BIOL 3316
Biometry Laboratory

Ken Aho
Idaho State University
Department of Biological Sciences
Idaho State University

April 19, 2024

```
niter = 150000
fern <- sample(1:4, niter, repl = TRUE, prob = c(.01, .07, .07, .85))
x <- numeric(niter+1); y <- numeric(niter + 1); x[1] <- 0; y[1] <- 0
for(i in 1:niter){x[i + 1] <- 0 * (fern[i]==1) + (0.2 * x[i] -
0.26 * y[i]) * (fern[i]==2) + (-0.15 * x[i] + 0.28 * y[i]) *
(fern[i]==3) + (0.85 * x[i] + 0.04 * y[i]) * (fern[i]==4)
y[i + 1] <- 0.16 * y[i] * (fern[i]==1) + (0.23 * x[i] +
0.22 * y[i] + 1.6) * (fern[i]==2) + (0.26 * x[i] + 0.24
* y[i] + .44) * (fern[i]==3) + (-0.04 * x[i] + 0.85
* y[i] + 1.6) * (fern[i]==4)}
plot(x, y, pch = 17, cex = .3,
axes = F, ann = F)
```

Contents

1

Introduction to Statistics

Lab 1 Topics

1. The purpose and importance of statistics
2. Variables and associated terms
3. Statistical inference and its limitations
    - Sampling design
    - Experimental design
4. Statistical software

The Purpose of Statistics

Unfortunately, many people (including, perhaps you) are apprehensive about statistics. As a consequence you may be conditioned to view the topic as irrelevant, tedious, and/or confusing. This is regrettable. The discipline of statistics plays a vital role in the **empirical sciences** (in which knowledge is gained by experience or observation), including biology. Statistics is important for at least four reasons.

1. **The field of statistics objectifies information and decreases bias**. When we make **inferences** we draw conclusions from available data. The human capacity to make inferences is impressive but imperfect. We are often fooled by our senses, and are preconditioned to arrive at particular decisions. Stare at the line intersections in Fig 1.1. Note that although the dots are actually white, it is difficult to determine whether they are black, white, or gray.

   As a physiological example, the perception of danger will often trigger a fight or flight response in animals. This will cause the autonomic nervous system to signal the body

to release adrenaline, increase heart rate, hyperventilate, and otherwise prepare for fighting or fleeing. Recent evidence suggests that the human fight or flight response may error on the side of caution (Rakison, 2009). That is, we may be hard-wired to see the world in a slightly paranoid way.

Science attempts to study and describe phenomena **objectively** (in a way that does not depend on the investigator). **Bias** is the tendency or inclination to choose certain answers and interpretations at the expense of other equal or more valid answers (Aho, 2014). Mathematical tools like statistics can help us to address our tendencies for non-objectivity that can lead to errors and biases.

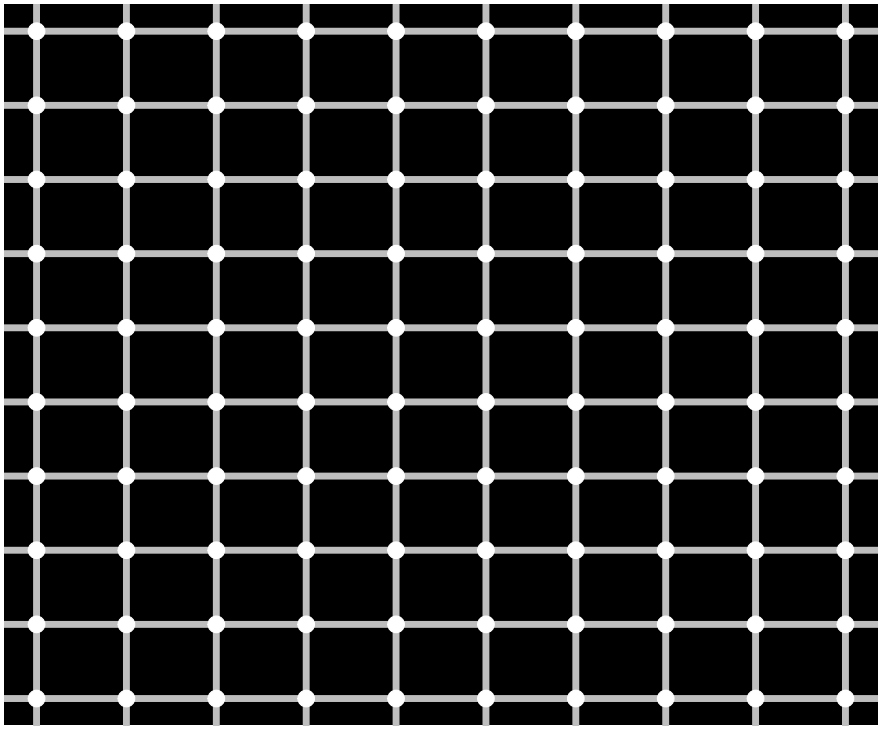


Figure 1.1. Are the dots black or white? Figure based on code from Yihui Xie's **R** package *animation* (Xie *et al.*, 2018).

2. **The field of statistics complements biology**

In biology, phenomena under study (e.g., deer size, metabolic rates for an enzyme, paramecium density, etc.) will often vary randomly from one observation to the next. Statistics allow us to measure this variability and to make valid inferences *despite* this variability.

Furthermore, in biological studies it is generally impossible to record all possible observations. As a result we must rely on statistical methods to make inference to true processes, using necessarily incomplete information.

Finally, the field of biology uses the scientific method, which requires that we propose

and test hypotheses. Statistical methods can be used to quantify the amount support that data provide for and against scientific hypotheses.

3. **Statistical knowledge allows you to distinguish information from misinformation**

   In today's world it is increasingly necessary to be able to filter misinformation. Statistics facilitates this process, allowing you to objectively consider claims. For instance,

   - When a nutrition supplier claims that its product causes significant weight loss, what does this mean, and are the claims valid?
   - When a climate scientist claims that the correlation between atmospheric CO_2 concentrations and global temperature is greater than 0.95, what does this mean, and are these claims valid?

4. **Statistics will make you rich and famous!**

   Joking aside, more than half of the jobs currently advertised on the American Statistical Association jobs website are for biostatisticians. Furthermore, in 2022 the Bureau of Labor Statistics reported the annual median pay for a statistician was $99,960 per year, and anticipated 30% growth in demand for statisticians from 2022-2032.

Variables and Experimental Units

A **variable** is simply a measurable phenomenon that varies. Statisticians deal with an important type of variable called a **random variable** whose outcomes cannot be known in advance, and whose propensities must be modeled probabilistically. Consider a fair die throw (Fig 1.2). Preceding a throw we will not know what the outcome will be: 1, 2, 3, 4, 5 or 6. Nonetheless, we assume that the probability of rolling a one is 1/6. We will explore probability in Lab 2.

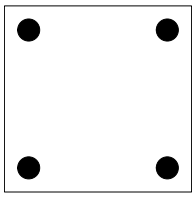


Figure 1.2. Die throw animation. To run, make sure this document is open in Adobe Reader and click on die image.

Data and Experimental Units

Scientific **data** consist of outcomes (e.g. measurements, observations) from variables with respect to **experimental units** (**EUs**) or **sampling units**. Experimental units are often called subjects if the EUs are human. It can often be surprisingly difficult to define what the EUs in a study actual are.

Quantitative, Categorical and Ordinal Variables

Variables can be classified as quantitative, categorical, or ordinal depending on the characteristics of their data. With **categorical variables** the magnitude of the data has no quantitative meaning. For instance, a record of whether a deer observed at feeding area is male or female could be recorded as an"M" (indicating male) or an "F" (indicating female) or as 1 and 0. Outcomes from both approaches are numerically meaningless, despite the fact that the second approach distinguishes categories using 1 and 0. **Ordinal variables** have data with *some* quantitative meaning, although it is imprecise. For instance, suppose a scientist records a qualitative soil water index (from 1 to 10 indicating dry to wet) of sites she is studying. While this gives us a relative measure of water in the soil (10 is wetter than 1), we don't have an exact idea of the meaning of outcomes with respect to each other (an index score of 10 probably does not indicate that the soil is 10 times wetter than a soil with a score of 1). Outcomes from **quantitative variables** have a precise quantitative meaning. For instance, if I record temperatures of 10° and 20° C, I know that the second record is exactly twice as warm as the first with respect to the baseline 0° C.

Discrete and Continuous Quantitative Variables

Two types of quantitative variables can be distinguished. **Discrete quantitative variables** have outcomes that are discontinuous. An example would be counts of mountain goats at a particular location and time. An outcome from this variable will be a natural integer: 0, 1, 2, etc. Conversely, **continuous quantitative variables** can be conceptualized as having no breaks. As a result the ability to distinguish outcomes depends on the resolution of the measuring device used to gather the data. An example of a continuous variable would be crop yield measured in kilograms from an agricultural experiment. Within the range of its support this variable does not have any breaks. That is, any interval bounded by two distinct outcomes would theoretically contain an infinite number of other distinct outcomes.

Explanatory and Response Variables

Scientists often want to make logical connections between cause and effect; i.e., **causality**. In this context we can distinguish two types of variables. The first type is called the **explanatory variable** (also called the X-variable, predictor variable, and independent variable). This variable is generally hypothesized to exert some influence on (cause) a second type of variable, called the **response variable** (also called the Y-variable and dependent variable). Biological examples include:

- Parental genotype (X) → Offspring genotype (Y).
- Temperature (X) → Enzymatic activity (Y).
- Mutagenic agent (X) → Number of mutations. (Y)

If two variables are **correlated**, then certain outcomes of one variable tend to occur with certain outcomes of the other variable. For instance, one variable will increase or decrease as

the other variable increases or decreases. Importantly, if two variables are correlated they may (or may not) be causally associated. That is,

$$\text{Correlation} \neq \text{Causation}$$

This idiom holds even if we are using terms like response variable and explanatory variable for the two variables being measured (Fig 1.3).

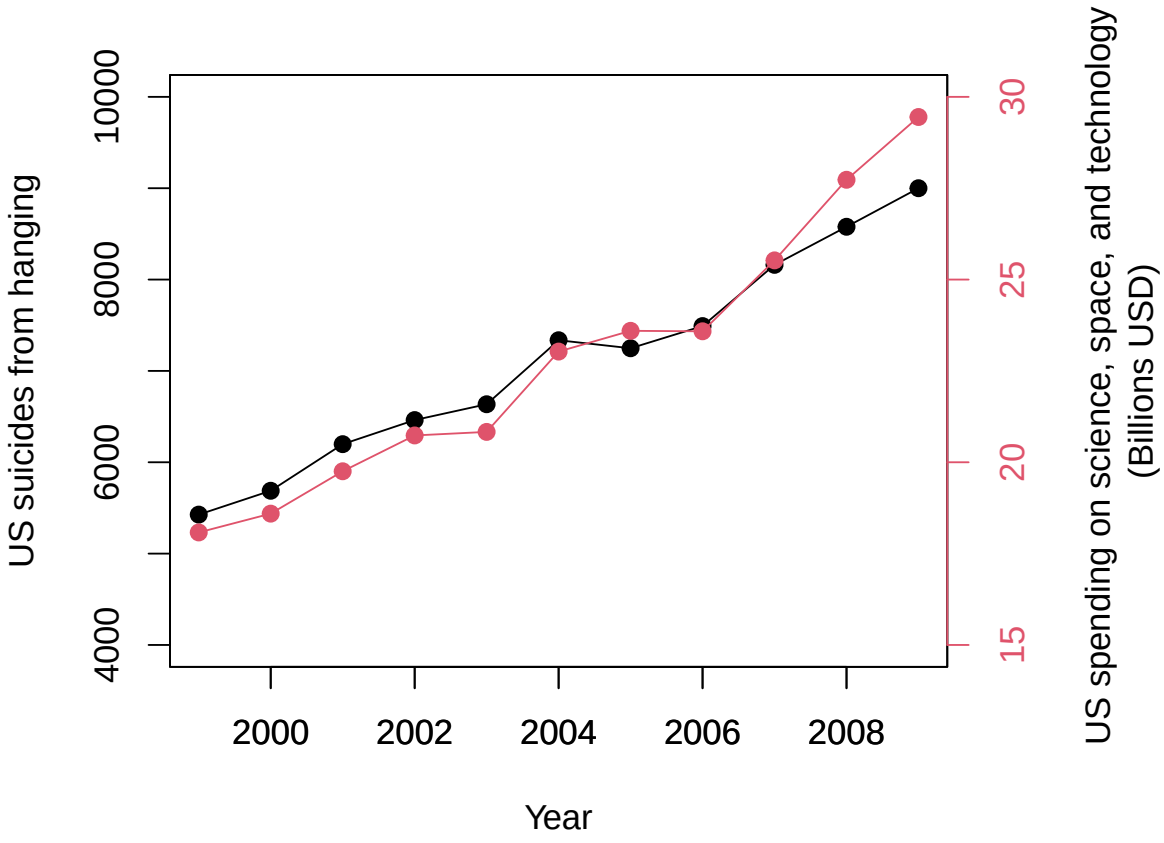


Figure 1.3. Association of US suicides by hanging and US spending on science, space, and technology from 1999-2009. While these variables are correlated, clearly one is not casuing the other to occur. Data: US Office of Management and Budget, US Centers for Disease Control and Prevention.

Sampling and Experimental Design

Recall that with inference we draw conclusions from available data. Statistics is generally concerned with two sorts of inferences: **inference to the population** and **causal inference**.

Inference to the Population

To a statistician, a **population** is a collection of all possible outcomes from a random variable. A biological example could be all possible weights from individual Dall's sheep (*Ovis dallii*) in the Wrangell-St. Elias Wilderness in Alaska at one point in time[1]. We cannot obtain weight measures of all the sheep: there are too many and they are difficult to capture and weigh in their native habitat. We can, however, attempt to make inference to the population of weights.

Causal Inference

With causal inference we are making inference of cause and effect. For instance, that an explanatory variable, X, is affecting a response variable, Y, in some way.

Sampling Design

Sampling design refers to way experimental units (for instance, individual Dall's sheep which will be tranquilized, and then weighed) are selected from a population. Sampling is necessary because of the general impossibility of observing all possible outcomes from a variable of interest. For instance, a soil scientist cannot possibly measure soil characteristics for the entire surface of a mountain range. Instead the researcher selects experimental units to represent a larger clearly defined population that he or she wishes to make inference to.

Experimental Design

In an **experimental design** we are concerned with how experimental treatments (for instance, high and low calorie diets) are assigned to EUs or vice versa.

Inferences Resulting from Sampling and Experimental Designs

Inference to the population and causal inference are constrained by the sampling design and experimental design of a study, respectively. Of great importance is whether an investigator incorporates randomization into his/her sampling and experimental designs. **Randomization** is "a selection or allocation process that produces outcomes that cannot be known in advance by the investigator" (Aho, 2014). Random outcomes can be obtained from random number generators or some other non-deterministic process e.g., die throws (see Fig. 1.2), coin tosses, etc.

Sampling Design → Inference to the Population

Inference to the population will depend on the sampling design. Inference to the population is done most effectively by acquiring data from the population using random sampling because

[1]Note that the statistical definition of a population differs from a conventional biological definition of a population: a group of interacting individuals from the same species.

this helps eliminate possible investigator bias. In a **Simple random sample** the investigator mixes up (randomizes) the population before selecting EUs. Consider Fig 1.4. Assume that individual highlighted squares in the grid are 20 EUs that are randomly sampled from a population of 400 (all squares). To get a new random sample, click on the figure.

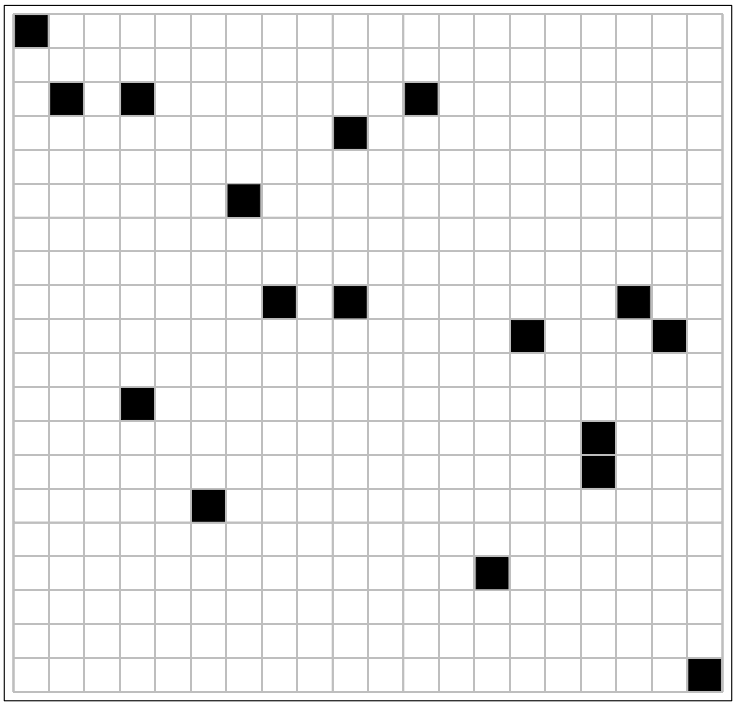


Figure 1.4. Simple random sampling animation. To run the animation, make sure this document is open in Adobe Reader and click on the image.

Replication is another an important concern in sampling designs. Clearly one must obtain a sufficient number of independent observations to adequately describe a population. **Independent observations** will not be affected by outcomes from other observations. Most statistical analyses are based on the assumption that sample observations are independent.

The answer to "how many samples is enough?" will depend on the distributional characteristics of the population. If a population is highly variable, then larger sample sizes will be needed for reliable description. Generally, a sample size of around 30 is considered large.

Experimental Design → Causal Inference

Causal inferences will depend on the experimental design. We call a scientific investigation in which a researcher randomly assigns treatments to experimental units a **randomized experiment**. Consider a situation where you have ten EUs and you wish to randomly assign these to two treatments. We make sure that the design is **balanced** by randomly assigning exactly five EUs to each treatment (Fig 1.5).

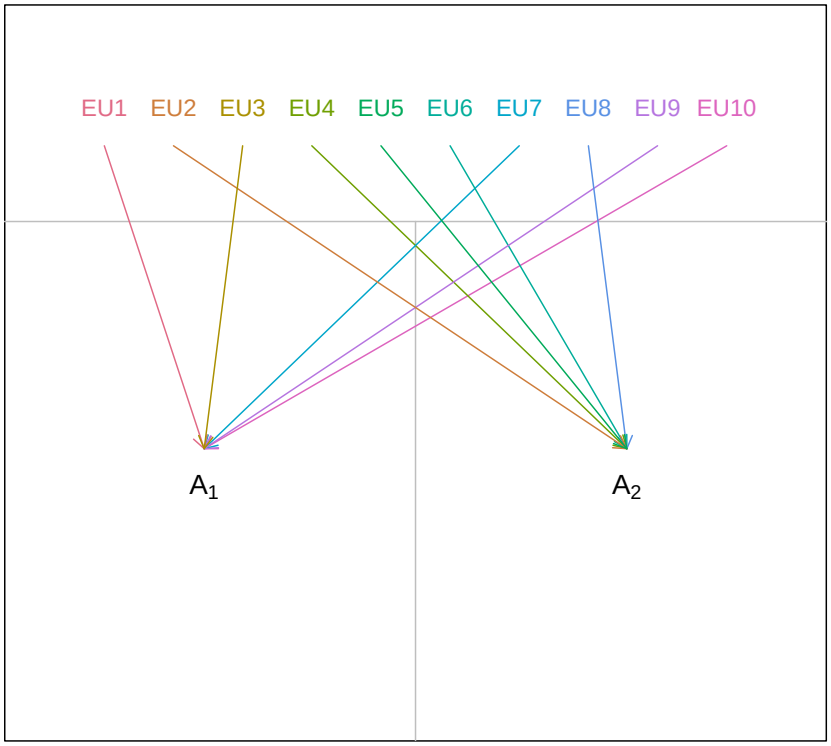


Figure 1.5. **Completely randomized design** (**CRD**) animation. A CRD is the simplest randomized experimental design because constraints like blocking and nesting are not considered. To run the animation, make sure this document is open in Adobe Reader and click on the image.

Randomization in experimental studies allows one to make causal inference with respect to the effect of a treatment because it helps to control **confounding variables** (those variables that may impede interpretation of the effect of the explanatory variable). Controlling confounding variables is easier in a lab setting where one can explicitly hold all other variables (temperature, humidity, etc.) constant while varying levels in the explanatory variable.

In an **observational study**, treatments are not randomly assigned to experimental units although the experimental units may be acquired (sampled) randomly. Many ecological field studies are observational because of the sheer impossibility of randomly assigning treatments. For instance, transporting populations of mountain goats to randomly selected mountains to ascertain mountain goat effects is unrealistic for most researchers. Causal inference are generally prevented in observational studies.

Constraints to inference due to experimental and sampling design are summarized in Table 1.1.

Table 1.1. Summary of statistical inferences permitted by sampling and experimental designs.

	Randomized Experiment	Observational Study
Random sample	Causal inference to population	Inference to population
Nonrandom sample	Causal inference to sample	Inference to sample

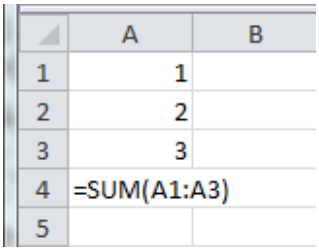
R and Excel

We will use two computer programs throughout the semester: Excel and **R**. Many of you have already used Excel in other classes (although I will assume initially that you haven't). **R** is much more robust and flexible than Excel, but it is command line based, and will feel alien for a while. **R** can be downloaded for free at http://www.r-project.org/.

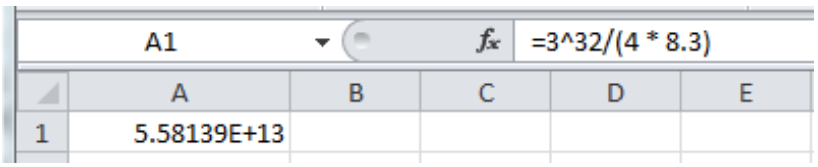
Excel

We can open Excel by clicking on the appropriate icon in your computer task bar.

To write a function in Excel we first type the equals sign "=" in a cell. Function names and arguments (in parentheses, separated by commas) will follow. For instance, I tell Excel to find the sum of numbers in cells A1 to A3 by supplying between the parentheses the one argument required by =SUM: a call to the cell addresses.

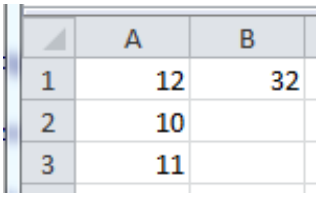


To find $3^{32}/(4 \times 8.3)$ I type in the code:

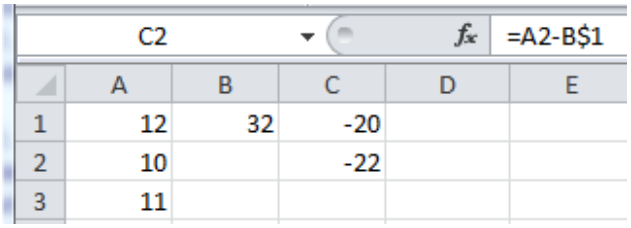


The answer, given in scientific notation, is 5.58E+13. This means 5.58×10^{13}.

What if I wanted to subtract the contents of cell B1 from cells A1, A2, and A3? Then I would type the command =(A1-B$1) into cell C1. I could then copy the contents of C1 by pulling down on the bottom right-hand side of C1 while left-clicking with the mouse. By pulling the contents of C1 into C2, the command becomes A2-B$1 which equals $10-32 = -22$. The cell address with the dollar sign would remain anchored. To anchor an address when pulling right/left (across columns) put the dollar sign in front of the anchor cell address, e.g., $B1. To anchor both up/down and right/left pulls, type dollar signs on both sides of the column name, e.g., B1



	A	B
1	12	32
2	10	
3	11	



C2 | =A2-B$1

	A	B	C	D	E
1	12	32	-20		
2	10		-22		
3	11				

R

The **R** icon is:

Function names for many **R** operations are similar to Excel although they don't require an equals sign. Also like Excel, arguments for **R** functions follow function names, are enclosed in parenthesis, and are separated by commas. To find the sum of 1, 2, and 3, I could simply type `1 + 2 + 3` at the **R command line** prompt and type Enter.

```
1 + 2 + 3

[1] 6
```

The `[1]` in the **R** output means: this is the first requested element. I could also create an **object** that contains these values by using the combine function, `c`, and give it some name, maybe `joe`.

```
joe <- c(1, 2, 3)
```

The <- command is the **assignment operator**, and is supposed to be an arrow. It means that `joe` contains the stuff to the right of the operator. I could take the sum of `joe` by typing:

```
sum(joe)

[1] 6
```

Of course we can do all sorts of other things with `joe`.

```
log(joe)

[1] 0.0000000 0.6931472 1.0986123

joe^2

[1] 1 4 9
```

To find $3^{32}/(4 \times 8.3)$ I type in the code:

```
3^32/(4 * 8.3)

[1] 5.581386e+13
```

How do we subtract 32 from 12, 10, and 11?

```
joe <- c(12, 10, 11)
joe - 32

[1] -20 -22 -21
```

These **R** functions, and several others required in Assignment 1, are summarized in the Appendix to this Lab.

Assignment 1

Answer all questions in one MS Word document and submit to Moodle. At the beginning of the document include the assignment number, the date, your name and section number.

Use complete sentences when appropriate, and make sure any tables, figures and computer output you include adhere to class standards (see Syllabus).

The purpose of statistics

1. (3 pts) Why is the field of statistics important? Don't let anyone tell you differently.

Variables and experimental units

2. (3 pts) Identify which are response and which are explanatory variables in the examples below.

a) Human body weight; calories in diet.

b) Mean parental phenotype; offspring phenotype

c) Presence or absence of aspirin; rate of myocardial infarction

3. (6 pts) In the table below identify the:

a) Experimental units.

b) Quantitative variable(s). For each identified quantitative variable note whether it is discrete or continuous.

c) Categorical variable(s).

d) Ordinal variable(s).

Table 1.2. Table for question 3

Site	Soil %Nitrogen	Soil %Carbon	Species richness	Soil texture	Elevation (m)	Soil water class (0 = dry, 10 = wet)
1	15.2	40.3	26	Silty	3000	7
2	14.2	30	22	Silty	2200	7
3	16.2	27.1	21	Silty	2220	6
4	13.1	24.2	20	Clayey	1900	7
5	10.2	20.4	21	Clayey	1850	5
6	15.5	26.6	20	Clayey	1970	5
7	11.1	30.5	25	Sandy	1400	3
8	14.9	24.1	20	Clayey	1900	5
9	12.3	23	13	Sandy	2000	1
10	10.1	15.1	10	Sandy	2200	2

Sampling and experimental design

4. (3 pts) Distinguish the terms "sampling design" and "experimental design."

5. (6 pts) Are the examples below observational studies or randomized experiments? Explain your answers.

a) A software company wants to compare the effectiveness of computer animation for teaching cell biology versus a textbook presentation of the same information. The company tests the biological knowledge of a group first year college students, and then randomly assigns the students to one of two groups. One group uses the computer animation software while the other learns using a conventional textbook. The company retests all the students using a continuously scaled index and compares the increase in knowledge of cell biology in the two groups.

b) In an 1898 lecture at Woods Hole, Massachusetts, Herman Bumpus, a professor of zoology, presented measurements on house sparrows (*Passer domesticus*) brought to the anatomical laboratory at Brown University after a severe winter storm. Some of the birds survived the storm while others had died. Bumpus measured physical characteristics

(e.g., humerus length) of the survivors, and the mortalities and drew comparisons between survivorship and physical characteristics.

c) A researcher is interested in how the weight of gray wolves (*Canis lupus*) changes with latitude. She determines the average adult weight from ten randomly selected packs of wolves situated from northern Alaska to Southern Canada. She then compares the pack weight to the average pack latitude.

6. (2 pts) In the context of statistics, what does the word "population" mean?

7. (3 pts) An investigator interested in levels of glycogen in Norway rat (*Rattus norvegicus*) livers, gathers data in three different ways. Answer the questions below related to this research.

a) The researcher obtains a single rat liver, samples it at unusually mottled spots in the liver surface, and makes glycogen measures. With this data she can make inference to (choose one):

i) The spots on this single liver.
ii) The entire rat liver.
iii) The livers of all Norway rats.

b) The researcher obtains a single rat liver and samples it at random locations with replication adequate to describe the variability in the liver, and measures glycogen. With this data she can make inference to (choose one):

i) The spots on this single liver.
ii) The entire rat liver.
iii) The livers of all Norway rats.

c) The researcher obtains a random sample of Norway rats with replication adequate to describe the population of Norway rats. She removes the liver of each rat, samples each liver at random locations for glycogen with replication adequate to describe the variability in the livers. With this data she can make inference to (choose one):

i) The spots on this single liver.
ii) The entire rat liver.
iii) The livers of all Norway rats.

8. (4 pts) Define the term inference. What are the two types of inference that statisticians are primarily interested in?

9. (2 pts) How is inference to the population established?

10. (2 pts) Define the term causality.

11. (2 pts) How is causal inference established?

R and Excel

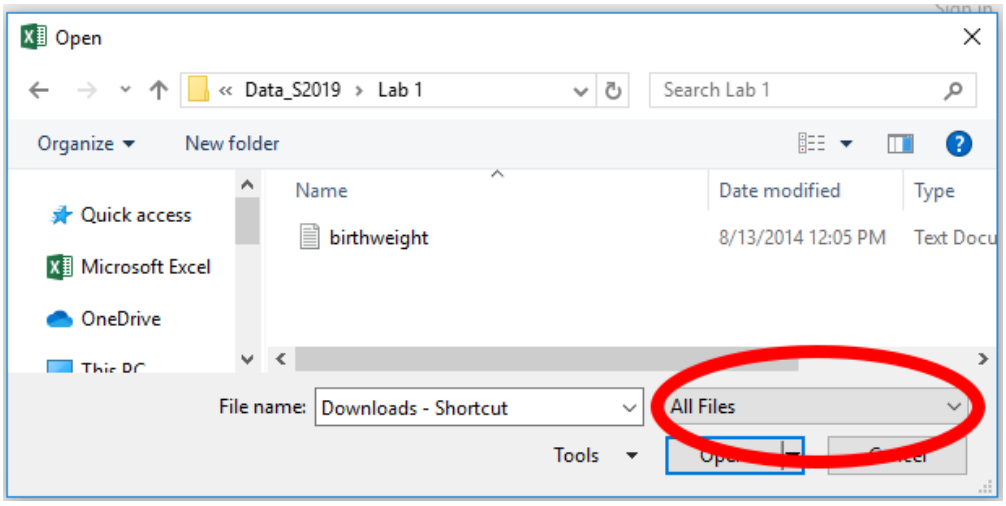
Spreadsheet programs like Microsoft Excel have serious limitations with respect to how many columns or rows of data they can handle. For instance, Excel can only handle 256 columns of data. This may seem like a lot, but for many modern datasets it will be insufficient. For this and other reasons, data are often not saved directly as .xls files, but as text files with character separators (e.g., commas, semicolons, etc.). One of the skills you will have to learn in this course is how to manipulate data and get it in a form you want. This will often involve handling data in non-.xls file formats.

12. For this question you will graphically examine a dataset that describes pregnancies occurring from 1960-1967 for women enrolled in the Kaiser Foundation Health Plan in the San Francisco-East Bay area. Baby birth weights (in ounces) were recorded along with information pertaining to whether the mother smoked during her pregnancy. The dataset is in a text file called `birthweight.txt`.

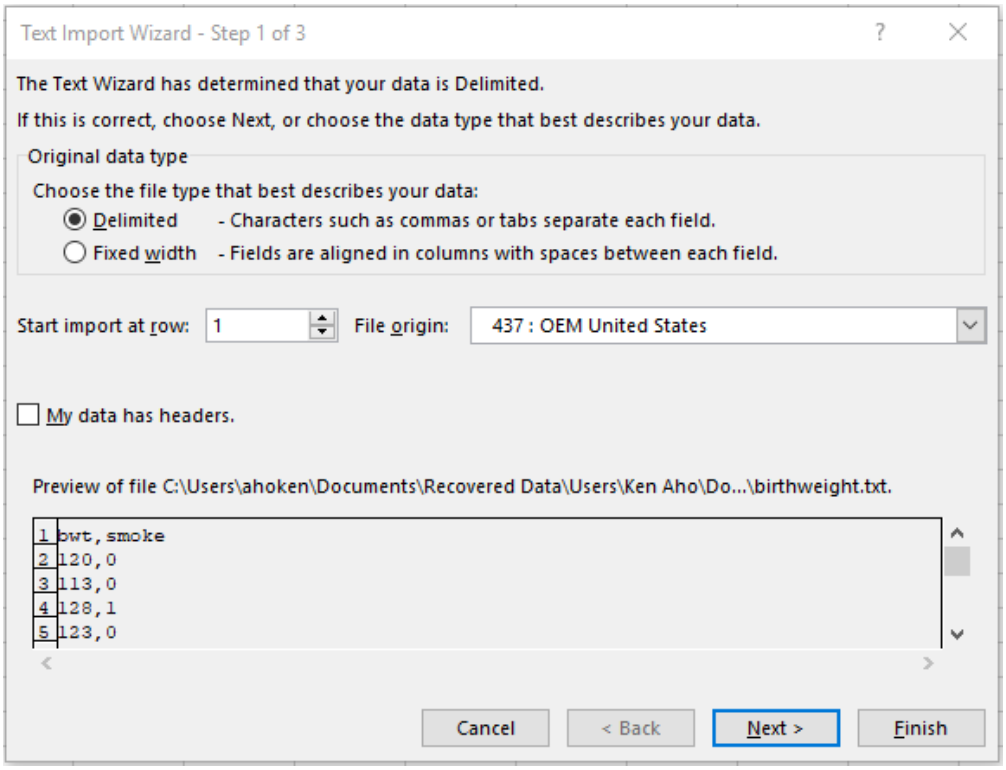
Your graphical analysis will use histograms. A **histogram** is a graph that depicts the distributional characteristics of dataset (e.g., its spread and symmetry). A histogram accomplishes this by binning data into categories and displaying the counts of observations in those categories. Additional useful information on histograms can be found here.

a) (3 pts) Save the `birthweight.txt` dataset from the class data directory onto your workstation.

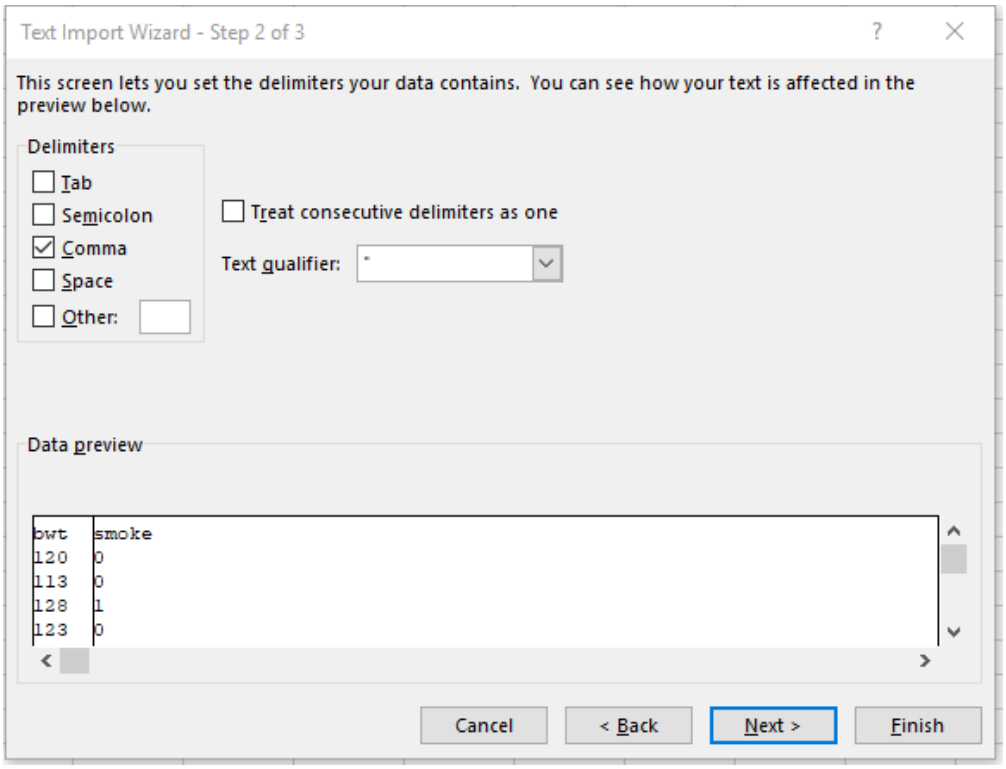
Open the dataset in Excel. You will need to open the file by searching for **All Files** (not just .xls files) in the directory in which you placed `birthweight.txt`.



The format of the `birthweight.txt` data is: birth weight (in ounces) and smoker/non-smoker designations (1 = smoker, 0 = non-smoker), separated by a comma. This is not typical .xls format. Thus, a Text Import Wizard will open asking you to define the column separator. Choose **Delimited** (the default).



Next, choose **Comma** and choose **Finish**.



If you pasted the `birthweight.txt` data into Excel, you can still use the comma as a column separator by selecting the column that contains the data (i.e., column **A**), and going to: **Data** > **Text to Columns** > **Delimited** > **Comma** > **Finish**.

Take a screenshot of the resulting spreadsheet and paste it into your homework with a appropriate caption. Close the `birthweight.txt` file without saving your changes.

b) (3 pts) We now examine the `birthweight.txt` data in **R**. To bring the data into **R**, type (or paste) the following code into R, and then find the file.

```
birthweight <- read.csv(file.choose())
```

The function `read.csv` allows import of.csv format data (data in which columns are separated by commas). The `birthweight.txt` dataset has this format. The function `file.choose` allows navigation to a file. To examine the object `birthweight` and verify that you brought the data in correctly, simply type:

```
birthweight
```

Now, type (or paste) the following into **R**:

```
smoke <- birthweight[,1][birthweight[,2] == 1]
no.smoke <- birthweight[,1][birthweight[,2] == 0]
```

This code subsets column 1 (the birth weights) based on outcomes in column 2 (the smoke/no smoke categorical assignments). See Appendix 1, Section 4 in this lab for more information on logical operators and Appendix 1, Section 5 for more information on subsetting.

Create histograms in **R** by typing (or pasting) the following code:

```
par(mfrow = c(2, 1))
hist(smoke, xlim = c(50, 180), ylim = c(0, 200), xlab =
       "Birth weight (oz)", ylab = "Frequency", main = "")
hist(no.smoke, xlim = c(50, 180), ylim = c(0, 200), xlab =
       "Birth weight (oz)", ylab = "Frequency", main = "")
```

The first line of code, `par(mfrow = c(2, 1))`, creates a graphical device with 2 rows and one column. It will hold a plot in each of its rows. The function `hist` creates histograms. Note that the first argument in `hist` calls either the `smoke` or `no.smoke` data. The `xlim` and `ylim` arguments define the upper and lower x and y axis limits. We hold these constant to make the plots for the smoker and non-smoker groups comparable. See Appendix 1, Section 6 in this lab for more information on plotting in **R**.

Right click on the figure and copy it as a metafile or bitmap. Paste the figure into your homework with an appropriate caption.

c) (3 pts) Histograms bin data into discrete categories. A histogram with two bins will separate data in to two groups, and so on. The number of bins will strongly affect histogram interpretations. By default, **R** chooses the number of bins based on an approach called Sturges' formula. Override this by adding an additional `breaks` argument, and define the number of bins to be 20. For instance:

```
par(mfrow = c(2, 1))
hist(smoke, xlim = c(50, 180), ylim = c(0, 200), xlab =
       "Birth weight (oz)", ylab = "Frequency", main = "",
     breaks = 20)
hist(no.smoke, xlim = c(50, 180), ylim = c(0, 200), xlab =
       "Birth weight (oz)", ylab = "Frequency", main = "",
     breaks = 20)
```

Right click on the figure and copy it as a metafile or bitmap. Paste the figure into your homework with an appropriate caption.

d) (2 pts) Briefly contrast the figures in 12b and 12c. How does the altered binning affect the histograms?

e) (4 pts) Consider the figures in 12b and 12c. What do the histograms suggest about the effect of maternal smoking on birth weight? Given what you know about experimental design, is it possible to make inferences of cause and effect? Why or why not?

13. To understand statistics and perform statistical analyses, you will need to be able to understand mathematical formulae, and use those formulae in Excel or **R**.

In this question you will calculate two important statistics, the **sample mean**, $\bar{X}$,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1.1}$$

and the **sum of squares**,

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{1.2}$$

where X_i represents the ith data outcome, $i = 1, 2, \ldots, n$. Thus, $\sum_{i=1}^{n} X_i$ means: take the sum of X_i outcomes, from $i = 1, 2, \ldots, n$. In Eq 1.2 the squared differences are calculated first, and the sum of those squared differences is calculated last. We will discuss the meaning and use of these statistics in Lab 3.

a) (4 pts) Calculate Eqs. 1.1 and 1.2 for the soil %N data in Table 1.2 using Excel. Use only the data, the Excel function =SUM, representing sum, and the operators /, -, and ^, representing division, subtraction, and exponentiation, respectively. Take a screenshot of your work and paste it into your homework with an appropriate caption.

b) (4 pts) Calculate Eq. 1.1 and 1.2 for the soil %N data in Table 1.2 using **R**. Use only data objects, the **R** function `sum`, representing sum, and the operators `/`, `-`, and `^`, representing division, subtraction, and exponentiation, respectively, and the assignment operator, `<-`. Take a screenshot of your work and paste it into your homework with an appropriate caption.

Q1 3pts, Q2 3pts, Q3 6pts, Q4 3 pts, Q5 6pts, Q6 2pts, Q7 3pts, Q8 4pts, Q9 2pts, Q10 2pts, Q11 2pts, Q12 12pts, Q13 8pts. **Total pts: 58.**

Appendix: R-code used in this lab

Elementary operators

Operator	Operation	To find	We type
`+`	addition	$2+2$	`2 + 2`
`-`	subtraction	$2-2$	`2 - 2`
`*`	multiplication	2×2	`2 * 2`
`/`	division	$\frac{2}{3}$	`2/3`
`^`	exponentiation	2^3	`2^3`

Creating objects

The function `c` allows one to assign data points to a single object. The operator `<-` is the assignment operator. Note that `=` can be used instead of `<-`. However we will save `=` to define arguments in functions.

