model working model.
- Interpret the regression model coefficients. You don't have to interpret all coefficients. Pick two coefficients associated with categorical and numerical predictor variables, respectively, to interpret.
- Assume there are two incoming patients with the following demographics and clinical characteristics:

male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke
1	60	1	0	0	0	0
0	37	2	1	5	0	0

prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose
1	0	278	160.5	96	26.4	55	75
0	0	185	100	68	18.38	70	72

Chapter 13

Contingency Table Analysis

We have introduced different modeling to solve the association and prediction analyses. The ANOVA is used when the response is a continuous normal random variable and the predictor variable is categorical. When the response variable is a continuous normal random variable and is also a continuous (non-random) variable, we use the simple linear regression model to address the associated problems. When predictor variables are hybrid (continuous and categorical), we have a multiple linear regression model (numerically coded categorical variables need to be specified during the modeling process). When the response variable is binary, we use logistic regression to study the association between the response and predictor variables.

The following table summarizes different models for different situations.

Response variable	Predictor variable	Type of Models
continuous, normal	single categorical	ANOVA
continuous, normal	single continuous	SLR
continuous, normal	continuous or categorical	MLR
binary, categorical	continuous or categorical	logistic model

- **ANOVA**: Analysis of variance model
- **SLR**: Simple linear regression model
- **MLR**: multiple linear regression model, ANOVA is a special MLR. MLR with both categorical and numerical predictor variables is also an Analysis of covariance (ANCOVA).
- **The logistic model** is also called the **logit model**.

We also discussed the relationship between two numerical variables using the Pearson correlation coefficients. When we use correlation coefficients to measure

the strength of the linear correlation between two numerical variables, we don't specify the response variable.

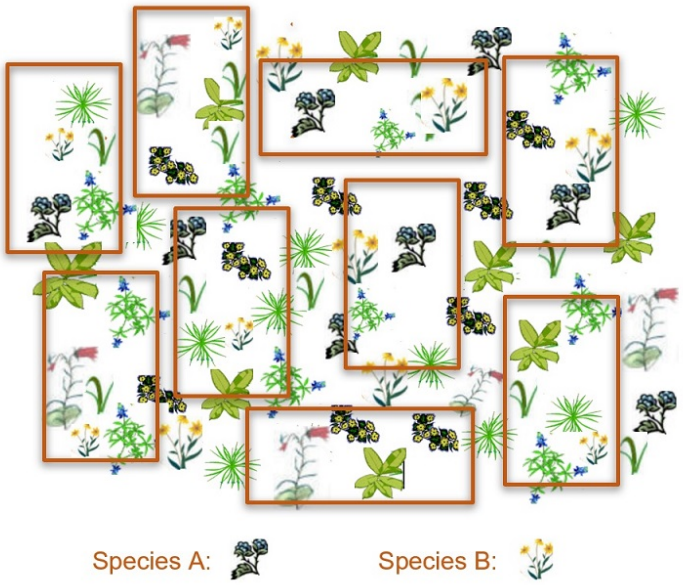
In this note, we will discuss the association between two categorical variables. We first assess the association through χ^2 test of independence of the two variables. If they are dependent, we then define measures for the association.

13.1 The Motivational Examples

Example 1: A biologist might want to determine if two species of organisms associate (are found together) in a vegetation community. Two hypotheses could be set up in the following:

Null Hypothesis (Ho): There is no significant association between Species A and Species B; the species are independent of each other. The location of Species A does not affect the location of Species B.

Alternative Hypothesis (Ha): There is a significant association between Species A and Species B; the species are dependent. Either Species A significantly associates with Species B or Species A does not significantly associate with Species B.



To test the hypotheses, we can use the quadrant random sampling method to collect the raw data. Assume we collect data in 9 randomly placed quadrants and classify the survey outcomes in the following:

- The number of quadrants with both species present
- The number of quadrants with Species A but not Species B

- The number of quadrants with Species B but not Species A
- The number of quadrants with neither species

The above classification is summarized in the following table.

..	Presence of A	Absence of A	**Total**
Presence of B	5	2	**7**
Absence of B	1	1	**2**
Total	**6**	**3**	**9**

The above table is called the contingency table. Since this table has two rows and two columns, it is also called 2 by 2 contingency table.

In practice, we have a more general m-by-n contingency table - this is called the two-way table. This week, we focus on analyzing two-way tables.

Example 2: Arrington et al. (2002) examined the frequency with which African, Neotropical and North American fishes have empty stomachs and found that the mean percentage of empty stomachs was around 16.2%. As part of the investigation, they were interested in **whether the frequency of empty stomachs was related to dietary items**. The data were separated into four major trophic classifications (detritivores, omnivores, invertivores, and piscivores) and whether the fish species had greater or less than 16.2% of individuals with empty stomachs. The number of fish species in each category combination was calculated and a subset of that (just the diurnal fish) was provided.

STOMACH Categorical listing of the proportion of individuals in the species with empty stomachs ($< 16.2\%$ or $> 16.2\%$).

TROPHIC Categorical listing of the trophic classification (DET = detritovore, OMN = omnivore, INV = invertivore, PISC = piscivore).

Then the above information is summarized in the following table.

Trophic classification	Stomachs empty ($<$ 16.2%)	Stomachs empty (>16.2%)	Total
DET	18	4	**22**
OMN	45	8	**53**
INV	58	15	**73**
PISC	16	34	**50**
Total	**137**	**61**	**198**

13.2 Two-way Contingency Tables and Analysis

A k-way contingency table is used to summarize the relationship between k categorical variables. It is a special type of frequency distribution table that k

variables are shown simultaneously.

In this note, We will introduce the general structure of two-way contingency tables and statistical tools for analyzing these tables. We will also introduce different study designs that generate two-way contingency tables containing different amounts of information.

13.2.1 The Structure of Two-way Contingency Table

Let X and Y be two categorical variables. Y is usually random (except in a case-control study) and is also called the response variable; X can be random or fixed, and usually acts as a predictor variable. Assume that X has I levels and Y has J levels. For ease of illustration, we consider the 4×3 contingency table. The general structure of this **observed** two-way contingency table (also called 4×3 table) is given by

X	Y			
	1	2	3	Row Total
1	n_{11}	n_{12}	n_{13}	n_{1+}
2	n_{21}	n_{22}	n_{23}	n_{2+}
3	n_{31}	n_{32}	n_{33}	n_{3+}
4	n_{41}	n_{42}	n_{43}	n_{4+}
Column Total	n_{+1}	n_{+2}	n_{+3}	n_{++}

Figure 13.1: Observed two-way table based on a random sample

Where the figures in the bottom and right-most column are called marginal totals. For examples,

- $n_{+2} = n_{12} + n_{22} + n_{32} + n_{42}$ is the column total of the second column.
- $n_{2+} = n_{21} + n_{22} + n_{23}$ is the row total of the second row.
- n_{++} = total of all frequencies in the table. This is also commonly called the grand total.

If the grand total is a simple random sample from a population, then we can easily turn the above-observed frequency table to an estimated **true joint probability distribution table** of X and Y that is below.

For illustration, we look at a few examples in the following.

- **Joint Probability**: $\hat{P}(X = 2, Y = 3) = \hat{p}_{23} = n_{23}/n_{++}$ is probability of observing $Y = j$ and $X = 2$. We can also estimate the marginal probability from the above model.

X	Y			
	1	2	3	Row Total
1	p_{11}	p_{12}	p_{13}	p_{1+}
2	p_{21}	p_{22}	p_{23}	p_{2+}
3	p_{31}	p_{32}	p_{33}	p_{3+}
4	p_{41}	p_{42}	p_{43}	p_{4+}
Column Total	p_{+1}	p_{+2}	p_{+3}	**1.0**

Figure 13.2: The true joint probability distribution of two categorical variables

- **Marginal Probability**: $\hat{P}(Y = 3) = \hat{p}_{+3} = n_{+3}/n_{++}$ is the probablity of observing $Y = 3$. The probability of observing $X = 2$ is estimated by $\hat{P}(X = 2) = \hat{p}_{2+} = n_{2+}/n_{++}$ is the probability of observing $X = 2$.

This means that we can estimate the **true joint distribution of** X **and** Y by using the observed table

13.2.2 Pearson χ^2 Test of Independence

As mentioned earlier, the null hypothesis of the Pearson χ^2 test of independence is the two categorical variables are independent. Since the general relationship between X and Y is completely determined by the above **joint probability distribution table**.

Under the null hypothesis H_o, the **joint probability distribution table of** X **and** Y has the following special structure.

X	Y			
	1	2	3	Row Total
1	$\hat{p}_{1+} \times \hat{p}_{+1}$	$\hat{p}_{1+} \times \hat{p}_{+2}$	$\hat{p}_{1+} \times \hat{p}_{+3}$	$\hat{p}_{1+}$
2	$\hat{p}_{2+} \times \hat{p}_{+1}$	$\hat{p}_{2+} \times \hat{p}_{+2}$	$\hat{p}_{2+} \times \hat{p}_{+3}$	$\hat{p}_{2+}$
3	$\hat{p}_{3+} \times \hat{p}_{+1}$	$\hat{p}_{3+} \times \hat{p}_{+2}$	$\hat{p}_{3+} \times \hat{p}_{+3}$	$\hat{p}_{3+}$
4	$\hat{p}_{4+} \times \hat{p}_{+1}$	$\hat{p}_{4+} \times \hat{p}_{+2}$	$\hat{p}_{4+} \times \hat{p}_{+3}$	$\hat{p}_{4+}$
Column Total	$\hat{p}_{+1}$	$\hat{p}_{+2}$	$\hat{p}_{+3}$	**1.0**

Figure 13.3: The true joint probability distribution of two categorical variables

This implies that, under the null hypothesis, the **estimated frequency** of any given cell, say $(X = 2, Y = 3)$, is given by $e_{23} = n_{++} \times \hat{p}_{2+} \times \hat{p}_{+3} = n_{2+}n_{+3}/n_{++}$. We can use the same formula to calculate all **estimated frequencies** - we call it the estimated table.

In other words, if the **observed** and **estimated** tables are close to each other, we intend to conclude the null hypothesis.

X	Y			
	1	2	3	Row Total
1	e_{11}	e_{12}	e_{13}	e_{1+}
2	e_{21}	e_{22}	e_{23}	e_{2+}
3	e_{31}	e_{32}	e_{33}	e_{3+}
4	e_{41}	e_{42}	e_{43}	e_{4+}
Column Total	e_{+1}	e_{+2}	e_{+3}	n_{++}

Figure 13.4: The true joint probability distribution of two categorical variables

To answer how close is close, we define a **statistical distance** between the **observed** and **expected** tables as follows

$$TS = \sum_{i=1}^{4}\sum_{j=1}^{2}\left(\frac{n_{ij}-e_{ij}}{\sqrt{e_{ij}}}\right)^2 \to \chi^2_{(4-1)(2-1)}$$

where $(4-1)(2-1)$ is the degrees of freedom of χ^2_3 distribution based on the above 4×2 contingency table. In general, for an $I\times J$ contingency table, the degrees of freedom are defined to be $(I-1)(J-1)$.