Operator	Operation	To	We type
`<-`	assignment operator	assign the object `y` the name `x`	`x <- y`
`c`	combine	place the numbers 1, 2, and 3 into the object `x`	`x <- c(1, 2, 3)`

Data summarization

- For the functions below let **x** be some collection of data. For instance,

```
x <- c(1, 2, 3)
```

Function	Operation	To find	We type
`sqrt(x)`	$\sqrt{x}$	$\sqrt{2}$	`sqrt(2)`
`sum(x)`	summation	$\sum_{i=1}^{n} x_i$	`sum(x)`
`exp(1)`	$e = 2.718282\ldots$	e	`exp(1)`
`exp(x)`	e^x	e^3	`exp(3)`
`log(x)`	$\log_e(x)$	$\log_e(20)$	`log(20)`
`log(y,x)`	$\log_x(y)$	$\log_{10}(20)$	`log(20, base = 10)`

Data import

While it is possible to enter data into **R** at the command line, this will generally be inadvisable except for small datasets. In general it will be much easier to import data. **R** can import data from many different kinds of formats including .txt, and .csv (comma separated) files, and files with space, tab, and carriage return datum separators. I generally organize my datasets using Excel or some other spreadsheet program, then save them as .csv files for import into **R**.

Operator	Operation	To	We type
`read.csv`	Read-in 'comma-separated value' files	Read in `data.csv` from the directory `Dir`. Name the resulting object `data`	`data <- read.csv("Dir/data.csv")`
`file.choose`	Choose a file interactively	Navigate to a .csv file and import the data. Name the resulting object `data`	`data <- read.csv(file.choose())`

Boolean (logical) operators

Operator	Operation	To find	We type
`==`	Logical "equals"	which elements of $x = y$?	`x == y`
$>$	greater than	which elements of $x > y$?	`x > y`
$<$	less than	which elements of $x < y$?	`x < y`
$>=$	greater than or equal to	which elements of $x \geqslant y$?	`x >= y`
$<=$	less than or equal to	which elements o $x \leqslant y$?	`x <= y`

Subsetting data using []

Operator	Operation	To find	We type
`x[y]`	subset `x` based on `y`	2nd outcome from `x`	`x[2]`
	subset `x` based on logical outcome in `y`	outcomes from `x` $\geqslant 2$	`x[x >= 2]`
`x[y,]`	subset rows of `x` based on `y`	2nd row of `x`	`x[2,]`
`x[,y]`	subset columns of `x` based on `y`	2nd column of `x`	`x[,2]`

Graphs

In this lab we are introduced to **R** graphs using histograms. The workhorse **R** function for plotting is `plot`, which we will use frequently later in the semester. Plotting function, e.g., `plot` or `hist` generally have similar arguments including:

- `xlab` x-axis label.
- `ylab` y-axis label.
- `xlim` The upper and lower limits of the x axis, specified as `xlim = c(lower, upper)`.
- `ylim` The upper and lower limits of the y axis, specified as `ylim = c(lower, upper)`.

The function `par`, when placed in front of a plotting function, can be used to specify additional optional graphical operations including multiple plots placed within a single graphical device.

Operator	Operation	To	We type
`plot`	make a plot	make a plot at x locations defined by `x` and y locations defined by `y`	`plot(x,y)`

`par`	graphical parameters	create a graphical device to hold two graphs of x and y configured as two rows and one column	`par(mfrow = c(2,1))`
			`plot(x, y)`
			`plot(x, y)`

`hist`	make a histogram	make a histogram of data in `x`	`hist(x)`

2

Probability

Lab 2 Topics

1. Random variables and sets
2. Conceptions of probability
   - Frequentist
   - Bayesian
3. Probability rules
4. Combinatorial analysis
5. Bayes theorem

Random Variables and Sets

We learned in Lab 1 that a **variable** is simply a phenomenon that varies. Statistics is concerned with **random variables** whose outcomes cannot be known preceding a measurement or trial. The behavior of random variables must be described using **probability**. Consider a fair coin flip (Fig 2.1). Toss outcomes will be unknown preceding a toss. Nonetheless, we assume that heads will occur for 50% of tosses. That is, the probability of a head = 0.5.

Figure 2.1. Coin toss animation. To run, make sure this document is open in Adobe Reader and click on the coin.

Set Theory

Probability is often introduced using notation from from a branch of mathematics called **set theory**. A **set** is simply a collection of distinct objects. Sets can be used to define the behavior of random variables with a finite number of distinct outcomes. A set is notated using capital letters, for instance, A, B, etc. Objects comprising sets are listed inside curly brackets. For instance, $A = \{\male, \female\}$ is a set defining biological genders, where $\male$ = male, and $\female$ = female. **Elements** are individual outcomes comprising a set. For the previous example, $\male$ and $\female$ are elements of A.

In set theory, a conceptual realization of a random variable is called an **experiment**, an iteration of an experiment is a **trial**, and the observed result of a trial is an **outcome** or an **event**. For instance, we wish to flip a coin once (an experiment), we conduct the experiment and flip the coin (a trial) and the observed result is a tail (outcome).

The so-called **universal set**, often denoted S, contains all possible outcomes from an experiment. Consider an experiment involving two coin tosses. We have:

$$S = \{HH, HT, TH, TT\}.$$

S can be considered an event that will always occur with probability 1.

Conceptions of Probability

We denote the probability of an event A as: $P(A)$. But what does this actually *mean*? Several conceptualizations of probability have arisen over time. The two most common are probabilities as the limit of relative frequency over many trials (i.e., **frequentist**) and **degrees of belief**.

Frequentist Paradigm

Under the **frequentist interpretation of probability** (the kind we will almost always use in this class) probability is the proportion of times a particular outcome will occur over an infinite number of trials. For instance, the outcome for the toss of a coin (head or tail)

cannot be predicted in advance, but there is nonetheless a regular pattern to the long-term results from an essentially infinite series of coin flips (Fig 2.2). Obviously this view is an idealization (we can't flip a coin an infinite number of times) but it is very useful one.

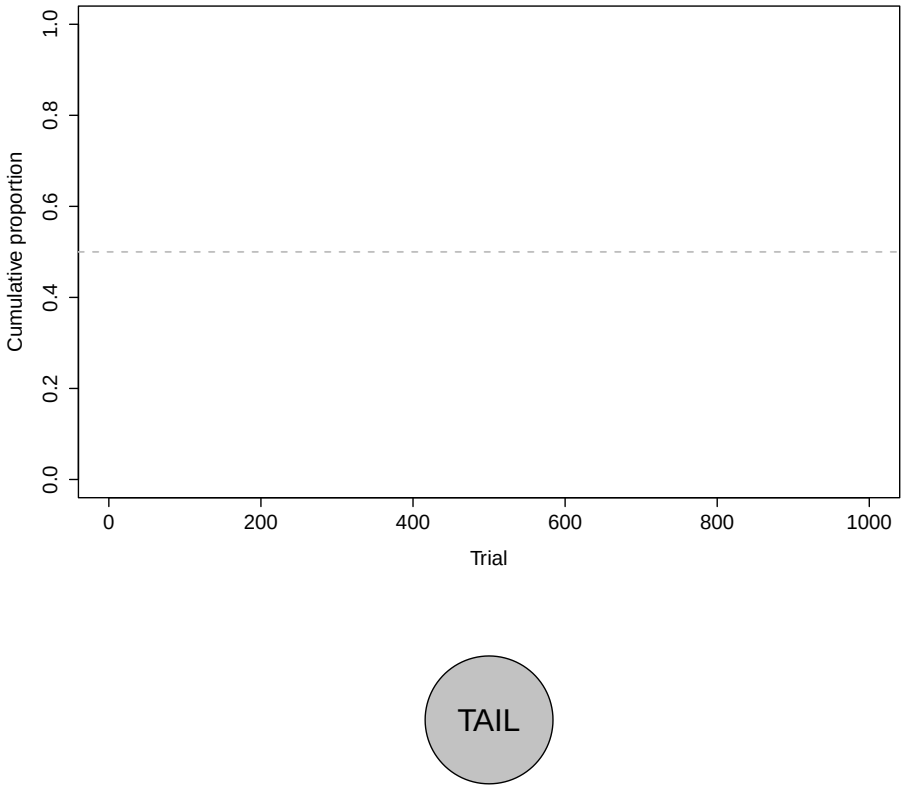


Figure 2.2. Coin toss realizations over many trials. As trials accumulate, the proportion of head outcomes approaches 0.5. To run the animation, click play while viewing in Adobe Reader.

Degrees of Belief: Bayesian Paradigm

The **degrees of belief interpretation of probability** is also called the Bayesian interpretation. It describes one's personal belief that an outcome is true. The degrees of belief interpretation is useful when making probabilistic statements concerning single events, or when considering current data in the context of past data.

Consider the statement (Aho, 2014):

"An explosion on the moon documented by Gervase of Canterbury in 1178, was

probably due to a meteorite impact, resulting in the lunar crater now known as Giordano Bruno."

Figure 2.3. Gervase of Canterbury (c. 1141 - c. 1210), an important chronicler of events in Medieval England.

This outcome is *possible* but not certain, and so can be considered probablistically. The crater Giordano Bruno probably formed around 1178. However the advent of the crater is not an outcome from an infinite frequentist distribution. Because it concerns a single event, this probabilistic statement requires a degrees of belief interpretation.

Comparison

The frequentist paradigm underlies most statistical methods, including the methods emphasized in this class. However, it is an idealization because of the impossibility of observing an infinite number of trials. The Bayesian paradigm has resulted in useful statistical methods. However, Bayesian personal probabilities can be subjective which is contrary to the goals of science. Differences in frequentist and Bayesian perspectives are most important in the context of inferential statistical procedures, e.g., parameter estimation and hypothesis testing. These topics will be addressed in upcoming labs, and we will address them using a conventional frequentist perspective. Bayesian and frequentist conceptual differences are irrelevant to the mathematical rules of probability described in the next section.

Probability Rules

For any event, A, its probability must be between 0 and 1. That is,

$$0 \leqslant P(A) \leqslant 1. \tag{2.1}$$

We call the event that A *does not* occur A **complement** and denote it as A'. Because of the limits set in Eq 2.1, it follows that:

$$P(A') = 1 - P(A). \tag{2.2}$$

Venn diagrams, provide a tool for visualizing sets and their relationships in sample space. Plots using this approach will generally be delimited by a rectangle, representing the universal set, S. The rectangle will contain one or more geometric shapes (usually circles) representing subsets of S. We can represent the probability of an event A with a shape whose proportional area in S will be equivalent to $P(A)$.

Example 2.1

Consider a sample space that compares O and B blood types in the U.S. (ignoring the two other blood type and all Rh possibilities). The probability of encountering an individual with blood type O, $P(O) = 0.44$, whereas the probability of encountering individual with blood type B, $P(B) = 0.1$ (Stanford Blood Center, 2020). A Venn diagram shows these probabilities within a rectangle of unit area that contains outcomes for all possible blood types, i.e. A, B, AB, and O for a single individual, although only $P(O)$ and $P(B)$ (probabilities for the blood types of interest) are shown as circles (Fig 2.4).

```
library(asbio)
Venn(.44, .1, labA = "O", labB = "B")
```

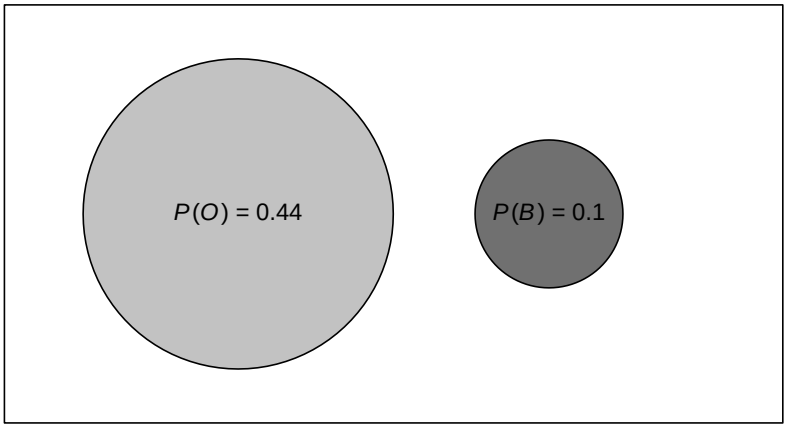


Figure 2.4. Venn Diagram of B and O blood types in the U.S in 2020. The white area in the diagram encompasses the probability of all blood types other than O and B, i.e., $P(O \cup B)'$. The figure was generated using the function `Venn` in library *asbio*.

■

Union

If two events are **mutually exclusive** or **disjoint** this means that they cannot occur simultaneously. One cannot get a head and a tail on a single coin flip. Thus, in a one flip

universe, head and tail outcomes are disjoint. Blood types for a single US citizen are also disjoint. An individual cannot have two distinct blood types. If events are disjoint, shapes representing their probabilities will not graphically overlap in a Venn diagram (Fig 2.4).

If, for two events (e.g. A and B), we wanted to know: "What is the probability of A *or* B?" we would express this using the set theory notation: $P(A \cup B)$. The term $\cup$ is called **union**. We can think of $\cup$ as representing the word "or." If A and B are mutually exclusive, then:

$$P(A \cup B) = P(A) + P(B). \tag{2.3}$$

Given this, what is $P(H \cup T)$ in a one coin toss universe?

Intersection

It is easy to imagine a situation where events are *not* mutually exclusive. For instance, assume that during a single round of feeding, an herbivore feeds on plant A with a probability of 0.3, plant B with a probability of 0.3, and plants A and B with a probability of 0.09 (Fig 2.5).

```
Venn(A = 0.3, B = 0.3, AandB = 0.09)
```

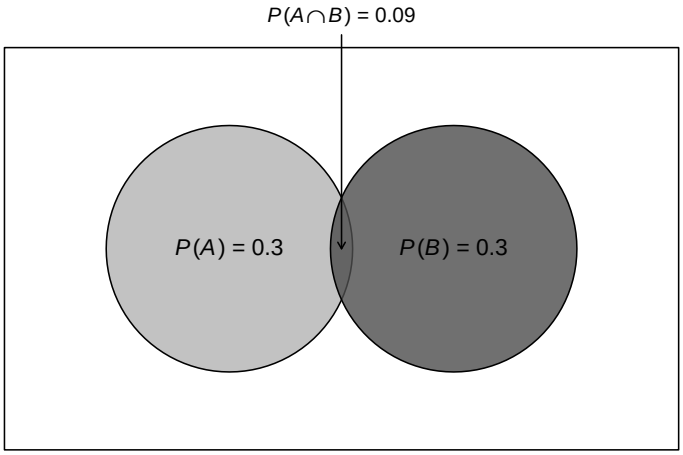


Figure 2.5. Venn diagram for the herbivore problem posed above. Note that A and B are not mutually exclusive.

We denote the probability of A *and* B as $P(A \cap B)$. The term $\cap$ is called **intersect**. We can think of $\cap$ as meaning "and." If events are not disjoint, then shapes representing their probabilities in a Venn diagram will overlap, and this overlap will represent $P(A \cap B)$ (Fig 2.5). Thus, if $P(A \cap B) > 0$, then A and B are not disjoint.

Even if A and B are not mutually exclusive we still know that:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{2.4}$$

Thus, in the previous example: $P(A \cup B) = 0.3 + 0.3 - 0.09 = 0.51$.

Note that Eq. 2.3 is not invalidated by Eq. 2.4. This is because, if A and B are mutually exclusive, then $P(A \cap B) = 0$, and $P(A \cup B) = P(A) + P(B) + 0$.

Independence

If, when A occurs it does not affect the probability of B occurring, then we say that A and B are **independent**. An example would be a head on a fair coin flip, A, and a head on a second consecutive fair coin flip, B. The probability of getting a head on the second flip will not be affected by the first outcome. Thus, fair coin flips are independent. If A and B are independent, then:

$$P(A \cap B) = P(A)P(B). \tag{2.5}$$

Thus, the probability of two consecutive heads in a two coin toss universe is $P(H \cap H) = 0.5 \cdot 0.5 = 0.25$. Note that here we assume that $P(T \cap H) = 0.25$ is distinguishable from $P(H \cap T) = 0.25$.

It is also easy to think of a situation where events are *not* independent. For instance, let A be the event that a student (let's call him Joe) passes a test on Monday. Let B be the event that Joe passes if he takes the same test on Tuesday. Clearly B is not independent of A. Joe will almost certainly do better on the test Tuesday if he took the same test Monday. It is important to note that if two non-zero probability events A and B are disjoint, then they cannot be independent. This is mathematically true because if A are B disjoint, then $P(A \cap B) = 0$, but under independence $P(A)P(B) = P(A \cap B)$ and, given this, we know that $P(A \cap B) > 0$ because $P(A)$ and $P(B)$ are both greater than zero.

Conditional Probability

If events A are B not independent, then we will need to use **conditional probability** to find their intersection. Specifically, we will need to find the probability of B "given" A or the probability of A "given" B. We denote the probability of B "given" A as: $P(B \mid A)$. It is important to note that:

$$P(A \mid B) \neq P(B \mid A).$$

For instance, whereas $P(spots \mid measles) = 1$, $P(measles \mid spots) \neq 1$.

We have the following relationships:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}, \tag{2.6}$$

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}. \tag{2.7}$$

Example 2.2

Let $P(A) = 0.4$, $P(B) = 0.5$, and $P(A \cap B) = 0.1$, we have:

$$P(B \mid A) = \frac{0.1}{0.4} = 0.25,$$

and

$$P(A \mid B) = \frac{0.1}{0.5} = 0.2.$$

■

Given Eq. 2.6 and Eq 2.7 we have:

$$\begin{aligned} P(A \cap B) = P(B \cap A) &= P(B \mid A)P(A) \\ &= P(A \mid B)P(B). \end{aligned} \tag{2.8}$$

Note Eq. 2.8 that does not invalidate Eq. 2.5. This is because if A and B are independent, then $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$. Thus, under independence, Eq. 2.8 becomes: $P(B \cap A) = P(A \cap B) = P(B)P(A) = P(A)P(B)$.

Because of the mathematical consequences of non-independence, it is vital for a scientist to be able to distinguish independent and non-independent events.

Example 2.3

Sudden infant death syndrome (SIDS) causes babies to suddenly die, often with no explanation. Deaths from SIDS have been reduced by doctor's recommendations that babies should be placed on their backs. Little progress, however, has been made in determining the casual factors behind SIDS.

In England parents were occasionally convicted of murder when more than one SIDS case occurred in a family. This is because the prosecution claimed that there was only a one in 73 million chance that two children could die from SIDS in the same family. The rationale was based on the observation that the rate of SIDS in non-smoking families was 1/8500. Given this perspective, the probability of two deaths would be:

$$\frac{1}{8500} \cdot \frac{1}{8500} = \frac{1}{72250000}.$$

However, the reasoning behind this premise was flawed. The prosecutors had no reason to believe that SIDS deaths in the same family were independent events. As the Royal Statistical Society stated:

> "There may well be unknown genetic or environmental factors that predispose families to SIDS, so that a second case becomes far more likely."

On this basis the British government decided to review the cases of 258 parents convicted of murdering their children, and 5,000 cases of children taken away from their parents (Aho, 2014).

■

Cominatorial Analysis

Counting methods are fundamentally tied to probability since they allow determination of the number of outcomes in a sample space, $N(S)$. While the enumeration of points in S will occasionally be straightforward, in many cases it will be extremely difficult, and require mathematical approaches. This branch of mathematics is called **combinatorial analysis**

A large number of counting methods are based on a concept called the multiplication principle, summarized in Theorem 1.

Theorem 1 (Multiplication Principle)**.** *If there are n outcomes for each of r trials in an experiment, then there are n^r possible outcomes in the sample space.*

Example 2.4

The litter size for domestic dogs (*Canis familiaris*) generally ranges between 6-10 pups (Society for the Prevention of Cruelty to Animals (SPCA), 2020). How many different sorts of litters (combinations of male and female pups) are possible for a litter size of six? How about a litter size of ten?

We have two possible outcomes for each pup (male or female). Thus, for a litter size of 6 we have $2^6 = 64$ possible outcomes. For instance, one possibility is three male pups followed by three female pups. Thus, for a litter size of ten we have $2^{10} = 1024$ possible outcomes.

■

Bayes Theorem

Bayes theorem is a very important conditional probability rule that allows us to find $P(A \mid B)$ from $P(B \mid A)$. Recall that these are not the same. It is very cool and mysterious these days to say: "I am a Bayesian."

Theorem 2 (Bayes Theorem)**.**

$$P(\theta \mid data) = \frac{P(data \mid \theta)P(\theta)}{P(data)}. \tag{2.9}$$

Proof.

$$\begin{aligned} P(A \mid B) &= \frac{P(A \cap B)}{P(B)} && \text{Eq. 2.7} \\ &= \frac{P(B \mid A)P(A)}{P(B)} && \text{Eq. 2.8} \end{aligned}$$

□

Bayes theorem (Theorem 2) has four components:

- **Likelihood function**: $P(data \mid \theta)$
- **Prior probability**: $P(\theta)$
- **Posterior probability**: $P(\theta \mid data)$
- **Total probability** or the **normalizing constant**: $P(data)$.

We use Bayes theorem to find posterior probabilities, $P(\theta \mid data)$. The priors, $P(\theta)$, are given as personal (potentially subjective) probabilities. This means that interpretation of the posterior requires a degrees of belief approach. Use of priors is the most contentious component of Bayes theorem. The likelihood function, $P(data \mid \theta)$, reflects the form of known conditional probabilities. The total probability, $P(data)$, will have (for our purposes) the form:

$$P(data) = \sum_{i=1}^{k} P(data \mid \theta_i)P(\theta_i)$$

Thus, for our purposes, Bayes theorem has the final form:

$$P(\theta_i \mid data) = \frac{P(data \mid \theta_i)P(\theta_i)}{\sum_{i=1}^{k} P(data \mid \theta_i)p(\theta_i)}. \tag{2.10}$$

The events $\theta_1, \ldots, \theta_k$ are called k **states of nature**. The prior probabilities correspond to these events.

Example 2.5

A meat inspector must decide if a meat sample contains *Escherichia coli* using a diagnostic test. For a perfect diagnostic test, a positive result (POS) would always indicate that *E. coli* is present and a negative test result (NEG) would always indicate that no *E. coli* is absent. However, as with most diagnostic tests, this test gives false positive and false negative results. Assessing the diagnostic procedure with 10,000 samples containing *E. Coli* (EC+) and 10,000 samples without *E. coli* (EC-) yielded the results shown in Table 2.1.

Table 2.1. Results of a diagnostic assessment of a test for the presence of *E. coli*.

		E Coli presence/absence in meat sample	
		EC+	EC-
Test Result	POS	9,500	100
	NEG	500	9,900
	Total	10,000	10,000

From Table 2.1 we have:

- True positive rate = $P(\text{POS} \mid \text{EC+}) = 9,500/10,000 = 0.95$
- False positive rate = $P(\text{POS} \mid \text{EC-}) = 100/10,000 = 0.01$
- True negative rate = $P(\text{NEG} \mid \text{EC-}) = 9,900/10,000 = 0.99$
- False negative rate = $P(\text{NEG} \mid \text{EC+}) = 500/10,000 = 0.05$

The question of greatest interest is: what is the probability that meat contains *E. coli* even if the test is negative? That is, what is $P(\text{EC+} \mid \text{NEG})$? This is the probability of a piece of *E. coli* meat showing up on your dinner table! We don't know this, although we do know the inverted conditional probability, $P(\text{NEG} \mid \text{EC+}) = 0.05$.

One more step is necessary. We have to provide prior probabilities. In this case the priors will concern the states of nature for the presence/absence of *E. coli* in the meat that the USDA will inspect. Ott *et al.* (2004) estimated that the proportion of all inspected US meat that contains *E. coli* is around 4.5% . Thus, we will let $P(\text{EC+}) = 0.045$, and $P(\text{EC-}) = 0.955$. These are the priors. We have:

$$\begin{aligned} P(\text{EC+} \mid \text{NEG}) &= \frac{P(\text{NEG} \mid \text{EC+})P(\text{EC+})}{P(\text{NEG} \mid \text{EC+})P(\text{EC+}) + P(\text{NEG} \mid \text{EC-})P(\text{EC-})} \\ &= \frac{0.05 \cdot 0.045}{0.05 \cdot 0.045 + 0.99 \cdot 0.955} \\ &= 0.002374169 \end{aligned}$$

If our priors are right, 0.24% (approximately 2/10 of 1%) of meat that passes inspection, and is available for consumption, has *E. coli*. Not too bad, unless of course, you are the stuck with meat representing the rare 0.24% infected group! We can facilitate the calculation process (and avoid errors) using **R**:

```
NEG.noEC <- 0.99
NEG.EC <- 0.05
EC <- 0.045
noEC <- 0.955

NEG.EC * EC/(NEG.EC * EC + NEG.noEC * noEC)

[1] 0.002374169
```

■

Assignment 2

Answer all questions in one MS Word document and submit to Moodle. At the beginning of the document include the assignment number, the date, your name and section number.

Use complete sentences when appropriate, and make sure any tables, figures and computer output you include adhere to class standards (see Syllabus).

Conceptions of Probability

- Open **R**
- Install the *asbio* package by typing `install.packages("asbio")`. The acronym *asbio* means Applied Statistics for Biologists.
- Load the *asbio* package by typing `library(asbio)` or by going to **Packages > Load packages > asbio**
- Type `book.menu()` in the **R** console.

1. (4 pts) Open the coin flip GUI (Graphical User Interface) by going to **Chapter 2 > Coin flips** in the *asbio* book menu. Run the function using the default values in the GUI.

a) Does the proportion of heads approach 0.5 after 5 coin flips? After 100 flips? After 1000 flips?

b) The convergence of the probability for an event to a single fixed number given many trials corresponds to which interpretation of probability?

2. (1 pt) Open the die toss GUI, by going to **Chapter 2 > Die throws** in the *asbio* book menu. Run the function using the defaults (a fair die). What exact probability will each die toss outcome converge to under the frequentist paradigm?

3. (2 pts) Provide a definition for the degrees of belief conception of probability.

Probability Rules

4. Open the probability self test GUI by going to **Chapter 2 > Self test questions > Probability** in the *asbio* book menu. Answer questions 1, 2, 5, and 8. Nothing needs to be handed in for this question.

5. (8 pts) Open the Venn diagram GUI, by going to **Chapter 2 > Venn diagrams** or by simply typing `Venn.tck()` in the console.

a) Create a Venn diagram representing $P(A) = 0.3$, $P(B) = 0.4$, $P(A \cap B) = 0$. Paste the figure into your homework with an appropriate caption, e.g., "Venn diagram for question 4." You will need to copy the figure as a bitmap to preserve the shading.

b) Are A and B disjoint? Why?

c) Are A and B independent? How do you know this mathematically?

d) Calculate $P(A \cup B)$.

6. (8 pts) Consider the situation: $P(A) = 0.6$, $P(B) = 0.3$, $P(A \cap B) = 0.3$

a) Create a Venn diagram representing the probabilities. Paste the figure into your homework with an appropriate caption.

b) Are A and B disjoint? Why?

c) Are A and B independent? Why?

d) Calculate $P(A \cup B)$.

7. (6 pts) Consider the situation: $P(A) = 0.4$, $P(B) = 0.3$, $P(A \cap B) = 0.12$

a) Create a Venn diagram representing the probabilities. Paste the figure into your homework with an appropriate caption.

b) Are A and B disjoint? Why?

c) Are A and B independent? Why?

8. (8 pts) The probability of finding a particular plant species at habitat $B = 0.4$. The probability of finding the species at habitat A, given the presence of habitat $B = 0.01$. The probability of finding the species at habitat B given the presence of habitat $A = 0.02$.

a) Calculate $P(A \cap B)$.

b) Calculate $P(A)$.

c) Are habitats A and B probabalistically disjoint? Why?

d) Are A and B independent? Why?

9. (6 pts) Let A and a be alleles for a gene from a population under the Hardy Weinberg equilibrium. The Hardy Weinberg equilibrium assumes the independence of alleles at a gene locus. Let $P(A) = p$ and $P(a) = q$, represent the relative frequency of the A and a in the general population, and hence the probability of those alleles occurring in a random selected individual.

a) What are the four possible allele combinations for the gene?

b) What is the probability (in terms of p and/or q) of the homozygote dominant genotype, AA, in the population, i.e., what is $P(A \cap A)$?

c) What is the probability (in terms of p and/or q) of the homozygote recessive genotype, aa, in the population, i.e., what is $P(a \cap a)$?

d) What is the probability (in terms of p and/or q) of the heterozygote genotype, Aa, in the population, i.e., what is $P(A \cap a)$?

10. (12 pts) The approximate litter size for brown spiny field mice (*Mus plantythrix*) on dry land sites in India is approximately 4. Assume that the sex of offspring in litters are independent.

a) How many outcomes are possible in terms of male and female offspring (see Theorem 1, and the associated Example)?

b) List all the possible outcomes in terms of male and female offspring for a litter size of four.

c) What is the probability of each outcome?

d) What is the probability of getting exactly three females in a litter of four?

e) What is the probability of getting at least one female in a litter of four?

f) What is the probability of getting no females in a litter of four?

11. (6 pts) For the spiny field mice in Q. 10, let A be the event of getting exactly 2 females, B be the event of getting at least 1 female, and C be the event of getting no females.

a) What is $P(B \mid A)$? Hint: try to think of this logically without using math.

b) What is $P(A \mid C)$?

c) Are A and B independent? What about A and C? Defend your answers.

Bayes Theorem

12. (7 pts) Rao *et al.* (1998) reported on the utility of using computerized tomography as a diagnostic test for patients with clinically suspected appendicitis. Traditional clinical methods of diagnosis, and diagnosis using the aid of computerized tomography (CT) were used on 100 patients. Whether patients actually had appendicitis was determined later by examining the appendix following an appendectomy. The results are shown in Table 2.2. The 1996 rate of appendicitis (APP+) was approximately 0.00108. Thus, we will use $P(\text{APP+}) = 0.00108$ and $P(\text{APP}-) = 1 - 0.00108$ as the priors. Note, a very similar example to this problem is given in the Bayes Theorem section of this lab.

a) How many states of nature are there for the presence of appendicitis?

b) Determine $P(\text{DA} \mid \text{APP+})$, $P(\text{DA} \mid \text{APP}-)$, $P(\text{NA} \mid \text{APP+})$, and $P(\text{NA} \mid \text{APP}-)$. Note that this does not require use of Bayes theorem.

c) Using Bayes theorem, find the probability that a patient did not have appendicitis (APP−) given that the radiological determination was definite appendicitis (DA). This may be higher than you might think.

Table 2.2. Results from Rao *et al.* (1998).

		Presence of appendicitis	
		APP+	APP−
CT determination	Definite appendicitis (DA)	50	1
	Possible appendicitis (PA)	2	2
	No appendicitis (NA)	1	44
	Total	53	47

Q1 4pts, Q2 1pt, Q3 2pts, Q4 0pts, Q5 8pts, Q6 8pts, Q7 6pts, Q8 8pts, Q9 6pts, Q10 12pts, Q11 6pts, Q12 7pts. **Total pts: 68.**

Appendix: R-code used in this lab

asbio Functions

We will rely on the package *asbio* for many applications in this course. **R**-packages are collections of functions and datasets that can be utilized in the **R**-environment. Packages can be **installed** using the function `install.packages`. Once a package is installed you shouldn't have to install it on a workstation again. An exception will be computers in ISU lab environments wherein computer memories are scrubbed nightly. Unless a package is part of the default **R** download it will need to be **loaded** using the function `library` for use in the current **R** session. Loading will require that the package has already been installed

Function	Operation	To	We type
`install.packages`	Install package(s)	Install package `asbio`	`install.packages("asbio")`
`library`	Load package(s) for use in current work session	Load package `asbio`	`library(asbio)`
`book.menu`	Open the the *asbio* textbook menu	Open the book menu	`book.menu()` Requires that *asbio* is loaded
`Venn`	Make Venn diagram	Make a Venn diagram with $P(A) = 0.2$, $P(B) = 0.2$, and $P(A \cap B) = 0.06$	`Venn(0.2, 0.2, 0.06)` Requires that *asbio* is loaded

3

Probability Density Functions

Lab 3 Topics

1. Probability distribution terms:
   - Probability density function (PDF)
   - Cumulative distribution function (CDF)
   - Inverse CDF (CDF^{-1})
2. Discrete PDFs
   - Bernoulli distribution
   - Binomial distribution
   - Poisson distribution
3. Continuous PDFs
   - Continuous uniform distribution

Probability Density Functions

It is often possible to express the probabilistic distribution of a quantitative random variable as a mathematical function. Functions that define probability distributions are called **probability density functions** or **PDFs**. PDFs are mathematically denoted $f(x)$, and vary with random variable outcomes, denoted x. The random variable itself is denoted X. The output generated by the function is called **density**.

For a discrete random variable density is equivalent to probability. As a result, a discrete

Discrete PDFs are often called probability mass functions, or PMFs.

PDF has the general form:

$$f(x) = P(X = x).$$

For a continuous random variable, the magnitude of a PDF provides insight into patterns of the relative frequencies for outcomes in X. However, unlike a discrete distribution, it will not represent $P(X = x)$. Integration is theoretically necessary to calculate probability for a continuous random variable. The probability $P(X = x)$ is uninformative in a continuous distribution because the results of integration at any single point on the number line will always be 0.

Instead, we calculate continuous probability for a *range* of outcomes. Specifically, let X be a continuous random variable with PDF $f(x)$, and let $\{a, b\} \in X$, where $a < b$. To obtain $P(a \leqslant X \leqslant b)$, we find:

$$P(a \leqslant X \leqslant b) = \int_a^b f(x)dx.$$

Valid PDFs have two characteristics:

1. $f(x) \geqslant 0$ for all x.
2. With regard to the densities of cumulative outcomes:
   - For discrete random variables, $\sum_x f(x) = 1$.
   - For continuous random variables, $\int_{-\infty}^{\infty} f(x)dx = 1$.

Example 3.1

Imagine that you are an alpine ecologist studying the demographics of mountain goats (*Oreomnos americanus*). You observe a certain ridge top for a long period of time and decide that the probabilities of goat counts in this area can be described by Table 3.1. The table here serves as a discrete PDF. Nothing would be gained by summarizing this variable mathematically. The table can also be expressed as a figure (Fig 3.1a). All PDFs (both discrete and continuous) give the y-axis "height" in a plot of $f(x)$ as a function of x. The y-axis outcomes of these figures will always represent density. A graph of a discrete PDF will always look similar to Fig 3.1a: a series of disconnected vertical lines.

Table 3.1. Example of a discrete PDF. Hypothetical mountain goat (*Oreomnos americanus*) counts, and their densities. Example taken from Aho (2014)

x = goat count	0	1	2	3	4
$f(x)$ = probability of a particular goat count	0.5	0.3	0.1	0.05	0.05

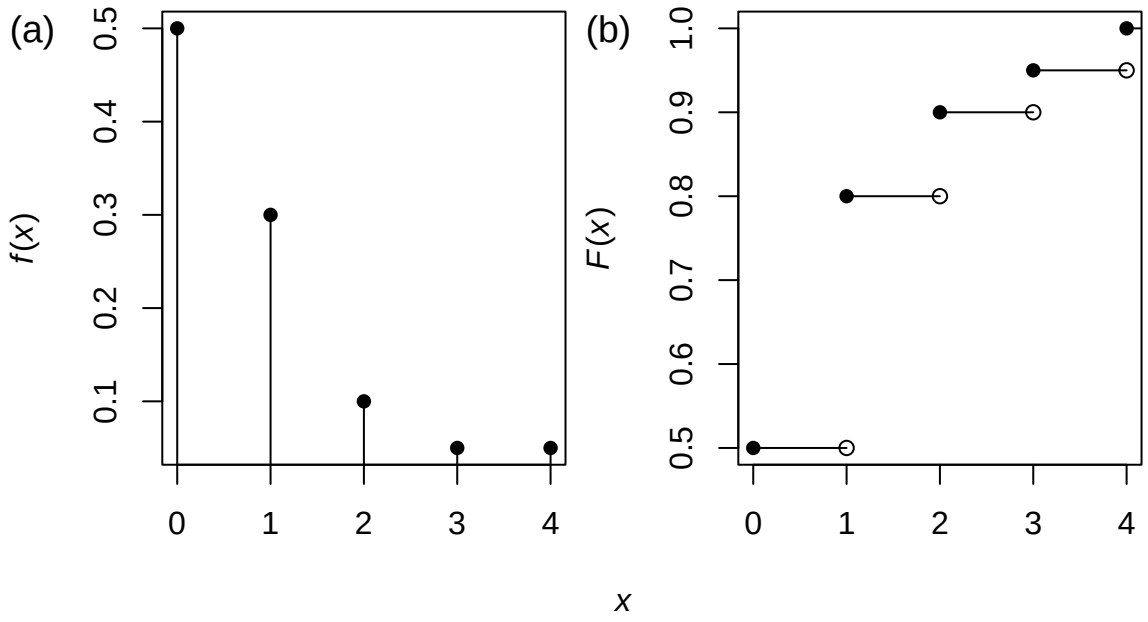


Figure 3.1. Table 3.1 re-expressed as a figure: (a) shows the PDF, (b) shows the CDF (see below). Filled dots indicate that the point is included in the density and/or cumulative distribution functions. Open dots indicate the point is not included in the CDF.

Because the densities of all outcomes in Table 3.1 are greater than or equal to 0, and the sum of all the densities equals 1, this appears to be a valid discrete PDF. We note that density, here is equivalent to probability. This will be true for all discrete PDFs. However, it will not be true for continuous PDFs.

■

Cumulative distribution function (CDF)

A **cumulative distribution** function or **CDF** gives the probability that a random variable, X is less than or equal to an outcome x. Thus, the CDF gives the **lower tail probability**. The CDF is denoted as $F(x)$. Thus,

$$F(x) = P(X \leqslant x) \tag{3.1}$$

Eq. 3.1 holds for both discrete and continuous random variables. We see that in Table 3.1 and Fig 3.1b, $P(X \leqslant 2) = F(2) = 0.9$.

Inverse CDF

The inverse CDF or quantile function is literally the inverse of the CDF. That is, it gives the outcome x for an associated lower tailed probability. The inverse CDF is sometimes denoted as F^{-1}. Because $F(2) = 0.9$ in Table 3.1, $F^{-1}(0.9) = 2$.

Common probability distributions

Conventional statistical methods use a handful of PDFs with well understood mathematical properties. Procedures built on these algorithms are called **parametric** because they use PDFs with known parametric forms. A **parameter** for a probability distribution can be defined as a fixed numeric characteristic of the distribution. Important PDF parameters include the mean and the variance. An important parameter for many discrete PDFs is the probability of success (see below). Infinitesimally changing the parameters for any PDF will result in an infinite number of distinct distributions. We will be introduced to four such PDFs today: the Bernoulli, binomial, Poisson, and uniform distributions.

Discrete PDFs

Bernoulli distribution

The **Bernoulli distribution** is arguably the simplest useful discrete PDF. It defines the probability of a success for a single random binary event, i.e., an event with only two possible outcomes. For instance, presence/absence, life/death, male/female, head/tail. Either outcome from any of these examples could be defined as a "success" depending on the focus of an investigation. If a random variable X follows a Bernoulli distribution, then its PDF has the form:

$$f(x) = p^x(1-p)^{1-x}. \tag{3.2}$$

- Here x is a Bernoulli outcome. This will be the number of successes for a single binary trial. Thus, x will be either a 0 or a 1 (0 successes or 1 success).

The Bernoulli distribution has one parameter, p.

- p is the probability of a success, i.e., the probability that $X = 1$. Because p is a probability, $0 \leqslant p \leqslant 1$.

Example 3.2

A single coin toss can be considered a Bernoulli random variable. Assuming the coin is fair, the probability of a single success (a head) is:

$$f(1) = 0.5^1(1-0.5)^0 = 0.5.$$

The probability of zero successes (a tail) is also 0.5:

$$f(0) = 0.5^0(1-0.5)^1 = 0.5.$$

The probability of seeing an outcome less than or equal to 1 (of seeing a head *or* a tail) is:

$$F(1) = f(0) + f(1) = 0.5 + 0.5 = 1.$$

To calculate Bernoulli densities (equal to probabilities because this is a discrete distribution) and lower-tailed (CDF) probabilities in Excel we use the function =BINOM.DIST. The function actually expresses the binomial distribution, of which the Bernoulli distribution is a special case (see below). The function =BINOM.DIST requires four arguments:

1. The number of successes, x.
2. The number of trials, n.
3. The probability of a success, p.
4. Whether or not you want the CDF (TRUE) or the PDF (FALSE).

Thus, the probability of getting a head on a single coin flip is:

$$=\text{BINOM.DIST}(1,\ 1,\ 0.5,\ \text{FALSE}) = 0.5$$

The probability of getting a tail on a single coin flip is:

$$=\text{BINOM.DIST}(0,\ 1,\ 0.5,\ \text{FALSE}) = 0.5$$

The probability of seeing an outcome less than or equal to 1 (of seeing a head *or* a tail) is:

$$=\text{BINOM.DIST}(1,\ 1,\ 0.5,\ \text{TRUE}) = 1$$

Bernoulli densities and lower-tailed probabilities are easily obtained in **R**. The **R**-function for the binomial PDF is `dbinom`. The `d` prefix indicates density. The function `dbinom` requires three arguments:

1. The number of successes, x.
2. The number of trials, n.
3. The probability of a success, p.

```
dbinom(1, 1, 0.5) #P(Head)

[1] 0.5

dbinom(0, 1, 0.5) #P(Tail)

[1] 0.5
```

The binomial CDF is contained in the function `pbinom`. The `p` prefix indicates lower-tailed probability. The function `pbinom` has the same three arguments as `dbinom`.

```
pbinom(1, 1, 0.5) #P(one or fewer heads)

[1] 1
```

■

Binomial distribution

The **binomial distribution** gives the probability for a particular number of successes, x, given a particular number of independent Bernoulli trials, n. Thus, like the Bernoulli, the binomial distribution is also used to depict dichotomous variables. If a random variable X follows a binomial distribution, then its PDF has the form:

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}. \tag{3.3}$$

- Here x is a binomial outcome. This defines some number of successes across n independent Bernoulli trials. Thus, $x = 0, 1, 2, \ldots, n$.

The binomial PDF has two parameters, n and p.

- n indicates the number of trials.
- p is the probability of a success for each Bernoulli trial. Because p is a probability, $0 \leqslant p \leqslant 1$.

The term $\binom{n}{x}$ is called the **binomial coefficeint**, and is pronounced "n choose x,"

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}. \tag{3.4}$$

where $n! = n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1$ is referred to as ***n* factorial**. The binomial coefficient gives the number of unique success/failure combinations that exist with respect to x successes and $n - x$ failures.

Example 3.3

Last week we painstakingly found that there would be four distinct litters of size four, with three females. Through trial and error, we found that these were: ♂♀♀♀, ♀♂♀♀, ♀♀♂♀, and ♀♀♀♂. We could have easily found this answer mathematically by using the binomial coefficient. We have:

$$\begin{aligned}\binom{4}{3} &= \frac{4!}{3!(4-3)!} \\ &= \frac{24}{6(1)} \\ &= 4.\end{aligned}$$

The **R** functions for factorial and the binomial coefficient are `factorial` and `choose`, respectively.

```
factorial(4)/(factorial(3) * factorial(1))

[1] 4

choose(4,3)

[1] 4
```

■

A PDF is generally denoted with a short acronym for its name, followed by its required parameters, given inside parentheses. A binomial distribution has two parameters, n and p. Thus, if a random random variable, X follows a binomial distribution this is summarized: $X \sim BIN(n, p)$. The symbol $\sim$ means "follows" or "follows in distribution." The distribution $BIN(n = 10, p = 0.5)$ is shown in Fig. 3.2.

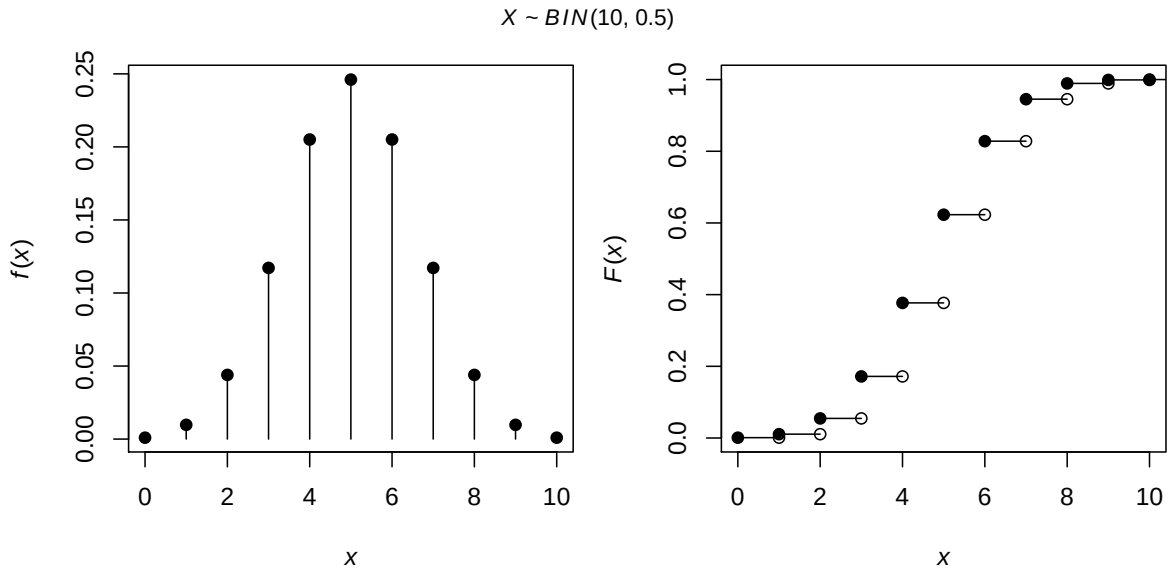


Figure 3.2. The PDF and CDF of the distribution $BIN(10, 0.5)$. Filled dots indicate that the point is included in the density and/or cumulative distribution functions. Open dots indicate the point is not included in the CDF.

If $n = 1$, then there will only be two possible outcome for x, zero successes and one

success, and in this case, $\binom{n}{x}$ becomes

$$\binom{1}{1} = \binom{1}{0} = 1.$$

Thus, if $n = 1$, then Eq 3.3 becomes Eq 3.2. Because of this, the Bernoulli is a special case of the binomial distribution. Specifically, it is the binomial distribution when $n = 1$. As a result, if a random variable X follows a Bernoulli distribution, this can be written as $X \sim BIN(1, p)$.

Example 3.4

To calculate the binomial probability of obtaining three females in a litter size of four, we need to define the binomial parameters for our particular problem. Because we have $n = 4$ (the total number of offspring) and $p = 0.5$ (we assume the probability of obtaining a male or a female equals 0.5), we have $X \sim BIN(4, 0.5)$. Our outcome of interest is $x = 3$. Thus, we have:

$$\begin{aligned} f(x) &= \binom{n}{x} p^x (1-p)^{n-x} \\ P(X = 3) = f(3) &= \binom{4}{3} 0.5^3 0.5^{4-3} \\ &= 4(0.5^4) \\ &= 4(0.0625) \\ &= 0.25. \end{aligned}$$

This is because:

$$\binom{1}{0} = \frac{1!}{0!(1-0)!} = \frac{1}{1(1)} = \binom{1}{1} = \frac{1!}{1!(1-1)!} = \frac{1}{1(1)} = 1.$$

The proof that $0! = 1$ is very interesting. By definition:

$$x! = x \cdot (x-1) \cdot (x-2) \cdot 3 \cdot 2 \cdot 1.$$

Manipulating our definition for $x!$, we have:

$$\begin{aligned} x! &= x \cdot (x-1)! \\ (x-1)! &= \frac{x!}{x}. \end{aligned}$$

Substituting $x = 1$ to obtain $0!$, we have:

$$\begin{aligned} (1-1)! &= \frac{1!}{1} \\ 0! &= \frac{1}{1} = 1. \end{aligned}$$

The Bernoulli distribution is a special case of the binomial distribution. Thus, to calculate binomial probabilities in Excel we again use the function =BINOM.DIST. Thus, for our current example we have:

$$=\text{BINOM.DIST}(3,\ 4,\ 0.5,\ \text{FALSE}) = 0.25$$

In **R** we have

```
dbinom(3, 4, 0.5)

[1] 0.25
```

■

Poisson distribution

Like the Bernoulli and binomial distributions, the **Poisson distribution** describes probability for some number of successes, x. Unlike the Bernoulli and binomial PDFs, however, there is no designated upper limit (at n) for the number of successes. If a random variable X follows a Poisson distribution, then its PDF is:

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}. \tag{3.5}$$

- Here x is a Poisson outcome, $x = 0, 1, 2, \ldots$, and e represents Euler's number, $e = 2.71828....$

The Poisson PDF has one parameter, λ.

- λ describes the rate of successful outcomes (e.g., the number of organisms encountered per unit time). By definition, $\lambda > 0$.

The Poisson is a unique distribution in that its mean will always equal its variance, and both of those parameters will be equal to its lone parameter, λ. If a random variable, X, follows a Poisson distribution this is denoted $X \sim POI(\lambda)$. The distribution $POI(5)$ is shown in Fig. 3.3.

Recall that e is the base of natural logarithms. It turns that e can be calculated as the sum of the infinite series:

$$e = \sum_{n=0}^{\infty} \frac{1}{n!} = 1 + \frac{1}{1} + \frac{1}{2 \cdot 1} + \frac{1}{3 \cdot 2 \cdot 1} + \cdots$$

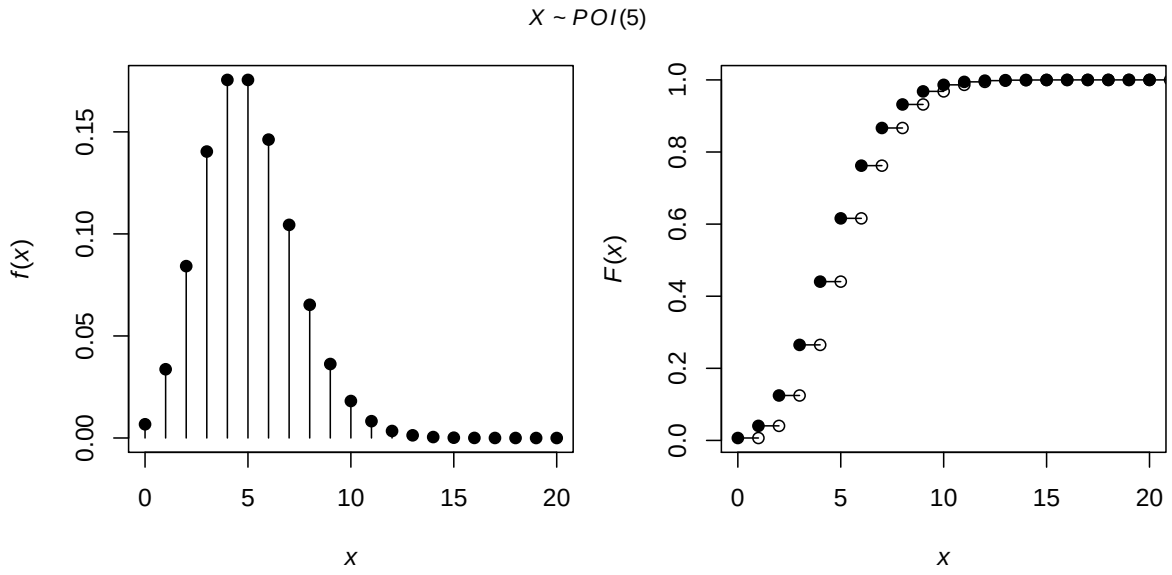


Figure 3.3. The PDF and CDF of the distribution $POI(5)$. Filled dots indicate that the point is included in the density and/or cumulative distribution functions. Open dots indicate the point is not included in the CDF.

Because of its properties, the Poisson distribution is most often used by biologists to represent spatial or temporal randomness of counts in space or time. We can gauge whether data come from a Poisson distribution by comparing the sample mean and the sample variance of the data. If these are approximately equal, this indicates that the data likely come from a Poisson distribution, and are thus randomly distributed in space or time. The equation for the sample mean was given in Eq. 1 in Lab 1. The sample variance is the sum of squares (Eq. 2 in Lab 1) divided by $n - 1$. We will discuss the sample mean and variance in greater detail in Lab 4.

Example 3.5

Assume that bald eagle sightings in Last Chance, Idaho, are Poisson distributed with a rate of 0.2/hr. What is the probability of seeing three eagles in an hour? We have:

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

$$P(X = 3) = f(3) = \frac{e^{-0.2}0.2^3}{3!}$$

$$= 0.00109.$$

To calculate Poisson densities (and lower tailed probabilities) in Excel we use the function =POISSON.DIST. The function requires three arguments.

1. The number of successes, x.
2. The rate, λ.
3. Whether or not you want the CDF (TRUE) or the PDF (FALSE).

For the current example we have:

$$=\text{POISSON.DIST}(3,\ 0.2,\ \text{FALSE}) = 0.00109$$

To calculate Poisson densities in **R**, we use the function `dpois`. The function requires two arguments.

1. The number of successes, x.
2. The rate, λ.

For the current example we have:

```
dpois(3, 0.2)

[1] 0.001091641
```

■

Continuous PDFs

Continuous uniform distribution

The simplest continuous distribution is the **continuous uniform distribution**. It is often used as a naïve model to represent processes in which all possible continuous outcomes have the same likelihood. If a random variable, X, follows a continuous uniform distribution then it will have the PDF:

$$f(x) = \frac{1}{b-a}. \tag{3.6}$$

The continuous uniform PDF has two parameters.

- a is the lower limit of the support for X (minimum possible value of X).
- b is the upper limit (maximum possible value of X).

Note that because density will be equal for all possible outcomes, and will depend only on the limits of the distribution, x is not required in the density function. By definition, $a \leqslant x \leqslant b$.

If a random variable X follows a continuous uniform distribution, we denote this as $X \sim UNIF(a, b)$. The distribution $UNIF(2.5, 3)$ is shown in Fig 3.4.

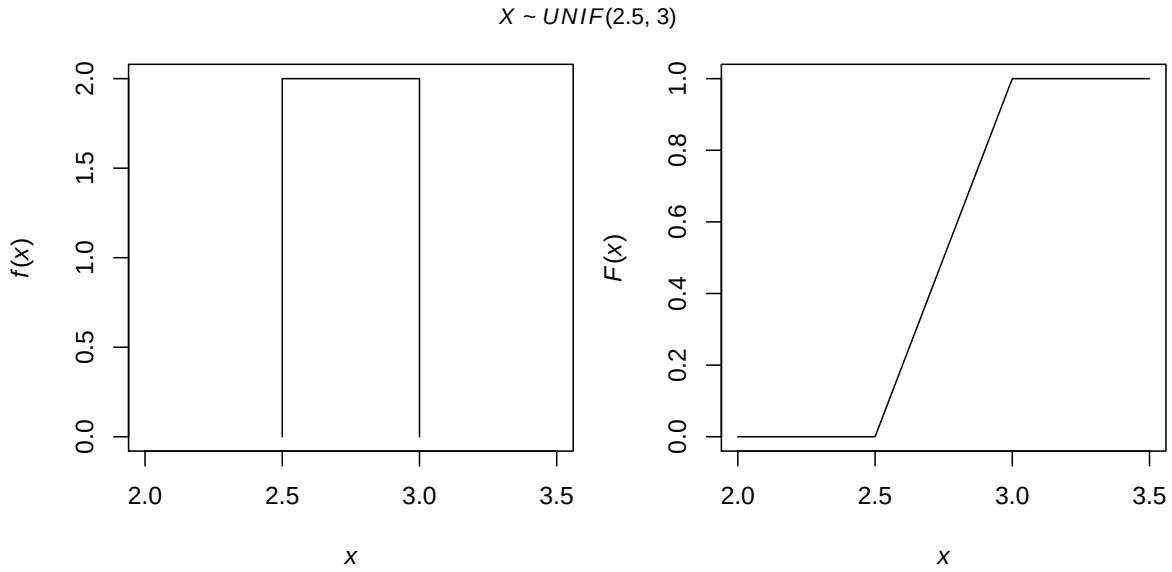


Figure 3.4. The PDF and CDF of the distribution $UNIF(2.5, 3)$.

Example 3.6

Unlike a discrete distribution, continuous PDFs do not directly provide probability. Instead, we have to find an area beneath a PDF corresponding to a range of defined outcomes in the PDF. We do this by integrating the PDF.

Let $X \sim UNIF(2.5, 3.0)$ (see Fig 3.4). Then, to find $P(2.5 \leqslant X \leqslant 2.8)$, we have:

$$\begin{aligned} P(2.5 \leqslant X \leqslant 2.8) &= \int_{2.5}^{2.8} \frac{1}{3 - 2.5} dx \\ &= \int_{2.5}^{2.8} 2dx \\ &= 2x\Big|_{2.5}^{2.8} \\ &= 2(2.8) - 2(2.5) \\ &= 0.6. \end{aligned}$$

The integration here is extremely straightforward because we are finding the area under a rectangle (Fig 3.4). Thus, to find the probability we could have found the difference of the x outcome of interest, and lower bound of the PDF $(2.8 - 2.5 = 0.3)$ and multiplied this by the

lone density of the PDF, 2. That is, $P(2.5 \leqslant X \leqslant 2.8) = 0.3 \cdot 2 = 0.6$. Calculating probabilities using integration, however, will work for *all* continuous PDFs, including non-rectangular ones.

Excel does not have a function for the continuous uniform distribution. **R**, however, allows us to integrate the continuous uniform PDF, `dunif`. For the current example we have:

```
integrate(function(x) dunif(x, min = 2.5, max = 3.0),
          lower = 2.5, upper = 2.8)

0.6 with absolute error < 6.7e-15
```

We can also use the uniform CDF, `punif`, directly

```
punif(2.8, min = 2.5, max = 3.0)

[1] 0.6
```

Assignment 3

Answer all questions in one MS Word document and submit to Moodle. At the beginning of the document include the assignment number, the date, your name and section number.

Use complete sentences when appropriate, and make sure any tables, figures and computer output you include adhere to class standards (see Syllabus).

PDFs and CDFs

1. (2 pts) Provide a definition for the term "probability density function" (PDF).

2. (2 pts) Provide a definition for the term "cumulative distribution function" (CDF).

3. (6 pts) Let X be a discrete random variable whose distribution is described by the function $f(x) = x/8$, if $x = 1, 2$, or 5, and $f(x) = 0$ otherwise.

a) Make a table similar to Table 3.1 to represent the problem. Include this in your homework with an appropriate caption.

b) What is $P(X = 3)$? That is, what is $f(3)$?

c) What is $P(X = 2)$? That is, what is $f(2)$?

d) What is $P(X \leqslant 2)$? That is, what is $F(2)$?

Binomial distribution

- Open **R**
- Load the *asbio* package by typing `library(asbio)` or by going to **Packages > Load packages > *asbio***.
- Type `book.menu()` in the **R** console.

4. (9 pts) To see a depiction of the binomial distribution go to **Chapter 3 > Pdf depiction**. Select **Binomial** and uncheck the **Show cdf** widget. Mac users, type: `see.bin.tck()` to access the GUI directly. Answer the following questions.

a) Is this distribution used to represent continuous or discrete random variables? Do you know this just by looking at the graph? Why?

b) How many parameters does the distribution have?

c) What does x on the x-axis represent?

d) What does $f(x)$ on the y-axis represent?

5. (5 pts) Alter the binomial distribution parameters in the binomial distribution GUI to create a Bernoulli distribution. Attach the resulting figure to your homework. Is the Bernoulli a special case of the binomial? Why?

6. (10 pts) You are working on a mark-recapture study of boreal toads, (*Bufo boreas*). You predict that there is a 60% chance of capturing a marked toad for each of 30 traps that you establish.

a) Define the binomial distribution represented by the problem. What are the parameter values for n and p? What is the x outcome of interest?

b) Use the binomial PDF to determine the probability that exactly 22 toads will be found in the 30 traps?

i) First, calculate the probability "by hand" using Excel or **R** to help. I recommend that you use **R** as this will facilitate later steps. Show your work by attaching screen shots.

- If you use Excel, you are allowed to use the function =FACT, which provides factorials, along with the basic mathematical operators for summation, subtraction, multiplication, and exponentiation.
- If you use **R** (recommended) you can use the function `choose(n, x)`, which gives $\binom{n}{x}$ directly, along with the basic mathematical operators for summation, subtraction, multiplication, and exponentiation.

ii) Confirm your calculation in (**i**) using =BINOMDIST (Excel), or `dbinom` (**R**). Show your work by attaching screen shots.

c) What is the probability that four or fewer toads will be found in the 30 traps?

i) First, calculate the probability "by hand" using Excel or **R** (recommended) applying the constraints mentioned above. Show your work by attaching screen shots.

ii) Confirm your calculation using =BINOMDIST (Excel), or `pbinom` (**R**). Show your work by attaching screen shots.

7. (1 pt) Use the binomial coefficient (Eq. 3.4) to calculate the total number of ways that 20 heads can occur in 50 coin tosses. Use **R** or Excel functions to help. Show work using snapshots.

Poisson distribution

8. (8 pts) To see a depiction of the Poisson distribution go to **Chapter 3 >Pdf depiction** in the *asbio* `book.menu`. Select **Poisson** and uncheck the **Show cdf** widget. Mac users, type: `see.pois.tck()` to access the GUI directly. Answer the following questions.

a) Is this distribution used to represent continuous or discrete random variables? Can you determine this by simply looking at the graph? Why?

b) How many parameters does the distribution have?

c) What does x on the X-axis represent?

d) What do $f(x)$ on the y-axis represent?

9. (15 pts) As noted earlier, the Poisson distribution is often used by biologists to quantify the randomness of organism counts in time and space. The following rules utilize the sample mean and sample variance (see the section describing the Poisson distribution in the this lab). We will explore the sample mean and variance with greater emphasis next week.

- random: counts occur at frequencies approximating the Poisson distribution (i.e. the sample variance equals the sample mean). Remember, in a Poisson distribution the variance will equal the mean, and both will equal the rate parameter, λ.
- clumped: The sample variance is greater than the sample mean. This usually indicates some sort of positive interaction between individuals (e.g., social insects), dispersal limitations, or patchiness of resources.
- regular: The sample variance is less than the mean. This is often interpreted as a negative or competitive interaction between individual organisms.

Clumped, random and regular spatial distributions are shown in Fig 3.5.

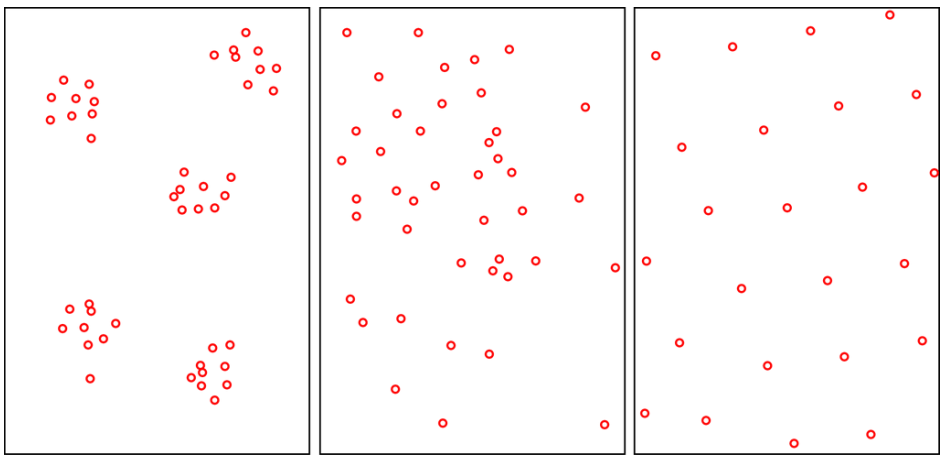


Figure 3.5. Potential spatial distributions of organisms: (a) clumped, (b) random, (c) regular.

You are studying a population of *Trigona dorsalis*, stingless bees that nest in trees in tropical dry forests. Hubbell & Johnson (1978) found that nests

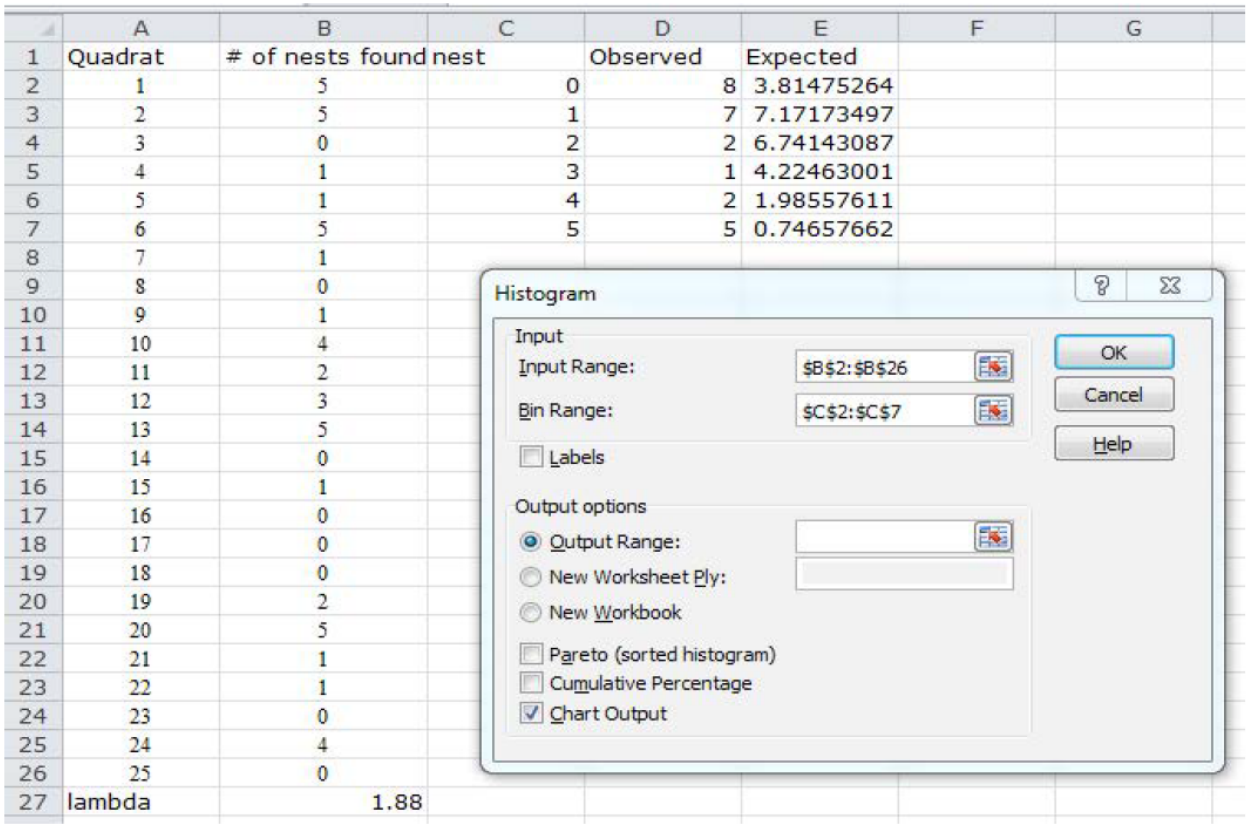
of the bees were randomly (Poisson) distributed in space. You wish to test if the nests in your study area are randomly distributed in space, or if there is evidence of territoriality (i.e., a repulsed (clumped) distribution). To test whether nests are random (Poisson) distributed, regular, or clumped, you lay out a regular grid of 25 equal-sized quadrats in Southeastern Venezuela dry-forest and count the nests in each quadrat. We will only use Excel for this Exercise. You should utilize the Excel spreadsheet provided for this question in the lab.

a) Use the function =AVERAGE in Excel to calculate the mean rate of counts based on your observed data. Calculate this value in the cell reserved for λ, B27, on your Excel spreadsheet (see Fig 3.6). We will use this sample mean as an estimate the Poisson distribution rate.

b) Assuming a Poisson distribution with the sample mean, 1.88, as the parameter value for λ, calculate the expected probability of the of occurrence 0 nests, 1 nests, 2 nests, up to the maximum number of nests observed in a quadrat, 5. This will require use of the function =POISSON.DIST. Multiply the resulting probabilities by 25 to get the expected counts of each number of nests, given 25 quadrats. Calculate these values in cells E2:E7 (see Fig 3.6).

c) Plot a histogram comparing the observed and expected frequencies.

- First, create a histogram showing frequencies of quadrats versus number of nests in a quadrat (e.g., 0, 1, 2, 3 ,4, and 5). This will require the Excel **Data Analysis** plug-in. Click on **Data** in the Excel pulldown menu. Does the **Data Analysis** toolbar show up on the right-side of the menu? If not, contact me.
  - Go to **Histogram** in the Data Analysis toolbar
  - In the **Input Range** put the observed number of nests (B2:B26).
  - In the **Bin Range** put the nest count categories (C2:C7)
  - Click on **Output Range** and put in some address on the worksheet
  - Click on **Chart Output** (see Fig 3.6).



	A	B	C	D	E	F	G
1	Quadrat	# of nests found	nest	Observed	Expected		
2	1	5	0	8	3.81475264		
3	2	5	1	7	7.17173497		
4	3	0	2	2	6.74143087		
5	4	1	3	1	4.22463001		
6	5	1	4	2	1.98557611		
7	6	5	5	5	0.74657662		
8	7	1					
9	8	0					
10	9	1					
11	10	4					
12	11	2					
13	12	3					
14	13	5					
15	14	0					
16	15	1					
17	16	0					
18	17	0					
19	18	0					
20	19	2					
21	20	5					
22	21	1					
23	22	1					
24	23	0					
25	24	4					
26	25	0					
27	lambda	1.88					

Figure 3.6. Creating a histogram of the observed number of *Trigona dorsalis* nests counts in Q 9.

  - Go to **Design**
  - Click on **Select Data** and click on **Add** in the **Select Data Source** dialog box.em Click on **Select Data** and click on **Add** in the **Select Data Source** dialog box.
  - Under **Series values**, insert the expected values you have calculated in cells E2:E7(Fig 3.7).
- Finally, insert the resulting histogram to the homework you will turn in. Make sure it is formatted to meet the course specifications.

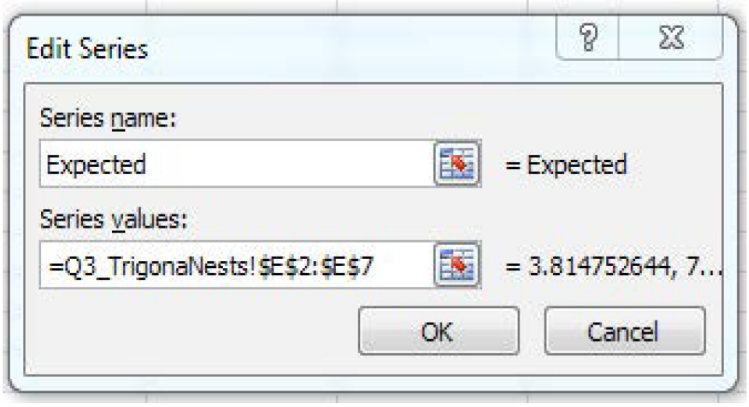


Figure 3.7. **Edit Series** dialog box in Excel.

d) Based on your figure, do *Trigona* nests in Venezuela appear to be clumped, regularly, or randomly distributed in space? Why?

e) Calculate the sample variance of the count data, i.e., cells B2:B26, using the Excel function =VAR. Compare the variance to the mean you calculated earlier, 1.88 nests, in support of your graphical analysis.

Continuous uniform distribution

10. (4 pts) To see a depiction of the continuous uniform distribution go to **Chapter 3 >Pdf** in the *asbio* `book.menu`. Select **Uniform** and uncheck the **Show cdf** widget. Mac users, type: `see.unif.tck()` to access the GUI directly. Answer the following questions.

a) How many parameters does the distribution have?

b) Is this a continuous distribution? Is it more difficult to calculate probability for continuous distributions? Why?

11. (8 pts) Assume $X \sim UNIF(1, 5.5)$.

a) Calculate $P(1 \leqslant X \leqslant 5.5)$ using calculus.

b) Verify your work in **R**.

c) Does your result make sense in the context of what you know about valid PDFs (see page 2 in this lab)? Why?

Q1 2pts, Q2 2pts, Q3 6pts, Q4 9pts, Q5 5pts, Q6 10pts, Q7 1pt, Q8 8pts, Q9 15pts, Q10 4pts, Q11 8pts. **Total pts: 70.**

Appendix: R-code used in this lab

Here are **R** functions for PDFs we will use this semester.

Name	Specification	Cont. or Discrete	**R** function	Parameter arguments
Binomial	$BIN(n, p)$	D	`dbinom(x, size, prob)`	$n =$ `size`, $p =$ `prob`
Chi-squared	$\chi^2(\nu)$	C	`dchisq(x, df)`	$\nu =$ `df`
F	$F(\nu_1, \nu_2)$	C	`df(x, df1, df2)`	$\nu_1 =$ `df1`, $\nu_2 =$ `df2`
Normal	$N(\mu, \sigma^2)$	C	`dnorm(x, mean, sd)`	$\mu =$ `mean`, $\sigma =$ `sd`
Poisson	$POI(\lambda)$	D	`dpois(x, lambda)`	$\lambda =$ `lambda`
t	$t(\nu)$	C	`dt(x, df)`	$\nu =$ `df`
Uniform	$UNIF(a, b)$	C	`dunif(x, min, max)`	$a =$ `min`, $b =$ `max`

4

Parameters and Statistics

Lab 4 Topics

1. Universal PDF Parameters
   - $E(X)$
   - $Var(X)$
2. Parameter Estimators
   - Location estimators: $\bar{X}$, Median
   - Scale estimators: S^2, S, IQR
   - Skewness and Kurtosis estimators: G_1, G_2
3. The effect of linear transformation on parameters and statistics

Parameters

A parameter is a fixed numeric characteristic describing an entire statistical population. Recall that a statistical population is a collection of all possible outcomes from a random variable.

$E(X)$

The **expected value** of X is a parameter denoted $E(X)$. $E(X)$ represents the arithmetic mean (see next section) of an entire population of outcomes. For a probability density function, $f(x)$, defining a discrete random variable X, the expected value of X is:

$$E(X) = \sum_x x f(x). \tag{4.1}$$

Recall that $f(x)$ means density, and will be equivalent to probability for discrete random variables (see Lab 3). Recall also that x represents an individual outcome from a random variable X. Higher order expectations can also be calculated using the approach given in Eq. 4.1. For example, to calculate $E(X^2)$ and $E(X^3)$ for a discrete random variable X, we would use:

$$\begin{aligned} E(X^2) &= \sum_x x^2 f(x), \\ E(X^3) &= \sum_x x^3 f(x), \\ &\textit{etc.} \end{aligned} \tag{4.2}$$

$Var(X)$

Another important parameter is the **variance of** X, denoted as $Var(X)$. This parameter quantifies the variability or the amount of "spread" in a distribution. For any discrete or continuous random variable X, the variance of X, is:

$$Var(X) = E(X^2) - E(X)^2. \tag{4.3}$$

The **standard deviation of** X is the positive square root of the variance of X. That is:

$$SD(X) = \sqrt{Var(X)}. \tag{4.4}$$

Example 4.1

Consider the PDF introduced in Lab 3 describing mountain goat (*Oreomnos americanus*) counts in an alpine meadow (Table 4.1).

Table 4.1. Example of a discrete PDF. Hypothetical mountain goat (*Oreomnos americanus*) counts, and their densities. Example taken from Aho (2014).

x = goat count	0	1	2	3	4
$f(x)$ = probability of a particular goat count	0.5	0.3	0.1	0.05	0.05

We have:

$$E(X) = \sum x f(x) = 0(0.5) + 1(0.3) + 2(0.1) + 3(0.05) + 4(0.05) = 0.85,$$

$$E(X^2) = \sum x^2 f(x) = 0(0.5) + 1(0.3) + 4(0.1) + 9(0.05) + 16(0.05) = 1.95.$$

Thus,

$$Var(X) = E(X^2) - E(X)^2 = 1.95 - 0.85^2 = 1.2275,$$

and

$$SD(X) = \sqrt{Var(X)} = 1.1079.$$

■

Going out two standard deviations from the mean in either direction on the number line will always create an interval that contains at least 75% of the population for any distribution (binomial, uniform, etc.). In a normal population distribution (introduced in lab 5), an interval ± two standard deviations from the mean will contain approximately 95% of the population. This normal distribution pattern is called the **empirical rule**.

Estimators

In general, we will be unable to observe all possible outcomes of a phenomenon we are interested in. Indeed, if we did we would not need statistics, at least not in an inferential capacity. We use estimating algorithms, i.e., **estimators**, to obtain estimates of population parameters. Individual **estimates** are called **statistics**. Thus, while we call the topic we are studying statistics, we also use this term for numerical summaries from samples. Through the use of statistics we attempt to describe an entire population using only sample data (Fig. 4.1). Occasionally our statistics will describe a clearly finite population (e.g., all students in BIOL 3316). If we have sampled this entire population, then statistics serve a descriptive instead of an inferential role. Specifically, they will define the behavior of the population instead of merely estimating it.

The usefulness of an estimator can be judged in three ways.

1. **Lack of bias:** If an estimator is **unbiased** then it will tend to neither over nor underestimate the parameter.

2. **Consistency:** If estimator is **consistent** then as the sample size increases, the precision of the estimator for a parameter increases.

3. **Efficiency:** If an estimator is **efficient** then it will provide a more precise (less variable) estimate of the parameter then other estimators for a particular sample size.

The can be shown using a mathematical construct called Chebyshev's inequality.

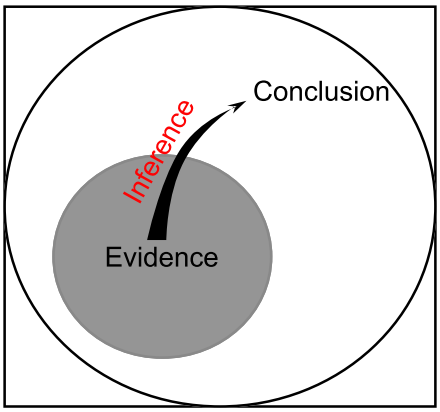


Figure 4.1. The process of making inference to a population (large white circle) using representative samples (smaller gray circle). In statistical analyses we often do this by calculaing sample statistics to estimate population parameters.

There are three types of **point estimators** (estimators that estimate a single value).

1. Measures of **location** estimate a typical or central value.
    - Examples include the sample mean, sample median and sample mode.
2. Measures of **scale** quantify data variability.
    - Examples include the sample variance, sample standard deviation and sample interquartile range.
3. Measures that quantify the **shape** of the data distribution.
    - Examples include the sample skew and sample kurtosis.

Sample Mean

We estimate $E(X)$, i.e., Eq. 4.1, using the **arithmetic mean**, also simply called the **sample mean**. The estimator is denoted $\bar{X}$, and pronounced "X bar." The sample mean will be unbiased and consistent for $E(X)$ for any distribution, and will be a maximally efficient estimator for $E(X)$ when the distribution underlying the data is normal.

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{4.5}$$

Conventionally capital letters are used for an estimator. For instance, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Conversely, lower case letters are used for a particular estimate, based on data. For instance, if data outcomes, x, are $6, 2$ and 1, then $\bar{x} = 3$.

Sample Variance

We estimate $Var(X)$, i.e., Eq. 4.3, using the **sample variance**, denoted S^2. The sample variance is an unbiased and consistent estimator for $Var(X)$ for any distribution.

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \tag{4.6}$$

As we learned in Lab 1, the numerator in Eq. 4.6 is called the **sum of squares**. The denominator, $n-1$, represents the **degrees of freedom**, i.e., the number of independent pieces of information we have to estimate the true variance. In the case that the entire population is sampled we would calculate the *population* variance by dividing by n instead of $n-1$.

We estimate $SD(X)$ using the **sample standard deviation**, denoted S.

$$S = \sqrt{S^2} \tag{4.7}$$

S is biased (low), but is consistent for $SD(X)$. S allows increased interpretability, compared to S^2. The units of S will be in the units of the original observations. Conversely, the units of S^2 will be the original units of measurement, squared. S also allows inferences using the empirical rule (Lab 5).

Example 4.2

As an example, assume that we don't know everything about the mountain goat count probability distribution for a meadow shown in Table 4.1, and that we need to rely on data (and statistics) to describe the distribution. We sample the meadow independently 15 times and get the following goat counts: 1, 0, 0, 0, 1, 0, 0, 2, 0, 0, 0, 1, 0, 2, 1.

We have:

$$\bar{x} = \frac{1+0+0+0+1+0+0+2+0+0+0+1+0+2+1}{15} = \frac{8}{15} = 0.533.$$

$$s^2 = \frac{(1-0.533)^2 + (0-0.533)^2 + (0-0.533)^2 + \ldots + (1-0.533)^2}{14} = \frac{7.733}{14} = 0.552.$$

$$s = \sqrt{s^2} = 0.743.$$

The Excel function function for the sample mean, sample variance and sample standard deviation are =AVERAGE, =VAR, and =STDEV. The latter two functions are equivalent to the functions =VAR.S and =STDEV.S, respectively. Thus, for the data above we could calculate the sample variance using the approach shown in Fig 4.2:

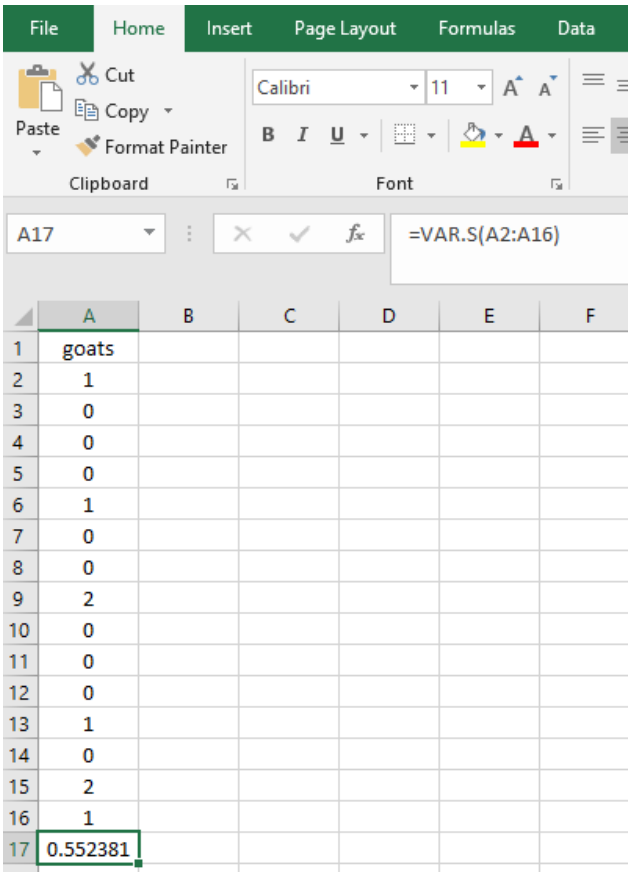


Figure 4.2. Calculating the sample variance in Excel.

The **R** functions for the sample mean, sample variance, and sample standard deviation are `mean`, `var` and `sd`, respectively. Thus, we could do something like:

```
goats <- goats <- c(1,0,0,0,1,0,0,2,0,0,0,1,0,2,1)
mean(goats)

[1] 0.5333333

var(goats)

[1] 0.552381

sd(goats)

[1] 0.7432234
```

■

Robust Estimators

Oftentimes data will contain **outliers** (unusual observations) that will negatively affect valid inferences. The mean, and measures that rely on the mean (i.e., the variance and standard deviation), are not **resistant**. That is, these statistics will be strongly affected by outliers. **Robust measures** of location and scale (those resistant to outliers) include the sample median and the interquartile range, respectively.

Sample Median

The **population median** is the **50th percentile** of its PDF. That is, 50% of a distribution will lie below and above its median. The **sample median** is the middle value from a set of n ordered responses (i.e., data arranged from low to high). If n is odd then the median is the middle response from a set of ordered data. If n is even, then the median is the average of the two middle ordered values. The median is only 64% as **efficient** as the arithmetic mean. That is, median estimates will vary more than mean estimates, when sampling from the same normal population (see Lab 5). However, $\bar{X}$ has a **breakdown point** of 0% because it can be made arbitrarily large with a single outlier. The same effect to the median would require that 50% of data were outliers. Thus, the median has a break down point of 50%.

Example 4.3

Consider nine counts of medium ground finches (*Geospiza fortis*) taken at Daphne Major Island in the Galapagos from 1976 to 1984 (Table 4.2).

Table 4.2. Medium ground finch (*Geospiza fortis*) counts from Daphne Major. Data from Gibbs & Grant (1987).

Year	Finch Count
1976	1220
1977	400
1978	380
1979	298
1980	280
1981	200
1982	297
1983	280
1984	1250

We first order the data from smallest to largest. The ordered data are:

$$200, 280, 280, 297, 298, 380, 400, 1220, 1250$$

Next, we find the middle of the data. This will be the $(n + 1)/2$ value. Because $n = 9$, we have:

$$(9 + 1)/2 = 5.$$

The fifth ordered value is 298. Thus, the sample median is 298.

The Excel function function for the sample median is =MEDIAN, so we could calculate sample median using the approach shown in Fig 4.3.

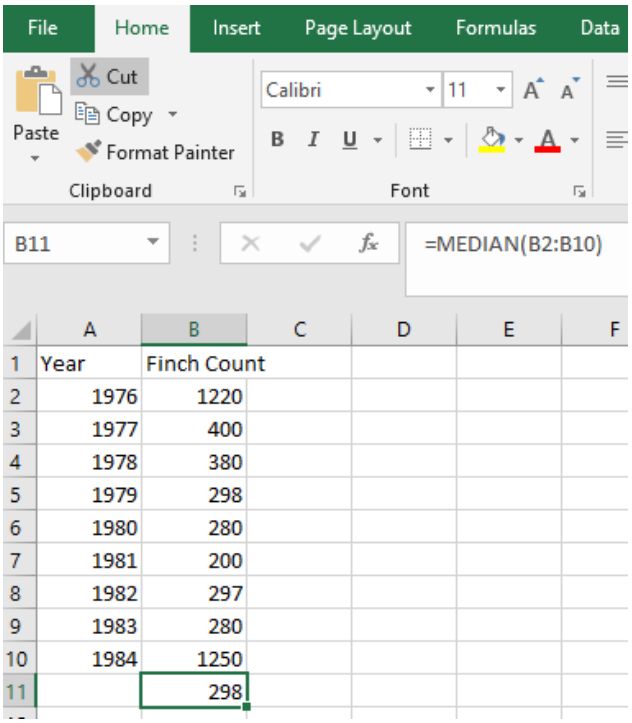


Figure 4.3. Calculating the sample median in Excel.