χ^2 is a distribution of a positive random variable such as the **distance** between **observed** and **estimated** tables. Different degrees of freedom define different χ^2 distributions. The following curve is a specific χ^2 distribution

Similar to normal and t-distributions, there are also R functions to find quantiles and tail probabilities.

```
pchisq(quantile, df)   # left-tail area of the chi-square distribution.
qchisq(left.tail.area, df) # quantile of the chi-square distribution.
```

The chi-squared test is essentially always right-tailed since the chi-squared test measures the discrepancy between two distributions. A large **distance** implies rejecting the null hypothesis. Therefore, the rejection is always on the right tail of the χ^2 density curve.

Example 2 (continued):

```
source("https://raw.githubusercontent.com/pengdsci/STA501/main/ref/w12-table2x2Calculat
## define column vector
less.16.2 = c(18, 45, 58, 16)
bigger.16.2 = c(4,8,15,34)
cont.table = cbind(less.16.2 = less.16.2 , bigger.16.2 = bigger.16.2)
chi.test = Pearson.chisq(cont.table)$inference
kable(chi.test, caption="Pearson chi-square test of independence")
```

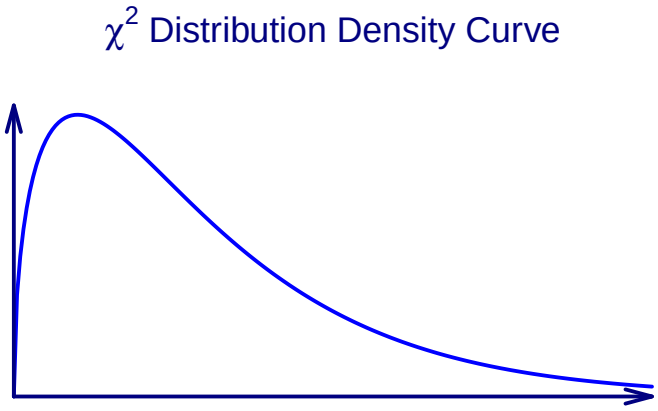


Figure 13.5: Density curve of a chi-sqaure distribution

Table 13.4: Pearson chi-square test of independence

	ts.stats	p.value	d.f	method
	43.8346	0	3	Pearson's Chi-squared test

Table 13.5: The expected table under the null hypothesis

	less.16.2	bigger.16.2
DET	15.22222	6.777778
OMN	36.67172	16.328283
INV	50.51010	22.489899
PISC	34.59596	15.404040

Table 13.6: The original observed frequency table

	less.16.2	bigger.16.2
DET	18	4
OMN	45	8
INV	58	15
PISC	16	34

The small p-value implies we rejected the null hypothesis. Therefore, the frequency of empty stomachs was related to dietary items. We can also extract the expected frequency table from the above χ^2 test

```
exp = Pearson.chisq(cont.table)$expected
row.names(exp)=c("DET", "OMN", "INV", "PISC")
kable(exp, caption = "The expected table under the null hypothesis")
```

The original observed table can also be retrieved from the above χ^2 test as follows.

```
obs = Pearson.chisq(cont.table)$observed
row.names(obs) = c("DET", "OMN", "INV", "PISC")
kable(obs, caption = "The original observed frequency table")
```

Caution of Using Pearson χ^2 Test: The Pearson χ^2 test assumes that the sample size must be large. The test result is not reliable if one or more of the cells of the contingency table is less than or equal to 5.

13.3 Measures of Association

Depending on the structure of the contingency table, we can define different measures for the association between the two categorical variables. In this section, we restrict our discussion to 2×2 tables. The general two-way or three-way tables are more complex and will not be discussed in this note.

A 2×2 table can be obtained from different study designs, hence, has a different amount of information.

13.3.1 2×2-table Under Different Study Designs

We use the following table as a template to discuss the different amount of information in the 2×2 tables based on the different study designs.

Characteristic A	Characteristic B		Total
	presence (D+)	absence (D-)	
presence (T+)	n_{11}	n_{12}	n_{1+}
absence (T-)	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

Figure 13.6: The true joint probability distribution of two categorical variables

We can think about **Characteristic B** to be the status of a disease and **Characteristic A** as the status of exposure to some risk.

- **Cross-sectional Study Design**

This method assumes a random sample of n subjects from a large population **followed** by the determination for each subject of the presence or absence of characteristic A and the presence or absence of characteristic B. **Only the total sample size n can be specified before the collection of the data**.

- **Stratified Sampling Design**

The stratified sampling (also called purposive sampling) assumes that the original population is stratified by either **characteristic A** or **characteristic B**.

1. *Stratification with Characteristic B.* This design stratifies the population into exposure and non-exposure sub-populations. Two simple random

samples were taken independently from the exposure and non-exposure populations respectively. This means that n_{1+} and n_{2+} are pre-determined sample sizes - the typical follow-up study is an example of this design.

2. *Stratification with Characteristic A*. This design stratifies the population into diseased and disease-free sub-populations. Two simple random samples were taken independently from the diseased and disease-free populations respectively. This means that n_{+1} and n_{+2} are pre-determined sample sizes - the typical case-control study is an example of this design.

- **Randomized Clinical Trials**

Of a total of n_{++} subjects, n_{1+} are selected at random to be treated with the control treatment (placebo), and the remaining n_{2+} to be treated with test treatment.

13.3.2 Measures of Association

We now define several measures of association. The inference of all three measures of association assumes that the sample size is large. **By convention, each cell of the observed frequency table must be bigger than or equal to 5 in order to yield reliable results**. For example, motivational example 1 has several small cells. Pearson χ^2 test and inference on measures of association are reliable. In order to perform the test of independence, we need to increase the sample size until all cell sizes are bigger than or equal to 5.

- **Relative Risk (also Risk Ratio)**: is defined to be the ratio of proportions of disease in exposure and non-exposure populations.

The relative risk (RR) is estimated by

$$\widehat{RR} = \frac{n_{11}/n_{1+}}{n_{21}/n_{2+}}$$

Note that the definition of RR requires fixed n_{1+} and n_{2+}. If $RR = 1$, there is no association between characteristics A and B. We will use the R program to calculate the confidence interval.

- **Risk Difference** is defined to be the difference of proportions of disease of exposure and non-exposure populations.

The risk difference (RD) can be estimated by

$$RD = n_{11}/n_{1+} - n_{21}/n_{2+}$$

Note also that the definition of RD requires fixed n_{1+} and n_{2+}. If $RD = 0$, there is no association between characteristics A and B.

- The **odds** of disease is the ratio of the probability of being disease divided by the probability of being disease-free.

For example, we can define odds of disease in exposure population to be $O(D+)_{T+} = [n_{11}/n_{1+}/[n_{12}/n_{1+}] = n_{11}/n_{12}$, and $O(D+)_{T-} = [n_{21}/n_{2+}/[n_{22}/n_{2+}] = n_{21}/n_{22}$ for the non-exposure population. **Note that the odds of an event are not a probability since it can be bigger than 1**.

From the definition of the odds in the above examples, we don't require fixed marginal totals.

- The **odds Ratio** of a disease is the ratio of the odds of disease in the exposed population and the odds of disease in the non-exposed population.

For example, $OR(disease) = O(D+)_{T+}/O(D+)_{T-} = n_{11}n_{22}/n_{12}n_{21}$. From this definition, we can see that there is no requirement for the marginal totals.

13.3.3 Validity of Measures of Association

We have defined several commonly used measures of association for 2×2 tables. Since the definition of different measures requires to have different fixed marginal totals. Therefore, one that is valid under one design might not be valid for another study design. The following is the summary of the validity of measures of association.

- Under the **cross-sectional design**, all measures of association (RD, RR, and OR) are valid.
- Under the **case-control study design** (fixed column totals), relative risk (RR) and risk difference (RD) are **invalid** since their definitions require fixed marginal totals of both exposure and non-exposure populations. The odds ratio (OR) is valid in the case-control study.
- Under the **follow-up study design** (fixed row totals), all three measures (RR, RD, and OR) are valid.
- Under the **randomized clinical trial (RCT)**, all three measures (RR, RD, and OR) are valid since the RCT is the combination of random sampling and (random) stratification for follow-up.

Study Design	Relative Risk	Risk Difference	Odds Ratio
cross-section	Valid	valid	valid
Stratified follow-up	valid	valid	valid
Stratified case-control	invalid	invalid	valid
RCT	valid	valid	valid

We will present examples in the next section of case studies using the two R functions **Pearson.chisq()** and **table2x2()** that can be sourced at source("https://raw.githubusercontent.com/pengdsci/STA501/main/ref/w12-table2x2Calculator.txt"). Whenever you use the function for the first time in the

current version, you simply place the one-line code **source("source("https://raw.githubusercontent.com/pengdsci/STA501/main/ref/w12-table2x2Calculator.txt")")** in the code chunk before calling the above two R functions. In the rest of the current session, you don't need to source the code!

13.4 Case Studies

When conducting the independence test, we need to check the assumption of sample size in order to make the correct decision. For the analyzing 2×2 table, we need to know the study designs since some measures may not be valid for all designs.

13.4.1 Case Study 1

Example 1. In 1992, the U.S. Public Health Service and the Centers for Disease Control and Prevention recommended that all women of childbearing age consume 400mg of folic acid daily to reduce the risk of having a pregnancy that is affected by a neural tube defect such as spina bifida or anencephaly. In a study by Stepanuk et al, 693 pregnant women called a teratology information service about their use of folic acid supplementation. The researchers wished to determine if preconceptional use of folic acid and race are independent. The data appear in the following table.

	Preconceptional Use of Folic Acid		
	Yes	**No**	**Total**
White	260	299	559
Black	15	41	56
Other	7	14	21
Total	282	354	636

Source: Kathleen M. Stepanuk, Jorge E. Tolosa, Dawneete Lewis, Victoria Meyers, Cynthia Royds, Juan Carlos Saogal, and Ron Librizzi, "Folic Acid Supplementation Use Among Women Who Contact a Teratology Information Service," *American Journal of Obstetrics and Gynecology, 187* (2002), 964–967.

Figure 13.7: Use of folic acid supplementation data table

We use the R function **Pearson.chisq()** to test the independence between the race and the use of folic acid.

```
source("https://raw.githubusercontent.com/pengdsci/STA501/main/ref/w12-table2x2Calculat
## 2-by-2 table must be defined as a data frame
case1 <- data.frame(Yes=c(260, 15, 7),No=c(299, 41,14))
```

```
rownames(case1) <- c("white","Black", "Other")
Pearson.chi.test = Pearson.chisq(case1)
##
pander(Pearson.chi.test$inference, caption="Summary of Pearson chi-square test of independence")
```

Table 13.8: Summary of Pearson chi-square test of independence

ts.stats	p.value	d.f	method
9.0913	0.0106	2	Pearson's Chi-squared test

The above Pearson χ^2 test independence indicates that the pre-conceptional use of folic acid and race are dependent with $\chi^2_2 = 9.0913$ with p-value = 0.0106.

To compare the observed frequency table and the expected frequency table, we can extract them from the above **Pearson.chisq()**.

```
pander(Pearson.chi.test$observed, caption = "The observed frequency table")
```

Table 13.9: The observed frequency table

	Yes	No
white	260	299
Black	15	41
Other	7	14

```
pander(Pearson.chi.test$expected, caption = "The expected frequency table")
```

Table 13.10: The expected frequency table

	Yes	No
white	247.9	311.1
Black	24.83	31.17
Other	9.311	11.69

We can see that the above two tables are different. The second row of the expected table is significantly different from that of the observed table.