The **R** function for the median is `median`:

```
finches <- c(1220,400,380,298,280,200,297,280,1250)
median(finches)

[1] 298

# Note, sorting by hand we have:
sorted.finches <- sort(finches)
sorted.finches

[1]  200  280  280  297  298  380  400 1220 1250

n <- length(finches)
(n - 1)/2

[1] 4

# The fifth obs. is 298
sorted.finches[5]

[1] 298
```

■

The mean and the median can be contrasted by the fact that the median is the central value in an ordered distribution, while the mean is the "center of gravity" for the distribution. Note that to "balance" the distribution in Fig. 4.4, the mean is pulled in the direction of the long right tail.

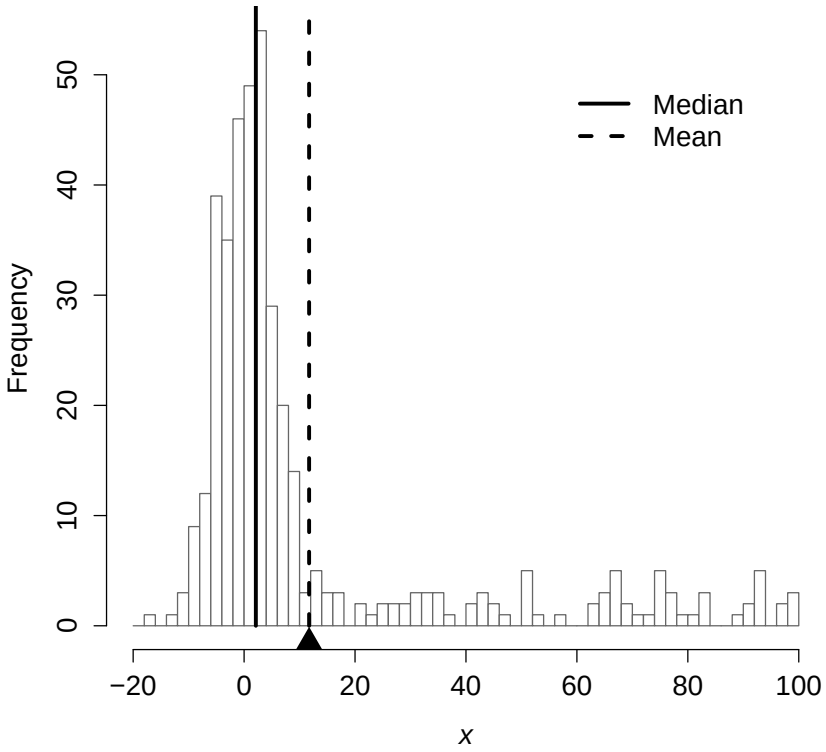


Figure 4.4. Comparison of the mean and the median for a right skewed distribution.

Sample IQR

Because they are calculated from squared deviations, S and S^2 are even more strongly affected by outliers than the sample mean. Although, like the mean, each has a breakdown point of 0%. A robust estimator of scale is the **sample interquartile range** (**IQR**).

For any probability distribution, the region between the **1st quartile** and **3rd quartile** contains the middle 50% of a distribution. One quarter of the distribution will be *below* the 1st quartile and one quarter of the distribution will be *above* the third quartile.

To calculate the sample interquartile range we find the medians of the ordered data that lie below and above the median. These are Q_1 and Q_3, which are estimators for the first and third population quartiles, respectively. The population median is the second population quartile, so its estimator is often referred to as Q_2. Q_1 is subtracted from Q_3 to obtain the sample interquartile range, IQR (Fig 4.5):

$$IQR = Q_3 - Q_1. \tag{4.8}$$

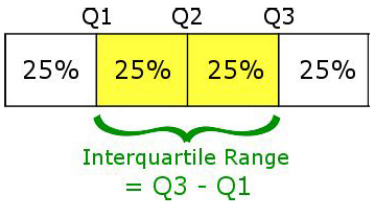


Figure 4.5. Heuristic for the interquartile range.

Excel does not have a function for IQR, but the **R** function for IQR is `IQR`.

```
IQR(finches)

[1] 120
```

Skewness and Kurtosis

Population Skew and Kurtosis

We can also describe distributions with respect to their shape. Two important shape descriptors are skewness and kurtosis. **Skew** describes the symmetry of a distribution whereas **kurtosis** describes its peakedness. The population skewness and kurtosis (i.e., parameters) are denoted as γ_1 and γ_2 respectively, and are derived from expected values:

$$\gamma_1 = \frac{E[(X - E(X))]^3}{[SD(X)]^3}, \tag{4.9}$$

$$\gamma_1 = \frac{E[(X - E(X))]^4}{[Var(X)]^2}. \tag{4.10}$$

For a **symmetric distribution**, skewness will equal zero; i.e., $\gamma_1 = 0$. In this case, the population mean and median will be equal because the tails of the distribution will be of equal length and symmetric. Conversely, if $\gamma_1 \neq 0$, the distribution is **asymmetric** and the mean will be drawn towards the long tail. That is, the mean will be to the right of the median in a distribution with a longer right tail (Fig. 4.4). A distribution with a "long" right-hand tail, and a squashed left-hand tail will be **positively-skewed**, resulting in $\gamma_1 > 0$. Conversely, a distribution with a "long" left-hand tail and a squashed right-hand tail will be **negatively-skewed**, resulting in $\gamma_1 < 0$. If a distribution is normal (Lab 5), it will be **mesokurtic**, and γ_2 will equal 3. Because of this, the parameter γ_{2_excess}, defined to be:

$$\gamma_{2\_excess} = \gamma_2 - 3. \tag{4.11}$$

Note that estimates from the function `IQR` are based on the **R** function `quantile` (type `?quantile` for more information). As a result, estimates from `IQR` may differ slightly from those obtained using Eq. 4.8.

The parameter γ_{2_excess} is generally used to define the kurtosis of distribution. Normal distributions will have $\gamma_{2\_excess} = 0$, strongly peaked (**leptokurtic**) distributions will have $\gamma_{2\_excess} > 0$, and flat-looking (**platykurtic**) distributions will have $\gamma_{2\_excess} < 0$.

Sample Skew and Kurtosis – Method of Moments

We can estimate γ_1 and γ_2 with the estimators G_1 and G_2, respectively. The sample skew and kurtosis, G_1 and G_2, are **method of moments** estimators. That is, they are based on Eq. 4.12, in which the jth sample moment is given by:

$$m_j = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^j \tag{4.12}$$

To obtain the first or central sample moment, m_1, we let $j = 1$ in Eq 4.12. Thus, m_1 is the average deviation of all observations from the sample mean. By definition $m_1 \equiv 0$. That is, the sum of the differences of observations from the sample mean must equal 0. To obtain second sample moment m_2 we let $j = 2$ in Eq. 4.12, and so on. Note that m_2 is a biased (low) version of the sample variance. We calculate G_1 as:

$$G_1 = \frac{m_3}{m_2^{3/2}}. \tag{4.13}$$

we calculate G_{2_excess} (the estimator for γ_{2_excess}) using:

$$G_{2\_excess} = \frac{m_4}{m_2^2} - 3. \tag{4.14}$$

Example 4.4

Table 4.3 and subsequent calculations show the steps necessary to calculate the sample skew, G_1, and sample excess kurtosis, G_{2_excess}, by hand.

Table 4.3. Calculations required for skewness and kurtosis estimates for the finch data (see Table 4.2).

	Finch counts	$(x_i - \bar{x})^1$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^3$	$(x_i - \bar{x})^4$
	200	-311.667	97136.11	-30274088	9.44E+09
	280	-231.667	53669.44	-12433421	2.88E+09
	280	-231.667	53669.44	-12433421	2.88E+09
	297	-214.667	46081.78	-9892222	2.12E+09
	298	-213.667	45653.44	-9754619	2.08E+09
	380	-131.667	17336.11	-2282588	3.01E+08
	400	-111.667	12469.44	-1392421	1.55E+08
	1220	708.3333	501736.1	355396412	2.52E+11
	1250	738.3333	545136.1	402492162	2.97E+11
$\sum_{i=1}^{n} x_i$	4605	0	1372888	679425793	5.69E+11
m_j		0	152543.1	75491755	6.32E+10
$\bar{x}$	511.666667				

$$G_2 = \frac{m_3}{m_2^{3/2}} = \frac{75491755}{152543.1^{3/2}} = 1.27,$$

$$G_{2\_excess} = \frac{m_4}{m_2^2} = \frac{6.3197 \times 10^{10}}{152543.1^2} = -0.284.$$

We could speed up these "by hand" calculations using **R**.

```
x.bar <- mean(finches)
m2 <- sum((finches-x.bar)^2)/9
m3 <- sum((finches-x.bar)^3)/9
m4 <- sum((finches-x.bar)^4)/9

G1 <- m3/m2^(3/2)
G1

[1] 1.267099

G2.excess <- m4/m2^2 - 3
G2.excess

[1] -0.2841179
```

The Excel functions for skew and kurtosis, =SKEW and =KURTOSIS, are calculated in a different way than shown in Eqs. 4.13 and 4.14. However, we can confirm our results by using the functions `skew` and `kurt` from the library *asbio*.

```
library(asbio)
skew(finches, method = "moments")

[1] 1.267099

kurt(finches, method = "excess")

[1] -0.2841179
```

■

For more information type `?skew` or `?kurtosis` after loading the package *asbio*. Excel calculates the *unbiased* estimators for γ_1 and γ_{2_excess}. By default, `kurt` and `skew` also calculate unbiased estimates, resulting from the default argument: `method = "unbiased"`.

Linear Transformations

When we perform a **linear transformation** on a random random variable or dataset we add or multiply constants to all outcomes from the random variable or dataset. Linear transformation will result in straightforward changes to the concomitant parameters and statistics.

If X is a random variable and a and b are constants, then

$$E(a + bX) = a + bE(X), \tag{4.15}$$

$$Var(a + bX) = b^2 Var(X), \tag{4.16}$$

$$SD(a + bX) = bVar(X). \tag{4.17}$$

The left side of Eqs. 4.15, 4.16 and 4.17 show the form of the linear transformation. Specifically, we are adding a constant a to the original random variable X and multiplying by a constant, b. Eq. 4.15 expresses the mean of the random variable following this transformation, Eq. 4.16 gives the variance, and 4.17 gives the standard deviation.

Let's consider the case that $a \neq 0$ and $b = 1$ in Eqs 4.15, 4.16 and 4.17. Adding or subtracting a from a random variable results in a new random variable with a mean equivalent to the old mean plus or minus a. The new variance and standard deviation, however, will be identical to the old variance and standard deviation.

Now, let's consider the case that $a = 0$ and $b \neq 1$ or 0 in Eqs. 4.15, 4.16 and 4.17. Multiplying a random variable by b results in a new random variable with a mean equivalent to the old mean times b. The new variance will be the old variance times b^2, and the the new standard deviation will be the old standard deviation times b.

These principles also hold for statistics (Table 4.4.)

Table 4.4. Sample mean, variance, and standard deviation after data are linearly transformed by adding or subtracting by a constant a or multiplying by a constant, b.

	Original data: x	$x \pm a$	$x \times b$
Sample mean	$\bar{x}$	$\bar{x} \pm a$	$\bar{x} \times b$
Sample variance	s^2	s^2	$s^2 \times b^2$
Sample standard deviation	s	s	$s \times b$

Example 4.5

Assume that over twenty years the mean maximum June temperature in a desert location is 95°F, with a variance of 5°F^2. What is the mean and variance in degrees C? Because:

$$\text{C}^\circ = \frac{5}{9}(\text{F} - 32)^\circ,$$

converting from Fahrenheit to Celsius is a linear transformation. We have:

$$\bar{x}_{\mathrm{C}} = \frac{5}{9}(95 - 32)^{\circ} = 35^{\circ}\mathrm{C},$$

$$s^2_{\mathrm{C}} = \left(\frac{5}{9}\right)^2 \times 5^{\circ} = 1.54^{\circ}\mathrm{C}^2.$$

■

Assignment 4

Answer all questions in one MS Word document and submit to Moodle. At the beginning of the document include the assignment number, the date, your name and section number.

Use complete sentences when appropriate, and make sure any tables, figures and computer output you include adhere to class standards (see Syllabus).

Parameters

1. (3 pts) What is a parameter? Give an example.

2. (3 pts) What is an estimator? Give an example.

3. (8 pts) Let X be a discrete random variable whose distribution is described by the function $f(x) = x/8$ if $x = 1, 2$, or 5, and $f(x) = 0$ otherwise. Find:

a) $E(X)$

b) $E(X^2)$

c) $Var(X)$

d) $SD(X)$

4. (10 pts) For the soil %N data from Table 1 from Lab 1, calculate summary statistics "by hand" using either Excel or **R** (recommended) to help. Include snapshots to show your work.

- In Excel you are allowed to use only sorting functionality built into Excel, the function =SUM, and the mathematical characters -, +, ^, /, and *,

- In **R** you can use only the functions `sort`, `sum`, and the mathematical characters `-`, `+`, `^`, `/`, and `*`. If you are using **R**, here is the %N data:

```
N <- c(15.2, 14.2, 16.2, 13.1, 10.2, 15.5, 11.1, 14.9, 12.3, 10.1)
```

a) Sample mean, $\bar{X}$

b) Sample median

c) Sample variance

d) Sample skew, G_1

e) Sample excess kurtosis, G_{2_excess}

5. (3 pts) Check your answer from Q. 4d, and 4e in **R**. To do this:

- Open **R**
- Load the **asbio** library by typing: `library(asbio)`
- For the method of moments skew, type:

```
skew(N, method = "moments")
```

- To get the excess estimate, G_{2_excess}, type:

```
kurt(N, method = "excess")
```

Take snapshots to show your work.

6. (4 pts) Interpret the skewness and kurtosis values in Q. 5.

7. (5 pts) In Excel, calculate summary statistics for the birth weights for the smokers and the nonsmoker groups separately (i.e., summarize each group) using the data from week 1. Calculate the sample mean, sample median, sample standard deviation, sample variance, sample skew and sample kurtosis.

NOTE: To speed this process up, use the Descriptive Statistics Wizard under **Tools** > **Data Analysis** > **Descriptive Statistics**, which allows calculation of all of these statistics simultaneously. Paste the Excel output into the document you will hand in.

8. (4 pts) Repeat question 7 using **R** by following the steps below.

- Download the `birthweight.csv` dataset from Lab 1 into **R**. To do this, type: (or paste) the following code into **R**, and then navigate to the file:

```
birthweight <- read.csv(file.choose())
```

- To get summary statistics with respect to levels in a categorical variable (e.g. smoker and non-smoker) we can use the function `tapply`. For example, to get the sample means for both smoker and nonsmoker groups individually and simultaneously, you could type:

```
tapply(birthweight$bwt, birthweight$smoke, mean)
```

- To get all of the stats for each group simultaneously, paste the following function into **R**.

```
stats <- function(x, digits = 5){
mean <- mean(x)
median <- median(x)
kurt <- kurt(x, "excess")
skew <- skew(x)
var <- var(x)
sd <- sd(x)
round(t(cbind(mean, median, kurt, skew, var, sd)), digits)
}
```

- Now type:

```
tapply(birthweight$bwt, birthweight$smoke, stats)
```

Take snapshots of the output to show your work.

9. (6 pts) In **R** add 10 to each birth weight and calculate summary stats. To do this, first type:

```
bwt10 <- birthweight$bwt + 10
```

Then calculate summary stats for `bwt10` by groups, as before, by typing:

```
tapply(bwt10, birthweight$smoke, stats)
```

a) How does this affect the mean of the smoker birth weight?

b) How does this affect the standard deviation of smoker birth weight?

c) How does this affect the variance for smoker birth weight?

10. (6 pts) In **R** multiply each birth weight by 10 and calculate summary stats. To do this, first type:

```
bwt.times.10 <- birthweight$bwt * 10
```

Then calculate summary stats for `bwt10` by groups, as before, by typing:

```
tapply(bwt.times.10, birthweight$smoke, stats)
```

a) How does this affect the mean of the smoker birth weight?

b) How does this affect the standard deviation of smoker birth weight?

c) How does this affect the variance for smoker birth weight?

11. (8 pts) This question relates to appropriateness of statistical measures.

a) Why might we want to use the median instead of the mean as a measure of the centrality for a sample?

b) Calculate the sample mean using made up data with and without an outlier using the function `mean` in **R**. Provide snapshots to show work.

c) Why might we want to use the interquartile range instead of the variance as a measure of sample variability?

d) Calculate the sample variance using made up data with and without an outlier using the function `var` in **R**. Provide snapshots to show work.

Q1 3pts, Q2 3pts, Q3 8pts, Q4 10pts, Q5 3pts, Q6 4pts, Q7 5pts, Q8 4pts, Q9 6pts, Q10 6pts, Q11 11pts. **Total pts: 63.**

Appendix: R-code used in this lab

- Simple statistical functions.

The functions `skew` and `kurt` require the package *asbio*.

Function	Acronym	Description
`mean(x)`	$\bar{X}$	Arithmetic mean of x
`median(x)`	Q_2	Median of x.
`sd(x)`	S	Standard deviation of x
`var(x)`	S^2	Variance of x.
`IQR(x)`	IQR	Interquartile range of x.
`skew(x)`	G_1	Skew of x
`kurt(x)`	G_2	Kurtosis of x

- Summary statistics by groups:

Operator	Operation	To	We type
`tapply(X, INDEX, FUN)`	Summarize data in `X` for levels in `INDEX`, with respect to stats in `FUN`	Calculate means of data in `y` for levels in `x`	`tapply(y, x, mean)`

5

Normal Distribution, Sampling Distributions, Confidence Intervals

Lab 5 Topics

1. Normal distribution
    - Standard normal distribution
    - Empirical rule
2. Adding and subtracting normal random variables
3. Sampling distributions
    - Central limit theorem
4. Confidence interval for μ
5. Sample size adequacy

The Normal Distribution

The most commonly used distribution in statistics is the **normal distribution**. It is popular for three reasons. First, normal distributions are symmetric Thus, given a distribution centered at 0, $P(X < x)$ will be equivalent to $P(X > x)$. This facilitates the interpretation and probability calculations for lower and upper tailed hypothesis tests (Lab 6). Second, normal distributions are simply very useful for describing many real biological variables (e.g., height, weight, girth, etc.). Third, the sampling distributions of many statistics become normally distributed given large sample sizes, regardless of the distributional characteristics of the sampled parent distribution. Sampling distributions are a central topic of this lab.

If a random variable, X, follows a normal distribution, it will have the PDF:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)} \tag{5.1}$$

where $\sigma > 0$, $\mu \in \mathbb{R}$, $x \in \mathbb{R}$.

The normal distribution has two parameters, μ and $\sigma\sigma$, which represent the mean and the standard deviation of the PDF, respectively. That is, if a random variable X is normally distributed, $E(X) = \mu$ and $Var(X) = \sigma^2$ (Lab 4). If X is normal, we denote this as: $X \sim N(\mu, \sigma^2)$. The distribution $N(0, 1)$ is shown in Fig. 5.1.

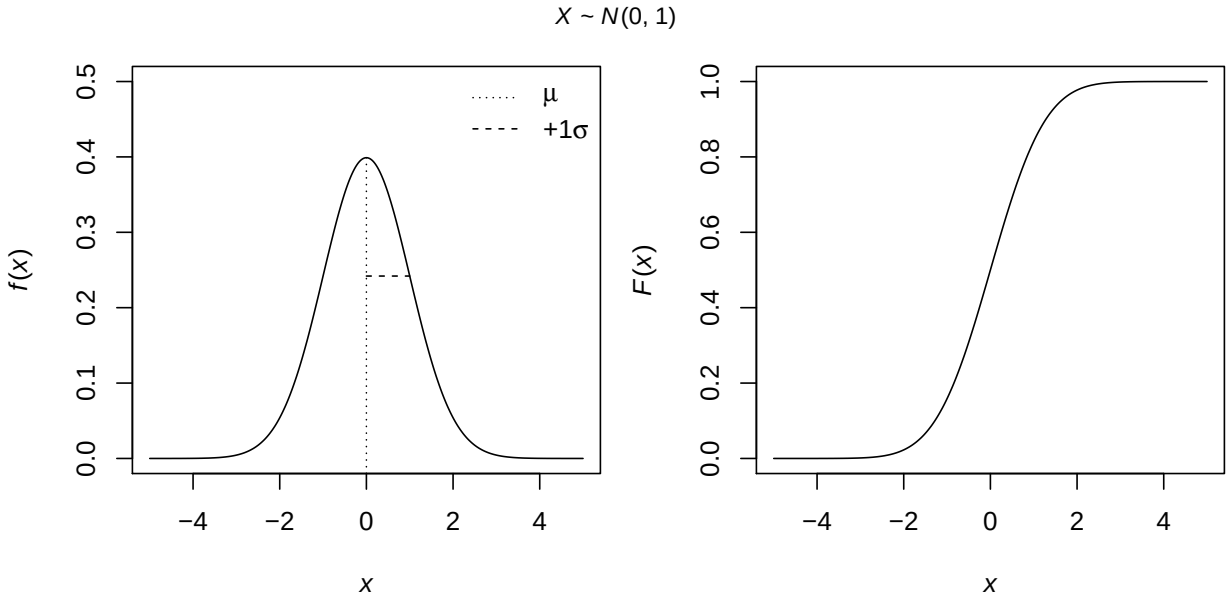


Figure 5.1. The PDF and CDF of the distribution $N(0, 1)$.

Note that the normal PDF has a symmetric bell-shaped appearance whereas the CDF is sigmoidal (S-shaped).

Example 5.1

What is the density $f(3)$, given $X \sim N(5, 4)$? We have:

$$f(3) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{3-5}{2}\right)} = 0.120985.$$

To calculate normal densities (and probabilities) in Excel we use the function =NORM.DIST. It requires four arguments:

1. The value of x.
2. The mean, μ.
3. The standard deviation, σ.
4. Whether or not you want the CDF (TRUE) or the PDF (FALSE).

Thus, we have:

=NORM.DIST(3, 5, 2, FALSE) = 0.120985

In **R** we use the function `dnorm` to get density and `pnorm` to get cumulative (left tailed) probabilities. The functions have the same first three arguments as the Excel function =NORM.DIST. The last argument in =NORM.DIST is not necessary, because `dnorm` gives density and `pnorm` gives cumulative probabilities.

```
dnorm(3, 5, 2)

[1] 0.1209854
```

■

The Empirical Rule

In a normal probability distribution, the interval $\pm 1\sigma$ from μ contains approximately 68% of the distributional area, the interval $\pm 2\sigma$ from μ contains approximately 95% of the area of the distribution, and the interval $\pm 3\sigma$ from μ contains 99.7% of the area (Fig. 5-2). This is known as the **empirical rule** (Fig 5.2).

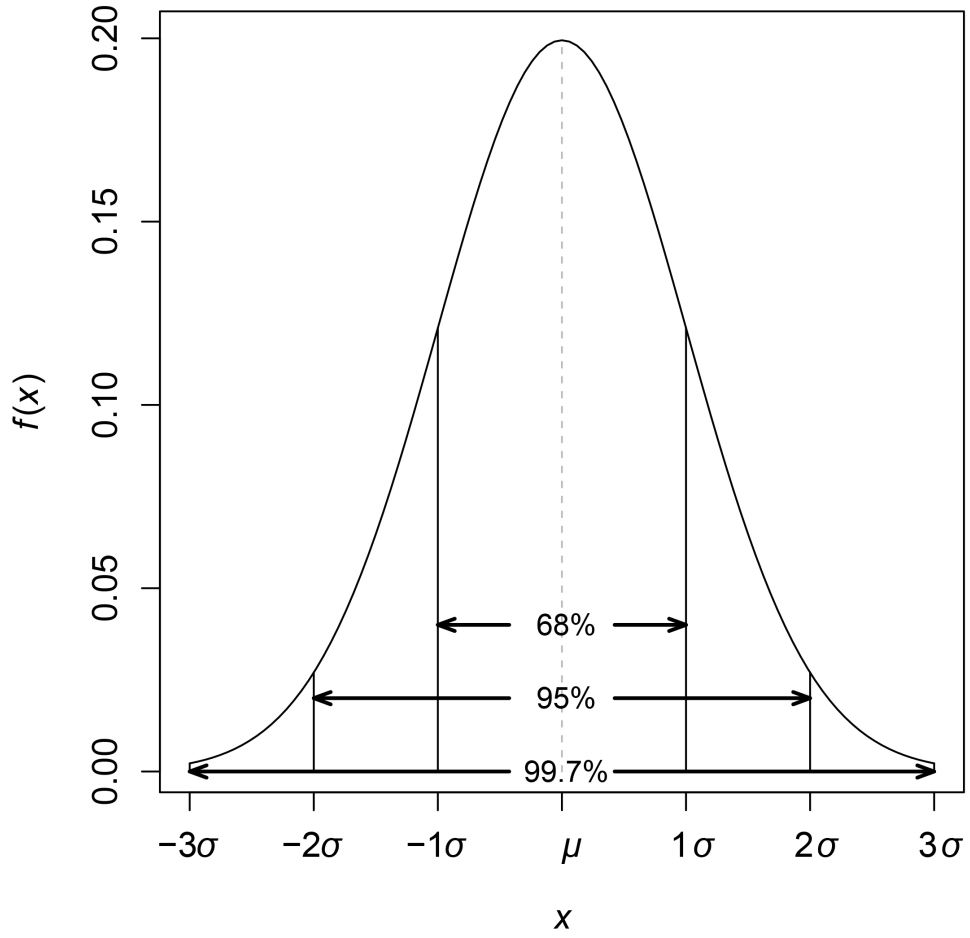


Figure 5.2. Standard deviations in a normal distrinition and the empirical rule.

Using the characteristics of a normal distribution we can find the area under a normal curve, corresponding to particular ranges of outcomes, to find the corresponding probabilities. For instance, the probability of outcomes greater than or equal to a particular value. x.

The Standard Normal Distribution

The **standard normal distribution** or Z **distribution** is a normal distribution in which $\mu = 0$, and $\sigma^2 = \sigma = 1$. Thus, if a random variable X follows a standard normal distribution, we write this as $X \sim N(0, 1)$ (see Fig. 5.1).

We can standardize any normal distribution, $X \sim N(\mu, \sigma^2)$, to be a standard normal distribution, $Z \sim N(0, 1)$, using Equation 5.2

$$Z = \frac{X - \mu}{\sigma} \tag{5.2}$$

where μ is the mean and σ is the standard deviation of the un-standardized normal variable, X. Once standardized, outcomes in the original variable are expressed as standard deviations away from the mean of the standard normal distribution.

An outcome, x, from any normal distribution becomes a standard normal outcome (***z*-score**) by applying:

$$z = \frac{x - \mu}{\sigma}.$$

Historically, the standard normal distribution arose because of the practical need for a single normal distribution in statistical textbooks and manuals, for use in inference. Given a single normal distribution, probabilities derived from integration could be placed in look-up tables for direct application or interpolation. The advent of modern statistical software has made calculation of probabilities for any normal distribution a simple endeavor, and has thus decreased the overriding need for the Z-distribution.

Example 5.2

The heights of young women (20-23 years of age) in the United States are normally distributed with $\mu = 64.5$ inches and $\sigma = 2.5$ inches. What proportion of women are less than or equal to 60 inches tall. That is, what is $P(X \leqslant 60)$, given $X \sim N(64.5, 6.25)$ (Fig. 5.3)? Importantly, the probability of women $\leqslant 60$ inches tall is equivalent to the proportion of women < 60 inches tall if X is continuous. This is because probabilities for discrete outcomes, $P(X = x)$, will equal zero.

```
library(asbio)
shade.norm(60, mu = 64.5, sigma = 2.5)
```

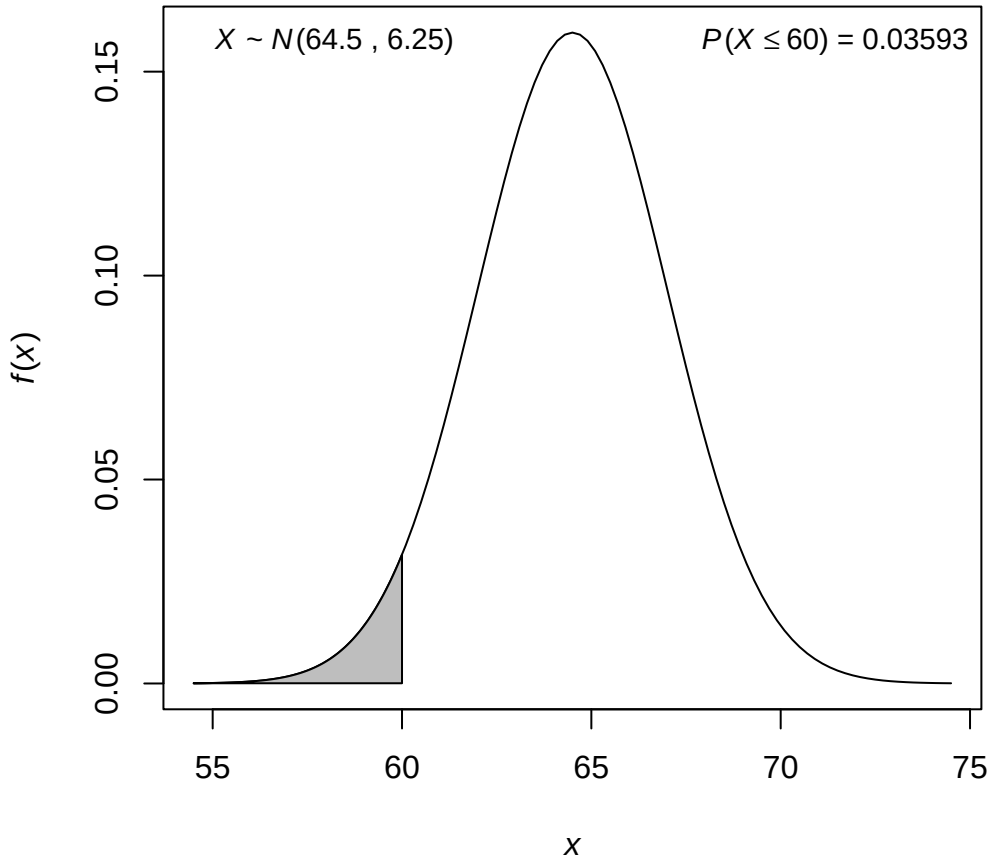


Figure 5.3. Distribution of young women's heights, $X \sim N(64.5, 6.25)$. It is always helpful to draw the distribution and denote the area of the curve you are interested in.

To calculate this probability, we can

1. Standardize X, with respect to the outcome 60.

$$P(X \leqslant 60) = P\left(Z \leqslant \frac{60 - 64.5}{2.5}\right) = P(Z \leqslant 1.8)$$

   Following the Z transformation, a height of 60 inches is equivalent to -1.8. This means that 60 inches is 1.8 standard deviations below the mean height of 64.5 inches.

2. We find the proportion of the Z distribution less than or equal to -1.8. We can do this by using either Excel or **R**.

=NORM.DIST(-1.8, 0, 1, TRUE) = 0.0359

```
pnorm(-1.8) # by default, pnorm uses a standard normal PDF

[1] 0.03593032
```

3. We find that $P(Z \leqslant -1.8) = 0.0359$. Therefore the proportion of young women less than or equal to 60 inches tall is 0.0359. That is, approximately 3.6% of women are less than or equal to 60 inches in height.

4. **Note:** we could have actually bypassed the whole Z-transformation process and used the original normal distribution to calculate the identical probability.

=NORM.DIST(60, 64.5, 2.5,TRUE) = 0.0359

```
pnorm(60, 64.5, 2.5)

[1] 0.03593032
```

■

Adding and Subtracting Normal Random Variables

Linear combinations of normal random variables will also be normally distributed. In particular, let X and Y be independent normal random variables: $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$, and let $Q = X + Y$, then,

$$Q \sim N(\mu_x + \mu_Y, \sigma_X^2 + \sigma_Y^2). \tag{5.3}$$

Further, if $Q = X - Y$

$$Q \sim N(\mu_x - \mu_Y, \sigma_X^2 + \sigma_Y^2). \tag{5.4}$$

Note that the variances are still added to get the variance of Q in Eq. 5.4, even though X and Y are being subtracted from each other.

Sampling Distributions

If you were to randomly sample a population many times with the same sized sample, say $n = 10$, and calculate a sample mean, $\bar{X}$, for each of those samples, those means would constitute a **sampling distribution**. It turns out that if the **parent distribution** (the one we sample from) has a mean, μ, and a variance, σ^2, then sampling distribution of $\bar{X}$ will

always have a mean of μ and a variance of σ^2/n, and the standard deviation for the sampling distribution of $\bar{X}$ is $\sigma/\sqrt{n}$. We call this the standard deviation $\bar{X}$, the **standard error** of the mean and denote it as $\sigma_{\bar{X}}$.

Central Limit Theorem

A very important tenet for inferential statistics is the **central limit theorem**. It concerns the parameters of the sampling distribution of $\bar{X}$. It is a fact that if the parent population has the distribution $N(\mu, \sigma^2)$, then $\bar{X}$ will also be normal: $\bar{X} \sim N(\mu, \sigma^2/n)$. The central limit theorem states that even if the shape of the parent population is *not* normal, if the sample size is sufficiently large, (i.e. $n \geqslant 30$) the the sampling distribution of the mean will be approximately normal. This can be stated summarily as:

$$\bar{X} \xrightarrow{d} N(\mu,\ \sigma^2/n), \tag{5.5}$$

where $\xrightarrow{d}$ mean "converges in distribution."

The implications of this are profound. The parent population can have essentially any distribution shape, but if the sample size is sufficiently large we can safely assume that the sampling distribution of $\bar{X}$ is approximately normally distributed. Consider Fig 5.4 in which an exponential parent distribution is sampled to obtain sampling distributions of $\bar{X}$ for six samples sizes. A sample size of one is used to create the first plot . Because each mean is equivalent to an individual observation from $EXP(1)$, the plot simply re-displays the parent distribution which is strongly positively skewed. As predicted by the central limit theorem, the sampling distribution of $\bar{X}$ becomes increasingly normal as sample size increases.

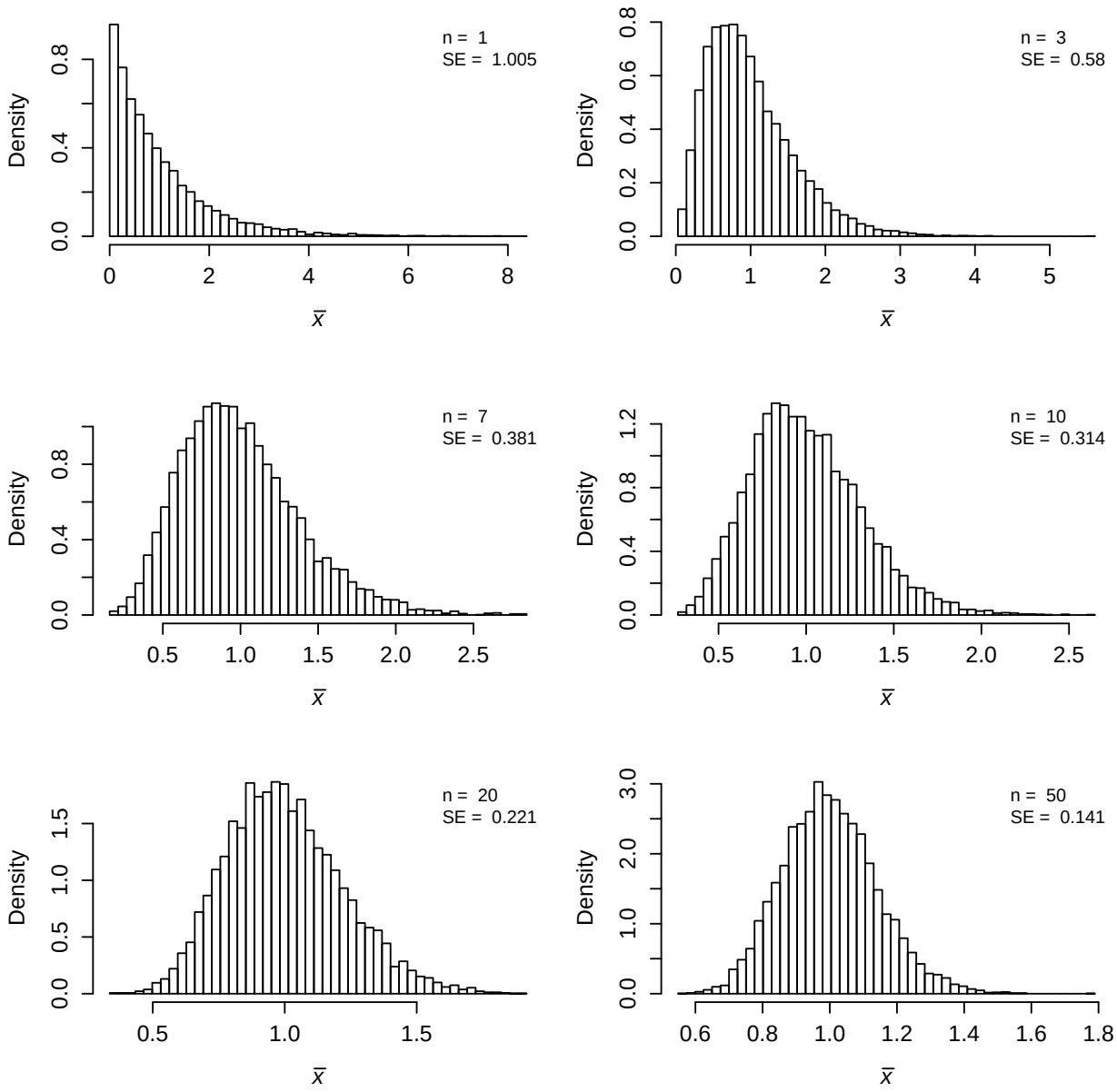


Figure 5.4. Empirical distributions of $\bar{X}$ for sample sizes of 1, 3, 7, 10, 20 and 50 from an exponentially distributed parent distribution, $EXP(1)$. Each histogram represents 10,000 random means.

Note that while the sample size $n = 30$ is often used as a cutoff for adequate sample size, this number will actually be determined by the shape of the parent distribution. The more normal the parent distribution, the smaller the sample size that will be required for a normal distribution of $\bar{X}$.

Confidence Interval for μ, σ^2 Known

We can use information concerning the sampling distribution of $\bar{X}$ to quantify how "confident" we are in a particular estimate of μ. We do this by calculating the **confidence interval for** μ. Confidence intervals are strongly tied to the concept of significance testing described in Lab 6. Confidence, is equal to $1-\alpha$, where α is the **significance level**: the probability of rejecting a null hypothesis when it is actually true. Conventional values of α, are 0.05 and 0.01, corresponding to 95% and 99% confidence intervals.

Here is the formula for a $(1-\alpha)100\%$ confidence interval for μ, if σ^2 is known.

$$\bar{X} \pm z_{1-(\alpha/2)} \cdot \frac{\sigma}{\sqrt{n}}. \tag{5.6}$$

The script $z_{1-(\alpha/2)}$ indicates the inverse CDF (cumulative distribution function) of the standard normal distribution for the probability $1-(\alpha/2)$. It literally means: "find the value of the Z-distribution such that the proportion $1-(\alpha/2)$ lies to the left of the value." Given the conventional case that $\alpha = 0.05$, we have:

$$1-(\alpha/2) = 1-(0.05/2) = 1-0.025 = 0.975.$$

Example 5.3

An alpine vegetation study using 25 samples at alpine late snowbank sites found that the mean cover of the grass *Agrostis variabilis* was 14.6%. Assume that we know $\sigma = 4$. Calculate the 95% confidence interval for μ. We have:

$$\begin{aligned}
\bar{x} &\pm z_{1-(\alpha/2)} \cdot \frac{\sigma}{\sqrt{n}} \\
14.6 &\pm z_{1-(0.05/2)} \cdot \frac{4}{\sqrt{25}} \\
14.6 &\pm z_{0.975} \cdot \frac{4}{\sqrt{25}} \\
14.6 &\pm 1.959964 \cdot \frac{4}{5} = (13.03203, 16.16797)
\end{aligned}$$

We must obtain the inverse CDF outcome $1.959964 \approx 1.96$ using Excel or **R**.

To calculate normal inverse CDF quantiles in Excel we use the function =NORM.INV. It requires three arguments:

1. A lower tailed probability of interest. In our case, this be $1-\alpha/2 = 0.975$.
2. The mean, μ of the normal distribution of interest. Recall that we are using a Z-distribution in current example, so $\mu = 0$.

3. The standard deviation, σ, of the normal distribution of interest. Because we are using a Z-distribution, $\sigma^2 = \sigma = 1$.

=NORM.INV(0.975, 0, 1) = 1.95996.

In **R** we use the function `qnorm` to get normal quantiles. The function has the same arguments as =NORM.INV.

```
qnorm(0.975, 0, 1)

[1] 1.959964
```

Note that between the quantiles -1.96 and 1.96, we have the central 95% of a Z-distribution (2.5% is in each tail). This is why the lower tailed probability 0.975 corresponds to a 95% confidence interval (Fig. 5.5).