13.4.2 Case Study 2

Example 2: The food-frequency questionnaire is widely used to measure dietary intake. A person specifies the number of servings consumed per day of

each of many different food items. The total nutrient composition is then calculated from the specific dietary components of each food item. One way to judge how well a questionnaire measures dietary intake is by its reproducibility. To assess reproducibility the questionnaire is administered at two different times to 50 people and the reported nutrient intakes from the two questionnaires are compared. Suppose dietary cholesterol is quantified on each questionnaire as high if it exceeds 300 mg/day and as normal otherwise. The contingency table in the following table is a natural way to compare the results of the two surveys. Notice that this example has no natural denominator. We simply want to test whether there is some association between the two reported measures of dietary cholesterol for the same person. More specifically, we want to assess how unlikely it is that 15 women will report high dietary cholesterol intake on both questionnaires, given that 20 of 50 women report high intake on the first questionnaire and 24 of 50 women report high intake on the second questionnaire. This test is called a test of independence or a test of association between the two characteristics.

	Second food-frequency questionnaire		
First food-frequency questionnaire	High	Normal	Total
High	15	5	20
Normal	9	21	30
Total	24	26	50

Figure 13.8: dietary intake data table

```
source("https://raw.githubusercontent.com/pengdsci/STA501/main/ref/w12-table2x2Calculat
## 2-by-2 table must be defined as a data frame
case2 <- data.frame(High=c(15, 9),Normal=c(5,21))
rownames(case2) <- c("High","Normal")
Pearson.chi.test02 = Pearson.chisq(case2)
##
pander(Pearson.chi.test02$inference, caption="Summary of Pearson chi-square
      test of independence")
```

Table 13.11: Summary of Pearson chi-square test of independence

ts.stats	p.value	d.f	method
8.0162	0.0046	1	Pearson's Chi-squared test(with correction)

The Pearson χ^2 test indicates that the first and second questionnaires are dependent ($\chi^2_1 = 8.0163$, p-value $= 0.0046$). Next, we use measures of association to assess the strength of the association.

```
source("https://raw.githubusercontent.com/pengdsci/STA501/main/ref/w12-table2x2Calculator.txt")
##
pander(table2x2(case2), caption="Measures of association")
```

Table 13.12: Measures of association (continued below)

	measure	SE	Lower(95%).CI	Upper(95%).CI
relative.risk	2.5	0.3073	1.3689	4.5658
risk.difference	0.45	0.128	0.1991	0.7009
odds.ratio	7	0.6522	1.9496	25.1334

	significance(0.05)
relative.risk	Yes
risk.difference	Yes
odds.ratio	Yes

Since the above 2×2 table was generated based on cross-sectional design (assuming the random sampling). All measures of association are valid. The above results based on measures of association are consistent with that of the Pearson χ^2 test. Moreover, we can explicitly interpret the strength of the association between the two survey results.

- **relative risk**: the percentage of those who reported **high-dietary-intake** in the first survey also reported the **same** in the second survey is 2.5 times the percentage of those who reported **low-dietary-intake** in the first survey but reported **high-dietary-intake** in the second survey. The corresponding 95% of the confidence interval is $[1.37, 4.57]$.
- **risk difference**: the percentage of those who reported **high-dietary-intake** in the first survey also reported the **same** in the second survey is 45% higher than the percentage of those who reported **low-dietary-intake** in the first survey but reported **high-dietary-intake** in the second survey. The corresponding 95% of confidence interval is $[19.9\%, 70.1\%]$.
- **odds ratio**: The odds of reporting **high-dietary-intake** among those who reported **high-dietary-intake** in the first survey is 7 times that among those who reported **low-dietary-intake** in their first survey.

13.5 Practice Problems

In this assignment, we focus on the Pearson χ^2 test and the measures of association. The level of detail should be similar to what you have seen in the case study (Section 5 of the class note). You can also use the two R functions that

were used in the class note by running the following one-line code in the code chunk.

```
source("https://raw.githubusercontent.com/pengdsci/STA501/main/ref/w12-table2x2Calculat
```

Problem 1.

In a prospective, randomized, double-blind study, Stanley et al. examined the relative efficacy and side effects of morphine and pethidine, drugs commonly used for patient-controlled analgesia (PCA). Subjects were 40 women, between the ages of 20 and 65 years, undergoing total abdominal hysterectomy. Patients were allocated randomly to receive morphine or pethidine by PCA. At the end of the study, subjects described their appreciation of nausea and vomiting, pain, and satisfaction by means of a three-point verbal scale. We only focus on the association between the severity of nausea and the use of the drug. The observed table is given below.

	Nausea			
Drug	**Unbearable/ Severe**	**Moderate**	**Slight/ None**	**Total**
Pethidine	5	9	6	20
Morphine	7	8	5	20
Total	12	17	11	40

Source: Data provided courtesy of Dr. Balraj L. Appadu.

Is there any association between appreciation of nausea and satisfaction? Use the Pearson χ^2 test to justify your conclusion and interpret the test results.

Problem 2

The following data table was collected to study the obesity status of children ages 5–6 years and the smoking status of the mother during the pregnancy, also reported on another outcome variable: whether the child was born premature (37 weeks or fewer of gestation). The following table summarizes the results of this aspect of the study. The same risk factor (smoking during pregnancy) is considered, but a case is now defined as a mother who gave birth prematurely.

	Premature Birth Status		
Smoking Status During Pregnancy	**Cases**	**Noncases**	**Total**
Smoked throughout	36	370	406
Never smoked	168	3396	3564
Total	204	3766	3970

Source: A. M. Toschke, S. M. Montgomery, U. Pfeiffer, and R. von Kries, "Early Intrauterine Exposure to Tobacco-Inhaled Products and Obesity," *American Journal of Epidemiology*, *158* (2003), 1068–1074.

Use the chi-square test of independence to determine if one may conclude that there is an association between smoking throughout pregnancy and premature birth. Let $\alpha = 05$. Compute the odds ratio to determine if smoking throughout pregnancy is related to premature birth.

Chapter 14

Analysis of Counts and Rates

This module considers the relationship between a discrete response variable and other numeric or categorical predictor variables. The analysis of frequency counts and rates is also one of the important statistical tools in life science. The appropriate model we will discuss is a small family of generalized linear models - The Poisson regression model. It has wide applications for both laboratory experimental data and field data which involve count or rate data.

We add this model to the summary table of models in the previous module as follows

Response variable	Predictor variable	Type of Models
continuous, normal	single categorical	ANOVA
continuous, normal	single continuous	SLR
continuous, normal	continuous or categorical	MLR (ANCOVA)
binary, categorical	continuous or categorical	logistic model
numeric, discrete	continuous or categorical	Poisson model

14.1 Motivational Examples

Example 1: Aerial counts of harbor seals (Phora vitulzna concoZw) on ledges along the Maine coast were conducted during the pupping season in 1981, 1986, 1993, 1997, and 2001 to study the changes in abundance of harbor seals. Detailed information on the study can be found in the published work of Gilbert et al (2005) see the link of the article.

We are interested in whether the counts of harbor seal counts changed significantly over the years. The data used to answer this question is taken from the first half of Table 6.

530 MARINE MAMMAL SCIENCE, VOL. 21, NO. 3, 2005

Table 6. Number of ledge and island sites occupied by harbor seals in Maine from 1981 to 2001.

Region	1981	1986	1993	1997	2001
Sites with harbor seals					
South of Cape Elizabeth	13	11	16	18	18
Casco Bay	26	22	41	33	43
Boothbay region	15	15	23	32	26
Muscongus Bay	28	21	44	44	47
Penobscot Bay	80	72	148	138	125
Blue Hill Bay	75	54	123	113	107
Frenchman's Bay	23	10	26	25	28
Narraguagus region	24	24	38	33	36
Western Bay	19	18	30	28	36
Eastern Bay	9	13	29	27	35
Machias region	14	15	37	34	41
Cobscook Bay	10	10	19	16	25
Total	336	285	574	541	566

We construct the R data set in the following based on the above data table that is suitable for modeling.

Example 2: This example is based on a study that investigated the cytotoxic and genotoxic effects of soluble and particulate hexavalent chromium in sperm whale skin fibroblasts. The data were extracted from a line plot in the published work of Wise et al (Fig. 1). In the first experiment, Particulate Cr(VI) induced a clear concentration-dependent decrease in cell survival over a range of 0.05 to 0.5 lg/cm2. Concentrations of 0.05, 0.1, 0.5, 1, and 5 lg/cm2 lead chromate induced 85%, 86%, 63%, 56%, and 26% relative survival. This information is given in the following figure 1 of the published paper.

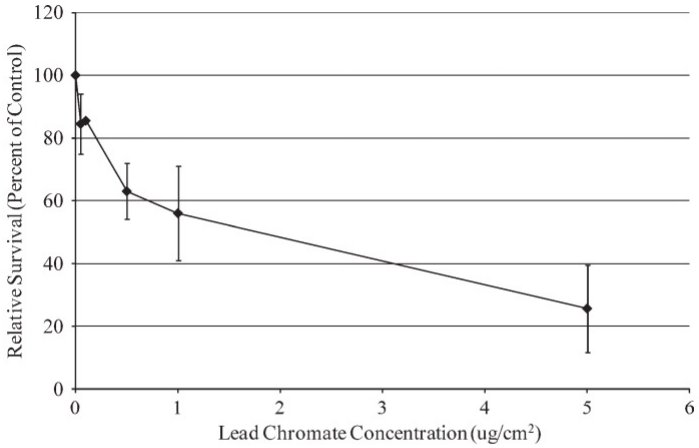


Fig. 1. Particulate Cr(VI) cytotoxicity in sperm whale skin cells. This figure shows the cytotoxic responses in sperm whale skin cells following exposure to particulate Cr(VI). No statistically significant difference ($P > 0.05$) was observed. Data represent the average of relative cell survival $\pm$ S. E.; $n = 4$.

Dose level	survived cells	total cells
0	100	100
0.05	85	100
0.1	86	100
0.5	63	100
1	56	100
5	26	100

The question is whether the survival rates are associated with the dose level.

14.2 Poisson Regression for Counts and Rates

The Poisson regression model assumes the random response variable to be a frequency count or a rate of an uncommon event such as COVID-19 positivity rates, COVID-19 death mortality, etc. As in the linear and logistic regression models, we also assume that predictor variables are non-random.

14.2.1 Structure and Interpretations

Let Y be the response variable that takes on frequency counts as values and X be the set of predictor variables such as demographics and social determinants. Further, let $\mu = E[Y]$ be the mean of the response variable.

Poisson Regression for Counts

The Poisson regression model is defined in the following analytic expression.

$$\log(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p,$$

where $\beta_0, \beta_1, \cdots, \beta_p$ are coefficients of the Poisson regression model.

Poisson Regression for Rates

The Poisson regression model for rates is defined in the following analytic expression.

$$\log(\mu/t) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p,$$

where $\beta_0, \beta_1, \cdots, \beta_p$ are coefficients of the Poisson regression model. t is called the **offset** variable. The offset variable serves to normalize the fitted cell means per some space, grouping, or time interval in order to define the meaningful rates.

Interpretation of Regression Coefficients

The interpretation of the regression coefficient β_i is as follows

- β_0 = the baseline logarithm of the mean of Y, $\log(\mu)$, when all predictor variables $x_i = 0$, for $i = 1, 2, \cdots, p$. As usual, we are not interested in the inference of the intercept parameter.
- β_i = is the change of the **logarithm of the mean count** due to one unit increases in x_i with all other x_j being fixed, for $j \neq i$.

Estimation of Regression Coefficients

Estimating Poisson regression coefficients requires numerical optimization. We will not go into detail about how to estimate the regression coefficients and perform model diagnostics in this module. Instead, we will focus on data analysis, in particular, the interpretation of regression coefficients.

14.2.2 Assumptions and Goodness-of-fit

Like other statistical models, there are some assumptions for the Poisson regression model:

- The response variable is a frequency count (or rate variable) that follows the Poisson distribution.
- The mean of the and the variance are equal.
- The relationship between the mean of the response and the predictor variables is correct.

- For a given value of the predictor variable, the mean and variance of the response variable are **equal**. - Unfortunately, this assumption is frequently violated. This type of violation is serious since it produces wrong estimates of the standard errors and, hence, yields wrong p-values.

14.2.3 Dispersion Issue and Remedies

The first three assumptions mentioned above are regular for all regression models. The violation of the last assumption is directly associated with the distribution of the response variable. Different causes lead to the violation. For example, the data is not from Poisson distribution or a mixture distribution of Poisson and other distributions. Therefore, there are different ways we can consider to fix the problem.

Although the detailed discussion of remedies is not the focus of this course, it is useful to know some of the available remedies for dispersion (either over-dispersion or under-dispersion).

- **quasi-Poisson regression** sticks to the simple structure of the Poisson regression and adjusts the dispersed standard error to obtain the valid p-values. We will use this approach in this class.
- **negative binomial regression**, another family regression model for the discrete response variable, relaxes Poisson's assumption on equality of mean and variance. In the **negative binomial regression**, the variance is a function of the mean. It could also have dispersion problems if the variance function is not correct. The **negative binomial regression** is implemented in R.
- **Zero-inflated family of regression models** assumes the data come from the mixture distribution of Poisson and other distributions such as binomial or negative binomial and binomial distributions. R also has libraries to fit several zero-inflated regression models.
- The **Hurdle model** also handles the issue of excess zeros in the data but assumes the sources of zeros in the data are different. Hurdle assumes structural zero and the regular zero-inflated model assumes sampling zeros.

We will use **Poisson regression** to detect potential **dispersion** and then decide whether to use the regular Poisson regression model to report and implement in practical applications.

14.2.4 Data Structure of Poisson Regression

The Poisson regression is a subfamily of generalized linear regressions(GLM). The logistic regression is also a member of GLM. Similar to the structure used in the logistic regression, Poisson regression also requires the same data structure usually called the **long table**.

The data table in the first motivational example is not a **long table**. It is actually a **wide table**. The table in example 2 is a **long table**. When using R to build models in GLM, the data should always be in the form of a **long table**. Therefore, the data table in motivational example 1 **cannot** be used to build GLMs. The code for turning the wide table into a long table is given in Section 2. The resulting **long table** (partial table) is shown below.

```
y1981=c(13, 26, 15, 28, 80, 75, 23, 24, 19, 9, 14, 10)
y1986=c(11, 22, 15, 21, 72, 54, 10, 24, 18, 13, 15, 10)
y1993=c(16, 41, 23, 44, 148, 123, 26, 38, 30, 29, 37, 19)
y1997=c(18, 33, 32, 44, 138, 113, 25, 33, 28, 27, 34, 16)
y2001=c(18, 43, 26, 47, 125, 107, 28, 36, 36, 35, 41, 25)
seal.vec = c(y1981, y1986, y1993, y1997, y2001)
time.vec = sort(rep(c("y1981", "y1986", "y1993", "y1997", "y2001"), 12))
pois.data= data.frame(cbind(seal=seal.vec, time=time.vec))
```

```
n=dim(pois.data)[1]
partial.long.table = pois.data[sample(1:n, n, replace = FALSE), ][1:10,]
pander(partial.long.table, caption="Part of the converted long table from the harbor se
```

Table 14.3: Part of the converted long table from the harbor seal data table

	seal	time
23	15	y1986
12	10	y1981
2	26	y1981
14	22	y1986
46	27	y1997
6	75	y1981
54	107	y2001
33	30	y1993
24	10	y1986
36	19	y1993

14.3 Case Studies

We will use the two motivational examples in this section. As mentioned earlier, the Quasi-Poisson Regression is a generalization of the Poisson regression and is used when modeling an overdispersed count variable.

The Poisson model assumes that the variance is equal to the mean, which is not always a fair assumption. When the variance is greater than the mean, a Quasi-Poisson model, which assumes that the variance is a linear function of the mean, is more appropriate.

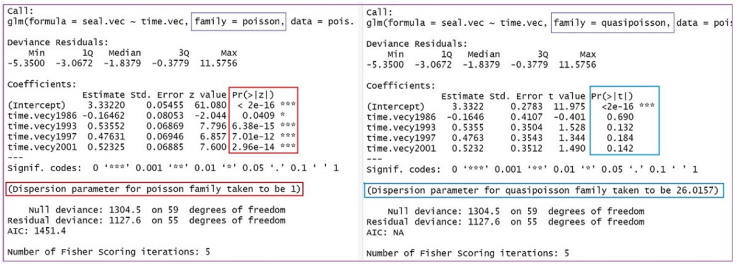
For each example, we will fit both Poisson and quasi-Poisson regression models.

14.3.1 Harbor Seal Data

We will use the **long table** to fit the two models. Which is to choose to report will depend on the dispersion parameter. Next, we use the R function **glm()** to fit Posson and quasi-Poisson models in the following.

```
pois.model = glm(seal.vec ~ time.vec, family=poisson, data = pois.data)
summary.table.pois = summary(pois.model)
##
quasi.pois.model = glm(seal.vec ~ time.vec, family=quasipoisson, data = pois.data)
summary.table.quasi.pois = summary(quasi.pois.model)
```

The complete outputs of the two models from R function **glm()** are given in the following figure.



```
Call:
glm(formula = seal.vec ~ time.vec, family = poisson, data = pois.

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.3500  -3.0672  -1.8379  -0.3779  11.5756

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    3.33220    0.05455  61.080  < 2e-16 ***
time.vecy1986 -0.16462    0.08053  -2.044   0.0409 *
time.vecy1993  0.53552    0.06869   7.796 6.38e-15 ***
time.vecy1997  0.47631    0.06946   6.857 7.01e-12 ***
time.vecy2001  0.52325    0.06885   7.600 2.96e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1304.5  on 59  degrees of freedom
Residual deviance: 1127.6  on 55  degrees of freedom
AIC: 1451.4

Number of Fisher Scoring iterations: 5
```

```
Call:
glm(formula = seal.vec ~ time.vec, family = quasipoisson, data = pois

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.3500  -3.0672  -1.8379  -0.3779  11.5756

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.3322     0.2783  11.975  <2e-16 ***
time.vecy1986 -0.1646     0.4107  -0.401   0.690
time.vecy1993  0.5355     0.3504   1.528   0.132
time.vecy1997  0.4763     0.3543   1.344   0.184
time.vecy2001  0.5232     0.3512   1.490   0.142
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 26.0157)

    Null deviance: 1304.5  on 59  degrees of freedom
Residual deviance: 1127.6  on 55  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

We can observe several pieces of information from the above figure.

- The regression coefficients in both Poisson and quasi-Poisson regression models are identical.
- The standard errors in the Poisson regression model are less than their corresponding standard errors in the quasi-Poisson regression model.
- The dispersion parameter is forced to be 1 in the Poisson regression model. However, the dispersion parameter is calculated through the quasi-likelihood that yields the value of dispersion parameter 26.0157. This is much bigger than 1 (for the Poisson regression model). Therefore, the p-values in the output of the regular Poisson regression model are not reliable. The inference should be based on the output of the quasi-Poisson model.

The p-values in the quasi-Poisson regression can be extracted in the following table.

```
example.coef.table = summary.table.quasi.pois$coef
pander(example.coef.table, caption="The summary statistics based on
      the quasi-Poisson regression model")
```

Table 14.4: The summary statistics based on the quasi-Poisson regression model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.332	0.2783	11.98	0.00000000000000005951
time.vecy1986	-0.1646	0.4107	-0.4008	0.6901
time.vecy1993	0.5355	0.3504	1.528	0.1321
time.vecy1997	0.4763	0.3543	1.344	0.1843
time.vecy2001	0.5232	0.3512	1.49	0.1419

All p-values associated with the survey year are insignificant. The baseline year is 1981 (which is not in the output), this means that counts of harbor seals in any of the survey years were **not significantly** from the baseline year (1981).

14.3.2 Toxicity Study

This case study is based on the second motivational example. The analysis in the original article involves errors. The statistical problem is a rate regression, the analysis in the original paper used ANOVA which led to a wrong conclusion. Two different methods will be used to assess the association between the concentration and the survival of cells.

14.3.2.1 Poisson Regression

We can simply use the summarized table extracted from the original article given in Section 2. This is a small data set with only 6 observations and three variables: dose(continuous), survival(discrete - count), and total (discrete - count, which can only be used as an offset variable in the model).

```
dose =   c(0, 0.05,  0.1, 0.5, 1,   5)
survived=c(100, 85,  86,  63,  56,  26)
total =  c(100, 100, 100, 100, 100, 100)
pois.data = data.frame(cbind(survived=survived, dose=dose, total = total))
```

Next, we fit the quasi-Poisson to the above data and check the dispersion parameter to see whether the regular Poisson regression is appropriate.

```
quasi.pois=glm(survived ~ dose + offset(total), family = quasipoisson, data = pois.data
disp = summary(quasi.pois)$dispersion
pander(cbind(dispersion=disp), caption = "The dispersion paramter of the Poisson regres
```

Table 14.5: The dispersion paramter of the Poisson regression The value of the dispersion is slightly bigger than 1 (if there is no dispersion, the dispersion parameter = 1). We only need to fit the Poisson regression to the data and use the fitted Poisson regression model to address the association between the concentration level and the survival rate.

dispersion
1.579

```
pois=glm(survived ~ factor(dose) + offset(total), family = poisson, data = pois.data)
coef=summary(pois)$coef
pander(coef, caption = "Summary statistics of the regression coefficients")
```

Table 14.6: Summary statistics of the regression coefficients Contrary to what was concluded in the original article, the concentration of lead chromate significantly affects the survival of the cells (p-value ≈ 0). As the concentration level increases, the survival rate decreases significantly. To be more specific, we can exponentiate the coefficient of the Poisson regression associated with **dose** and obtain $exp(-0.2625) = 0.7691 = 1 - 0.2309$. This means that, as the concentration increases by one unit, the **survival rate** of the skin cells **decreases** by 23.09%.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-95.39	0.1	-953.9	0
factor(dose)0.05	-0.1625	0.1475	-1.102	0.2706
factor(dose)0.1	-0.1508	0.1471	-1.026	0.3051
factor(dose)0.5	-0.462	0.1609	-2.872	0.004073
factor(dose)1	-0.5798	0.1669	-3.474	0.0005129
factor(dose)5	-1.347	0.2201	-6.119	0.0000000009406

14.3.2.2 Logistic Regression Approach (optional)

Since the logistic regression requires the response to be binary, we need to perform some data management to create a suitable data set for the logistic regression model.

Data Preparation

The data reported in the experiment are the percentage of survival of cells with the given total number of cells that died and survived respectively at different levels of concentration. We assume 100 cells were used at each concentration. So we retrieved the data table from the chart in the original paper and summarized

it in Section 2. Next, we create a **long table** and fit the Poisson model to the data. The idea is each column will record the information of each cell. The layout of the long table to be used in the Poisson and quasi-Poisson regression model is depicted in the following:

cell ID	dose	survival	total
1	0	1	1
2	0	1	1
...	...	...	...
100	0	1	1
101	0.05	1	1
...	...	...	...
185	0.05	1	1
186	0.05	0	1
187	0.05	0	1
...	...	...	...
...	...	...	...
501	5	1	1
502	5	1	1
...	...	...	...
526	5	1	1
527	5	0	1
526	5	0	1
...	...	...	...
599	5	0	1
600	5	0	1

We define the long table in the following code.