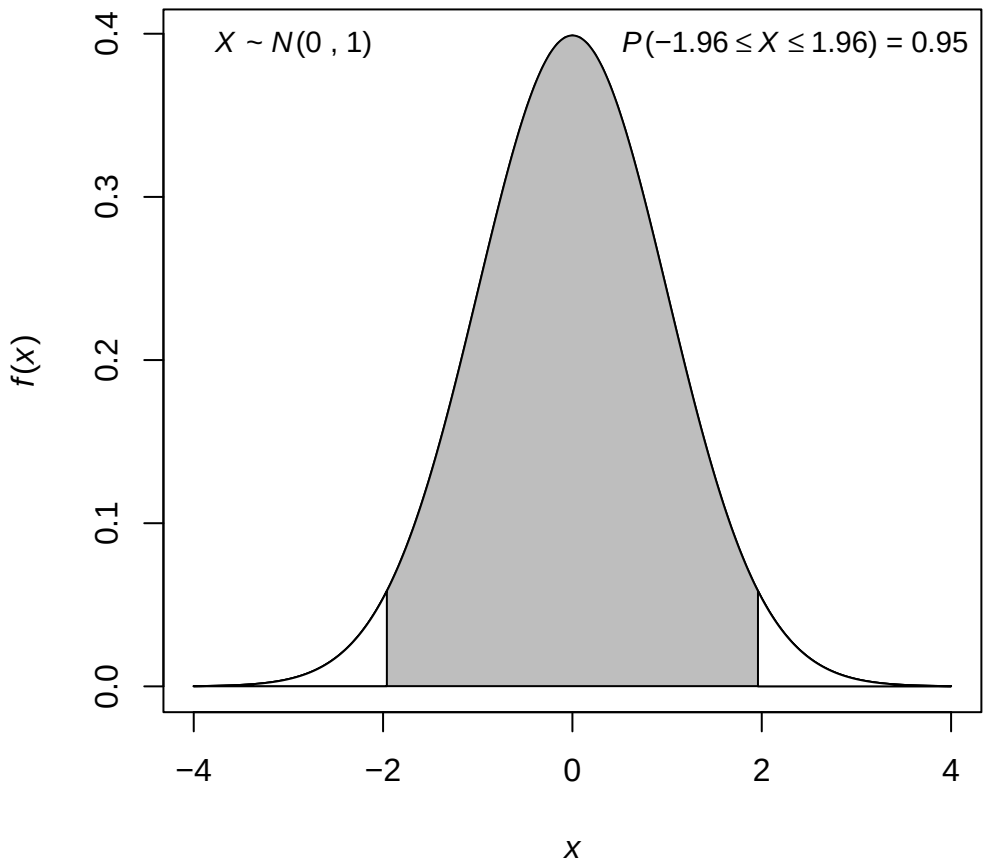


Figure 5.5. Central 95% of the Z-distrubtion.

We can check our confidence interval result using the function `ci.mu.z` from *asbio*.

```
ci.mu.z(xbar = 14.6, sigma = 4, n = 25, conf = 0.95, summarized = TRUE)

95% z Confidence interval for population mean
Estimate     2.5%    97.5%
14.60000 13.03203 16.16797
```

■

Correct interpretations of confidence intervals

There are correct and incorrect ways to interpret confidence intervals. The following are *correct* interpretations for the previous example.

1. By definition, we are 95% confident that the true mean vegetation cover (i.e. μ) of *A. variabilis* lies in the interval $(13.032, 16.168)$.

2. The confidence interval for μ comprises the central 95% of the estimated sampling distribution of $\bar{X}$, for a sample size of 25 (Fig 5.6).

3. Assume that we sampled the *A. variabilis* parent population an infinite number of times, with a sample size 25, and calculated an infinite number of 95% confidence intervals for μ from these samples. Then, 95% of those intervals will contain μ (Fig 5.7). This interpretation clearly shows that confidence intervals fall under the frequentist paradigm for probability.

Incorrect interpretations of confidence intervals

The following are common *incorrect* confidence intervals interpretations applied to the previous example.

1. There is a 95% probability that the confidence interval contains μ. This interpretation is incorrect because, under the frequentist paradigm μ is a constant. Therefore, once a confidence interval for μ has been calculated, it either contains μ or it doesn't; i.e., $P(\mu \text{ in interval}) = 1$ or $P(\mu \text{ in interval}) = 0$.

2. We are 95% confident that the sample mean cover is in the confidence interval. This is also incorrect. We are completely certain that the sample mean is in the center of interval because we used it to obtain the confidence interval.

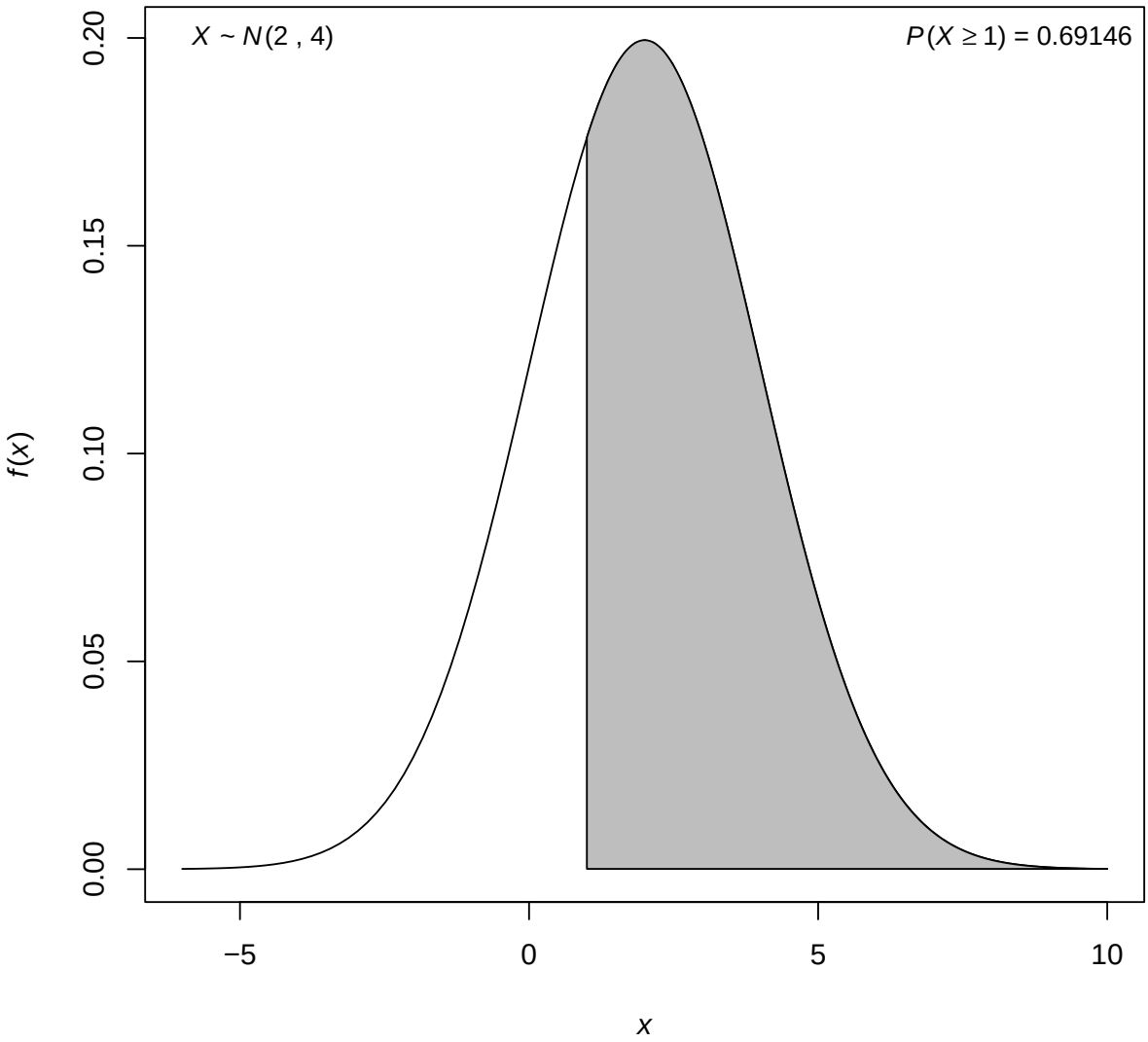


Figure 5.6. Central 95% of the the estimated sampling distribution of $\bar{X}$ for the previous problem. Note that the distribution mean is the sample mean, 14.6, and the variance = 16/25.

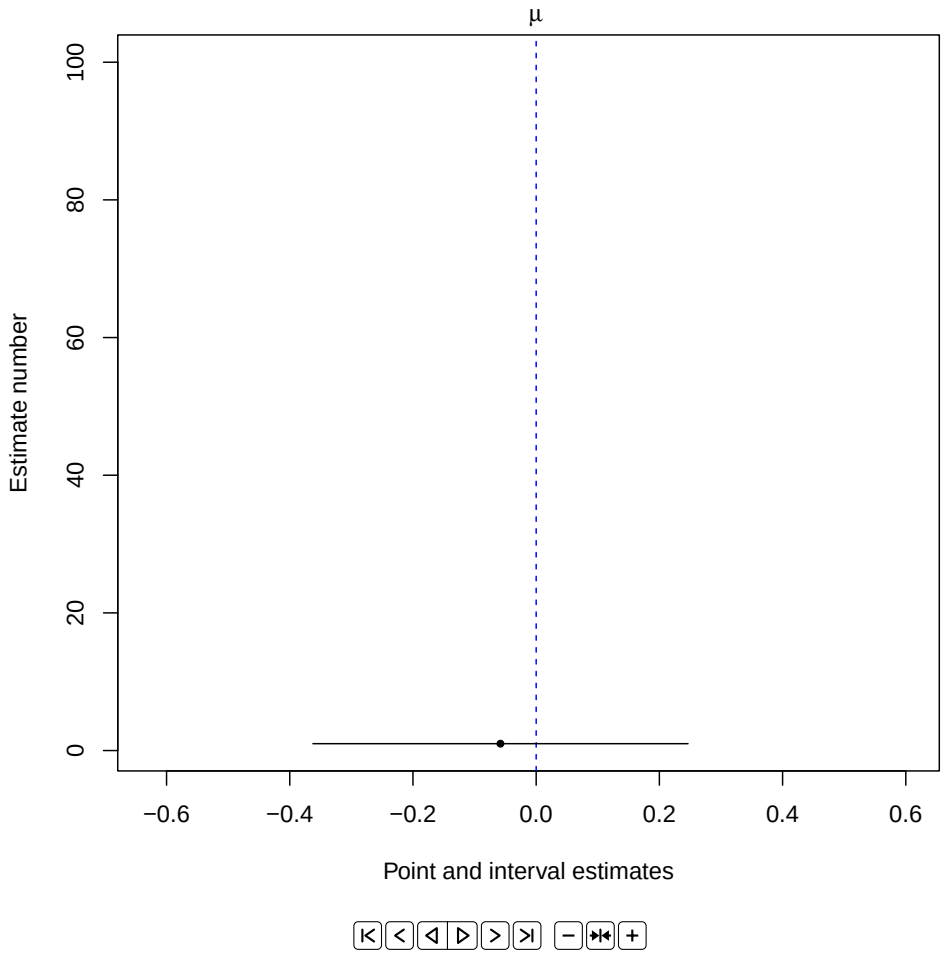


Figure 5.7. Animated depiction of a frequentist heuristic for a confidence interval for μ. To run the animation, click play while viewing in Adobe Reader.

Sample Adequacy

We can also use our formula for the confidence interval for μ to compute the correct sample size for a given margin of error and confidence level, $(1-\alpha)$. Because:

$$\begin{aligned} m &= z_{1-(\alpha/2)}\frac{\sigma}{n} \\ \sqrt{n}m &= z_{1-(\alpha/2)}\sigma \\ \sqrt{n} &= \frac{z_{1-(\alpha/2)}\sigma}{m} \\ n &= \left(\frac{z_{1-(\alpha/2)}\sigma}{m}\right)^2. \end{aligned}$$

Assignment 5

Answer all questions in one MS Word document and submit to Moodle. At the beginning of the document include the assignment number, the date, your name and section number.

Use complete sentences when appropriate, and make sure any tables, figures and computer output you include adhere to class standards (see Syllabus).

The normal distribution

- Open **R**
- Load the *asbio* package by typing `library(asbio)` or by going to **Packages > Load packages > *asbio***.
- Type `book.menu()` in the **R** console.

1. (5 pts) To see a depiction of the normal distribution go to **Chapter 3 > Pdf depiction** in the *asbio* book menu. Select **normal** and uncheck the **Show cdf** widget. Mac users, type: `see.norm.tck()` to access the GUI directly. Answer the following questions.

a) Is this distribution used to represent continuous or discrete random variables? Do you know this just by looking at the graph? Why?

b) How many parameters does the distribution have?

c) What parameter determines the location of the mode (peak of the curve)?

d) What parameter determines the thickness of the curve tails on either side of the mode?

2. (2 pts) Given $X \sim N(2, 4)$, find $f(4)$ "by hand" using **R** or Excel to help. Specifically, you can use =NORM.DIST in Excel or `dnorm` in **R**.

3. (8 pts) Let $X \sim N(2, 4)$ and let $Z \sim N(0, 1)$, find the following probabilities using =NORM.DIST in Excel or `pnorm` in **R**. Indicate if the probability is incalculable or if Excel or **R** are unneeded for calculating the probability.

a) $P(X < 2.57)$

b) $P(X > 1)$

c) $P(X \geqslant 1)$

d) $P(X = 1)$

e) $P(Z > 1)$

f) $P(-1.96 > Z > -0.5)$

g) $P(-1.96 < Z < -0.5)$

4. (10 pts) Draw the probabilities from Q. 3 using `shade.norm` in **R** to make the figures. Attach figures to homework with appropriate captions. It is not necessary to provide a figure if the probability is redundant (from a previous answer in Q 3) or incalculable. Type `?shade.norm` to get help making figures. In particular, see example code at the bottom of help page for normal probabilities.

5. (2 pts) Given $X \sim N(2, 4)$ and $P(X \leqslant x) = 0.2$, find x using functions for the inverse normal CDF in Excel or **R**. In particular, use =NORM.INV in Excel or `qnorm` in **R**.

Adding and subtracting normal random variables

6. (8 pts) Bob and Jimmy Joe are middle distance runners. The 1600 meter (metric mile) times for Bob can be represented by a random variable, B, which is normally distributed with a mean of 260 seconds and a variance of 20 seconds^2; i.e., $B \sim N(260, 20)$. The 1600 m times for Jimmy Joe can be represented by a random variable, J, which is normally distributed with a mean of 265 seconds and a variance of 17 seconds^2; i.e., $J \sim N(265, 17)$.

a) What are the parameters values of the distribution, $J - B$? That is, if $J - B \sim N(\mu, \sigma^2)$ what are μ and σ^2? **Hint:** see the section on Adding and Subtracting Normal Random Variables.

b) Given the distribution from **a)**, find the probability that Jimmy Joe beats Bob in a 1600 meter race? i.e., find $P(J - B < 0)$.

c) Draw a picture of the problem using `shade.norm`.

Sampling distributions

7. (8 pts) Go to **Chapter 5 > Sampling distribution basics** in the *asbio* book menu. Mac users, and others who wish to access the GUI directly can type `samp.dist.mech.tck()`.

In this demonstration, mountain goat weights (in kg) will be randomly obtained from a normal parent distribution, $N(90.5, 225)$. The goats are sampled with a sample size of ten, and a mean weight is calculated for the sample, $\bar{x}$. The sample mean weight is then added to an overall distribution of mean weights.

a) Choose one iteration.

- **i)** Were the randomly chosen goats the same weight?
- **ii)** What was the sample size?
- **iii)** What statistic was calculated?

b) Choose 100 iterations. Include the resulting figure in your assignment with an appropriate caption.

- **i)** Does the sample size (number of goats) stay consistent from iteration to iteration?
- **ii)** What statistic is calculated at each iteration?
- **iii)** Does average goat weight change from sample to sample?

8. (9 pts) Go to **Chapter 5 > Sampling distribution** in the *asbio* book menu. For the type of depiction choose **mean** and **snapshot**. Mac users, and others who want to access the correct GUI directly can type `samp.dist.snap.tck1()`. Note that the default parent population is exponential, a strongly positively-skewed distribution.

a) Run the function using the GUI defaults. Include resulting figure in your assignment.

b) What are the distributions that the histograms are depicting? That is, what are the individual outcomes making up the distributions?

c) What shape is the parent distribution? Symmetric? Platykurtic?

d) How do the histograms demonstrate the central limit theorem?

e) Show a sampling distribution other than than the sampling distribution of the mean by typing `samp.dist.method.tck()` and choosing a statistic that is not the sample mean. Include as a figure, and in the figure caption describe what is happening as sample size increases.

Confidence interval for μ

9. (6 pts) Go to **Chapter 5 > Confidence intervals** and run the GUI using the default values. Mac users, and others who want to run the application directly, should type `anm.ci(par.val = 0, par.type = "mu")`.

a) What proportion of the random calculated intervals contain the true value for μ, 0? (include figure with an appropriate caption).

b) According to interpretation three in the subsection entitled Correct interpretations of confidence intervals, what proportion of calculated intervals would contain μ if we had an infinite number of samples (not just 100)?

10. (6 pts) The citrus rust mite (*Phyllocoptruta oleivora*) is a major pest of citrus in Florida. The arthropod punctures the cells of leaves of fruit, causing considerable damage to citrus crops. Recently, more citrus growers have gone to a program of "preventative maintenance spraying for rust mites." In evaluating the effectiveness of the program, a random sample of 60 10-acre plots is taken. These plots show an average yield, $\bar{x}$, of 850 boxes of fruit with a standard deviation, σ, of 100 boxes.

a) Calculate a 98% confidence interval for the true mean yield, μ by hand, using function `qnorm` in **R** or NORM.INV in Excel to help when necessary. See Example 3 for guidance. Show work using snapshots when necessary.

b) Interpret your results correctly.

c) Verify your results using the function `ci.mu.z`. See code from Example 3 shown here.

Sample adequacy

11. (4 pts) A biologist wishes to estimate the effect of an antibiotic on the growth of a particular bacterium by examining the mean amount of bacteria present per cm^2 when a fixed amount of an antibiotic is applied. Previous experimentation with the antibiotic on this bacterium indicates that the population standard deviation, σ, is $11/cm^2$.

a) Determine the number of cultures that need to be developed to estimate the mean number of bacteria, with 99% confidence, given a margin of error of $m = 4/cm^2$. **Hint:** see the Sample Adequacy section.

b) Interpret your result correctly.

Q1 5pts, Q2 2pts, Q3 8pts, Q4 10pts, Q5 2pts, Q6 8pts, Q7 8pts, Q8 9pts, Q9 6pts, Q10 6pts, Q11 4pts. **Total pts: 68.**

Appendix: R-code used in this lab

Here are **R** functions for evaluating the normal distribution, $N(\mu, \sigma^2)$.

Function	What it does
`dnorm(x, mean, sd)`	Evaluates the normal PDF at `x` given `mean` $= \mu$ and `sd` $= \sigma$.
`pnorm(q, mean, sd)`	Evaluates the normal CDF at `q`.
`qnorm(p, mean, sd)`	Evaluates the normal inverse CDF at `p`.
`rnorm(n, mean, sd)`	Generates n pseudo-random samples from the normal distribution.

6

Hypothesis Testing

Lab 6 Topics

1. Deduction
    - *Modus tollens*
    - Affirming the consequent
2. The null hypothesis
    - Significance testing
3. One sample z-test
4. Type I and type II error

Deduction

Statistical hypothesis testing is based on a type of logical reasoning called **deduction**. Deductive arguments have two distinguishing characteristics:

1. General **premises** lead to a more specific **conclusion**.
2. If the premises are true then the conclusion from a valid deductive argument must also be true.

Argument one is a simple biological example of deduction.

Argument 1	
Bacterial cells do not have nuclei.	**Premise 1**
Escherichia coli (*E. coli*) is a bacterium.	**Premise 2**
E. coli does not have nuclei.	**Conclusion**

Next, let's consider a pair of deductive arguments in the context of data. To do this, let H be a hypothesis, and let I be an implication of H (the outcome I will always occur if H is true).

Argument 2	
If H is true, then so is I	**Premise 1**
Available evidence shows I is not true	**Premise 2**
H is not true.	**Conclusion**

Argument 3	
If H is true, then so is I	**Premise 1**
Available evidence shows I is true	**Premise 2**
H is true.	**Conclusion**

In Argument 2 we reject the hypothesis H using a logically correct form of deduction called ***modus tollens***. Argument 2 is deductive because if the premises are true then the conclusion must be true as well. This form of argument is also called **denying the consequent** because the consequence of H is denied, resulting in refutation of H.

Conversely, Argument 3 is a fallacious form of deduction called **affirming the consequent**. It is fallacious because the conclusion may be false even if the premises are true. As in Argument 2, the first premise in Argument 3 indicates that I is dependent on H, not the converse. However the conclusion is based on a reversal of the conditionality of premise one, given in premise two. As a result, the truth of I (suggested in the second premise) may not signify the truth of H. Additionally, the second premise is inconclusive because it consists of necessarily incomplete empirical evidence. Because at least some information concerning H is unknown, we cannot prove that hypothesis H is true. At best we can say that we have failed to reject H.

> *The implication is that we can only deductively reject or fail to reject a hypothesis whose premises include empirical data.*

The Null Hypothesis

By rejecting a hypothesis we have taken decisive action: we have eliminated a particular line of reasoning. However this does not clarify how one would *support* a hypothesis. One way around this difficulty is the **null hypothesis**, denoted $\mathbf{H}_0$. The null hypothesis generally represents a default position, or a statement of *no effect* or *no difference*. The **alternative hypothesis**, denoted $\mathrm{H_A}$, is generally constructed to encompass all possible outcomes other than those stated in H_0. Indeed, whereas H_0 defines no effect, $\mathrm{H_A}$ generally represents the *expected* effect. Thus, $\mathrm{H_A}$ is often a mathematical distillation of a **research hypothesis**. Because $\mathrm{H_A}$ is the opposite of H_0, rejection of H_0 provides conceptual support for $\mathrm{H_A}$.

We concern ourselves with H_0 and not with a research hypothesis directly for two reasons. First, as noted above, we cannot prove a hypothesis is true, however it may be possible to prove it is false, and H_0 is often a hypothesis we do not mind rejecting. Second, it is simply easier to consider statistical evidence from the perspective of H_0. This is because the research hypothesis will only suppose that there is "some effect." Exact effects (or the exact meaning of "no effect") can be specified in H_0.

A large number of null hypothesis testing procedures have been developed. Despite this variety all such methods take the same approach. They all ask the question:"How probable are the data if H_0 is true?"

Example 6.1

Hansen *et al.* (2011) vaccinated twenty-four rhesus monkeys against a powerful form of SIV (a simian cousin of HIV). The researchers believed that the vaccine would provide at least some additional protection from SIV. Thus, their null hypothesis was that the vaccine would provide no additional protection. The vaccine was found to protect half of the tested

For instance, the research hypothesis in Example 1 is that the new vaccine would provide additional protection from SIV. A mathematical distillation of the research hypothesis could be $\mathrm{H_A}$: The true mean effect of the new vaccine $\neq 0$. Note, is often easier to define $\mathrm{H_A}$ first, and then specify H_0 as the opposite of $\mathrm{H_A}$.

monkeys. This result would be highly unlikely if H_0 were true. Thus, the investigators rejected H_0, and concluded that their results supported the efficacy of the new vaccine.

■

Test statistics and P-values

In a null hypothesis we might predict that a parameter describing the difference between two populations is zero, or that a parameter equals a particular number (often zero). Thus, when H_0 is true we would expect an estimate of the parameter to be near the value specified in H_0 (e.g., zero). An estimator called a **test statistic** is used to quantify the difference between the parameter value specified in H_0 and a parameter estimate based on data. Generally speaking, the sampling distribution of the test statistic under H_0 will be known. As a result an investigator can calculate probabilities based on test statistic outcomes assuming that H_0 is true. These are called **probability values** or ***P*-values**. Specifically, a P-value is the probability that a test statistic would be "as or more extreme" than the one calculated, given that H_0 is true. Thus, P-value have the conditionality: $P(data \mid H_0)$. Smaller P-values provide stronger evidence against H_0.

Significance testing

In a widely-used approach called **significance testing** the choice between rejecting and failing to reject H_0 is based on a **decision rule** which considers the magnitude of the P-value. Significance testing defines an outcome that would be extremely unusual under H_0 – that is, an outcome with a very small P-value – as **statistically significant**. A statistically significant outcome allows probablistic rejection of H_0.

The demarcation value for statistical significance is called the **significance level** and is denoted α. Recall from Lab 5 that α is a user-defined probability for type I error, i.e., the probability of rejecting a null hypothesis when it is actually true. If we let $\alpha = 0.05$ (this is the most common significance level) we are requiring that a significant test statistic would occur no more than 5% of the time if H_0 were true. If we chose $\alpha = 0.01$, we are insisting on even stronger evidence for rejection of H_0, i.e., that a significant test statistic would occur no more than 1% of the time if H_0 were true.

If the P-value is less than or equal to α, then we say that a hypothesis test is statistically significant at level α. Thus, we can also define a P-value as the smallest possible significance level at which H_0 can be rejected.

The structuring of hypothesis statements

Both H_0 and H_A are generally expressed in terms of a parameter of interest. For instance, the mean of a normal distribution, μ. The alternative and null hypotheses are generally expressed as mathematical opposites. The value of μ stipulated by H_0 is denoted μ_0. Generally, we let $\mu_0 = 0$.

See Aho (2014, Ch 6) for additional comments and criticisms on significance testing.

The most common framework for H_0 and H_A corresponds to a **two-tailed test** in which the not-μ_0 effect is specified in H_A. In this case, the hypotheses for μ would would have the following form:

$$H_0 : \mu = \mu_0$$
$$H_A : \mu \neq \mu_0$$

It may be possible to anticipate directionality in a phenomenon under study, prompting the specification of **one-tailed tests**. For instance, we might expect plants given a nutrient supplement would grow larger than control plants given no supplement. There are two types of one-tailed tests: **lower-tailed tests** and **upper-tailed tests**. Directionality in one-tailed tests is specified in the alternative hypothesis, H_A. In a lower-tailed test the null and alternative hypotheses would have the form:

$$H_0 : \mu \geqslant \mu_0$$
$$H_A : \mu < \mu_0$$

whereas, an upper tailed test would have the form:

$$H_0 : \mu \leqslant \mu_0$$
$$H_A : \mu > \mu_0$$

Let Z be the distribution of the test statistic under H_0, that is, the **null distribution**, and let z^* be the observed test statistic. Then,

- A two-tailed P-value is calculated as $2 \cdot P(Z \geqslant |z^*|)$.
- A lower-tailed P-value is calculated as $P(Z \leqslant z^*)$.
- A upper-tailed P-value is calculated as $P(Z \geqslant z^*)$.

Example 6.2

Assume the null distribution is $Z \sim N(0, 1)$, and we find the test statistic $z^* = -1.2$ (Fig 6.1).

- The two-tailed P-value is $2 \cdot P(Z \geqslant |-1.2|) = 2 \times P(Z \geqslant 1.2) = 0.23104$.

```
2 * pnorm(1.2, lower.tail = F)
```

```
[1] 0.2301393
```

- The lower tailed P-value is $P(Z \leqslant -1.2) = 0.11507$.

In some texts H_0 is given as $H_0 : \mu = \mu_0$ for two-tailed, lower-tailed and upper-tailed tests. This effectively defines the form of the null distribution, but does not elucidate how P-values are calculated (see next).

```
pnorm(-1.2)

[1] 0.1150697
```

- The upper-tailed P-value is $P(Z \geqslant -1.2) = 0.88493$.

```
pnorm(-1.2, lower.tail = F)

[1] 0.8849303
```

- **Note, in practice only one type of test should be analyzed, and test specification should be made *a priori*.**

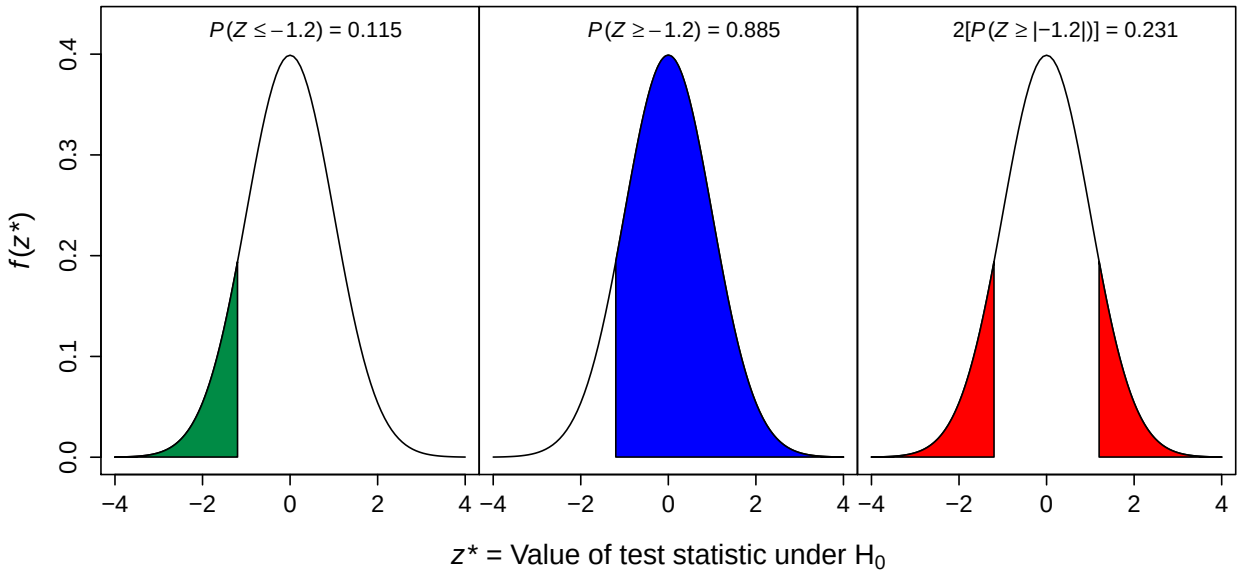


Figure 6.1. Calculating P-values for a standard normal null distribution and a test statistic of -1.2, given lower-tailed, upper-tailed, and two-tailed tests.

■

Non-significant Results

If we have insufficient evidence to reject H_0 then we will have very little evidence to support H_A. In order to not waste further time with this line of reasoning it is reasonable to conclude that H_A (and its underlying research hypothesis) are false. Given non-significance we could revise our hypotheses, recheck our underlying model(s), re-gather data, and thus restart the steps of the scientific method.

Significant Results

If we have sufficient evidence to reject H_0 this does not mean that H_0 is untrue or impossible, but it does mean that the outcome from our experiment would be unlikely if H_0 were true. Most introductory texts posit that if H_0 represents all other outcomes except H_A then rejection of H_0 corroborates H_A. From this perspective, statements following a significant result have the form: "We reject H_0 and conclude in favor of H_A." It should be emphasized, however, that just as one cannot identify H_0 as true given an insignificant result, one assuredly cannot define H_A as true (despite this claim in some introductory texts) given significance. After a significant test result an investigator can apply the research hypothesis to other biological scenarios or datasets to provide additional support and expand the original scope of inference.

Procedure for null hypothesis testing

The most commonly-used procedure for null hypothesis testing is a four-step compromise between two classic approaches, those of R. A. Fisher and Neyman-Pearson (see (Aho, 2014), Ch. 6).

1. Specify H_0, H_A, and the significance level, α, to be used.
2. Calculate the test statistic
3. Calculate the P-value.
4. State a conclusion based on the following decision rule.
    - If the P-value is greater than the specified significance level, conclude there is insufficient evidence to reject H_0, and retain H_0.
    - If the P-value is less than or equal to the specified significance level, reject H_0 and conclude in favor of H_A.

One Sample z-test

Our first formal demonstration of hypothesis testing will use a **one sample z-test**. The test statistic is calculated as:

$$z^* = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \tag{6.1}$$

Note that: 1) the test assumes that σ is known, and 2) the denominator in Eq. 6.1 is the standard error of the sampling distribution of $\bar{X}$ (Lab 5).

If test assumptions hold (see below), and if H_0 is true, z^* will be a random outcome from a standard normal distribution. The test statistic will indicate the number of standard deviations that an observed mean, x, is away from its hypothesized expectation, μ_0.

Test assumptions

All statistical testing procedures, including the one sample z-test have particular assumptions. Meeting these assumptions will determine if the test results are trustworthy, and allow valid inferences. We have the following assumptions for the one sample z-test:

1. **The underlying parent population is normally distributed with mean μ and variance, σ^2.** This assumption is to insure that the distribution of $\bar{X}$ is normal. Recall that the distribution of $\bar{X}$ will always be normal if the parent population is normal.

   Importantly, the assumption of parent population normality is only important for small sample sizes (i.e., $n < 30$). In all other cases the sampling distribution of $\bar{X}$ will be approximately normal regardless of the distributional form of the parent population, because of the tenets of the central limit theorem (Lab 5).

2. **σ^2 is known.**

3. **Observations are independent**. That is, data are from random samples of the parent population.

Example 6.3

The mean systolic blood pressure for males 35-44 years of age is approximately normal with a mean of 128 mm Hg, and a standard deviation of 15 mm Hg. The medical director of a major university looks at the medical records of 72 randomly selected professors from this age class and finds that $\bar{x} = 126.07$. The director wonders: "Do middle-aged professors in my university have a different average blood pressure than the general population?"

1. State H_0, H_A, and α. We will use a conventional significance level of 0.05 and let $\mu_0 = 128$, to reflect the research hypothesis of the medical director. Thus, we have $\alpha = 0.05$,

   $$H_0 : \mu = 128$$
   $$H_A : \mu \neq 128.$$

   The "$\neq$" sign in H_A indicates a two-tailed test. That is, we are interested in *any* difference between the true mean blood pressure of middle-aged university professors, μ, and the null mean, μ_0. A "$<$" sign in H_A would indicate that we believe that the actual true mean blood pressure of professors, μ, is *less* than 128 mm Hg, and would require a lower-tailed test. Conversely, a "$>$" sign would indicate that we believe that the mean blood pressure is *greater* than 128 mm Hg, and would require an upper-tailed test.

2. Calculate the test statistic using Eq. 6.1.

   $$\begin{aligned} z^* &= \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \\ &= \frac{126.07 - 128}{15/\sqrt{72}} \\ &= -1.091773. \end{aligned}$$

3. Calculate the P-value. Calculating the P-value for a two-sided alternative hypothesis requires finding the proportion of the null distribution that is above $|z^*|$, and below its additive inverse, $-|z^*|$. Because the standard normal distribution is symmetric, we can find the probability associated with either tail and multiply it by 2 to get the two tailed probability.

$$\begin{aligned} 2 \cdot P(Z \geqslant |z^*|) &= 2 \cdot P(Z \geqslant 1.091773) \\ &= 2 \cdot (1 - 0.8621) \\ &= 0.2757131. \end{aligned}$$

We can use Excel or **R** to find the P-value. In Excel we have:

=2*(1 - NORM.DIST(1.091773, 0, 1, TRUE)) = 0.2749329.

In **R** we have:

```
2 * (1 - pnorm(1.091773))

[1] 0.2749329

# or
2 * pnorm(1.091773, lower.tail = F)

[1] 0.2749329
```

4. State a conclusion. Because $P = 0.2757$ is > 0.05 we fail to reject H_0 and conclude that the blood pressure of middle aged professors at the university is not different from that of the general population.

We can use the function `one.sample.z` from *asbio* to do all of the z-test calculations for us:

```
z.test <- one.sample.z(n = 72, null.mu = 128,
                       sigma = 15, xbar = 126.07,
                       alternative = "two.sided")
z.test

One sample z-test
        z*   P-value
 -1.091773 0.2749329
```

Fig 6.2 provides a graphical representation of the P-value.

```
shade.norm(-1.091773, tail = "two")
```

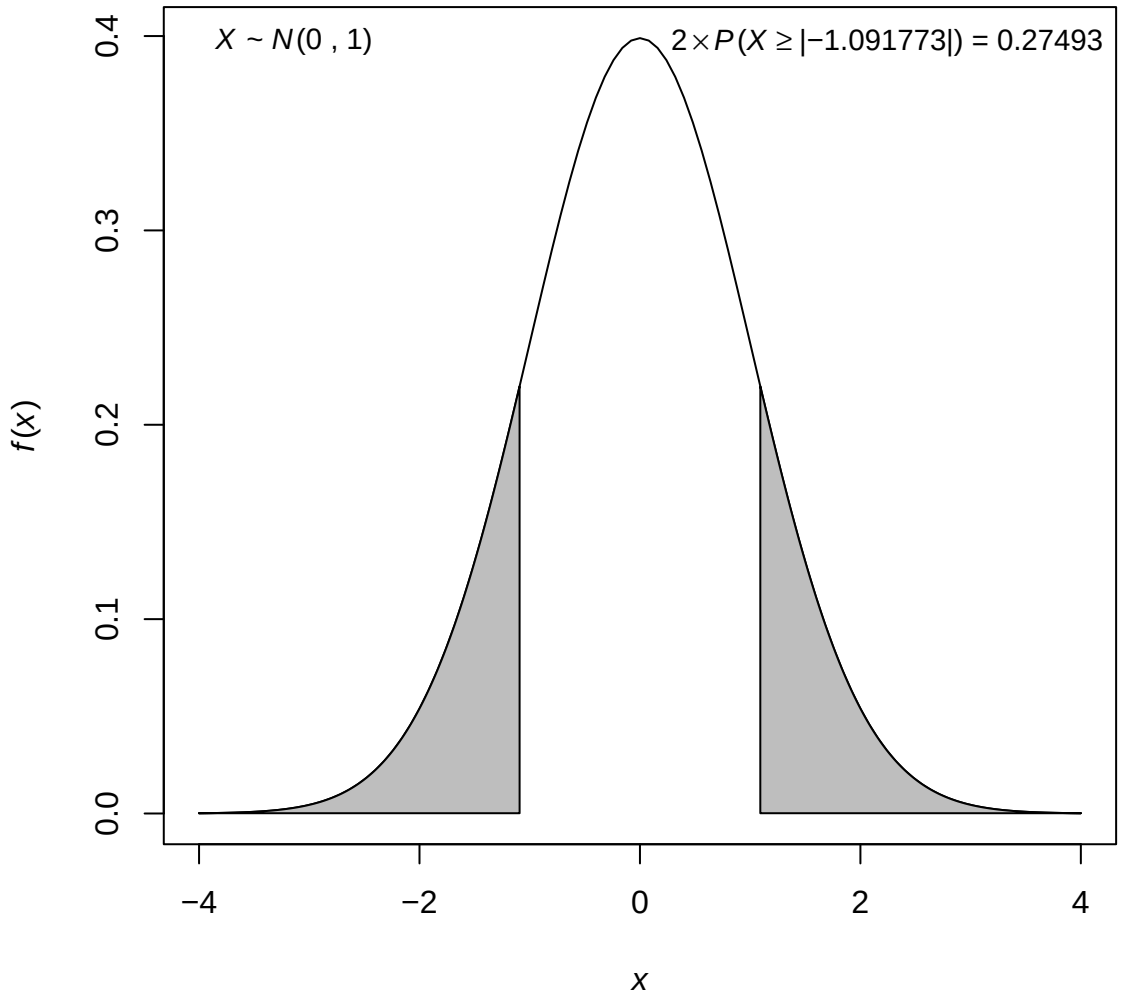


Figure 6.2. Depiction of the P-value as the area under a standard normal distribution for Example 3.

■

Two-sided significance tests and confidence intervals

There is a fundamental connection between confidence intervals using the confidence level $(1-\alpha)100\%$, and two-sided significance tests using a significance level of α. Specifically, If we can reject H_0 for a two-sided test using a significance level of α, then μ_0 will not be

contained in a confidence interval for μ using a confidence level of $(1-\alpha)100\%$. The function `one.sample.z` can be used to generate a 95% confidence interval for μ.

```
z.test$confidence

95% z Confidence interval for population mean
Estimate     2.5%    97.5%
126.0700 122.6052 129.5348
```

Note that for the systolic blood pressure example, we fail to reject H_0 at $\alpha = 0.05$, and μ_0 (i.e., 128 mm Hg), is contained in the 95% confidence interval for μ. Conversely, if the P-value was $\leqslant \alpha$, μ_0 would *not* be in the interval.

Type I and Type II Error

The significance level and the confidence level are frequentist concepts that say how reliable the method is given repeated (infinite) sampling. By definition, if we use $\alpha = 0.05$ repeatedly when H_0 is in fact true, the test will reject H_0 incorrectly 5% of the time, and will fail to reject a true H_0 (correct decision) 95% of the time.

If we reject H_0 when is H_0 is true, this is a called a **type I error**. It will occur with a probability equal to the significance level of a test, α (Table 6.1). On the other hand if we fail to reject H_0 when in fact H_0 false, this is a **type II error**. Type II errors will with a probability denoted as β. Type I error is generally considered more serious than type II error, hence its emphasis in significance testing.

Table 6.1. Correct and incorrect decisions in significance testing, with associated probabilities.

	H_0 True	**H_0 False**
Reject H_0	Type I error α	Power $1-\beta$
Fail to Reject H_0	$1-\alpha$	Type II error β

The probability that we will reject H_0 at a fixed α for a particular value of the alternative hypothesis (a particular effect size) is called **power**. Power has the probability $1-\beta$ (Table 6.1). Just as $\alpha = 0.05$ is a conventional value for type I error, $\beta = 0.2$ is becoming a conventional value for type II error, resulting in a power of 0.8.

Example 6.4

Many nutrients are essential to organisms but are fatal at higher dosages. Let's say that extensive testing indicates that the correct RDA of selenium (an essential nutrient, but one that is fatal at high doses) for an endangered animal is 87.5 mcg (1000 mcg = 1 mg). We want to test if the RDA of selenium is greater than 87.5 mcg. We quantify the optimal

amount of selenium by finding the fitness (number of offspring of the animals) at different dosages. We have the hypotheses:

$$H_0 : \mu \leqslant 87.5\text{mcg}$$
$$H_A : \mu > 87.5\text{mcg}.$$

What if we reject H_0, but H_0 was true? That is, the RDA should be $\leqslant 87.5$ mcg, but our test indicates that the RDA is > 87.5. In this case a type I error would occur and we could poison all the animals. *Solution:* use a very conservative α. On the other hand, what if we fail to reject null but H_0 was false? That is, the RDA should be > 87.5 but our test indicates that the RDA is $\leqslant 87.5$. In this case a type II error occurs and all the animals could die from selenium deprivation. *Solution:* use a larger α, or a smaller value of β (larger value of $1 - \beta$).

■

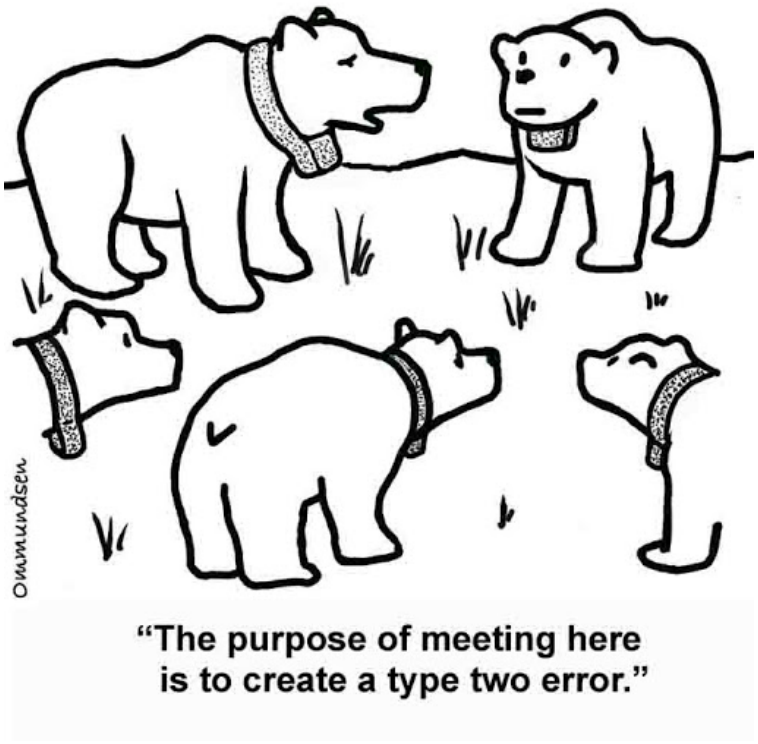


Figure 6.3. Bears selfishly preventing valid inferences.