```
cell.id = 1:600                      # cell ID, not a meaningful variable!
total = rep(1,600)                   # the "total" of each cell is simply equal to 1.
dose.0 = rep(1, 100)                 # all 100 cells survived with concentration level
dose.005 = c(rep(1,85), rep(0,15))   # 85 cells survived and 15 died at concentration l
dose.0.1 = c(rep(1,86), rep(0,14))   # 86 cells survived and 14 died at concentration l
dose.0.5 = c(rep(1,63), rep(0,37))   # 63 cells survived and 37 died at concentration l
dose.1 = c(rep(1, 56), rep(0,44))    # 56 cells survived and 44 died at concentration l
dose.5 = c(rep(1,26), rep(0,74))     # 26 cells survived and 74 died at concentration l
survival = c(dose.0, dose.005, dose.0.1, dose.0.5, dose.1, dose.5)
# next line of code defines an indicator telling the concentration level of each cell
dose = c(rep(0, 100), rep(0.05, 100), rep(0.1, 100), rep(0.5, 100), rep(1, 100), rep(5
## The long table is defined in the following one line of code
toxic.data = data.frame(cbind( cell.id = cell.id, survival = survival, dose=dose, total
```

Model Building

This is a typical Poisson rate regression problem. As usual, we will fit both

Poisson rate and quasi-Poisson rate models to the data set and then look at the dispersion parameter to decide which model should be used.

Based on the extracted data, the sample size is 600. The following quasi-Poisson regression model indicates that there is no significant dispersion issue for the Poisson regression model.

```
logit.model = glm(survival ~ dose, family = binomial, data = toxic.data)
logit.summary = summary(logit.model)$coef
pander(logit.summary, caption = "Summary statistics on the regression coefficients")
```

Table 14.8: Summary statistics on the regression coefficients

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.519	0.1199	12.67	9.23e-37
dose	-0.5635	0.05662	-9.952	0.00000000000000000000002462

The regression coefficient associated with **dose** is negative meaning that as the dose increases the log-odds of survival decrease. To be more specific, we exponentiate the coefficient of *dose* and obtain $exp(-0.5635) = 0.5692 = 1 -43.08 $. This means that, as the dose increase by one unit, the **odds of survival** of the skin cells **decreases** about 43.08%.

Conclusion

In summary, the logistic regression yields the same result as that from the Poisson regression. The concentration of lead chromate is negatively associated with the survival of the skin cells of the sperm whale.

14.3.2.3 Pearson χ^2 Test of Independence Approach

We can also answer the original question by testing the hypothesis that the concentration does not affect survival (independence test). This method only provides whether there is an association between concentration and survival but not the direction and magnitude of the association. To prepare for the Pearson χ^2 test, we need to construct a $2 \times k$ table in the following.

surv.status	level-0.0	level-0.05	level-0.1	level-0.5	level-1	level-5
survived	100	85	86	63	56	26
died	0	15	14	37	44	74

We next construct the observed table in R and then perform the χ^2 test.

```
source("https://raw.githubusercontent.com/pengdsci/STA501/main/ref/w12-table2x2Calculator.txt")
survived=c(100, 85,  86,  63,  56,  26)
```

```
died = 100 - survived
obs.table = rbind(survived = survived, died = died)
colnames(obs.table) = as.character(c(0, 0.05, 0.1, 0.5, 1, 5))
chi.test = Pearson.chisq(obs.table)$inference
pander(chi.test, caption="Pearson chi-square test of independence of concentration and
```

Table 14.10: Pearson chi-square test of independence of concentration and survival

ts.stats	p.value	d.f	method
167.4018	0	5	Pearson's Chi-squared test

The Pearson χ^2 test indicates that the concentration level of lead chromate is **NOT** independent of the survival of the skin cell of the sperm whale. Although the test itself does not give the direction of the association, we can find the information plot plotting the concentrations and survival rates.

14.4 Concluding Remarks

Comparing multiple unrelated proportions (rates) is one of the important methods in analyzing laboratory data. We have summarized several different methods above to address different types of comparison questions that may arise in the actual research hypotheses. It is not as straightforward as the comparison for multiple population means since it requires different and more advanced statistical tools to address the specific comparison questions.

There are several other recently developed procedures in the literature. Some of these new methods have been implemented in software packages.

- we can also use the command **prop.test()** to test the equality of multiple proportions.
- The Marascuilo procedure enables us to simultaneously test the differences of all pairs of proportions when there are several populations under investigation
- One most recently (2017) developed one-way ANOVA-like method uses the idea of the likelihood ratio test.

14.5 Practice Problems

We have assessed the association between the harbor seal count observed in Maine's coastal regions and time (different survey years between 1981 and 2001) using both Poisson and quasi-Poisson regression. In this assignment, you assess the association between the counts of harbor seal pubs and the time. The data

table is given below (the frequency counts are in the red box of the following table)

Table 6. Number of ledge and island sites occupied by harbor seals in Maine from 1981 to 2001.

Region	1981	1986	1993	1997	2001
Sites with harbor seals					
South of Cape Elizabeth	13	11	16	18	18
Casco Bay	26	22	41	33	43
Boothbay region	15	15	23	32	26
Muscongus Bay	28	21	44	44	47
Penobscot Bay	80	72	148	138	125
Blue Hill Bay	75	54	123	113	107
Frenchman's Bay	23	10	26	25	28
Narraguagus region	24	24	38	33	36
Western Bay	19	18	30	28	36
Eastern Bay	9	13	29	27	35
Machias region	14	15	37	34	41
Cobscook Bay	10	10	19	16	25
Total	336	285	574	541	566
Sites with harbor seal pups					
South of Cape Elizabeth	6	6	10	8	17
Casco Bay	13	18	32	26	37
Boothbay region	13	9	17	21	22
Muscongus Bay	17	12	32	27	40
Penobscot Bay	49	45	112	92	113
Blue Hill Bay	45	39	97	91	101
Frenchman's Bay	13	8	23	19	27
Narranguagus region	12	22	28	26	34
Western Bay	10	15	25	26	33
Eastern Bay	5	11	23	26	34
Machias region	5	7	23	26	36
Cobscook Bay	4	7	12	9	19
Total	186	193	424	389	496

To save your time, define a vector for each survey year in the following code chunk.

```
y.1981=c(6, 13, 13, 17, 49, 45, 13, 12, 10, 5, 5, 4)
y.1986=c(6, 18, 9, 12, 45, 39, 8, 22, 15, 11, 7, 7)
y.1993=c(10, 32, 17, 32, 112, 97, 23, 28, 25, 23, 23, 12)
y.1997=c(8, 26, 21, 27, 92, 91, 19, 26, 26, 26, 26, 9)
y.2001=c(17, 37, 22, 40, 113, 101, 27, 34, 33, 34, 36, 19)
```

Please refer to the case study in the class note to analyze the data and draw conclusions to address the research question. To be more specific, you are expected to answer the following question.

1. Fit both regular and quasi-Poisson regression models.
2. Pick a model as the final model and justify your choice.
3. Comment on the dispersion and interpret the output of the final model.
4. Write a separate paragraph to summarize the results and draw a conclusion.

Chapter 15

Power Calculation and Sample Size Determination

Estimating the sample size for a prospective study is one of the first and most important tasks in classical statistics. In clinical studies, researchers need to determine the number of subjects to enroll in a clinical trial to guarantee the reliability of the study results. If the sample size is too small, it will not be able to achieve the power of the analysis. If the size is excessively larger than needed, it will waste resources.

Estimating the optimal number of patients is necessary for developing a Statistical Analysis Plan (SAP) in all prospective studies. In retrospective studies, we can determine how much power of the analysis (for example, the power of detecting a difference if one existed).

This chapter focuses on how to

1) estimate sample size for a prospective study (e.g., randomized controlled trial)

2) determine how much power we have to detect a difference if one existed (post hoc analysis)

We also restrict our discussion within two-sample tests.

15.1 Two-sample Proportion Problems

Before enrolling patients in a clinical trial, we want to make sure that we have the optimal number of patients for the study. Different statistical procedures require different optimal sample sizes to attain the target power. Next, we discuss sample size determination for several commonly used statistical procedures.

Before moving forward, let's recall two error probabilities associated with a statistical test of hypothesis H_0:

α = P[type I error] = P[reject Ho | Ho is true] = false positive rate

β = P[type II error] = P[fail to reject Ho | Ho is false] = false negative.

power of a test = $1 - \beta$ = P[reject Ho | Ho is false]

Several observations about α and β:

- *alpha* and β are negatively dependent on each other (not linearly dependent).
- Estimated *alpha* and β are dependent on the sample size.

  In general, if more than one test for testing a given hypothesis Ho under the same α, the one that achieves the lowest β is the best. This tells the relationship among estimated α, β, power, and the sample size n.

15.1.1 Sample size estimations for two proportions

This is a common sample size calculation when you have two groups and the outcome is dichotomous (e.g., Yes / No) Assume that we have a study that aims to estimate the efficacy of Treatment A versus Treatment B. The main endpoint is a dichotomous variable: "Response" or "No response."

The null hypothesis is:

H_0: There is no difference in the proportion of responders between Treatment A and Treatment B

The alternative hypothesis is:

H_a: There is a difference in the proportion of responders between Treatment A and Treatment B

To determine the sample size needed to detect a meaningful difference in the proportion of responders between Treatment A and Treatment B, we need the following information:

- α **level** (type I error) is the threshold where we determine whether something is statistically significantly different. We normally use an alpha level of 0.05 (type I error).
- **Effect size** is the standardized difference one would expect to see between treatment groups. This is an estimate and something that is usually based on past studies.

- **Power (1 - β)** is the ability to correctly reject the null hypothesis if the actual effect of the population is equal to or greater than the specified effect size.

We'll set the two-tailed alpha level to 0.05 estimate size h using the following equation:

$$h = \varphi_1 - \varphi_2$$

where $\varphi_i = 2 * arcsine(\sqrt{p_i})$ and p_i denotes the proportion in treatment i that were classified as "responders." In general, the power is set to 80%.

With these pieces of information, we can estimate the sample size for a study with two proportions as the outcome using the following formula:

$$n_i = \frac{(Z_{\alpha/2} + Z_{1-\beta})^2 + (p_1(1-p_1)) + (p_2(1-p_2))}{(p_1 - p_2)^2}$$

where n_i is the sample size for one group.

Use the `pwr.2p.test()` function in `library("pwr")` to find out the optimal sample size that meets the requirements on the alpha level (`alpha = 0.05`), effect size h, and power level (`power = 80%`).

Example 1: We consider Treatment A and Treatment B in a survey study. Set Treatment B as the reference (e.g., control) with a 50% response rate. Assume that Treatment A is slightly better with a response rate of 60%. With these pieces of information, we can estimate the sample size required to detect a difference in the "response" rate between Treatment A and Treatment B that is 10% (p_1 - p_2) with an alpha of 5% and power of 80%.