Assignment 6

Answer all questions in one MS Word document and submit to Moodle. At the beginning of the document include the assignment number, the date, your name and section number.

Use complete sentences when appropriate, and make sure any tables, figures and computer output you include adhere to class standards (see Syllabus).

Deduction

1. (2 pts) Define deduction.

2. (4 pts) Provide an example of *modus tollens*. Is this a valid or fallacious form of deduction? Why?

Null hypotheses and P-values

3. (2 pts) Why do we create and test null hypotheses?

4. (2 pts) What is a P-value?

5. (1 pts) When using significance testing you find that the P-value is less than α. Choose the correct decision.

a) H_0 is false.

b) H_A is true.

c) Reject H_0 in favor of H_A.

d) Reject H_A in favor of H_0.

One sample z-tests

6. (8 pts) Heights of female high school students are assumed to be normally distributed with a mean of 64.5 inches and a standard deviation, $\sigma = 5$ inches. You take a random sample of 20 female high school freshman at Marsh Valley High (in Downey ID) and find that $\bar{x} = 62.3$ inches. Test the hypothesis that the true mean height of female freshman at Marsh Valley High does not equal 64.5 inches. Use a significance level of 0.05.

a) State H_0, H_A, and α.

b) Calculate the one-sample z-test test statistic.

c) Calculate the P-value.

d) State your conclusions correctly.

7. (2 pts) Check your answers for Q. 6 using `one.sample.z` in **R**. Show output using snapshots.

8. (2 pts) Using `shade.norm` draw a picture of the distribution under H_0 for Q. 6, and overlay the correct P-value. Include the graph in your assignment.

9. (9 pts) Bring the Creosote data in the Lab 6 folder into **R** by saving the file onto your computer and navigating to it using

```
creosote <- read.csv(file.choose())
```

Calculate the sample standard deviation in **R** by typing:

```
sd(creosote[,1])
```

And calculate the mean by typing:

```
mean(creosote[,1])
```

The code `cresoste[,1]` indicates that the data are in the first column of `creosote`. Substitute the sample standard deviation for σ to allow use of a one-sample z-test.

Test the hypothesis that creosote productivity is less than 17.5 m^2 yr^{-1}. Use a significance level of 0.05.

a) State H_0, H_A, and α.

b) Calculate the one-sample z-test test statistic.

c) Calculate the P-value.

d) State your conclusions correctly.

10. (2 pts) Check your answers for Q. 9 using `one.sample.z` in **R**. Show output using snapshots.

11. (4 pts) Make a histogram showing the distribution of the creosote data by typing the code below. Include the graph in your assignment.

```
hist(creosote[,1], main = "", xlab =
       expression(paste("Productivity (g ", m^2,y^{-1},")")))
```

Note that most of the code above specifies superscripts on x-axis for the units of productivity.

Are the assumptions for the one sample z-test in Q. 9 valid? Why or why not? Before answering, go to the section concerning z-test assumptions and reread the section on the central limit theorem from Lab 5.

Type I, type II error and power

- Open **R**
- Load the *asbio* package by typing `library(asbio)` or by going to **Packages > Load packages > *asbio***.
- Type `book.menu()` in the **R** console.

12. (3 pts) Go to **Chapter 6 >Type I and II error**. Mac-users and others wishing to obtain the GUI directly can type: `see.typeI_II()`. Click the **More info** widget to learn more about type I and II error and power (you will need to click at the edge of the widgets).

a) What is considered a more serious type of error, type I or II?

b) What are the conventional (most frequently used) levels for α, β, and $1 - \beta$?

13. (5 pts) Go to **Chapter 6 >Power**. Mac-users and others wishing to obtain the GUI directly can type: `see.power.tck()`. The top graph shows a distribution which assumes that H_0 is true. This is the one we use to compute P-values. The lower graph shows power, i.e., the probability of rejecting H_0 when H_0 is false.

a) Does power, i.e., $1 - \beta$, equal $1 - \alpha$?

b) Does decreasing α increase or decrease power?

c) Does decreasing σ increase or decrease power?

d) Does increasing n (i.e., sample size) increase or decrease power?

e) Does increasing effect size (i.e., $|\mu - \mu_0|$) increase or decrease power?

Q1 2pts, Q2 4pts, Q3 2pts, Q4 2pts, Q5 1pt, Q6 8pts, Q7 2pts, Q8 2pts, Q9 9pts, Q10 2pts, Q11 4pts, Q12 3pts, Q13 5pts. **Total pts: 46.**

7

t-tests

Lab 7 Topics

1. *t*-distribution
2. The family of *t*-tests
   - Confidence interval for μ, σ^2 unknown
   - Pooled variance *t*-test
   - Welch *t*-test
   - Paired *t*-test

t-distribution

If the variance of an underlying normal parent distribution is unknown (and this will generally be true), then we must estimate σ^2 with the sample variance, S^2. Now, however, it will be impossible to derive confidence intervals or test hypotheses concerning the population mean, μ, using the standard normal distribution. Instead, we must approximate the Z-distribution using the *t*-distribution.

Like the standard normal distribution, the ***t*-distribution** is symmetric and centered at zero. In fact, the *t*-distribution asymptotically converges to the standard normal distribution as its lone parameter ν (commonly called the degrees of freedom) approaches ∞. For smaller values of ν the *t*-distribution is platykurtic (flatter) compared to the Z-distribution (Fig 7.1).

If a random variable X follows a *t*-distribution, its PDF is:

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\nu\pi}} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{7.1}$$

where $\nu > 0$, Γ is the so-called **gamma function** which is a generalization of the factorial

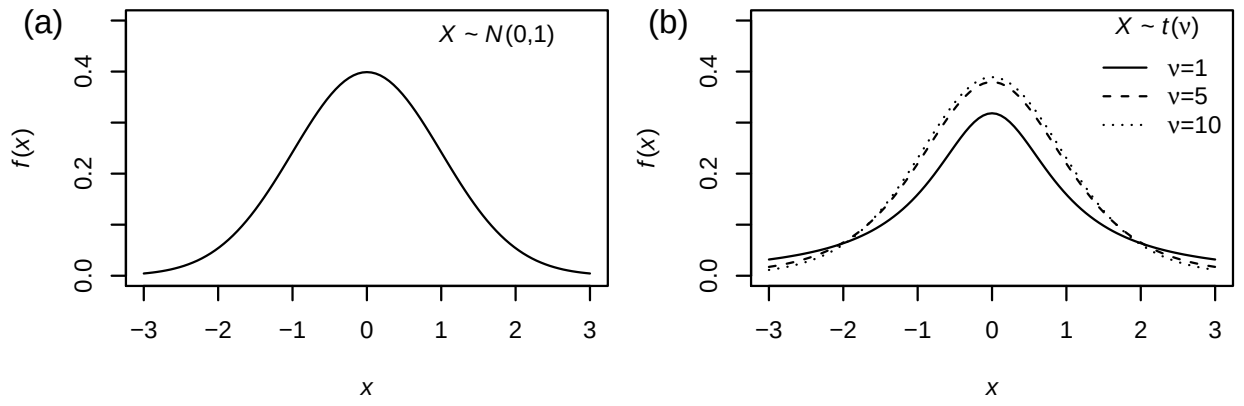


Figure 7.1. A comparison of (a) the standard normal distribution and (b) t-distributions with varying values of ν.

function for non-integers, and $x \in \mathbb{R}$. If a random variable X follows a t-distribution, we write this as $X \sim t(\nu)$.

The t-distribution arises as the sampling distribution of the mean, following z-transformation, when the standard deviation of the parent distribution is unknown. Specifically, if the parent distribution X is normal with mean, μ, and an unknown variance, and X is sampled with a sample size of n, then t^* will be a random outcome from t-distribution with $n-1$ degrees freedom, when

$$t^* = \frac{\bar{X} - \mu}{S/\sqrt{n}}. \tag{7.2}$$

The denominator of Eq. 7.2, $S/\sqrt{n}$, is called the **sample standard error**. It is an estimator for the standard deviation of sampling distribution of the mean, $\sigma_{\bar{X}}$.

Confidence interval for μ, σ^2 unknown

What if we want to calculate a confidence interval for the mean of a normal distribution μ, but we don't know σ^2? In this case, a $(1-\alpha)100\%$ confidence interval for μ can be calculated as:

$$\bar{X} \pm t_{(1-(\alpha/2),n-1)} \frac{S}{\sqrt{n}}. \tag{7.3}$$

where $t_{(1-(\alpha/2),n-1)}$ indicates the inverse CDF (cumulative distribution function) of a t-distribution with $n-1$ degrees of freedom at the lower tailed probability $1-(\alpha/2)$. We multiply this quantile by the sample standard error (Eq. 7.2) to get the margin of error for the confidence interval.

For any positive integer, n, the gamma function has the form: $\Gamma(n) = (n-1)!$. More generally, $\Gamma(z) = \int_0^\infty t^{z-1}e^{-1}dt$, where $z > 0$.

Example 7.1

We obtain 25 random samples of vascular plant cover at alpine late snowbank sites and find that the mean cover of the grass *Agrostis variabilis* is 14.6%. We do not know σ, but from our sample we find that the sample standard deviation, s, equals 12.25%. We calculate a 95% confidence interval for μ as follows:

$$\begin{aligned}
\bar{x} \pm t_{(1-(\alpha/2),n-1)} \frac{s}{\sqrt{n}} &= \\
14.6 \pm t_{(1-(0.05/2),24)} \frac{12.25}{\sqrt{25}} &= \\
14.6 \pm t_{(0.975,24)} \frac{12.25}{5} &= \\
14.6 \pm 2.063899 \cdot \frac{12.25}{5} &= \\
14.6 \pm 5.056551 &= \\
(9.543449, 19.656551)
\end{aligned}$$

We must obtain the inverse CDF outcome 2.063899 using Excel or **R**.

To calculate t inverse CDF quantiles in Excel we use the function =T.INV. It requires two arguments:

1. A lower tailed probability of interest. In our case, this be $1 - \alpha/2 = 0.975$.
2. The degrees of freedom, ν. For the current example this will be $n - 1 = 24$.

=T.INV(0.975, 24) = 2.063899.

In **R** we use the function `qt` to get t quantiles. The function has the same arguments as =T.INV.

```
qt(0.975, 24)

[1] 2.063899
```

We can use the **R** function `ci.mu.t` to complete all of the calculations for us.

```
ci.mu.t(xbar = 14.6, n = 25, sd = 12.25, conf = 0.95, summarized = T)

95% t Confidence interval for population mean
 Estimate      2.5%     97.5%
14.600000  9.543449 19.656551
```

We are 95% confident that the true mean, μ, is between 9.543449 and 19.656551.

■

The mean difference of two normal random variables

If X and Y are normally distributed random variables, with $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$, then, from Lab 5, we know that:

$$X - Y \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2).$$

Additionally, if X is sampled with sample size n_X and Y is sampled with sample size n_Y, it follows, from our knowledge of the sampling distribution of the mean (Ch. 5), that:

$$\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}\right). \tag{7.4}$$

Further, given the tenets of the central limit theorem (Lab 5), it follows that if X and Y are sampled with sufficiently large sample sizes, then, regardless of the form of the parental distributions,

$$\bar{X} - \bar{Y} \xrightarrow{d} N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}\right). \tag{7.5}$$

Finally, because

$$\sigma_{\bar{X}-\bar{Y}}^2 = \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}, \tag{7.6}$$

it follows that

$$\sigma_{\bar{X}-\bar{Y}} = \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}. \tag{7.7}$$

The Family of t-tests

The so-called **family of t-tests** constitute the most common set of approaches for making inference to mean difference of two normal random variables when the variances of those distributions are unknown. Eqs. 7.2 - 7.7 form the mathematical basis for all t-tests for comparing μ_X to μ_Y.

Structuring of t-test hypotheses

We can test any one of three sets of hypotheses below using the family of t-tests. For each of the hypothesis sets D_0 denotes the hypothesized (null) difference for $\mu_X - \mu_Y$. Generally, $D_0 = 0$.

1. For a two-tailed test we have:

$$\begin{aligned} \mathrm{H}_0 &: \mu_X - \mu_Y = D_0 \\ \mathrm{H}_\mathrm{A} &: \mu_X - \mu_Y \neq D_0 \end{aligned}$$

   Under the usual case that $D_0 = 0$ we have:

$$\begin{aligned} \mathrm{H}_0 &: \mu_X = \mu_Y \\ \mathrm{H}_\mathrm{A} &: \mu_X \neq \mu_Y \end{aligned}$$

2. For an upper-tailed test ($\mu_X > \mu_Y$), we have:

$$\begin{aligned} \mathrm{H}_0 &: \mu_X - \mu_Y \leqslant D_0 \\ \mathrm{H}_\mathrm{A} &: \mu_X - \mu_Y > D_0 \end{aligned}$$

   Under the usual case that $D_0 = 0$, we have:

$$\begin{aligned} \mathrm{H}_0 &: \mu_X \leqslant \mu_Y \\ \mathrm{H}_\mathrm{A} &: \mu_X > \mu_Y \end{aligned}$$

3. For an lower-tailed test ($\mu_X < \mu_Y$), we have:

$$\begin{aligned} \mathrm{H}_0 &: \mu_X - \mu_Y \geqslant D_0 \\ \mathrm{H}_\mathrm{A} &: \mu_X - \mu_Y < D_0 \end{aligned}$$

   Under the usual case that $D_0 = 0$, we have:

$$\begin{aligned} \mathrm{H}_0 &: \mu_X \geqslant \mu_Y \\ \mathrm{H}_\mathrm{A} &: \mu_X < \mu_Y \end{aligned}$$

Pooled variance t-test

If the variances of the two populations under comparison, X and Y, can be assumed to be approximately equal, then we would use a **pooled variance t-test** (also called Student's t-test) to make inference to $\mu_X - \mu_Y$.

A natural estimator for the mean of the distribution of mean differences in Eqs. 7.4 and 7.5 is simply the difference in sample means:

$$\widehat{\mu_X - \mu_Y} = \bar{X} - \bar{Y}, \tag{7.8}$$

As noted in Lab 5, in some texts H_0 is given as $\mathrm{H}_0 : \mu = \mu_0$ for lower-tailed and upper-tailed tests.

where the hat sign $\widehat{\mu_X - \mu_Y}$ indicates: "the estimator for $\mu_X - \mu_Y$." If we can assume that the variances of X and Y are equal. That is, if we can assume $\sigma_X^2 = \sigma_Y^2 = \sigma_{pool}^2$ in Eq. 7.7, then

$$\begin{aligned}\sigma_{\bar{X}-\bar{Y}} &= \sqrt{\frac{\sigma_{pool}^2}{n_X} + \frac{\sigma_{pool}^2}{n_Y}} \\ &= \sqrt{\sigma_{pool}^2 \left(\frac{1}{n_X} + \frac{1}{n_Y}\right)} \\ &= \sigma_{pool}\sqrt{\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}.\end{aligned} \tag{7.9}$$

The estimator for the **pooled variance**, σ_{pool}^2, is the called the **mean squared error** or **MSE**:

$$\widehat{\sigma_{pool}^2} = MSE = \frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2}. \tag{7.10}$$

Thus, the estimator for σ_{pool} is $\sqrt{MSE}$. Combining the t-statistic framework of Eq. 7.2 with Eqs. 7.8, 7.9, and 7.10 we obtain the formula for the pooled variance t-test test statistic:

$$t^* = \frac{(\bar{X} - \bar{Y}) - D_0}{\sqrt{MSE}\sqrt{\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}}. \tag{7.11}$$

where, as noted above, D_0 is the hypothesized (null) difference for $\mu_X - \mu_Y$. Generally, $D_0 = 0$.

Calculating P-values for the pooled variance t-test

If H_0 is true, and assumptions for the test hold, then t^* will be a random outcome from a t-distribution with $n_X + n_Y - 2$ degrees of freedom.

- For a two-tailed test the P-value is: $2 \cdot P(T \geqslant |t^*|)$.
- For an upper-tailed test the P-value is: $P(T \geqslant t^*)$.
- For a lower-tailed test the P-value is: $P(T \leqslant t^*)$.

where $T \sim t(n_X + n_Y - 2)$.

Example 7.2

An experiment was conducted to evaluate the effectiveness of a treatment for tapeworm in the stomachs of sheep. A random sample of 24 worm-infected lambs was randomly divided into two groups. Twelve of the lambs were injected with am anti-tapeworm drug, and the other 12 remained untreated. After a six month period the lambs were slaughtered and the

worm counts were recorded (Table 7.1). Test whether the mean number of tapeworms in the treated sheep is less than the number in the untreated sheep. Use $\alpha = 0.05$

Table 7.1. Data for treated and untreated sheep in Example 2.

Drug-treated sheep	18	43	28	50	16	32	13	35	38	33	6	7
Untreated sheep	40	54	26	63	21	37	39	23	48	58	28	39

1. We designate $\alpha = 0.05$. We have the following hypotheses:

$$\mathrm{H}_0 : \mu_T \geqslant \mu_U$$
$$\mathrm{H}_\mathrm{A} : \mu_T < \mu_U$$

   where μ_T and μ_U denote the population means for the treated and untreated sheep, respectively.

2. To calculate the test-statistic, we first compile summary statistics. We have: $\bar{x}_T = 26.58$; $\bar{x}_U = 39.67$; $S_T^2 = 206.07$; $S_U^2 = 192.06$. For MSE we have:

$$\sqrt{MSE} = \sqrt{\frac{(n_T - 1)s_T^2 + (n_U - 1)s_U^2}{n_T + n_U - 2}} = \sqrt{\frac{11 \cdot 206.27 + 11 \cdot 192.06}{22}} = 14.113$$

   Thus, the test statistic is:

$$t^* = \frac{(\bar{x}_T - \bar{x}_U) - D_0}{\sqrt{MSE}\sqrt{\left(\frac{1}{n_T} + \frac{1}{n_U}\right)}} = \frac{(26.58 + 39.67) - 0}{14.113 \cdot \sqrt{1/6}} = -2.2719$$

3. To calculate the P-value, we need to first consider the form of the alternative hypothesis and the degrees of freedom in the null $t-$distribution. We have a two-tailed test. According to the Calculating P-values subsection for this test we calculate the P-value as $P(T \leqslant t^*)$ where $T \sim t(n_T + n_U - 2)$. Because $n_T = n_U = 12$, $T \sim t(22)$. We can calculate the P-value using either the function T.DIST from Excel or `pt` from **R**.

   =T.DIST requires three arguments:

   - A t-distribution outcome. For the current example this will be the test statistic outcome $t^* = -2.2719$
   - The degrees of freedom, ν. For the current example this will be $n_T + n_U - 2 = 22$.
   - Whether or not you want the CDF (TRUE) or the PDF (FALSE).

   Thus, we have: =T.DIST(-2.2719, 22, TRUE) = 0.01661.

The function `pt` represents the t-distribution CDF. It has two arguments which are identical to the first two arguments of =T.DIST. Thus, we have:

```
pt(-2.2719, 22)

[1] 0.01661012
```

4. Because $P < 0.05$ we reject H_0 and conclude that the treatment reduces tapeworms compared to untreated sheep.

Below is code to run the entire test "by hand" using **R**. First we bring in the data.

```
sheep <- read.csv(file.choose())
```

In the next code chunk we calculate the test statistic. Note that the fist column in the `sheep` dataset (`sheep[,1]`) contains tapeworm counts, and the second column (`sheep[,2]`) contains corresponding treatment assignments (`T` and `U`). Thus, the first two columns serve as response and explanatory variables for the experiment, respectively.

```
means <- tapply(sheep[,1], sheep[,2], mean)
means

       T        U
26.58333 39.66667

vars <- tapply(sheep[,1], sheep[,2], var)
vars

       T        U
206.2652 192.0606

ns <- tapply(sheep[,1], sheep[,2], length)
ns

 T  U
12 12

MSE <- (vars[1]*(ns[1]-1) + vars[2]*(ns[2]-1))/(sum(ns)-2)
t.star <- (means[1] - means[2])/(sqrt(MSE)*sqrt(1/ns[1]+1/ns[2]))
t.star

        T
-2.270857
```

Here we calculate the P-value:

```
pt(t.star, ns[1]+ns[2]-2)

         T
0.01664659
```

Note that `means[1] - means[2]` in the calculation of `t.star` corresponds to $\bar{X}_T$ - $\bar{X}_U$, and thus agrees with the ordering of μ_T and μ_U in the hypotheses.

We can use the **R** function `t.test` to do everything for us.

```
t.test(sheep[, 1] ~ sheep[, 2], var.equal = T, alternative = "less")

Two Sample t-test

data:  sheep[, 1] by sheep[, 2]
t = -2.2709, df = 22, p-value = 0.01665
alternative hypothesis: true difference in means between group T and group U is less than 0
95 percent confidence interval:
      -Inf -3.190165
sample estimates:
mean in group T mean in group U
       26.58333        39.66667
```

The code `sheep[,1] ~ sheep[,2]` means that tapeworm numbers in `sheep[,1]` are assumed to be a function of the of the categorical levels in `sheep[,2]`. We can use this sort of statement in `t.test` when the response and explanatory variables are in separate columns.[1] We specify `var.equal = T` to get a pooled variance t-test. By default `var.equal = F`. We define the alternative hypothesis with the argument `alternative`. Choices are `"less"`, `"greater"`, and `"two.sided"`. By default `alternative = "two.sided"`.

■

---

[1] The function `t.test` can also be run by inputting vectors of observations from hypothesized populations separately as the first two arguments. For example, if `sheep[,1]` and `sheep[,2]` represented tapeworm counts for the treated and untreated sheep, respectively, we would specify the test as: `t.test(sheep[,1], sheep[,2],....)`.

Care must be taken in `t.test` to insure that the test corresponds to the original hypotheses. When using $\sim$ to separate response and explanatory variables in one-tailed tests **R** will automatically order hypothesized populations by the alphanumeric order of their names in the explanatory variable. Thus, for our example, **R** will assume that the alternative hypothesis has the form: H_A: $\mu(1st\_alphanumeric\_name) < \mu(2nd\_alphanumeric\_name)$. Recall that the treatment names in `sheep[,2]` were `T` and `U`. We are okay because our alternative was: $H_A : \mu_T < \mu_U$, and T comes before U in the alphabet.

Also in the output is a one-side (lower-tailed) confidence interval the true mean difference, $\mu_T - \mu_U$. Thus, we are 95% confident that the true mean difference is less than -3.190165. We will not explicitly consider one-tailed confidence intervals in this lab.

Pooled variance t-test assumptions

The pooled variance t-test has three assumptions which we will formally consider in Lab 8.

1. Parental distributions X and Y are normally distributed.
2. Observations are independent.
3. Parental distributions have equal variances. **Heteroscedasticity**(unequal population variances) may result in untrustworthy pooled variance t-test results, particularly if sample sizes are unequal.

Welch t-test

If we cannot assume that the variances of X are Y equal, we can still test for hypothesized differences in $\mu_X - \mu_Y$ using the **Welch t-test**. The Welch test statistic is:

$$t^* = \frac{(\bar{X} - \bar{Y}) - D_0}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}} \tag{7.12}$$

The denominator in Eq 7.12 is an estimator for Eq. 7.7. Unlike the pooled variance test statistics, Welch test statistics will not exactly follow a t-distribution under H_0. Instead we identify an approximate null t-distribution using the **Satterthwaite procedure** to compute the degrees of freedom. This will be ν in Eq. 7.13.

$$\nu = \frac{\left(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}\right)^2}{\frac{\left(S_X^2/n_X\right)^2}{n_X - 1} + \frac{\left(S_Y^2/n_Y\right)^2}{n_Y - 1}} \tag{7.13}$$

The Satterthwaite procedure will generally produce a non-integer solution for the degrees of freedom.

Calculating P-values for the Welch t-test

If H_0 is true, and assumptions for the test hold, then t^* will be a random outcome from a t-distribution with approximately ν degrees of freedom.

- For a two-tailed test the P-value is: $2 \cdot P(T \geqslant |t^*|)$.
- For an upper-tailed test the P-value is: $P(T \geqslant t*)$.
- For a lower-tailed test the P-value is: $P(T \leqslant t*)$.

where $T \sim t(\nu)$.

Example 7.3

An agricultural experimental station is testing the effect of pesticides on insect counts (Table 7.2). A researcher wants to know if there is any difference in insect counts between

sprays A and E. We know that we cannot assume equal variances for A and E and will use $\alpha = 0.05$.

Table 7.2. Data for sprays A and E in Example 3.

Spray A	10	7	20	14	14	12	10	23	17	20	14	13
Spray E	3	5	3	5	3	6	1	1	3	2	6	4

1. We designate $\alpha = 0.05$. We have the following hypotheses:

$$\mathrm{H}_0 : \mu_A = \mu_E$$
$$\mathrm{H}_\mathrm{A} : \mu_A \neq \mu_E$$

   where μ_A and μ_E denote the population means for spray A and E, respectively.

2. To calculate the test-statistic, we first compile summary statistics. We have: $\bar{x}_A = 14.5$; $\bar{x}_E = 3.5$; $S_A^2 = 22.27$; $S_E^2 = 3$. Thus, the test statistic is:

$$t^* = \frac{(\bar{x}_A - \bar{x}_E) - D_0}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_E^2}{n_E}}} = \frac{(14.5 - 3.5) - 0}{\sqrt{\frac{22.27}{12} + \frac{3}{12}}} = 7.5798.$$

3. To calculate the P-value, we need to first consider the form of the alternative hypothesis and the degrees of freedom in the null $t-$distribution. We have a two-tailed test. According to the Calculating P-values subsection for this test we calculate the P-value as $2 \cdot P(T \geqslant |t^*|)$ where $T \sim t(n_T \nu)$. We calculate ν using the Satterthwaite procedure. We have:

$$\nu = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\frac{(s_X^2/n_X)^2}{n_X - 1} + \frac{(s_Y^2/n_Y)^2}{n_Y - 1}} = \frac{\left(\frac{22.7+3}{12}\right)^2}{\frac{(22.7/12)^2 + (3/12)^2}{11}} = 13.91046.$$

   Thus, the null distribution is $T \sim t(13.91)$. The P-value is:

```
2 * pt(7.579791, 13.91046, lower.tail = FALSE)

[1] 2.654553e-06
```

4. Because $P < 0.05$ we reject H_0 and conclude that the mean insect counts from the spray types differ.

Here is code to run the entire test "by hand" using **R**. First we bring in the data.

```
insect <- read.csv(file.choose())
```

```
#test statistic
means <- tapply(insect[,1], insect[,2], mean)
means

   A    E
14.5  3.5

vars <- tapply(insect[,1], insect[,2], var)
vars

       A        E
22.27273  3.00000

ns <- tapply(insect[,1], insect[,2], length)
ns

 A  E
12 12

t.star <- (means[1] - means[2])/sqrt(vars[1]/ns[1]+vars[2]/ns[2])
t.star

       A
7.579791
```

```
#Degrees of freedom
num <- (vars[1]/ns[1] + vars[2]/ns[2])^2
den <- (vars[1]/ns[1])^2/(ns[1]-1) + (vars[2]/ns[2])^2/(ns[2]-1)
nu <- num/den
nu

       A
13.91046
```

```
#P-value
2 * pt(t.star, nu, lower.tail = FALSE)

           A
2.654548e-06
```

Once again we can use `t.test` to do everything for us.

```
t.test(insect[,1] ~ insect[,2])

Welch Two Sample t-test

data:  insect[, 1] by insect[, 2]
t = 7.5798, df = 13.91, p-value = 2.655e-06
alternative hypothesis: true difference in means between group A and group E is not equal to 0
95 percent confidence interval:
  7.885546 14.114454
sample estimates:
mean in group A mean in group E
           14.5             3.5
```

■

Welch t-test assummptions

The Welch t-test has two assumptions which we will formally consider in Lab 8.

1. Parental distributions X and Y are normally distributed.
2. Observations are independent.

Paired t-test

If observations can be viewed as blocked or paired for the two hypothetical populations being compared, then we should test for hypothesized mean differences using a **paired t-test**. Pairing will result in a lack of independence in observations making application of other t-tests inappropriate.

In a paired t-test the response variable is the difference, D, between observations that are paired. Thus, if X and Y are normal, then $D \sim N(\mu_D, \sigma_D^2)$. . The test statistic for the paired t-test is:

$$t^* = \frac{\bar{X}_D - D_0}{S_D/\sqrt{n}} \tag{7.14}$$

where $\bar{X}_D$ and S_D indicate sample mean and sample standard deviation of the paired differences and n indicates the number of pairs.

Calculating P-values for the paired t-test

If H_0 is true, and assumptions for the test hold, then t^* will be a random outcome from a t-distribution with $n-1$ degrees of freedom.

- For a two-tailed test the P-value is: $2 \cdot P(T \geqslant |t^*|)$.

- For an upper-tailed test the P-value is: $P(T \geqslant t^*)$.
- For a lower-tailed test the P-value is: $P(T \leqslant t^*)$.

where $T \sim t(n-1)$.

Example 7.4

The ease or difficulty one has losing weight might be dependent on genetic makeup. To compare two different weight loss programs (X and Y), and to control for the confounding potential of genetics, 12 pairs of identical twins of similar weight were studied. Each pair of twins were randomly assigned to program X or Y (Table 7.3). Test to see if there is any difference in weight loss between the two programs. Use $\alpha = 0.05$.

Table 7.3. Data for weight loss programs X and Y in Example 4.

Program X	12.4	10.3	6.8	11.5	10.4	9.8	5.7	9.5	9.8	8.0	7.1	10.9
Program Y	12.8	10.0	8.7	11.9	10.6	9.7	7.9	10.8	11.6	8.8	9	11.1
D = Y - X	0.4	-0.3	1.9	0.4	0.2	-0.1	2.2	1.3	1.8	0.8	1.9	0.2

1. We designate $\alpha = 0.05$. We have the following hypotheses:

$$\mathrm{H}_0 : \mu_Y = \mu_X$$
$$\mathrm{H}_\mathrm{A} : \mu_Y \neq \mu_X$$

   where μ_X and μ_Y denote the population means for weight loss programs X and Y, respectively. Note that the hypothesis structure above is equivalent to testing the null H_0: $\mu_D = 0$.

2. To calculate the test-statistic we calculate the mean and variance of the differences, Y - X. We have: $\bar{x}_D = 0.89167$; $s_D^2 = 0.78083$; $n = 12$. Thus, the test statistic is:

$$t^* = \frac{\bar{x}_D - D_0}{s_D/\sqrt{n}} = \frac{0.89167}{0.88365/\sqrt{12}} = 3.496.$$

3. To calculate the P-value, we need to first consider the form of the alternative hypothesis and the degrees of freedom in the null $t-$distribution. We have a two-tailed test. According to the Calculating P-values subsection for this test we calculate the P-value as $2 \cdot P(T \geqslant |t^*|)$ where $T \sim t(n-1)$. Thus, the null distribution is $T \sim t(11)$. The P-value is:

```
2 * pt(3.496, 11, lower.tail = FALSE)

[1] 0.005005409
```

4. Because $P < 0.05$ we reject H_0 and conclude that weight loss differs for programs X and Y.

Below is code to run the entire test "by hand" using **R**. First we bring in the data.

```
weight <- read.csv(file.choose())
```

```
D <- weight$Y - weight$X
mean.D <- mean(D)
mean.D

[1] 0.8916667

var.D <- var(D)
var.D

[1] 0.7808333

n <- 12

t.star <- (mean.D)/(sqrt(var.D)/sqrt(n))
t.star

[1] 3.495538
```

Note, I can access a column of data in an **R** dataframe by giving the name of the dataframe followed by a dollar sign (`$`) and the column name. Here we calculate the P-value:

```
2 * pt(t.star, n-1, lower.tail = F)

[1] 0.005009484
```

Once again we can use `t.test` to do everything for us.

```
t.test(weight$Y, weight$X, paired = T)

Paired t-test

data:  weight$Y and weight$X
t = 3.4955, df = 11, p-value = 0.005009
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 0.3302237 1.4531097
sample estimates:
mean difference
      0.8916667
```

The columns in `weight` do not represent response and explanatory variables like the `sheep` and `insect` datasets. Thus, I specify the paired observations in `weight$Y` and `weight$X` as the first two arguments in `t.test`. I put `weight$Y` first and `weight$X` second in `t.test` to replicate the result I got "by hand". Specifically, I calculated D as `Y - X` (not as `X - Y`).

■

Paired t-test assummptions

The paired t-test has two assumptions which we will formally consider in Lab 8.

1. The underlying distribution of paired differences is normally distributed.
2. paired differences are independent, although raw observations will not be independent.

Assignment 7

Answer all questions in one MS Word document and submit to Moodle. At the beginning of the document include the assignment number, the date, your name and section number.

Use complete sentences when appropriate, and make sure any tables, figures and computer output you include adhere to class standards (see Syllabus).

t-distribution

- Open **R**
- Load the *asbio* package by typing `library(asbio)` or by going to **Packages > Load packages > *asbio***.
- Type `book.menu()` in the **R** console.

1. (2 pts) From the book menu in *asbio* go to **Ch 3 >Pdf depiction > t**. Mac-users and others who wish to obtain the GUI directly can type `see.t.tck()`.

a) How many parameters does the t-distribution have?

b) As the degrees of freedom increase does the t-distribution converge to the standard normal distribution?

The family of t-tests

2. (8 pts) From the book menu in *asbio* go to **Ch 6 >t-test mechanics** or type `see.ttest.tck()`.

a) What do you think the two normal distributions represent?

b) What do think the numbers (ones and twos) inside the distributions are?

c) Why are the degrees of freedom for the t-distribution non-integers when you click off **Variance equal widget**?

d) Set the populations to have EQUAL means. In this case the null hypothesis is true. Resample from the populations repeatedly (> 30 times) by clicking on the **Refresh** button. Is it still possible to reject H_0 at $\alpha = 0.05$? What is this called?

Pooled variance t-test

3. (4 pts) A pollution control inspector suspected that a riverside community was releasing semi-treated sewage into a river. He suspects that the nutrients from the dumping are causing the river to become eutrophic. He records dissolved O_2 readings in ppm for 15 random locations above and below the riverside community. The riverside O_2 data are in the Moodle data folder.

a) Calculate 95% confidence intervals for the true mean dissolved O_2 levels for both above and below locations using the **R** function `ci.mu.t` from *asbio*. To accomplish this you will need to import the riverside O_2 dataset and subset the `dO2` data (column 1) using the `location` column (column 2). The code below will do this and calculate the above and below town confidence intervals. Use snapshots to show results.

```
diss02 <- read.csv(file.choose())
above <- diss02[,1][diss02[,2] == "above"]
below <- diss02[,1][diss02[,2] == "below"]
ci.mu.t(above)
ci.mu.t(below)
```

b) Correctly interpret the result from (a)

4. (12 pts) For the riverside O_2 data, test if the O_2 below the town is lower (more eutrophic) than the O_2 above the town. Assume that the variances are equal. Use $\alpha = 0.05$.

a) State H_0, H_A, and α.

b) Go through the steps necessary to calculate the pooled variance t-test test statistic.

- Calculate sample means, variances, and sample sizes for both hypothesized populations. Use snapshots to show results.
- Calculate MSE, show results using shapshots.
- Calculate t^*, show results using snapshots.

c) Go through the steps necessary to calculate the p-value.

- Calculate the degrees of freedom.
- Calculate the P-value, show results using snapshots.

d) State your conclusions. Do you reject or fail to reject H_0?

e) Verify your answer in **R** using `t.test`. Show a snapshot of your result.

Welch t-test

5. (12 pts) PCB's (polychlorinated biphenyls) are a group of synthetic oil-like chemicals whose toxicity was first recognized in the 1970's. Until then they were widely used as insulation in electrical equipment, particularly transformers. PCB concentrations in heron eggs helps researchers quantify bioaccumulation of PCBs in ecosystems. Thirteen sites in the Great Lakes were selected for a study to quantify PCB concentrations in 1982 and 1996. At each site 9-13 heron eggs were randomly collected and tested for PCBs. Test to see if 1996 levels were lower than 1982 levels. Use a Welch test, and use $\alpha = 0.01$. The data are in the Moodle data folder.

a) State H_0, H_A, and α.

b) Go through the steps necessary to calculate the Welch t-test test statistic.

- Calculate sample means, variances, and sample sizes for both hypothesized populations. Use snapshots to show results.
- Calculate t^*, show results using snapshots.

c) Go through the steps necessary to calculate the p-value.

- Calculate the Satterthwaite degrees of freedom, show results using snapshots.
- Calculate the P-value, show results using snapshots.

d) State your conclusions. Do you reject or fail to reject H_0?

e) Verify your answer in **R** using `t.test`. Show a snapshot of your result.

f) Importantly, the same nests were used in 1982 and 1996. Does the independence assumption for this test appear to be violated? Why? What can you do about it?

Paired t-test

6. (12 pts) Ten hypertensive patients (diastolic blood pressure between 90 - 115 mmHg) were studied before and after 18 months on an antihypertensive treatment. Salt sensitivity (SENS) of the ten patients was evaluated at these two times. Test to see if the salt sensitivity was different after treatment. Use $\alpha = 0.05$. The data are in the Moodle data folder.

a) Are the before and after observations independent? Why or why not?

b) State H_0, H_A, and α.

c) Go through the steps necessary to calculate the paired t-test test statistic.

- Calculate sample mean, variance, and sample size for the paired differences. Use snapshots to show results.
- Calculate t^*, show results using snapshots.

d) Go through the steps necessary to calculate the p-value.

- Calculate the degrees of freedom.
- Calculate the P-value, show results using snapshots.

e) State your conclusions. Do you reject or fail to reject H_0?

f) Verify your answer in **R** using `t.test`. Show a snapshot of your result.

Appendix: R-code used in this lab

This lab focused on t-tests and frequently used the function `t.test`. Arguments from `t.test` can be modified to address situations in which:

- Populations variances are assumed to be approximately equal: $(1/2 < s_X/s_Y < 2)$ `var.equal = T`, or unequal: (default) `var.equal = F`.
- Samples are paired: `paired = T`, or unpaired: (default) `paired = F`.
- Specific alternative hypotheses are required. For example, lower-tailed: `alternative = "less"`, upper-tailed: `alternative = "greater"`, and two-tailed (default): `alternative = "two.tailed"`.

The `t.test` function can be used with two data formats: 1) data from hypothesized population are in separate columns (Table 7.4) and 2) data are in columns representing a quantitative response variable and a categorical explanatory variable with two categorical levels (Table 7.5). Care must be taken in both situations to insure that the correct hypotheses are being tested.

Table 7.4. Illustrative dataset 1. Plant biomass given high and low nitrogen treatments.

High N	Low N
10.2	12.1
21.2	12.3
13.3	11.1
12.5	8.2
10	7.6

Table 7.5. Illustrative dataset 2. Plant biomass given high and low nitrogen treatments.

Biomass	Treatment
10.2	High N
21.2	High N
13.3	High N
12.5	High N
10	High N
12.1	Low N
12.3	Low N
11.1	Low N
8.2	Low N
7.6	Low N

In data format one, data from the hypotheses are specified as the first two arguments `t.test`. The function then assumes that the first and second argument constitute the arrangement of treatments in hypotheses. For instance, to consider the lower-tailed alternative hypothesis $H_A : \mu_{\text{Lo N}} < \mu_{\text{Hi N}}$, I would specify:

```
with(data1, t.test(Lo.N, Hi.N, alternative = "less"))
```

Whereas, to consider the upper-tailed alternative hypothesis $H_A : \mu_{\text{Hi N}} > \mu_{\text{Lo N}}$ (the equivalent test), I would specify:

Q1 2pts, Q2 8pts, Q3 4pts, Q4 12pts, Q5 12pts, Q6 12pts. **Total pts: 50.**

```
with(data1, t.test(Hi.N, Lo.N, alternative = "greater"))
```

In data format two, data are specified as a formula, `Y ~ X`, where `Y` and `X` are response and explanatory data objects, respectively. In this case, the alphanumeric ordering of levels in `X` will constitute the arrangement of treatments in hypotheses. For instance, to consider the upper-tailed alternative hypothesis $H_A : \mu_{\text{Hi N}} > \mu_{\text{Lo N}}$, I would specify:

```
with(data2, t.test(Biomass ~ Treatment, alternative = "greater"))
```

The equivalent test, $H_A : \mu_{\text{Lo N}} < \mu_{\text{Hi N}}$ cannot be specified because Hi N will be ordered before Lo N alphanumerically (because H occurs before L in the alphabet).

8

Assumptions and Diagnostics for *t*-tests

Lab 8 Topics

1. Assumptions and Diagnostics for t-tests.
    - Normality of Parent Distributions.
    - Homoscedaticity
    - Independence of Observations
2. Log-transformation

Last week we were introduced to three kinds of t-tests: the pooled variance t-test, the Welch t-test and the paired t-test. These tests have the following assumptions:

1. All t-test procedures assume normal distributions for the parent populations under consideration. For practical purposes, however, this assumption is only really important for small sample sizes ($n < 30$). This is true because t-tests are concerned with the sampling distribution of mean differences. This distribution will converge to normality under the central limit theorem if sample sizes are large.

2. The pooled variance t-test assumes equal variances (homoscedasticity). The term for unequal variances is heteroscedasticity.

3. All three t-tests assume independence of analyzed responses. For the paired t-test, the independence assumption concerns the differences of paired (blocked) observations. Raw paired observations in a matched pair experimental design are unlikely to be independent. We will not formally address independence diagnostics in this lab, but we will consider this topic in upcoming labs.

P-values from t-tests will be exact if the parent populations are normal and sample sizes from those populations are identical. t-procedures are reasonably robust against both non-normality and unequal variances violations when the samples are nearly the same size, particularly if sample sizes are large. Thus, whenever possible, treatments should have equal (balanced) sample sizes. When the parent populations being compared have different distributional shapes, larger sample sizes will be needed for reliable P-values.

Diagnostics for Homoscedasticity

F-distribution

the **F-distribution** is often used as a null distribution in null hypothesis tests for homoscedasticity. If a random variable X follows an F-distribution, it will have the PDF:

$$f(x) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \cdot \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} \cdot x^{\left(\frac{\nu_1}{2}-1\right)} \cdot \left(1+\frac{\nu_1}{\nu_2}x\right)^{-\left(\frac{\nu_1+\nu_2}{2}\right)} \tag{8.1}$$

where $x > 0$, $\nu_1 > 0$, $\nu_2 > 0$, and $\Gamma(.)$ is the gamma function (Lab 7). The F- distribution has two parameters, ν_1 and ν_2. These are called the **numerator degrees of freedom** and **denominator degrees of freedom**, respectively. Thus, if a random variable X follows an F-distribution, we would denote this as $X \sim F(\nu_1, \nu_2)$. The distribution $F(3, 2)$ is shown in Fig 8.1.