```
### alpha = sig.level option and is equal to 0.05
### power = 0.80
### p1 = 0.60
### p2 = 0.50
power1 <-pwr.2p.test(h = ES.h(p1 = 0.60, p2 = 0.50), sig.level = 0.05, power = .80)
power1
```

```
##
##      Difference of proportion power calculation for binomial distribution (arcsine transformat
##
##               h = 0.2013579
##               n = 387.1677
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: same sample sizes
```

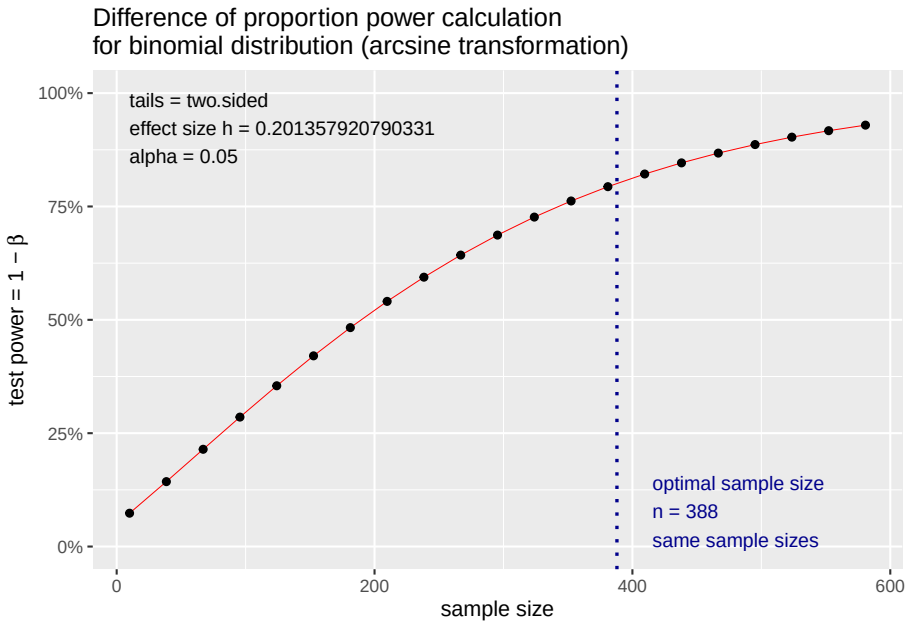
The effect size `h` is 0.201, and the sample `n` is 387.2 or 388 rounded to the nearest whole number. This `n` is only for one group. Hence, based on the parameters of our study, we need approximately 388 patients in Treatment A and 388 patients in Treatment B to detect a difference of 10% response or greater with an alpha of 0.05 and power of 80%.

15.1.2 Power analysis for two proportions

Power (1 - β) is the ability to correctly reject the null hypothesis if the actual effect of the population is equal to or greater than the specified effect size. In other words, if you conclude that there is no difference in the sample, then there is no true difference in the population conditioned on the specified effect size

We can also use the plot feature to see how the power level changes with varying sample sizes. As the sample size goes up, power increases. As the sample size goes down, power decreases. This is important to understand. As we increase our sample size, we reduce the uncertainty around the estimates. By reducing this uncertainty, we gain greater precision in our estimates, which results in greater confidence in our ability to avoid making a type II error.

```
### We can plot the power relative to different levels of the sample size.
plot(power1)
```

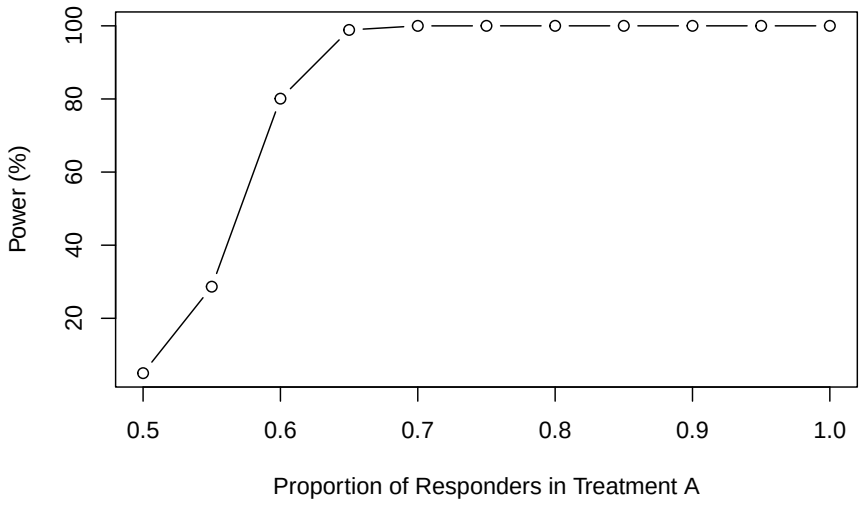


Let's change the p_i and see how the power level changes; we are fixing our sample size at 388 for each group with an alpha of 0.05.

We create a sequence of values by varying the proportion of "responders" for

Treatment A. We will change these from 50% to 100% in intervals of 5%.

```
p1 <- seq(0.5, 1.0, 0.05)
power1 <-pwr.2p.test(h = ES.h(p1 = p1, p2 = 0.50),
                     n = 388,
                     sig.level = 0.05)
powerchange <- data.frame(p1, power = power1$power * 100)
plot(powerchange$p1,
     powerchange$power,
     type = "b",
     xlab = "Proportion of Responders in Treatment A",
     ylab = "Power (%)")
```



We can also write a function for this:

```
iteration <- function(p_i, P_i, i_i, p2, n, alpha) {
              p1 <- seq(p_i, P_i, i_i)
              power1 <-pwr.2p.test(h = ES.h(p1 = p1, p2 = p2),
                                   n = n,
                                   sig.level = alpha)
              powerchange <- data.frame(p1, power = power1$power * 100)
              powerchange
}

iteration(0.5, 1.0, 0.05, 0.50, 388, 0.05)
```

```
##      p1     power
```

```
## 1  0.50   5.00000
## 2  0.55  28.65038
## 3  0.60  80.08415
## 4  0.65  98.88117
## 5  0.70  99.99190
## 6  0.75 100.00000
## 7  0.80 100.00000
## 8  0.85 100.00000
## 9  0.90 100.00000
## 10 0.95 100.00000
## 11 1.00 100.00000
```

15.2 Two-sample Mean Problems

15.2.1 Sample size estimations for two averages

This is another common sample size estimation for two groups when the outcome is a continuous variable.

Now let's estimate the sample size for a study where we are comparing the averages between two groups. Let's suppose that we are working on a randomized controlled trial that seeks to evaluate the difference in the average change in hemoglobin A1c (HbA1c) from baseline between Treatment A and Treatment B.

The null hypothesis is:

H_0: There is no difference in the average change in HbA1c from baseline between Treatment A and Treatment B

The alternative hypothesis is:

H_a: There is a difference in the average change in HbA1c from baseline between Treatment A and Treatment B

To determine the sample size needed to detect a meaningful difference in the average HbA1c change from baseline between Treatment A and Treatment B, we can use the following formula:

$$n_i = \frac{2\sigma^2 * (Z_{\alpha/2} + Z_{1-\beta})^2}{(\mu_1 - \mu_2)^2}$$

where n_i is the sample size for one of the groups, μ_1 and μ_2 are the average changes in HbA1c from baseline for Treatment A and Treatment B, respectively.

As usual, we'll set the two-tailed alpha level to 0.05. The effect size, also known as Cohen's d, is estimated using the following equation:

$$d = \frac{\mu_1 - \mu_2}{\sigma_{pooled}}$$

The pooled standard deviation is estimated using the following formula:

$$\sigma_{pooled} = \sqrt{\frac{sd_1^2 + sd_2^2}{2}}$$

Once again, the R `pwr` package can make this task easy for us. However, we'll need to estimate the Cohen's d.

Example 2: Since we haven't started the study, we have to make some assumptions about each treatment strategy's change in HbA1c. Let's assume that the expected average change in HbA1c from baseline for Treatment A was 1.5% with a standard deviation of 0.25%. Additionally, let's assume that the expected average change in HbA1c from baseline for Treatment B was 1.0% with a standard deviation of 0.20.

First, we'll calculate the pooled standard deviation (σ_{pooled}):

```
sd1 = 0.25
sd2 = 0.30
sd.pooled = sqrt((sd1^2 +sd2^2) / 2)
sd.pooled
```

```
## [1] 0.276134
```

Once we have the σ_{pooled}, we can estimate the Cohen's d:

```
mu1 = 1.5
mu2 = 1.0
d = (mu1 - mu2) / sd.pooled
d
```

```
## [1] 1.810715
```

Cohen's d is 1.81, which is considered a large effect size.

Now, we can take advantage of the `pwr` package and estimate the sample size needed to detect a difference of 0.5% (1.5% - 1.0%) in the average HbA1c change from baseline between Treatment A and Treatment B with 80% power and a significance threshold of 5%.

```
### d = Cohen's d
### power = 0.80
### alpha = 0.05
```

```
n.i <- pwr.t.test(d = d, power = 0.80, sig.level = 0.05)
n.i
```

```
##
##      Two-sample t test power calculation
##
##               n = 5.921286
##               d = 1.810715
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Based on our study parameters, we need 6 patients in each group to detect a difference of 0.5% or greater with 80% power and a significance threshold of 5%.

15.2.2 Power analysis for two averages

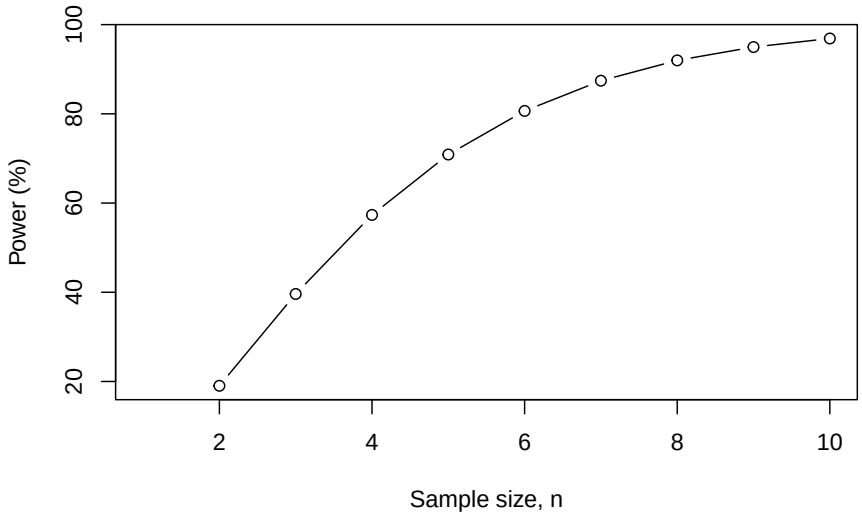
Power (1 - β) is the ability to correctly reject the null hypothesis if the actual effect of the population is equal to or greater than the specified effect size.

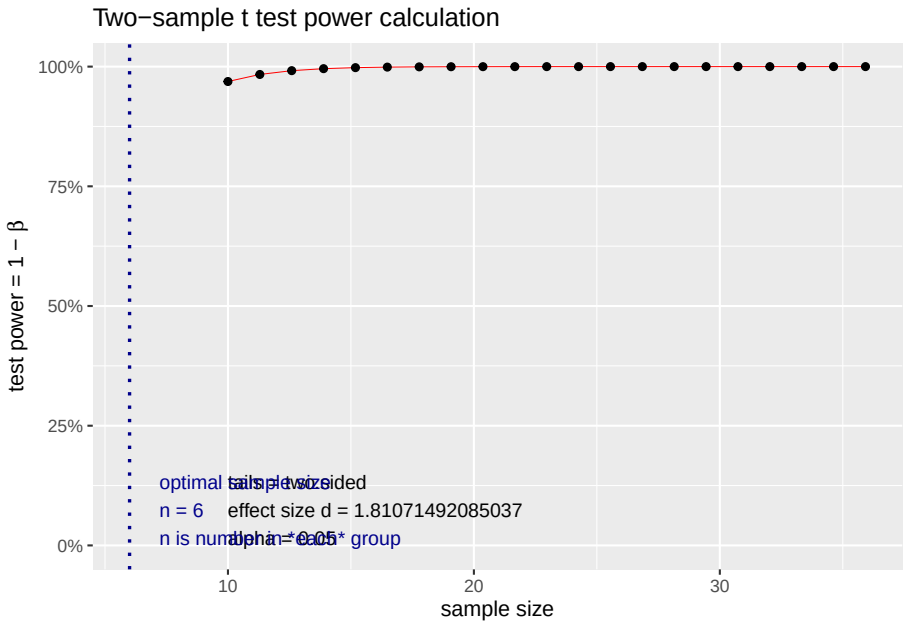
We can plot how the power will change as the sample size changes. As the sample size increases, power increases. This should make sense. Like our previous example with two proportions, as we increase our sample size, we reduce the uncertainty around the estimates. By reducing this uncertainty, we gain greater precision in our estimates, which results in greater confidence in our ability to avoid making a type II error.

```
### We can plot the power relative to different levels of the sample size.
n <- seq(1, 10, 1)
nchange <- pwr.t.test(d = d, n = n, sig.level = 0.05)
nchange.df <- data.frame(n, power = nchange$power * 100)
nchange.df
```

```
##     n    power
## 1   1      NaN
## 2   2 19.03307
## 3   3 39.61785
## 4   4 57.33850
## 5   5 70.87945
## 6   6 80.64997
## 7   7 87.42531
## 8   8 91.98145
## 9   9 94.96979
## 10 10 96.88938
```

```
plot(nchange.df$n,
     nchange.df$power,
     type = "b",
     xlab = "Sample size, n",
     ylab = "Power (%)")
```





As the sample size increases, our power increases. This makes sense because we have more patients to detect differences that may be smaller. But we fixed our effect size (Cohen's d), so as we increase the sample size, our power to detect that difference ultimately increases.

Let's change the μ_i and see how the power level changes; we are fixing our sample size at 6 for each group with an alpha of 0.05.

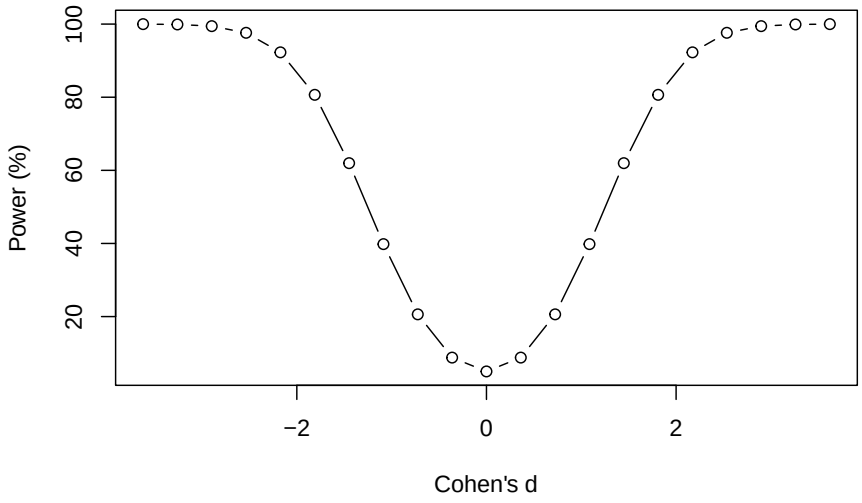
We create a sequence of values by varying the average change in HbA1c from the baseline for Treatment A. We will change these from 0% to 2% in intervals of 0.1%.

```
mu1 <- seq(0.0, 2.0, 0.1)

d <- (mu1 - mu2) / sd.pooled

power1 <- pwr.t.test(d = d, n = 6, sig.level = 0.05)
powerchange <- data.frame(d, power = power1$power * 100)
powerchange
```

```
##            d     power
## 1  -3.621430 99.986261
## 2  -3.259287 99.898978
## 3  -2.897144 99.437136
## 4  -2.535001 97.615372
## 5  -2.172858 92.271971
## 6  -1.810715 80.649971
```

```
## 7  -1.448572 61.960800
## 8  -1.086429 39.817661
## 9  -0.724286 20.610368
## 10 -0.362143  8.785855
## 11  0.000000  5.000000
## 12  0.362143  8.785855
## 13  0.724286 20.610368
## 14  1.086429 39.817661
## 15  1.448572 61.960800
## 16  1.810715 80.649971
## 17  2.172858 92.271971
## 18  2.535001 97.615372
## 19  2.897144 99.437136
## 20  3.259287 99.898978
## 21  3.621430 99.986261
```

```
plot(powerchange$d,
     powerchange$power,
     type = "b",
     xlab = "Cohen's d",
     ylab = "Power (%)")
```



This figure shows how the power changes with Cohen's d. It has a symmetrical pattern because of the negative and positive range associated with Cohen's d. But the story is the same. As the effect size increases (negative and positive signs do not matter; we only care about the absolute values), the power increases.

This makes sense because we only have enough power to detect large differences with the current sample size (which is fixed in this case). If the differences are small, then we do not have enough power with the current sample size of 6.

15.3 Paired-sample Problems

15.3.1 Sample size estimation for paired data (before and after)

So far, we discussed how to perform sample size estimations for "between-groups" comparisons. However, many studies investigate the "within-group" changes. These are paired data, which means that the observation for one data point is dependent on another observation. A common study design where paired data is collected is longitudinal studies. These types of studies involve repeated measures. For example, a pretest-posttest study design will measure a data point for a patient at baseline and then repeat that measurement at another point in time. In the figure, Patient A has two measurements at t_0 and t_f. Since these measurements were made in the same person, the change is "within" the patient. Alternatively, we can think of this as a "repeated" measure since the patient had the measurement performed twice.

Figure 15.1: Figure caption: Pretest-Posttest (repeated measures) framework

When you have a study where you are performing a paired t-test, you can use the same `pwr.t.test()` function for a two-sample test from the `pwr` package.

Let's assume that we want to conduct a prospective study to measure the weight change of a cohort of patients who started a diet. You want to enroll enough patients to detect a 5 lb reduction in weight 3 weeks after the diet started. Let's assume that at baseline the expected average weight for the cohort was 130 lbs with a standard deviation of 11. After 3 weeks of diet, the expected average weight was 125 lbs with a standard deviation of 12.

We can estimate the effect size (Cohen's d_z) for a paired t-test.

$$d_z = \frac{|\mu_z|}{\sigma_z} = \frac{|\mu_x - \mu_y|}{\sqrt{\sigma_x^2 + \sigma_y^2 - 2\rho_{x,y}\sigma_x\sigma_y}}$$

where x denotes "before" (or baseline), y denotes "after", d_z denotes Cohen's d for paired analysis ρ denotes the correlation between the measures before and after the diet. (For simplicity, I use 0.50 if I don't have prior information about this correlation.)

```
### Parameters for paired analysis or a pretest-posttest study design

mu_x <- 130     ### Average weight before the diet (baseline)
mu_y <- 125     ### Average weight after the diet

sd_x <- 11      ### Standard deviation before the diet
sd_y <- 12      ### Standard deviation after the diet

rho <- 0.5      ### Correlation between measures before and after the diet

sd_z <- sqrt(sd_x^2 + sd_y^2 - 2*rho*sd_x*sd_y)

d_z <- abs(mu_x - mu_y) / sd_z
d_z
```

```
## [1] 0.433555
```

The Cohen's d_z is 0.433. We can input this into the `pwr.t.test()` function.

```
n.paired <- pwr.t.test(d = d_z, power = 0.80, sig.level = 0.05, type = "paired")
n.paired
```

```
##
##      Paired t test power calculation
##
##               n = 43.71557
##               d = 0.433555
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number of *pairs*
```

We need 44 patients with two measurements (before and after) they implement their diet to detect a difference of 5 lbs or greater with 80% power and a significance level of 0.05.

15.3.2 Power analysis of paired samples (paired t-test)

We can plot how the power will change as the sample size changes for the paired t-test analysis. As the sample size increases, power increases. This should make sense. Like our previous examples, as we increase our sample size, we reduce the uncertainty around the estimates. By reducing this uncertainty, we gain greater precision in our estimates, which results in greater confidence in our ability to avoid making a type II error.

```
### We can plot the power relative to different levels of the sample size for paired a
n_z <- seq(1, 80, 5)
n_z.change <- pwr.t.test(d = d_z, n = n_z, sig.level = 0.05, type = "paired")
```

```
## Warning in qt(sig.level/tside, nu, lower = FALSE):   NaNs
```

```
n_z.change.df <- data.frame(n_z, power = n_z.change$power * 100)
n_z.change.df
```

```
##    n_z    power
## 1    1      NaN
## 2    6 14.03624
## 3   11 25.58334
## 4   16 36.84309
## 5   21 47.26307
## 6   26 56.56985
## 7   31 64.66154
## 8   36 71.54769
## 9   41 77.30572
## 10  46 82.04980
## 11  51 85.90929
## 12  56 89.01478
## 13  61 91.48950
## 14  66 93.44465
## 15  71 94.97744
## 16  76 96.17076
```

```
plot(n_z.change.df$n,
     n_z.change.df$power,
     type = "b",
     xlab = "Sample size, n",
     ylab = "Power (%)")
```

As the sample size increases, we generate more power to detect a difference of 5 lbs with a significance level of 0.05 and a fixed sample size of 44 patients with two measurements (before and after) they implement their diet.