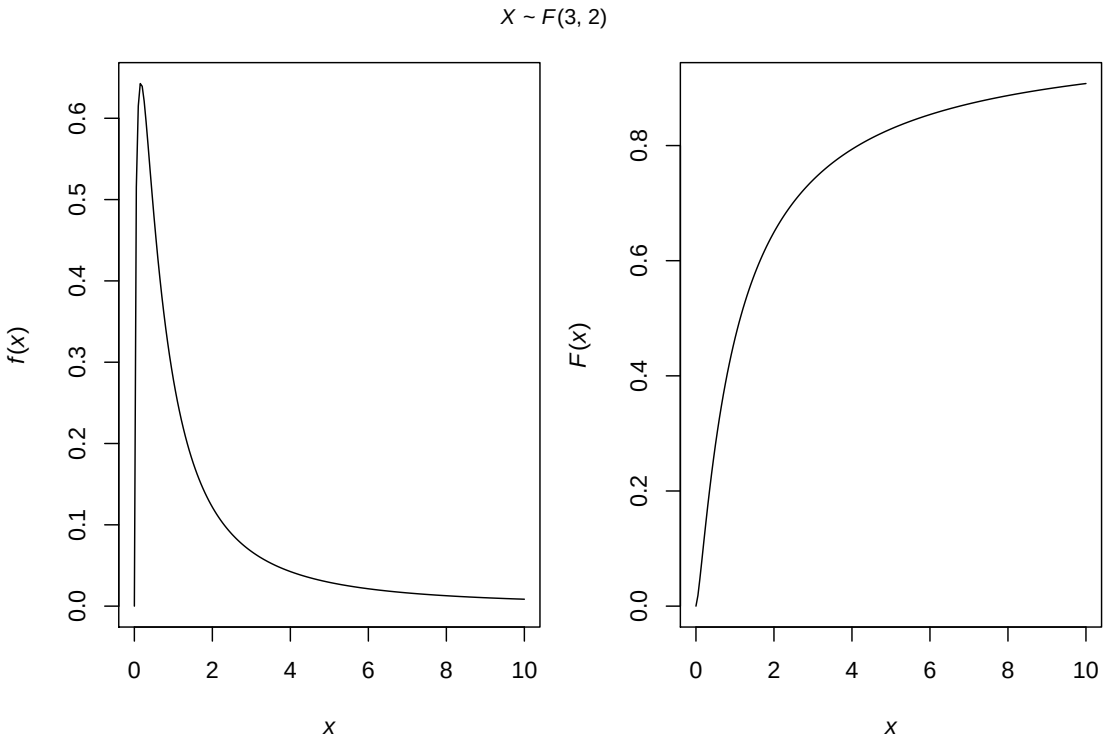


Figure 8.1. The PDF (left) and CDF (right) of the distribution $F(3, 2)$.

Note that unlike the normal and t-distributions, which are defined between $-\infty$ and ∞ the F-distribution is only defined between 0 and ∞.

F-test

An **F-test** can be used to test for the equality of variances, σ_2^2 and σ_1^2 from two normal distributions. The test has the hypotheses:

$$\mathrm{H_0} : \sigma_1^2 = \sigma_2^2$$
$$\mathrm{H_A} : \sigma_1^2 \neq \sigma_2^2$$

The test statistic for the F-test is:

$$F^* = \frac{S_1^2}{S_2^2} \tag{8.2}$$

where S_1^2 is the larger of the sample variances taken from the two parent populations. Under this framework $F^* \geqslant 1$. If $\mathrm{H_0}$ is true, and assumptions for the test are valid, F^* will follow an F-distribution with $n_1 - 1$ numerator degrees of freedom and $n_2 - 1$ denominator degrees of freedom. The P-value is calculated as $2 \cdot P(X \geqslant F^*)$ where $X \sim F(n_1 - 1, n_2 - 1)$. F-statistics much greater than one would suggest that $\mathrm{H_0}$ is false.

The F-test assumes normality of the parent distributions being compared and is highly sensitive to violations of this assumption.

Levene's test

Several null hypothesis tests can assess homoscedasticity without the assumption of normality for the the parent distributions. One example is the **modified Levene's test**. The test also allows comparison two or more variances at a time. Given r population variances, $i = 1, 2, 3, \ldots, r$, we are testing the hypotheses:

$$\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 \ldots = \sigma_r^2$$
$$\mathrm{H}_\mathrm{A} : \text{At least one } \sigma_i^2 \text{ not equal to the others.}$$

The modified Levene's test is considered a better test for homoscedasticity than the F-test because of its lack of normal assumptions for the parent distributions. We will not worry about the mechanics of the modified Levene's test here.

Example 8.1

An agricultural field station is curious about the efficacy of two types of spray in controlling insects (Table 8.1). The investigators would like to use a pooled variance t-test because it has more power than the Welch t-test. Thus, they would like to know whether their data meet the pooled variance t-test assumption of homoscedasticity.

Table 8.1. Data for sprays 1 and 2 in Example 1.

Spray 1	4	5	3	4	2	1	6	2	3	4	3	4	5	2
Spray 2	1	1	3	1	2	2	3	3	5	2	4	4	3	2

We will consider the F-test first. We proceed with the four steps of H_0 hypothesis testing.

1. State H_0, H_A and α. We will use $\alpha = 0.05$, and test the following hypotheses:

$$\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2$$
$$\mathrm{H}_\mathrm{A} : \sigma_1^2 \neq \sigma_2^2$$

2. Calculate the test statistic. To calculate F^* we bring in the spray data.

```
spray <- read.csv(file.choose())
```

and calculate the sample variances and sample sizes.

The F-test can also be extended to cases with more than two variances.

```
vars <- tapply(spray[,1], spray[,2], var)
vars

       1        2
1.956044 1.494505

ns <- tapply(spray[,1], spray[,2], length)
ns

 1  2
14 14
```

We find:

$$F^* = \frac{1.956044}{1.494505} = 1.308824.$$

3. We calculate the P-value using two times the upper tail of $F(13, 13)$.

```
F.star <- vars[1]/vars[2]
2 * pf(F.star, 13, 13, lower.tail = F)

        1
0.6346121
```

We can verity our results using the function `var.test`.

```
var.test(spray[,1] ~ spray[,2])

F test to compare two variances

data:  spray[, 1] by spray[, 2]
F = 1.3088, num df = 13, denom df = 13, p-value = 0.6346
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4201633 4.0770320
sample estimates:
ratio of variances
          1.308824
```

4. We fail to reject H_0 and conclude that the underlying variances, σ_1^2 and σ_2^2, are equal.

Next, we will apply the modified-Levene's test. We consider the same hypotheses as the F-test with the modified Levene's test. The function `modlevene.test` from *asbio* can be used to run the modified Levene's test in **R**:

```
library(asbio)
modlevene.test(spray[,1], spray[,2])

Modified Levene's test of homogeneity of variances

df1 = 1,  df2 = 26,  F = 0.29213,  p-value = 0.59346
```

The test here provides additional support for H_0. Thus, the investigators conclude that use of the pooled variance t-test is justified, given validity of the assumption of normality for the underlying populations.

■

Diagnostics for Normality

Normal probability plot

One tool for diagnosing population normality is the **normal probability plot** also called a **normal quantile plot**. If data are sampled from a normal distribution, then there will be a certain pattern to the data distribution. For instance, according to the empirical rule approximately 68% of the data should be one standard deviation from the mean, 95% of the data should be two standard deviations from the mean, and 99.7% of the data should be within three standard deviations of the mean. A normal probability plot examines the relationship between the way the data are actually distributed (the **sample quantiles**) and the way the data would be distributed under normality (the **theoretical quantiles**). If data *are* normally distributed, there should be a strong linear relationship between these outcomes (all points should be near a linear fit line). The `qqnorm` and `qqline` provide normal probability plots and normal probability plot fits in **R**. The function `qq.Plot` from *asbio* can provide normal probability plotting for multiple datasets.

Shapiro-Wilk test

We can also formally test the null hypothesis of normality.

$$H_0: \text{ The underlying population is normally distributed.}$$
$$H_A: \text{The underlying population is } \textit{not} \text{ normally distributed.}$$

The most frequently used test for these hypotheses is the **Shapiro-Wilk test**. We will not address the mechanics for this test here. The function `shapiro.test` runs the Shapiro-Wilk test in **R**.

Example 8.2

We will apply normal diagnostics to the insect spray data from Example 1. A normal probability plot for both spray types in shown in Fig 8.2.

```
qq.Plot(spray[,1], spray[,2], col = c("black","gray"), pch = c(1,2))
```

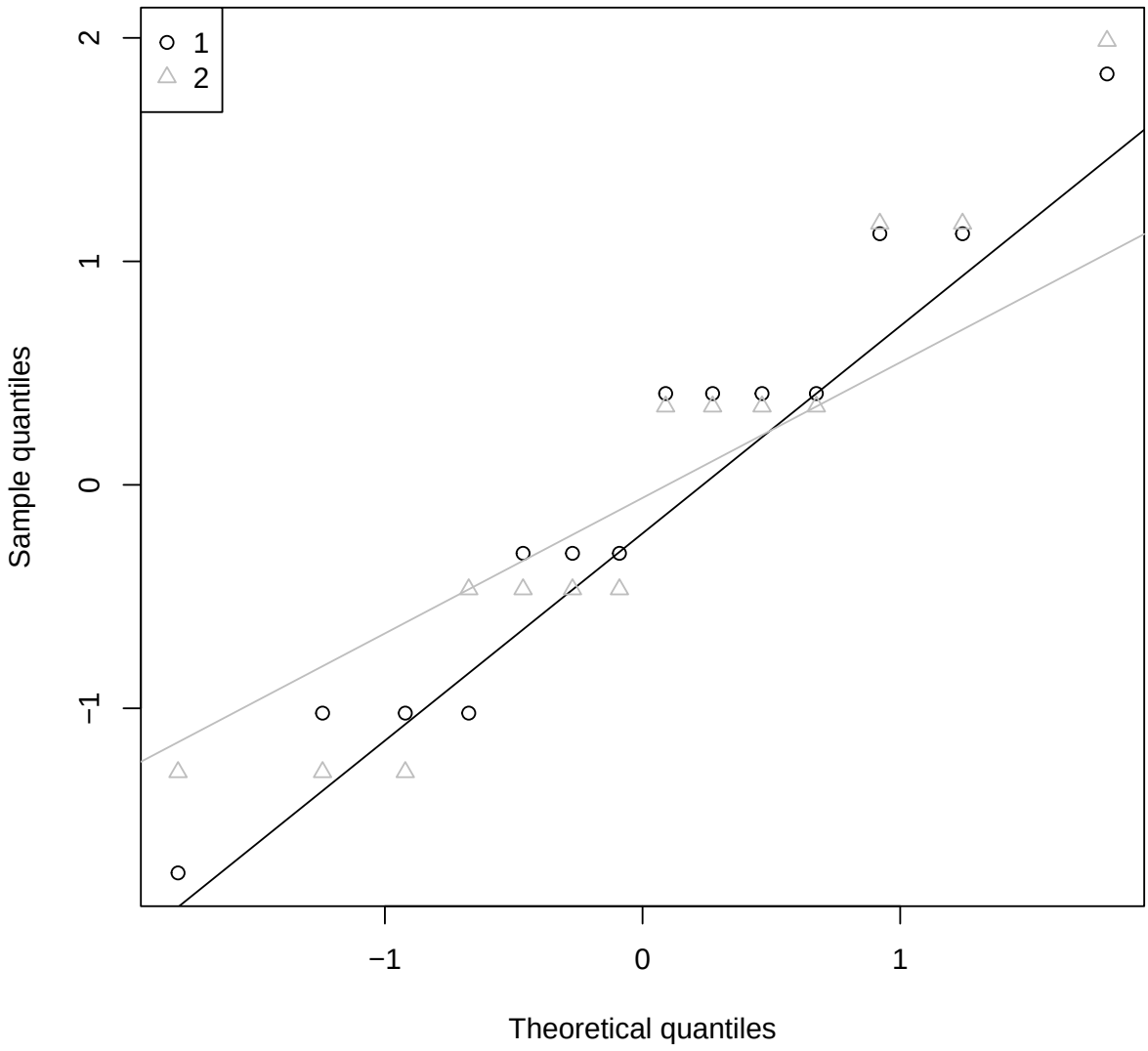


Figure 8.2. Normal quantile plots for treatments in the insect counts data set.

Here we apply the Shapiro-Wilks test to observations from spray 1 and 2

```
shapiro.test(spray[,1][spray[,2]==1])

Shapiro-Wilk normality test

data:  spray[, 1][spray[, 2] == 1]
W = 0.9595, p-value = 0.7148
```

```
shapiro.test(spray[,1][spray[,2]==2])

Shapiro-Wilk normality test

data:  spray[, 1][spray[, 2] == 2]
W = 0.92276, p-value = 0.2409
```

For both treatments we fail to reject H_0 at $\alpha = 0.05$ but, as with the homoscedasticity tests, this is what we generally want to do! It looks like the investigators will be able to perform a pooled variance t-test.

■

Log-transformation

If sample sizes are small, sample sizes are highly unequal, and the data are non-normal, then t-tests should not be used. In addition, if variances are unequal, a pooled variance t-test cannot be used.

However, data can often be transformed to get it into a required distributional form. Homoscedasticity is often violated because the treatment variances are proportional to the treatment means. That is, variances increase with increasing treatment means, and are hence not equal. An easy fix to this predicament is to **log-transform** data from the treatments. This will preserve mean differences between treatments, if any, while equilibrating the treatment variances.

For the exponential function $y = a^x$, we call a the **base** and x the **exponent**. Logarithms are the inverse of exponential functions. That is, if $y = a^x$, then $\log_a(y) = x$. For example, $\log_{10} 1 = 0$, because $10^0 = 1$. In a log-transformation, we take the log (generally or $\log_e$ or $\log_{10}$) of all the observations and use those outcomes as the new observations. We can perform log transformations in **R** using `log`. By default, `log` performs $\log_e$ transformations.

Example 8.3

```
data <- c(2,4,3,5,20,70,1000)
log(data)

[1] 0.6931472 1.3862944 1.0986123 1.6094379 2.9957323 4.2484952 6.9077553
```

■

Assignment 8

Answer all questions in one MS Word document and submit to Moodle. At the beginning of the document include the assignment number, the date, your name and section number.

Use complete sentences when appropriate, and make sure any tables, figures and computer output you include adhere to class standards (see Syllabus).

t-test assumptions and diagnostics

- Open **R**
- Load the *asbio* package by typing `library(asbio)` or by going to **Packages > Load packages > *asbio***.
- Type `book.menu()` in the **R** console.

1. (11 pts) Go to **Ch. 5 > Sampling distributions**. Mac-users and others who wish to obtain the GUI directly can type `samp.dist.method.tck()`. Under **Statistic** choose $\mathbf{t}^*$ **(2 sample)**; this will depict the sampling distribution for the test statistic from a pooled variance t-test.

a) Choose **Snapshot**, and click **Submit**. The histograms represent the sampling distributions of pooled variance t-test statistics. The gray line is $t(n_1 + n_2 - 2)$. This is the correct H_0 distribution if H_0 is true, and assumptions for the test are met. The black dotted line is $N(0, 1)$. The randomly generated test statistic distribution should closely resemble $t(n_1 + n_2 - 2)$, because H_0 is true (by default, both parent populations are standard normal, and hence have the same mean, 0) and the assumptions for the test are valid (the parent distributions are normal with the same variance). Paste the figure into your assignment with an appropriate caption.

b) Change the first parent distributions to $UNIF(-6, 6)$ by pasting or typing `expression(runif(s.size, -6, 6))` in one of the two parental distribution widgets. Note that the mean of this parent distribution will be zero. Let the the other parent distribution remain $N(0, 1)$.

Thus, the difference in the population means remains $0 - 0 = 0$, and H_0 remains true. To facilitate comparisons, change the sample sizes in the GUI to be `c(3,3,10,20)` and `c(3,5,10,20)`. Run the function by clicking **Submit**. Once again, the resulting histograms represent the sampling distributions of the t-test test statistics, the gray line is $t(n_1 + n_2 - 2)$ and the black dotted line is $N(0, 1)$. Paste the figure into your assignment with an appropriate caption.

c) Which test statistic sampling distribution in (b) appears to follow $t(n_1 + n_2 - 2)$ most closely? Which test statistic sampling distribution fits $t(n_1 + n_2 - 2)$ most poorly? Your possible choices for each question are: $n_1 = n_2 = 3$; $n_1 = 3, n_2 = 5$; $n_1 = n_2 = 10$; and $n_1 = n_2 = 20$.

d) Which assumptions for the pooled variance t-test are being violated by the altering one of the parental distributions in (b)?

e) Do these violations of assumptions appear to be relatively unimportant for larger, and equal, sample sizes? Why do you think that this is true?

2. (5 pts) Go to **Ch. 3 > Pdf depiction** and choose ***F*-distribution**. Mac-users and others who wish to obtain the GUI directly can type: `see.F.tck()`.

a) What is the lower limit and upper limit to the distribution?

b) What is the distribution shape?

c) How many parameters does the distribution have?

3. Soil nitrogen can have a strong effect on plant biomass. Download the dataset `biomass.csv` from Moodle. The data consider plant biomass as a function of high and low soil nitrogen treatments. You want to eventually run a t-test for the alternative hypothesis that high N sites will have higher plant biomass than low N sites. First, however, you will have to run some diagnostics to identify the correct t-test to use (pooled variance t-test or Welch test, depending on the validity of the homoscedasticity assumption) and to determine if you can even use a t-test (by diagnosing normality). For all components below, demonstrate your work, when necessary, using snapshots.

a) (10 pts) perform diagnostic checks for homoscedasticity for the high and low N treatments.

i) State hypotheses appropriate for null hypothesis tests for homoscedasticity for two parent populations (see section describing F-tests).

ii) Use the F-test to test the null hypothesis of homoscedasticity.

- Calculate the F-statistic "by hand" using **R** or Excel to help.
- Calculate P-value for the F-test "by hand" using `pf` in **R** or =F.DIST in Excel.
- State your conclusions.
- Check your result in **R** using `var.test` (see Example 1).

iii) Recheck your results from (ii) using the modified Levene's test. Run the test using the function `modlevene.test` from *asbio* (see Example 1).

iv) State your conclusions.

v) What advantage does the modified Levene's test have over the F-test?

b) (8 pts) Perform diagnostic checks for normality for the plant biomass data.

i) Make histograms and or normal quantile plots for both the high N and low N treatments.

ii) State hypotheses appropriate for the Shapiro-Wilks test for normality (see section describing the Shapiro-Wilks test).

iii) Run Shapiro's test using `shapiro.test`.

iv) State your conclusions.

c) (5 pts) Given results from (a) and (b) conduct the appropriate t-test.

i) State the correct t-test hypotheses (see Lab 7).

ii) Run the test using `t.test`.

iii) State your conclusions.

4. (5 pts) We want to test if three methods for measuring soil percent nitrogen differ with respect to their variability: "Are some methods more precise than others?" We have three methods, A, B and C. The data can be found in the data folder for this week in Moodle, under the name `N_methods.csv`.

a) State the correct null and alternative hypotheses. See example shown for

the modified Levene's test when there are more than two hypothesized populations.

b) Run the test in **R** using `modlevene.test`. Use snapshots to show work.

c) What are your conclusions?

Log-transformation

5. (7 pts) Download the dataset `height.csv` describing plant height (in cm) as a function of high and low soil N.

a) Provide hypotheses appropriate for null hypothesis tests for homoscedasticity for two parent populations.

b) Run Levene's test for plant height with respect to high and low N treatments. Use snapshots to show work.

c) What would be a good transformation so that the variances will be equal but mean differences will still be present?

d) Perform the transformation and run the Levene's test on the transformed data. Use snapshots to show work.

e) Discuss your results. Which dataset had more evidence of homoscedasticity?

Q1 11pts, Q2 5pts, Q3a 5pts, Q3b 5pts, Q3c 8pts, Q4 5pts, Q5 7pts. **Total pts: 51.**

Appendix: R-code used in this lab

This lab focused on diagnostics for heteroscedasticity and normality assumptions of t-tests.

Operator	Operation	Description
`var.test(y ~ x)`	F-test	Test H_0: $\sigma_1^2 = \sigma_2^2$, given quantitative responses in `y` with respect to levels 1 and 2 in `x`.
`asbio:modlevene.test(y, x)`	Modified Levene's test	Test H_0: $\sigma_1^2 = \sigma_2^2$, given quantitative responses in `y` with respect to levels 1 and 2 in `x`.
`shapiro.test(y)`	Shapiro-Wilks test	Test H_0: Underlying distribution is normal, given quantitative responses in `y`.
`qqnorm(y)`	Normal quantile plot	Create normal probability plot given quantitative responses in `y`
`qqline(y)`	Linear fit for normal quantile plot	Overlays linear fit for normal probability plot, given quantitative responses in `y`
`qq.Plot(y, x)`	Multiple normal quantile plot overlays, and associated linear fits	Creates multiple normal quantile plots overlays, given quantitative responses in `y` and associated categories in `x`.

9

Alternatives to *t*-tests

Lab 9 Topics

1. Rank-based permutation tests, including the Wilcoxon rank sum test
2. Strictly permutational tests
3. Comparison of nonparameteric approaches

If sample sizes are small, and/or sample sizes are highly unequal, and/or the data are highly non-normal, t-tests should not be used. Outliers will also have a very negative effect on valid inferences from t-tests. There are a number of different **nonparametric** alternatives to t-tests. Nonparametric tests don't rely on $a-priori$ distributions or parent populations. These methods are generally resistant to violations of t-test assumptions, particularly non-normality, and the presence of outliers.

Rank-Based Permutation – Wilcoxon Rank Sum Test

The Wilcoxon rank sum test can be considered a nonparametric analog of a pooled variance t-test. The test is equivalent to another nonparametric procedure called the **Mann-Whitney test**. The Wilcoxon test does not assume normal distributions for the populations being compared, and is resistant to outliers. However, it does assume that the distributions have similar shapes (thus, we ostensibly assume that underlying population variances are equal).

Let Δ be the true shift in location of one population with respect to a second population or hypothesized value. That is,

$$\Delta = \text{Location of population 1} - \text{Location of population 2.}$$

We are concerned with the following hypotheses:

- Two-tailed:
$$\mathrm{H}_0 : \Delta = 0$$
$$\mathrm{H}_\mathrm{A} : \Delta \neq 0$$

- Lower-tailed:
$$\mathrm{H}_0 : \Delta \geqslant 0$$
$$\mathrm{H}_\mathrm{A} : \Delta < 0$$

- Upper-tailed:
$$\mathrm{H}_0 : \Delta \leqslant 0$$
$$\mathrm{H}_\mathrm{A} : \Delta > 0$$

The Wilcoxon rank sum test is based on **ranked-transformed** data. The calculation of its test statistic is based on three steps.

1. Rank-transform data. That is, order the data, for both samples from smallest to largest values, then assign numbers from 1 to N (where $n_1 + n_2 = N$ is the total number of observations from both samples).

   - If there are ties (duplicated values), the ranks in the data are taken to be the average of the ranks for those observations.

2. Calculate W_1 and W_2:
$$W_1 = T_1 - \frac{n_1(n_1+1)}{2} \tag{9.1}$$
$$W_2 = T_2 - \frac{n_2(n_2+1)}{2} \tag{9.2}$$
   where n_1 = the sample size for population 1, n_2 = the sample size for population 2, T_1 is the sum of ranks for the sample from population 1, and T_2 is the sum of ranks for the sample from population 2.

3. Define the test statistic, W^*. For a two-tailed test, W^* will be whichever is smaller, W_1 or W_2. In an upper-tailed test, W^* will be W_1. In a lower tailed test, W^* will *also* be W_1. If H_0 is true, then W^* will be a random outcome from the Wilcoxon rank sum distribution, with parameter values n_1 and n_2.

   The Wilcoxon rank sum distribution is a two-parameter, discrete, bell-shaped, non-negative distribution, that describes the distribution of Wilcoxon rank sum test statistic, when comparing two unpaired samples drawn from the same arbitrary distribution.

   Let $W \sim RankSum(n_1, n_2)$.

   - For a two tailed test, the P-value is calculated as $2 \cdot P(W \leqslant W^*)$.
   - For a lower tailed test the P-value is $P(W \leqslant W^*)$.

- For an upper tailed test the P-value is $P(W \geqslant W^*)$.

  Note: care will be required to compute the upper-tailed P-value, because W is a discrete random variable. Specifically, while the Wilcoxon rank sum CDF always gives $P(W \leqslant W^*)$ (like all CDFs), it is also true that in the discrete case, $P(W \geqslant W^*) = P(W > (W^* - 1))$. Thus, unlike continuous distributions, $P(W \geqslant W^*) \neq 1 - P(W \leqslant W^*)$.

Example 9.1

A committee is studying the effect on alcohol on reaction time. They randomly assigned 20 subjects to either an alcohol treatment or a placebo treatment then measured reaction times of the subjects (Table 9.1).

Table 9.1. The effect of alcohol on reaction time in seconds.

Placebo	0.9	0.37	1.63	0.83	0.95	0.78	0.86	0.61	0.38	1.97
Alcohol	1.46	1.45	1.76	1.44	1.11	3.07	0.98	1.27	2.56	1.32

Histograms of the alcohol data are shown in Figure 9.1. Is a t-test appropriate? Why or why not?

```
## alcohol <- read.csv(file.choose())
qq.Plot(alcohol[,1], alcohol[,2], col = c("black","gray"), pch = c(1,2))
```

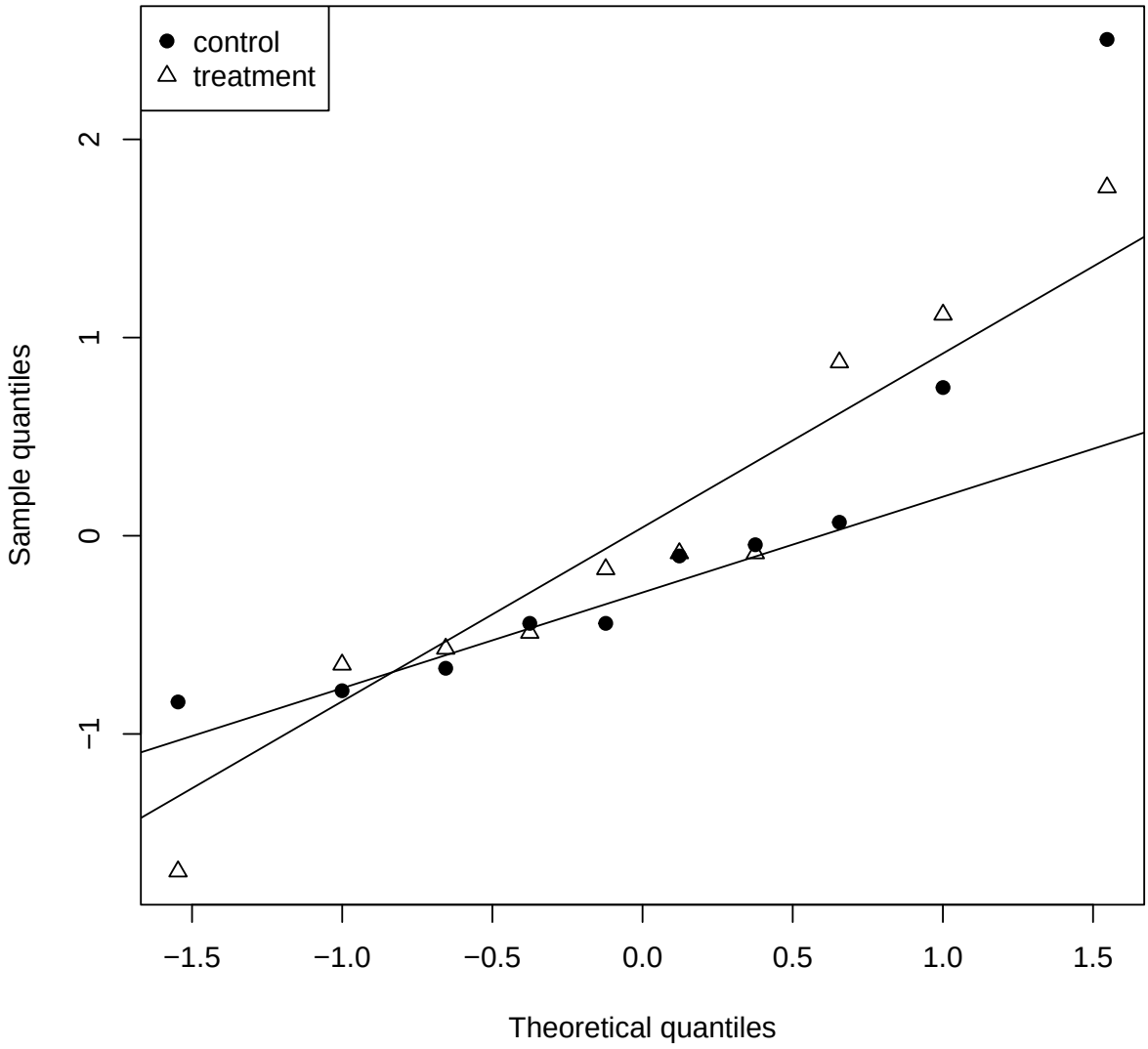


Figure 9.1. Normal quantile plots for the placebo and alcohol groups in Example 1. The data are strongly right skewed and thus non-normal for both groups.

We go through the four steps of null hypothesis testing.

1. We expect that reaction times for the alcohol group will be larger (slower). Thus, we define: Δ = Alcohol location − Placebo location, and test the hypotheses:

$$H_0 : \Delta \leqslant 0$$
$$H_A : \Delta > 0$$

We will use $\alpha = 0.05$.

2. Next we calculate the test statistic, W^*. This requires rank-transforming the data (Table 9.2).

Table 9.2. Raw and ranked data for Example 1. A = alcohol, P = Placebo.

Raw data		Ranked data		
Reaction time	Treatment	Reaction time	Treatment	Rank
0.9	P	0.37	P	1
0.37	P	0.38	P	2
1.63	P	0.61	P	3
0.83	P	0.78	P	4
0.95	P	0.83	P	5
0.78	P	0.86	P	6
0.86	P	0.9	P	7
0.61	P	0.95	P	8
0.38	P	0.98	A	9
1.97	P	1.11	A	10
1.46	A	1.27	A	11
1.45	A	1.32	A	12
1.76	A	1.44	A	13
1.44	A	1.45	A	14
1.11	A	1.46	A	15
3.07	A	1.63	P	16
0.98	A	1.76	A	17
1.27	A	1.97	P	18
2.56	A	2.56	A	19
1.32	A	3.07	A	20

To obtain the test statistic, we sum the ranks of the alcohol group (representing population 1, i.e., the 1st population in the hypotheses) and the sum of the ranks of the placebo group (representing population 2, i.e., the 2nd population in the hypotheses). We get:

$$T_1 = 9 + 10 + 11 + 12 + 13 + 14 + 15 + 17 + 19 + 20 = 140.$$
$$T_2 = 1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 16 + 18 = 70.$$

In **R** we could do something like:

```
ranks <- rank(alcohol[,1])
ns <- tapply(alcohol[,1], alcohol[,2], length)
n1 <- ns[1]; n2 <- ns[2]

T1 <- sum(ranks[alcohol[,2]=="Alcohol"])
T1


[1] 140


T2 <- sum(ranks[alcohol[,2]=="Placebo"])
T2


[1] 70
```

Calculating W_1 and W_2 we have:

$$W_1 = T_1 - \frac{n_1(n_1+1)}{2} = 140 - \frac{10(10+1)}{2} = 140 - 55 = 85.$$
$$W_2 = T_2 - \frac{n_2(n_2+1)}{2} = 70 - \frac{10(10+1)}{2} = 70 - 55 = 15.$$

We have a one-tailed test, so we use W_1 for the test statistic. We have:

$$W^* = W_1 = 85.$$

```
W1 <- T1 - n1 * (n1 + 1)/2
W1


Alcohol
     85


W2 <- T2 - n2 * (n2 + 1)/2
W2


Placebo
     15


W.star <- W1
```

3. To calculate the P-value, we first determine the null distribution. Because $n_1 = n_2 = 10$, our null distribution is $W \sim RankSum(10, 10)$. We were warned above that calculating an upper-tailed P-value using the Wilcoxon rank sum distribution requires that we find: $P(W > (W^* - 1))$. Thus, for the P-value we have:

```
pwilcox((W.star-1), 10, 10, lower.tail = F)


    Alcohol
0.003420728
```

4. We reject H_0 and conclude that reaction times for the alcohol group are slower than for the placebo group.

Using the function `wilcox.test` directly we have:

```
wilcox.test(alcohol[,1] ~ alcohol[,2], alternative = "greater")

Wilcoxon rank sum exact test

data:  alcohol[, 1] by alcohol[, 2]
W = 85, p-value = 0.003421
alternative hypothesis: true location shift is greater than 0
```

■

The sampling distribution of W^*

If sample sizes for both groups is greater than 10, the sampling distribution of W^* will be approximately normally distributed, with mean:

$$\mu_{W*} = \frac{n_1 n_2}{2}, \tag{9.3}$$

and variance:

$$\sigma^2_{W*} = \frac{n_1 n_2}{12} \cdot (n_1 + n_2 + 1). \tag{9.4}$$

In the presence of ties, the formula for the variance of W^* becomes more complicated:

$$\sigma^2_{W*} = \frac{n_1 n_2}{12} \cdot \left((n_1 + n_2 + 1) - \frac{\sum_{i-1}^{k} t_i(t_i^2 - 1)}{(n_1 + n_2)(n_1 + n_2 - 1)} \right). \tag{9.5}$$

where k = the number of ties, and t_i denotes the number of tied observations in the ith group of ties.

As we learned earlier (Lab 5), we can convert any normal distribution to a Z distribution by subtracting the mean of the distribution, and dividing by the standard deviation of the distribution. Thus we can calculate a z-test test statistic from a W^* test statistic using:

$$z^* = \frac{W^* - \mu_{W*}}{\sigma_{W*}}. \tag{9.6}$$

Notably, with most statistical software, the P-value for the Wilcoxon rank sum test is calculated using the normal approximation z-statistic, particularly in the case of ties.

Example 9.2

For the alcohol reaction time example we have:

$$\mu_{W*} = \frac{10 \cdot 10}{2} = 50.$$

Because there are no ties, we can use Eq 9.4 to calculate σ^2_{W*}. We have:

$$\sigma^2_{W*} = \frac{10 \cdot 10}{12} \cdot (10 + 10 + 1) = \frac{100}{12} \cdot 21 = 175.$$

The resulting z-score is

$$z^* = \frac{W^* - \mu_{W*}}{\sigma_{W*}} = \frac{85 - 50}{\sqrt{175}} = 2.645751.$$

The P-value is calculated as $P(Z \geqslant z^*)$, based on the upper-tailed form of H_A. Thus, we have:

```
pnorm(2.645751, lower.tail = F)

[1] 0.00407549
```

This P-value is very similar to the one we calculated in Example 1. The P-values would be identical if the sample sizes were extremely large.

■

Strictly Permutational Procedures

In a conventional **permutation test procedure**, a random distribution of test statistics is created by randomizing sample data. The observed test statistic is compared to this empirical distribution to calculate a P-value. For example, consider an experiment where two hypothetical treatment populations are being compared. The treatment populations are both sampled, and a t-statistic is calculated, denoted t_0. A permutational algorithm is then used to randomly reassign treatments to observations. At each round of permutations, a t-statistic is calculated, resulting (after many iterations) in an empirical distribution of test statistics. The observed test statistic, t_0, is compared to this distribution to find the permutational P-value. This is calculated as the number of times an outcome as or more extreme than t_0 occurs, divided by the number of permutations. Generally, to be conservative, we allow t_0 to be included in the number of observations equaling or exceeding the observed value. Thus, we have:

$$P\text{-value} = \frac{\text{No. of random trials } t_0 \text{ is equalled or exceeded} + 1}{\text{No. of random trials}} \tag{9.7}$$

Upper and lower-tailed tests are performed by finding the proportion of the empirical null distribution greater than or equal to, or less than or equal to t_0, respectively. A two-tailed test is performed by multiplying the proportion of the null distribution above the absolute value of t_0 by two.

The function `MC.test` from *asbio* performs upper, lower, and two-tailed tests using this approach. By default, a pooled variance t-test procedure is used to calculate test statistics.

Example 9.3

Revisiting the alcohol example we have:

```
library(asbio)
MC.test(alcohol[,1], alcohol[,2], alternative = "greater")

Monte Carlo t-test

Paired = FALSE, No. perms = 1000
Alternative: Alcohol greater Placebo

     Obs. test stat   Perms > test stat              P-val
           2.698941            6.000000           0.007000
```

■

A Comparison of Nonparametric Approaches

Rank based permutation procedures (e.g., the Wilcoxon rank sum test) have three advantages over parametric tests (e.g., t-tests) and non-ranked nonparametric procedures (e.g., strictly permutational tests). First, they are much less sensitive to outliers compared to parametric methods (and strictly permutation tests based on mean differences). Second, because of the fact that their empirical distributions include all possible outcomes, the scope of inference for rank based permutation procedures is generally considered to be less of an issue than for strictly permutational tests. Finally, rank-based permutation procedures are only slightly less powerful than parametric methods if their parametric assumptions hold, and may be more powerful than parametric methods if parametric assumptions do not hold (Pitman, 1949).

Strictly permutational tests are useful when underlying parental distributions are unknown, or when random sampling is not possible. This is because independence of observations is not required (Manly, 2006). A major difficulty with strictly permutational tests, however, is that inference will be hypothetically limited to the sample. Crowley (1992) asserted that the relevance of this problem is largely theoretical, because randomization tests result in similar P-values to parametric tests when parametric assumptions hold (Quinn & Keough, 2002;

Manly, 2006). Like rank-based permutation procedures, tests based strictly permutational tests are sensitive to differences in treatment variances (Boik, 1987). As a result, these procedures should not be looked upon as a cure-all for heteroscedasticity (Quinn & Keough, 2002).

Assignment 9

Answer all questions in one MS Word document and submit to Moodle. At the beginning of the document include the assignment number, the date, your name and section number.

Use complete sentences when appropriate, and make sure any tables, figures and computer output you include adhere to class standards (see Syllabus).

Wilcoxon rank sum test

1. Murakami *et al.* (1997) studied the effect of drug treatments on levels of serum β-2 microglobulin patients with multiple myeloma. Serum β-2 microglobulin is produced in the body as a result of myelomas, and thus can be used as an indicator of the severity of disease. The researchers randomly assigned twenty patients to treatment and control groups (Table 9.3). The treatment patients received two types of drugs: malphalan and sumerifon while the control group received only sumerifon. We will test the hypothesis that the control will have elevated levels of β-2 microglobulin. Use $\alpha = 0.05$.

Table 9.3. Effect of drug treatments on levels of serum β-2 microglobulin in patients with multiple myeloma.

Treatment	2	2.7	3.9	2.7	2.1	2.6	2.2	4.2	5	0.7
Control	3.5	2.5	3.8	8.1	3.6	2.2	5	2.9	2.3	2.9

a) (6 pts) Examine the data with normal quantile plots and histograms.

i) Insert the figures into your document.

ii) Does a t-test appear to be a good idea? Why or why not?

b) (10 pts) Conduct a Wilcoxon rank sum test.

i) State your hypotheses

ii) Rank the data. Provide snapshots to show work.

iii) Calculate the test statistic. Provide snapshots to show work.

- Calculate T_1 and T_2.
- Calculate W_1 and W_2 . Which one is your test statistic?

iv) Calculate the P-value in **R** using `pwilcox`. This will take some care for upper-tailed alternative hypotheses because the rank sum distribution is discrete. Show work.

v) State your conclusions, i.e., can we reject H_0?

c) (2 pts) Run the test using `wilcox.test`. Use the argument `correct = FALSE`. This prevents the application of Yate's correction for discontinuity, and makes it easier to replicate `wilcox.test` results by hand. Provide snapshots to show work.

Normal approximation

2. (7 pts) Calculate a normal approximation P-value for the Wilcoxon test. Show work.

a) Calculate the parameters for the sampling distribution of W^*, μ_{W^*} and $\sigma^2_{W^*}$. You will have to use Eq. 9.5 to calculate $\sigma^2_{W^*}$ because there are ties in the data.

b) Calculate a z-score from your Wilcoxon test statistic

c) Calculate the P-value, under the asymptotic normality of W^*.

d) Does the resulting P-value equal the one provided by `wilcox.test` in (**1biv**) (given rounding error)? Why?

Strictly permuational tests

3. (8 pts) Repeat the test of the Wilcoxon test hypotheses from **Q1** using the permutational algorithm `MC.test` from *asbio*.

a) Run the test. Provide snapshots to show work.

b) Interpret your results. Do they agree or disagree with the results from **Q1** and **Q2**? Why or why not?

c) Run the randomization test again and attach the results.

d) Are your results from your randomizations the same? Why or why not? If they are different, how different are they?

4. (4 pts) What are the advantages and disadvantages of rank-based permutation tests and strictly permutational tests?

Q1 18pts, Q2 7pts, Q3 8pts, Q4 4pts. **Total pts: 37.**

10

Regression I

Lab 10 Topics

1. Simple linear regression
    - Linear regression model
    - Parameter estimation
    - Hypothesis testing

The Simple Linear Regression Model

In a simple linear regression we study the relationship of a single quantitative response variable, Y, and a single quantitative explanatory variable, X. For instance, to measure the degree to which plant height is passed on from parent plant to offspring, we could measure seedling height at maturity, Y, and mean parental height, X, and then "regress" Y on X. A regression analysis provides a model that allows predictions of Y given X. The model can be graphically expressed as a **regression line** that can be overlain on a bivariate scatterplot of the observed Y and X outcomes (Fig 10.2). In a regression plot, X outcomes are positioned with respect to the **ordinate** (horizontal axis) and Y outcomes are positioned along the **abscissa** (vertical axis).