Let's see how power changes when we change the effect size. Let's change the average weight after the patients implement their diet. Instead of an average of 125 lbs, let's see how the power will change when we reduce that to 100 lbs.

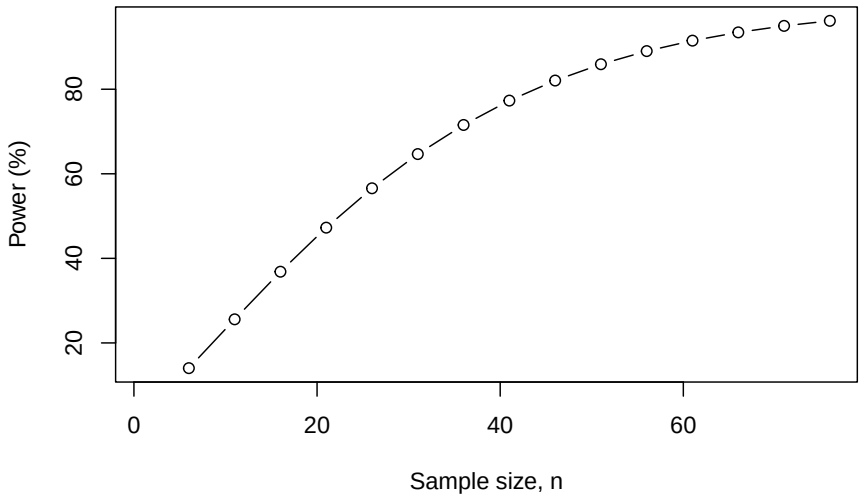


Figure 15.2: We increase the sample size from 1 to 80 at 5-unit intervals.

```
### Vary the mu_y from 50 lbs to 130 lbs in intervals of 5 lbs.
mu_y <- seq(50, 130, 5)

d_z <- abs(mu_x - mu_y) / sd_z

n_z.change <- pwr.t.test(d = d_z, n = 44, sig.level = 0.05)
n_z.change.df <- data.frame(d_z, power = n_z.change$power * 100)
n_z.change.df
```

```
##         d_z     power
## 1  6.936880 100.00000
## 2  6.503325 100.00000
## 3  6.069770 100.00000
## 4  5.636215 100.00000
## 5  5.202660 100.00000
## 6  4.769105 100.00000
## 7  4.335550 100.00000
## 8  3.901995 100.00000
## 9  3.468440 100.00000
## 10 3.034885 100.00000
## 11 2.601330 100.00000
## 12 2.167775 100.00000
```

```
## 13 1.734220 100.00000
## 14 1.300665  99.99766
## 15 0.867110  98.03639
## 16 0.433555  52.03146
## 17 0.000000   5.00000
```

```
plot(n_z.change.df$d_z,
     n_z.change.df$power,
     type = "b",
     xlab = "Cohen's d_z",
     ylab = "Power (%)",
     xlim = c(0, 2))
```

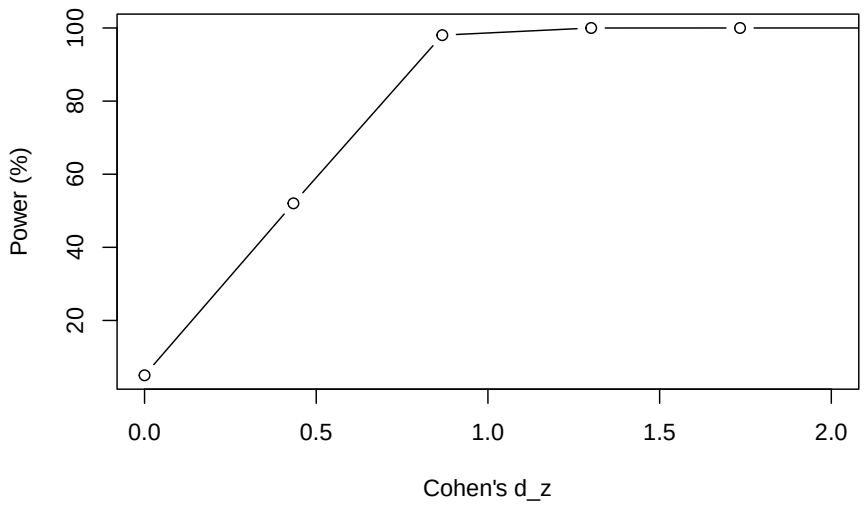


Figure 15.3: We vary the average weight μ_y between 50 lbs and 130 lbs in intervals of 5 lbs.

When we increase the effect size (Cohen's d_z), our power goes up; recall that the sample size is fixed at 44 and the significance level is 0.05. But when the effect size gets smaller (or when the average weight loss shrinks), we lose the power to detect a difference because our sample size is too small. We'll need to increase our sample size to have a reasonable power to detect small differences.

We can also see how power changes when we vary ρ. If we set $\rho = 0$, then the Cohen's $d_z = 0.307$. If we set $\rho = 1$, then the Cohen's $d_z = 5$.

```
mu_x <- 130     ### Average weight before the diet (baseline)
mu_y <- 125     ### Average weight after the diet

sd_x <- 11      ### Standard deviation before the diet
sd_y <- 12      ### Standard deviation after the diet

sd_z_1 <- sqrt(sd_x^2 + sd_y^2 - 2*1*sd_x*sd_y)
sd_z_0 <- sqrt(sd_x^2 + sd_y^2 - 2*0*sd_x*sd_y)

d_z_1 <- abs(mu_x - mu_y) / sd_z_1
d_z_0 <- abs(mu_x - mu_y) / sd_z_0

d_z_1
```

```
## [1] 5
```

```
d_z_0
```

```
## [1] 0.3071476
```

So, higher ρ results in large d_z and smaller ρ results in small d_z values.

Let's see how power changes when we change the ρ range from 0 to 1 in intervals of 0.1 units.

```
rho <- seq(0.0, 1.0, 0.1)

sd_z <- sqrt(sd_x^2 + sd_y^2 - 2*rho*sd_x*sd_y)

d_z <- abs(mu_x - mu_y) / sd_z

rho.change <- pwr.t.test(d = d_z, n = 44, sig.level = 0.05)
rho.change.df <- data.frame(d_z, power = rho.change$power * 100)
rho.change.df
```

```
##          d_z     power
## 1  0.3071476  29.65480
## 2  0.3236941  32.35129
## 3  0.3432395  35.66281
## 4  0.3668151  39.80733
## 5  0.3960280  45.10546
## 6  0.4335550  52.03146
## 7  0.4842743  61.25990
## 8  0.5583195  73.54735
## 9  0.6816774  88.52181
## 10 0.9552009  99.32401
## 11 5.0000000 100.00000
```

```
plot(rho.change.df$d_z,
     rho.change.df$power,
     type = "b",
     xlab = "Cohen's d_z",
     ylab = "Power (%)",
     xlim = c(0, 1.5))
```

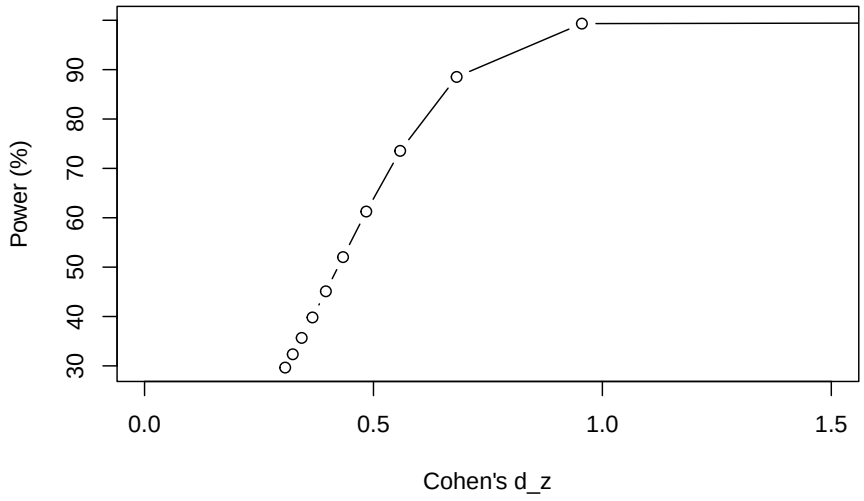


Figure 15.4: We vary ρ from 0 to 1 at intervals of 0.1 unit.

As ρ increases, our power increases. This makes sense because we are nearing "perfect" correlation, which would require less sample to detect a difference if one existed. As the correlation becomes less "perfect" our power drops suggesting that we need to increase our sample size to make up for this poor correlation.

15.4 Unequal Sample Sizes Problems

15.4.1 Power analysis with unequal sample sizes

It is common for the sample size to be different. The `pwr.t2n.test()` is a useful tool to help estimate the power given the sample sizes of the study.

It is common to perform power analysis on a study where the sample sizes between groups are different.

Suppose you have a retrospective study where the patients were prescribed

Treatment A and Treatment B. There were 130 patients in Treatment A ($n_A = 130$) and 120 patients in Treatment B ($n_B = 120$). The average change in HbA1c was 1.5% with a standard deviation of 1.25% in Treatment A, and the average change in HbA1c was 1.4% with a standard deviation of 1.01% in Treatment B.

First, we'll calculate the pooled standard deviation (σ_{pooled}):

```
sd1 <- 1.25
sd2 <- 1.01
sd_pooled <- sqrt((sd1^2 +sd2^2) / 2)
sd_pooled
```

```
## [1] 1.136354
```

Once we have the σ_{pooled}, we can estimate the Cohen's d:

```
mu1 <- 1.5
mu2 <- 1.4
d <- (mu1 - mu2) / sd_pooled
d
```

```
## [1] 0.08800076
```

Now, we can estimate the power with the different sample sizes across the groups ($n_A = 130$, $n_B = 120$).

```
n1 <- 130
n2 <- 120

power.diff_n <- pwr.t2n.test(d = d, n1 = n1, n2 = n2, sig.level = 0.05)
power.diff_n
```

```
##
##      t test power calculation
##
##              n1 = 130
##              n2 = 120
##               d = 0.08800076
##       sig.level = 0.05
##           power = 0.1064836
##     alternative = two.sided
```

Since the average HbA1c change from baseline for Treatment A is 1.5% and 1.4% for Treatment B, the average difference in the HbA1c change from baseline is 0.1%. This is a difference (difference between the groups) of the differences (difference from baseline within the group) calculation.

You only have 11% power to detect a difference of 0.10% or greater in the HbA1c change from baseline. This means that you are underpowered to detect a difference of 0.10% or greater in the HbA1c change from baseline with $n_A = 130$, $n_B = 120$, and a significance level of 0.05. When studies are underpowered,

there is a high potential for type II error. The only way to address this problem is to enroll more patients or expand the sample by relaxing inclusion criteria. However, this may increase the threats to the study's internal validity.

15.5 Conclusions

Sample size estimations and power analysis are very useful tools to determine how many patients you need in your study and how confident you are that you didn't make a type II error. Depending on the type of study, you will need to use different functions from the `pwr` package. I highly encourage you to explore the other functions of the `pwr` package to see if those fit the study design you have planned.