In the context of regression, the X variable is often called the **predictor.**

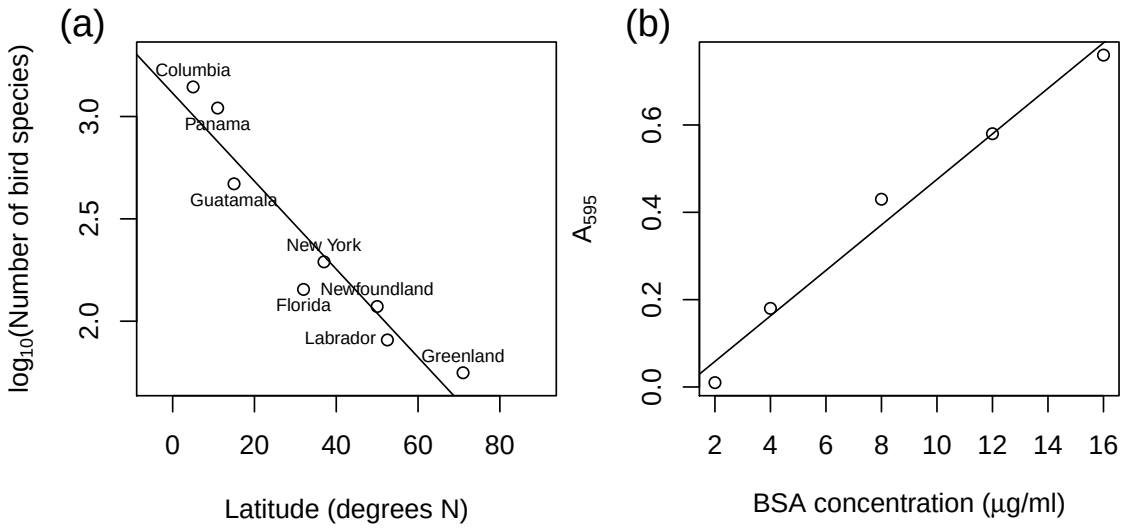


Figure 10.1. Biological applications for regression. Regression lines in plots show the response variable as a linear function of the explanatory variable. The plot in (a) shows $\log_{10}$ numbers of breeding species of birds as a function of latitude, based on the meta-analysis of Dobzhansky (1950). The plot in (b) is an example of a Bradford colorimetric protein assay (Bradford, 1976). Absorbance units at 595 nanometers are plotted as a function of the concentration of Bovine Serum Albumin (BSA) which has been mixed with the Bradford reagent. Figure from Aho (2014).

The simple linear regression model has the form:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i. \tag{10.1}$$

where β_0 and β_1 are parameters that define Y as a function of X. Thus, the model in Eq. 10.1 is an idealization called the **population regression model** or **true regression model**, based on *all possible* X, Y pairs.

- β_0 is the **true Y-intercept**, the true mean value of Y when $X = 0$ (Fig 10.2). As a result the units for β_0 will be the same as the units of the Y variable. When the scope of the regression model includes $X = 0$, then the intercept term has interpretable meaning. However, when the scope of the model does not cover $X = 0$ (the value of Y given $X = 0$ is **extrapolated**), then β_0 will not be interpretable.

- β_1 is the **true slope**, the average linear change in the response variable as the result of a one unit increase in the explanatory variable (Fig 10.2). The units for β_1 are the ratio of the units of Y and X.

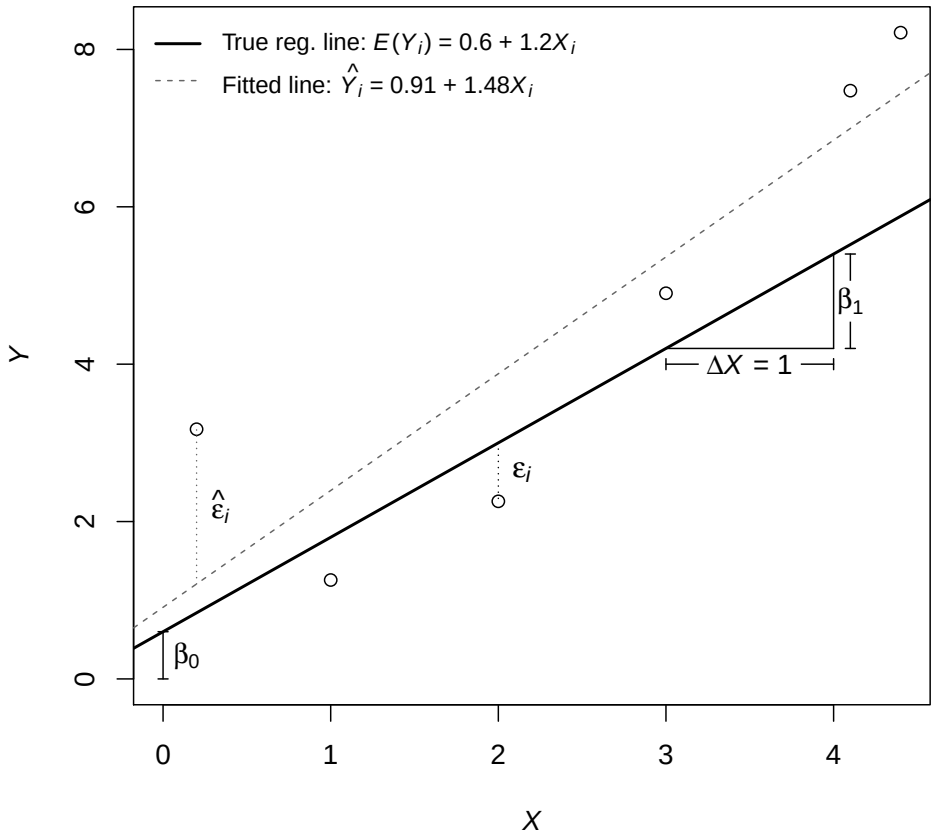


Figure 10.2. Graphical representation of parameters from the true regression model and a fitted model that estimates the true model. The fitted line is the best possible linear fit for six random points generated from $0.6 + 1.2X_i + \varepsilon_i$, where $\varepsilon_i \sim N(0, 1)$. Figure from Aho (2014).

The term ε_i in Eq. 10.1 represents a random variable describing the variability of the response given the ith value of the predictor. For inferential purposes, we assume $\varepsilon_i \sim N(0, \sigma^2)$, where σ^2 (the **error term variance**) is the true variance of the difference of observed values and fitted values (Fig 10.3). Because the errors are normally distributed with mean zero, the regression model is a mean function with fits occurring at the mean of normal distributions, whose mean, $E(Y_i)$, is $\beta_0 + \beta_1 X_i$, and whose variance is the error term variance, σ^2. That is,

$$E(Y_i) = \beta_0 + \beta_1 X_i, \tag{10.2}$$

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2). \tag{10.3}$$

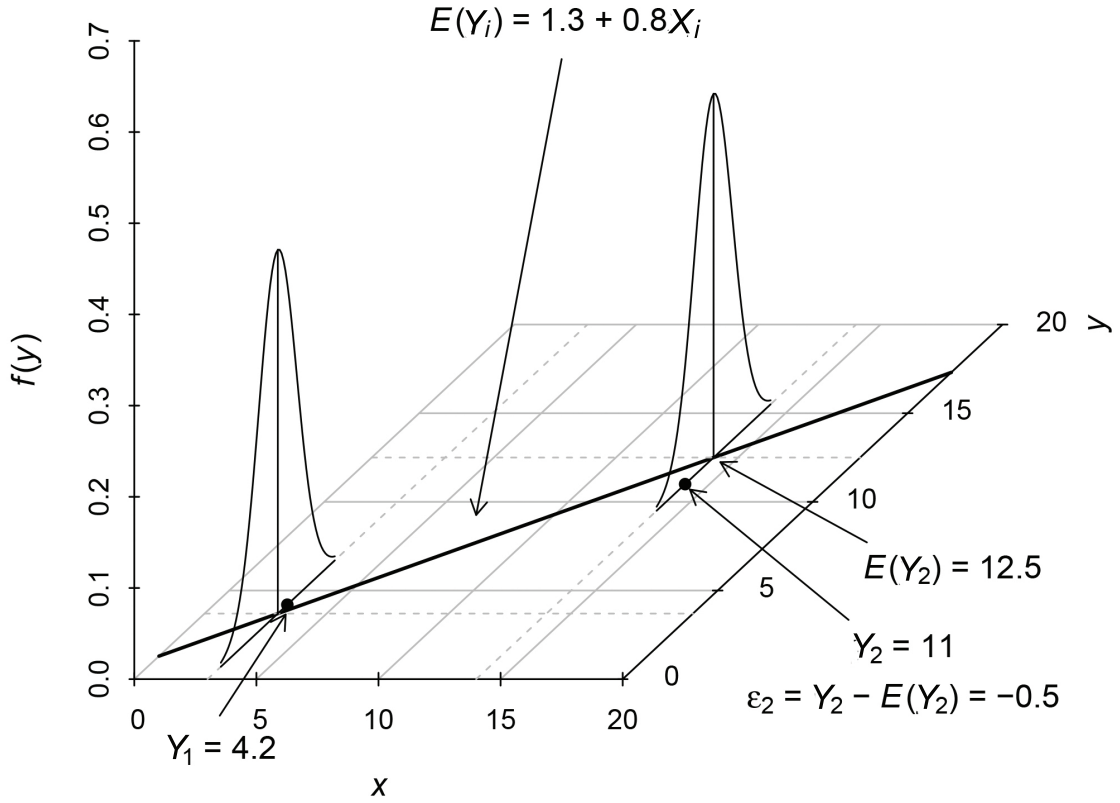


Figure 10.3. Example of a population (true) regression line for simple linear regression. The mean response is a straight-line function of the explanatory variable. The true model is $E(Y_i) = 1.3 + 0.8X_i$. We have two random observations of Y_i given X. These are $Y_1 = 4.2$ and $Y_2 = 11$. These responses correspond to the X outcomes 3 and 14. The expectation of Y given $X = 14$ is 12.5. Thus, we have the error, $\varepsilon_2 = -0.5$. Figure follows Aho (2014).

Regressions are in a class of algorithms known as **general linear models**. These models allow Y to be expressed as a linear transformation of X and assume model errors have the same distribution: $\varepsilon_i \sim N(0, \sigma^2)$. Another type of general linear model is Analysis of Variance (ANOVA). We will be introduced to ANOVAs over the last several weeks of lab.

Parameter Estimation

The terms β_0 and β_1 in Eq. 10.1 are the Y-intercept and slope of the true regression model, based on all observable X, Y pairs. It is very unlikely that we will be able to record all observable X, Y pairs. Thus, we must estimate β_0 and β_1 with $\hat{\beta}_0$ and $\hat{\beta}_1$, using sample data. Inferential interpretations of the sample Y-intercept and sample slope follow the interpretations given earlier for the parameters β_0 and β_1, respectively. Implementation of $\hat{\beta}_0$ and $\hat{\beta}_1$ results in the **fitted model:**

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{10.4}$$

where $\hat{Y}_i$ is the ith fitted value and $\hat{\beta}_0$ and $\hat{\beta}_1$ are the **sample Y-intercept** and the **sample slope**, respectively. We denote the ith model **residual** as the difference between ith fitted value and the ith observed response:

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i. \tag{10.5}$$

Residuals, $\hat{\varepsilon}_i$, serve as estimates for the true model errors, ε_i. $\hat{\beta}_0$ and $\hat{\beta}_1$ are the **ordinary least squares estimators** for β_0 and β_1. As a result, the sum of squared residuals for the model, $\sum_{i=1}^{n} \hat{\varepsilon}_i^2$, is guaranteed to be minimized, compared to a regression line based on any other possible estimates for the true slope and true Y-intercept. We calculate $\hat{\beta}_1$ using:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = r\frac{S_Y}{S_X}. \tag{10.6}$$

Utilizing the result from Eq. 10.6, the equation for $\hat{\beta}_0$ can be written as:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}. \tag{10.7}$$

r and r^2

The second version of Eq 10.6 allows computation of $\hat{\beta}_1$ by multiplying the ratio of the sample standard deviations for Y and X, $\frac{S_Y}{S_X}$, by the **Pearson correlation coeffient**, r . The correlation coefficient varies from -1 to 1, and represents the strength of the straight-line association of Y and X. Negative values of r indicate that X and Y are negatively associated (Y decreases as X increases, and *vice versa*), resulting in a negative value for the slope coefficient. Positive values indicate that Y and X are positively associated (Y and X increase together), resulting in a positive value for the slope coefficient. Values of r near 1 and -1 indicate that X and Y are *strongly* positively and negatively correlated, respectively. Values of r near zero indicate that Y and X are not linearly correlated. The Pearson correlation coefficient can be calculated as:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{S_X}\right)\left(\frac{Y_i - \bar{Y}}{S_Y}\right). \tag{10.8}$$

Squaring r gives r^2, the **coefficient of determination**. The coefficient of determination ranges from 0 to 1 and can be interpreted, in a regression context, as the proportion of variability in Y explained by the regression model (i.e., the proportion of variability in Y explained by X). Values of r^2 near 1 indicate that Y can be predicted effectively with X.

Hypothesis Testing

Our primary interest in regression analysis is whether Y changes linearly with X. To address this, the following hypotheses are generally used:

$$\mathrm{H}_0 : \beta_1 = 0$$
$$\mathrm{H}_\mathrm{A} : \beta_1 \neq 0$$

If H_0 is true, and β_1 does in fact equal 0, then clearly Y will not change linearly with X. While less typical (and not the default of most software packages), we can also test one-tailed (upper-tailed and lower-tailed) alternative hypotheses for β_1, i.e., $\mathrm{H}_\mathrm{A} : \beta_1 > 0$ and $\mathrm{H}_\mathrm{A} : \beta_1 < 0$, respectively.

The underlying sampling distribution of $\hat{\beta}_1$ will be normally distributed under model assumptions and will be asymptotically normal otherwise. The variance for sampling distribution will generally require estimation, requiring specification of a t-distribution for the null distribution. We calculate the test statistic as:

$$t^* = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}, \tag{10.9}$$

where

$$\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{MSE}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}, \tag{10.10}$$

The mean squared error, MSE, is analogous to the pooled-variance estimator in pooled variance t-tests. It is calculated as:

$$MSE = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{n-2} \tag{10.11}$$

The numerator in Eq. 10.11, i.e., the sum of squared residuals, is often called the **sum of squares error**, or SSE. We divide SSE by its degrees of freedom to obtain MSE. We use $n-2$ degrees of freedom instead of $n-1$ because, in a simple linear regression there are two parameters that need to be estimated, β_0 and β_1. As we add more explanatory variables in a format called multiple regression this will further decrease the degrees of freedom.

If H_0 is true, and assumptions for the test are met, then t^* will be a random outcome from the null distribution: $T \sim t(n-2)$. For upper-tailed tests we calculate the P-value as: $P(T \geqslant t^*)$. For lower-tailed tests we calculate the P-value as $P(T \leqslant t^*)$. For two tailed tests, the most common application, we calculate the P-value as: $2 \cdot P(T \geqslant t^*)$.

Example 10.1

We obtain height and DBH, i.e., diameters at breast height (i.e. 4' 6") data for twenty randomly chosen trees at a cherry tree farm (Table 10.1). We want to know if DBH is influencing the height of trees at the farm. While we would not ordinarily do this, the entire linear regression is calculated below by hand.

Table 10.1. DHB and height of cherry trees.

Observation	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
X = DBH (inches)	8.3	8.6	8.8	10.5	10.7	10.8	11	11	11.1	11.2	11.3	11.4	11.4	11.7	12	12.9	12.9	13.3	13.7	13.8
Y = Height (feet)	70	65	63	72	81	83	66	75	80	75	79	76	76	69	75	74	85	86	71	64

We wish to test the hypotheses:

$$\mathrm{H}_0 : \beta_1 = 0$$
$$\mathrm{H}_\mathrm{A} : \beta_1 \neq 0$$

We will use $\alpha = 0.05$. To test these hypotheses, we must first estimate the regression model parameters, β_1 and β_0, using Eqs. 10.6 and 10.7. This requires calculation of summary statistics for X and Y. We find: $\bar{x} = 11.32$, $\bar{y} = 74.25$, $s_X = 1.55279$, and $s_Y = 6.827768$. To calculate $\hat{\beta}_1$ we first calculate the correlation coefficient using Eq. 10.8. This requires centering (subtracting the mean) and scaling (dividing by the standard deviation) the data for X and Y and finding the sum of the product of these operations (Table 10.2).

Table 10.2. Centering and scaling data for X and Y, and finding the sum of the product of these operations, in order to calculate r.

Obs.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
$a = \frac{X_i - \bar{X}}{S_X}$	-1.94	-1.75	-1.62	-0.53	-0.40	-0.33	-0.21	-0.21	-0.14	-0.08	-0.01	0.05	0.05	0.24	0.44	1.02	1.02	1.28	1.53	1.60	
$b = \frac{Y_i - Y}{S_Y}$	-0.62	-1.35	-1.65	-0.33	0.99	1.28	-1.21	0.11	0.84	0.11	0.70	0.26	0.26	-0.77	0.11	-0.04	1.57	1.72	-0.48	-1.50	
$a \cdot b$	1.21	2.37	2.67	0.17	-0.39	-0.43	0.25	-0.02	-0.12	-0.01	-0.01	0.01	0.01	-0.19	0.05	-0.04	1.60	2.19	-0.73	-2.40	$\sum = 6.22$

We find:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{S_X} \right) \left(\frac{Y_i - \bar{Y}}{S_Y} \right) = \frac{1}{20-1} \cdot 6.215756 = \frac{6.215756}{19} = 0.3271.$$

Thus, $r^2 = 0.3271^2 = 0.107$. This means that approximately 11% of variation in height data can be explained by DBH. Calculating $\hat{\beta}_1$ we have:

$$\hat{\beta}_1 = r\frac{s_Y}{s_X} = 0.3271 \cdot \frac{6.83}{1.55} = 1.438488.$$

Thus, for every inch increase in DBH we would expect an increase of 1.438 feet in tree height. Calculating $\hat{\beta}_0$ using the result for $\hat{\beta}_1$ we have:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} = 74.25 - 1.438(11.32) = 57.96632.$$

Thus, our fitted model is: $\hat{Y}_i = 57.966 + 1.438X_i$. The fitted values and residuals for this model are shown in Table 10.3.

Table 10.3. Fitted values and residuals for the cherry tree model.

Obs.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
$\hat{Y}_i = 57.97 + 1.44X_i$	69.9	70.3	70.6	73.1	73.4	73.5	73.8	73.8	73.9	74.1	74.2	74.4	74.4	74.8	75.2	76.5	76.5	77.1	77.7	77.8
$\hat{\varepsilon}_i = \hat{Y}_i - Y_i$	0.1	-5.3	-7.6	-1.1	7.6	9.5	-7.8	1.2	6.1	0.9	4.8	1.6	1.6	-5.8	-0.2	-2.5	8.5	8.9	-6.7	-13.8

To calculate the test statistic, t^*, we must first calculate the standard error for $\hat{\beta}_1$, $\hat{\sigma}_{\hat{\beta}_1}$. We find:

$$\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{MSE}{\sum_{i=1}^n (X_i - \bar{X})^2}} = \frac{\sqrt{\frac{1}{n-2}\sum_{i=1}^n \hat{\varepsilon}_i^2}}{\sqrt{45.812}} = \frac{6.6289}{\sqrt{45.812}} = 0.9793764$$

Thus, the test statistic is:

$$t^* = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{1.438}{0.979} = 1.468779$$

To calculate the P-value, we find $2 \cdot P(T \geqslant |t^*|)$ where $T \sim t(n-2)$. We have:

```
2 * pt( 1.468779, 18, lower.tail = F)

[1] 0.1591527
```

We fail to reject H_0 at $\alpha = 0.05$. Surprisingly, perhaps, tree height does not appear to be a linear function of DBH.

■

Below we calculate results for Example 1 "by hand" using **R** to help. Here we calculate summary statistics:

```
height <- read.csv("height.csv")

x <- height[,2]; y <- height[,3]
x.bar <- mean(x); y.bar <- mean(y)
s.x <- sd(x); s.y <- sd(y)
n <- 20
```

Here we calculate r, $\hat{\beta}_1$ and $\hat{\beta}_0$.

```
r <- 1/(n - 1) * sum((x - x.bar)/s.x * (y - y.bar)/s.y)
r

[1] 0.3271449

beta.hat1 <- r * s.y/s.x
beta.hat1

[1] 1.438488

beta.hat0 <- y.bar - beta.hat1 * x.bar
beta.hat0

[1] 57.96632
```

Below we calculate the test statistic and P-value.

```
y.hat <- beta.hat0 + beta.hat1 * x
residuals <- y.hat - y

MSE <- sum(residuals^2)/(n - 2)
se.beta.hat1 <- sqrt(MSE/sum((x - x.bar)^2))
se.beta.hat1

[1] 0.9793764

t.star <- beta.hat1/se.beta.hat1
t.star

[1] 1.468779

2 * pt(t.star, n - 2, lower.tail = F)

[1] 0.1591526
```

We can easily create a plot of the regression using **R** (Fig 10.4)

```
plot(x, y, xlab = "DBH (inches)", ylab = "Height (ft)")
abline(beta.hat0, beta.hat1)
```

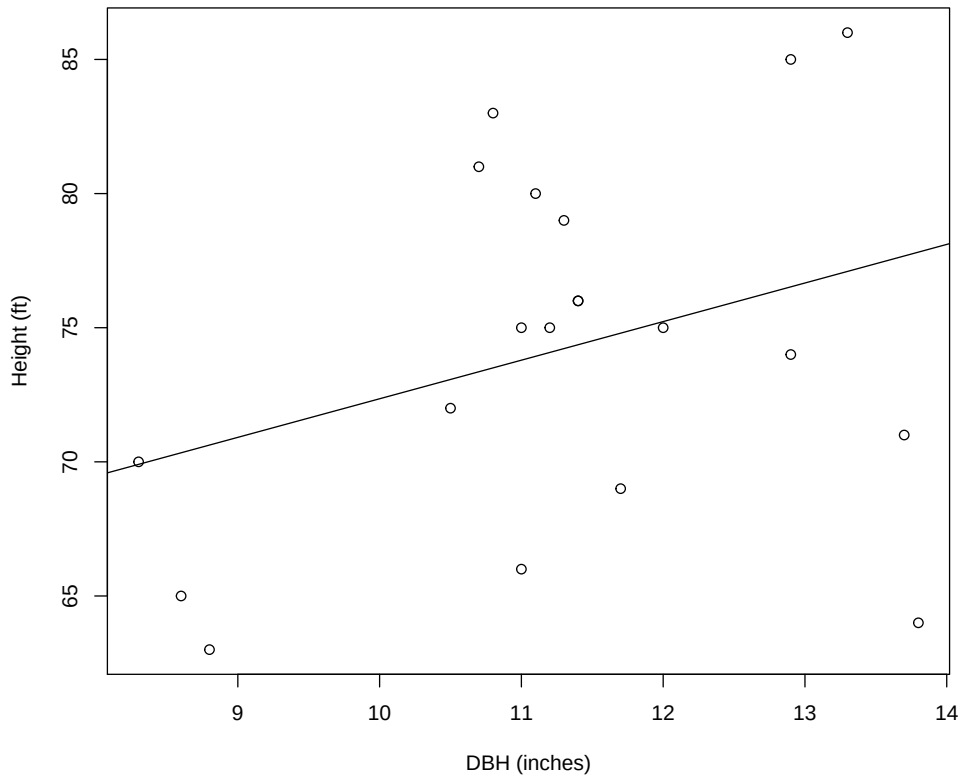


Figure 10.4. Plot of regression model from Example 1.

We could have created the regression model and run the hypothesis test in **R** using very little code.

```
model <- lm(y ~ x)
summary(model)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-13.8174  -5.4521   0.5084   5.1007   9.4980

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  57.9663    11.1852   5.182 6.27e-05 ***
x             1.4385     0.9794   1.469    0.159
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.629 on 18 degrees of freedom
Multiple R-squared:  0.107,Adjusted R-squared:  0.05741
F-statistic: 2.157 on 1 and 18 DF,  p-value: 0.1592
```

Assignment 10

Answer all questions in one MS Word document and submit to Moodle. At the beginning of the document include the assignment number, the date, your name and section number.

Use complete sentences when appropriate, and make sure any tables, figures and computer output you include adhere to class standards (see Syllabus).

The regression model

- Open **R**
- Load the *asbio* package by typing `library(asbio)` or by going to **Packages > Load packages > *asbio***.
- Type `book.menu()` in the **R** console.

1. (9 pts) Go to the Ch. 9 pulldown menu. Click on **Regression(Add/delete points)**. Mac-users and others who wish to obtain the GUI directly can type: `see.adddel()`.

a) By adding points create a poor explanatory model (one with a small slope, but with a non-zero slope). Make sure the slope is poor enough so that the model will be non-significant (P-value > 0.05). Insert an the figure into your document. This will require a screenshot or a snip tool.

b) Add a whole bunch of points. Try not to change the slope from (a) too much. Is it possible to get a significant P-value by simply adding more points?

c) What does this indicate with respect to the effect of sample size on P-values?

d) Does this suggest that investigators should potentially modify the significance level or sample size based on the potential effect size (slope)? Why?

2. (4 pts) Click on **Regression(Move points)**. Mac-users and others who wish to obtain the GUI directly can type: `see.move()`.

a) Which points have the greatest effect on the regression slope, those near the center with respect to the ordinate (X-axis), or those far from the center?

b) Insert illustrative example(s) into your document. This will require a screenshot or snip tool.

Parameter estimation and hypothesis testing

3. Table 10.4 shows the public domain cherry tree data for DBH and volume in cubic feet (the data are also available on Moodle as `volume.csv`). Test if the diameter of the trees determines their volume. Use snapshots when necessary to show work.

Table 10.4. Cherry tree volume data.

Observation	DBH (inches)	Volume (ft^3)
1	8.1	10.3
2	8.6	10.3
3	8.8	10.2
4	10.5	16.4
5	10.7	18.8
6	10.9	19.7
7	10.2	15.6
8	10.7	18.2
9	11.3	22.6
10	11.4	19.9

a) (2 pts) Correctly state your hypotheses.

b) (4 pts) Calculate r, r^2, $\hat{\beta}_1$, and $\hat{\beta}_0$, in that order, by hand, using **R** to help, see Eqs. 10.8, 10.6, and 10.7. Do not use `lm`.

c) (4 pts) Correctly interpret the meaning of your results for r^2 and $\hat{\beta}_1$.

d) (3 pts) Calculate $\hat{\sigma}_{\hat{\beta}_1}$ "by hand" using **R** to help, see Eq. 10.10. Do not use `lm`

e) (2 pts) Calculate t^* "by hand" using **R** to help, see Eq 10.9.

f) (2 pts) Calculate the P-value associated "by hand" using **R** to help.

g) (2 pts) What are your conclusions for the null hypothesis test concerning β_1?

h) (2 pts) Create a plot of the regression model using `plot`.

i) (2 pts) Verify your results by running the regression using `lm`.

j) (2 pts) Calculate $\hat{Y}_9$, i.e., the fitted value corresponding to the 9th observed X outcome, i.e., X_9, "by hand", using **R** to help, see Eq. 10.4.

k) (2 pts) Calculate $\hat{\varepsilon}_9$, i.e., the residual corresponding to the 9th observed X outcome, i.e., X_9, "by hand", using **R** to help, see Eq 10.5.

Q1 9pts, Q2 4pts, Q3 27 pts. **Total pts: 40.**

Appendix: R-code used in this lab

This lab focused on regression models which can be created using the function `lm`. The acronym `lm` stands for general linear model. General linear models, including regression, can be run using the framework:

```
lm(Y ~ X)
```

where `Y` and `X` represent response and explanatory variables, respectively. Let `weight` and `calories` be columns in a dataframe called `diet`. These variables can be called by name directly if the `data` argument in `lm` is used. For instance,

```
lm(weight ~ calories, data = diet)
```

Linear models generated by `lm` can be summarized using `summary.lm`:

```
model <- lm(weight ~ calories, data = diet)
summary(model)
```

11

Regression II

Lab 11 Topics

1. Simple linear regression assumptions and diagnostics
2. Regression confidence intervals and prediction intervals

Regression Assumptions

There are four assumptions for simple linear regression:

1. Observations are independent.
2. The error terms distributions are normally distributed.
3. The error terms distributions are homoscedastic.
4. The true relationship of Y and X is actually linear.

Most of these assumptions concern the form of the error term distribution. Recall that the simple linear regression model is:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{11.1}$$

where $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, and *iid* means independent and identically distributed.

To summarize: the errors (noise distributions around the regression line) for every value of Y given X are assumed to be independent and identically normally distributed, with mean 0, and a constant variance, σ^2. Because the errors are normally distributed with mean zero, the regression model is a mean function with fits occurring at the mean of normal distributions, whose mean, $E(Y_i)$, is $\beta_0 + \beta_1 X_i$, and whose variance is the error term variance, σ^2. That is,

$$E(Y_i) = \beta_0 + \beta_1 X_i, \tag{11.2}$$

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2). \tag{11.3}$$

These ideas are summarized in Fig 11.1.

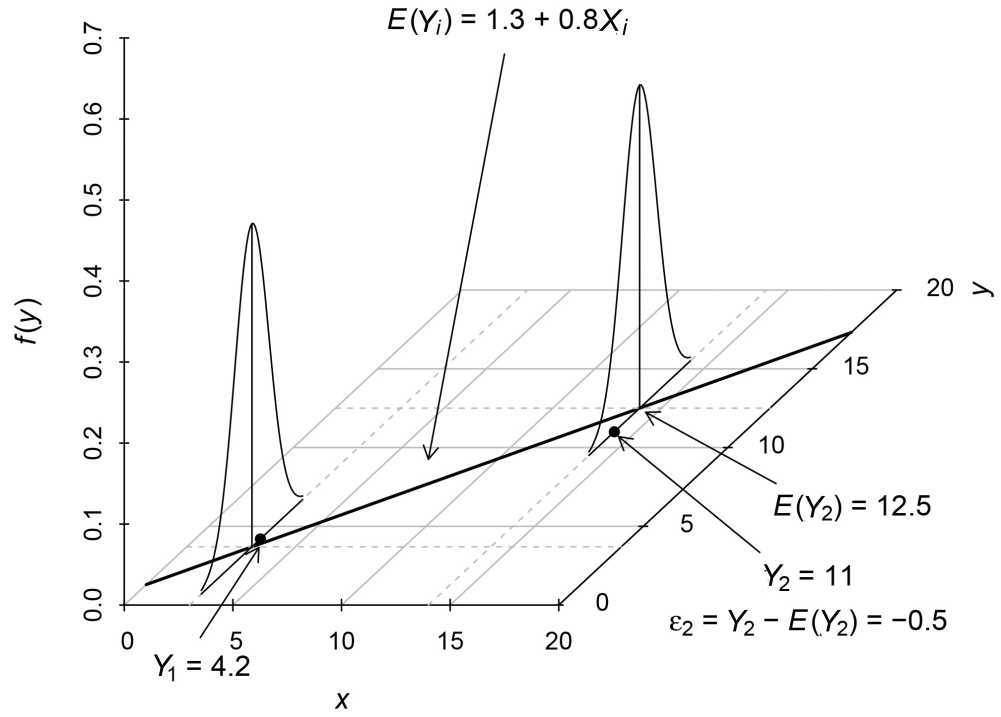


Figure 11.1. Example of assumptions for the population regression line for simple linear regression. The mean response is a straight-line function of the explanatory variable. The true model is $E(Y_i) = 1.3 + 0.8X_i$. We have two random observations of Y_i given X. These are $Y_1 = 4.2$ and $Y_2 = 11$. These responses correspond to the X outcomes 3 and 14. The expectation of Y given $X = 14$ is 12.5. Thus, we have the error, $\varepsilon_2 = -0.5$. Figure follows Aho (2014).

When model-checking, we remain attentive to the presence of outliers. This is because estimators in general linear models (including $\hat{\beta}_0$ and $\hat{\beta}_1$ in simple linear regression) are not resistant to outliers. Because our assumptions generally concern the characteristics of the true errors, we generally check model assumptions by by examining the model residuals. This because the residuals estimate the ε_is.

Example 11.1

Table 11.1 contains the cherry tree data that you used for Assignment 10. Recall that we fit the volume of trees as a function of their diameter (DBH).

Table 11.1. Cherry tree volume data.

Observation	DBH (inches)	Volume (ft^3)
1	8.1	10.3
2	8.6	10.3
3	8.8	10.2
4	10.5	16.4
5	10.7	18.8
6	10.9	19.7
7	10.2	15.6
8	10.7	18.2
9	11.3	22.6
10	11.4	19.9

We have the following hypotheses:

$$\begin{aligned} \mathrm{H}_0 &: \beta_1 = 0 \\ \mathrm{H}_\mathrm{A} &: \beta_1 \neq 0 \end{aligned}$$

We run the regression in **R** with the following result:

```
volume <- read.csv("volume.csv")
model <- lm(volume ~ DBH, data = volume)
summary(model)

Call:
lm(formula = volume ~ DBH, data = volume)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2034 -0.9944 -0.2142  0.5789  2.0420

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -21.1755     3.4947  -6.059 0.000303 ***
DBH           3.6932     0.3432  10.760  4.9e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.218 on 8 degrees of freedom
Multiple R-squared:  0.9354,Adjusted R-squared:  0.9273
F-statistic: 115.8 on 1 and 8 DF,  p-value: 4.898e-06
```

Based on the P-value for the slope hypotheses, $P = 1.29 \times 10^{-6}$, we can reject H_0 at $\alpha = 0.05$. This conclusion, however, is only valid if our assumptions are valid.

Independence

We first check the assumption of independence. There are two main ways that independence can be violated: **temopral dependence** and **spatial dependence**. Temporal dependence occurs when a serial process causes outcomes to be predictable based on how close together they are in time. Similarly, spatial dependence occurs when outcomes are predictable based on how close together they are in space. To check for temporal independence we can plot model residuals against the order that observations were taken. If independence is satisfied there should be no patterns to points in this plot. To get the model residuals for the current example, I could type:

```
e.hat <- resid(model)
```

The observation order of the `volume` data set is given in the 1st column. A plot of residuals as a function of the order of the observations is shown in Fig 11.2.

```
plot(volume[,1], e.hat, xlab = "Order of observations",
     ylab = "Residuals", type = "o")
```

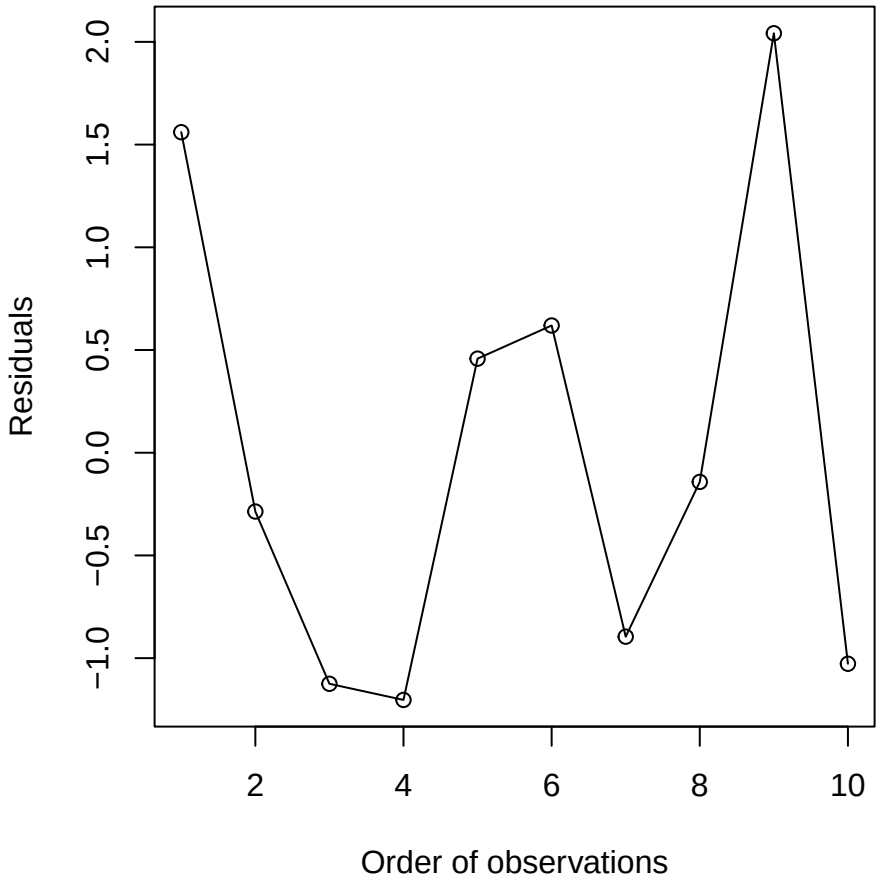


Figure 11.2. Residuals as a function of order in Ex. 1.

There seems to be little pattern to points in the plot, supporting the assumption of independence.

Normality

We expect noise in the regression model have the distribution $N(0, \sigma^2)$ for each Y given X (Fig. 11.1). We can check this assumption with a of residuals (Fig. 11.3), and the Shapiro-Wilk test for normality.

```
qqnorm(e.hat, main = "")
qqline(e.hat)
```

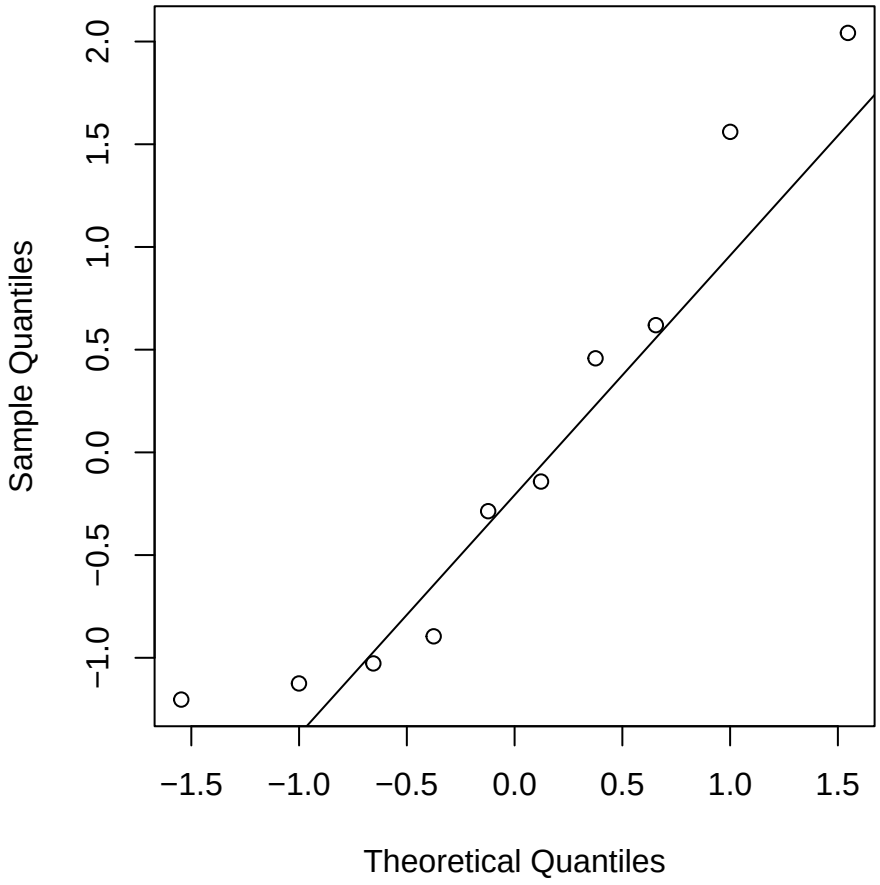


Figure 11.3. Normal quantile plot of residuals in Ex. 1.

In a Shapiro-Wilk test of the residuals the hypotheses are:

H_0 : The underlying distribution of errors is normal

H_A : The underlying distribution of errors is not normal

```
shapiro.test(e.hat)

Shapiro-Wilk normality test

data:  e.hat
W = 0.90091, p-value = 0.2242
```

Although the normal quantile plot raises some concerns, the Shapiro-Wilk test supports the assumption of normality.

Homoscedasticity

Because we assume errors have same the distribution, $N(0, \sigma^2)$, for each Y given X, we assume that the variance, σ^2 is constant along the regression line (Fig. 11.1). To check for equal error variances we can plot residuals as a function of fitted values (Fig 11.4). If the assumption of homoscedasticity is valid there should be no pattern to the points in the plot. A common heteroscedastic pattern is a shotgun scatter of points that become more dispersed as fitted values increase.

```r
fits <- fitted(model)
plot(fits, e.hat, xlab = "Fitted values", ylab = "Residuals")
abline(h = 0, lty = 2)
```

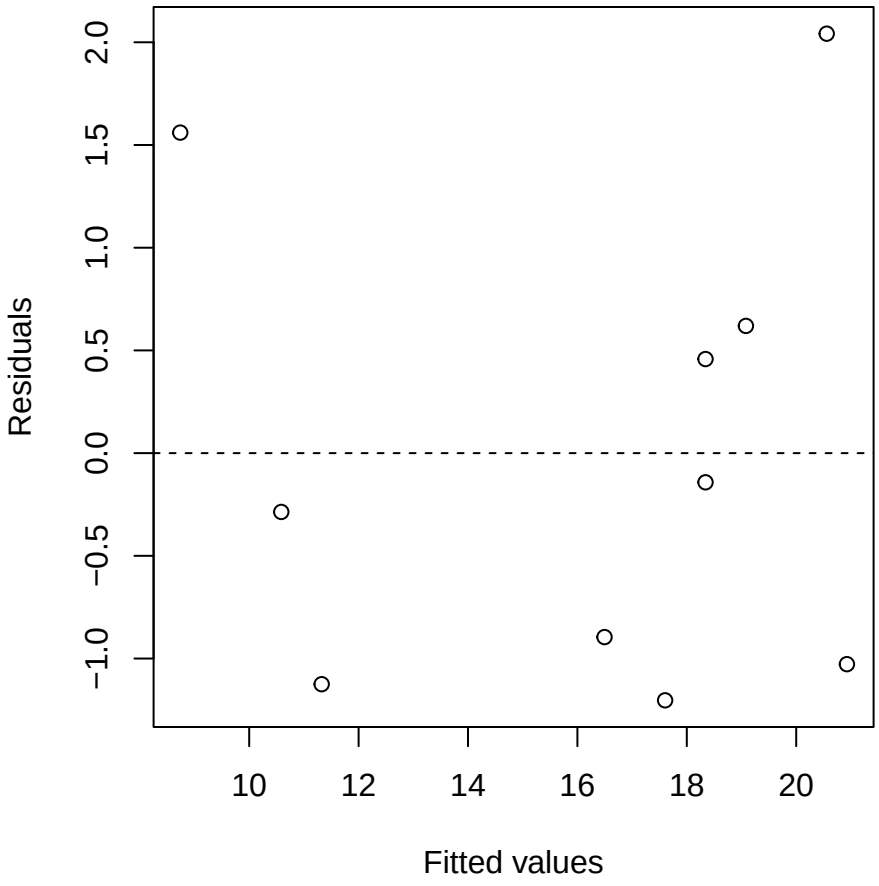


Figure 11.4. Plot of residuals as a function fits in Ex. 1.

There appears to be little evidence of heteroscedasticity in Fig 11.4.

Linearity

A fundamental assumption of general linear regression models is that the relationship between Y and X is linear. Non-linear relationships can be modeled in a number of ways. Given concave or convex associations, the simplest approach is to use a linear model of the polynomial relationship: $E(Y_i) = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$. We will not address **polynomial regression** in this lab.

We can check the appropriateness of the simple linear regression model by examining a scatterplot of the X, Y data with the regression line overlaid (Fig 11.5).

```
plot(volume[,2], volume[,3], xlab = "DBH (inches)", ylab =
       expression(paste("Volume (", feet^3,")")))
abline(coef(model))
```

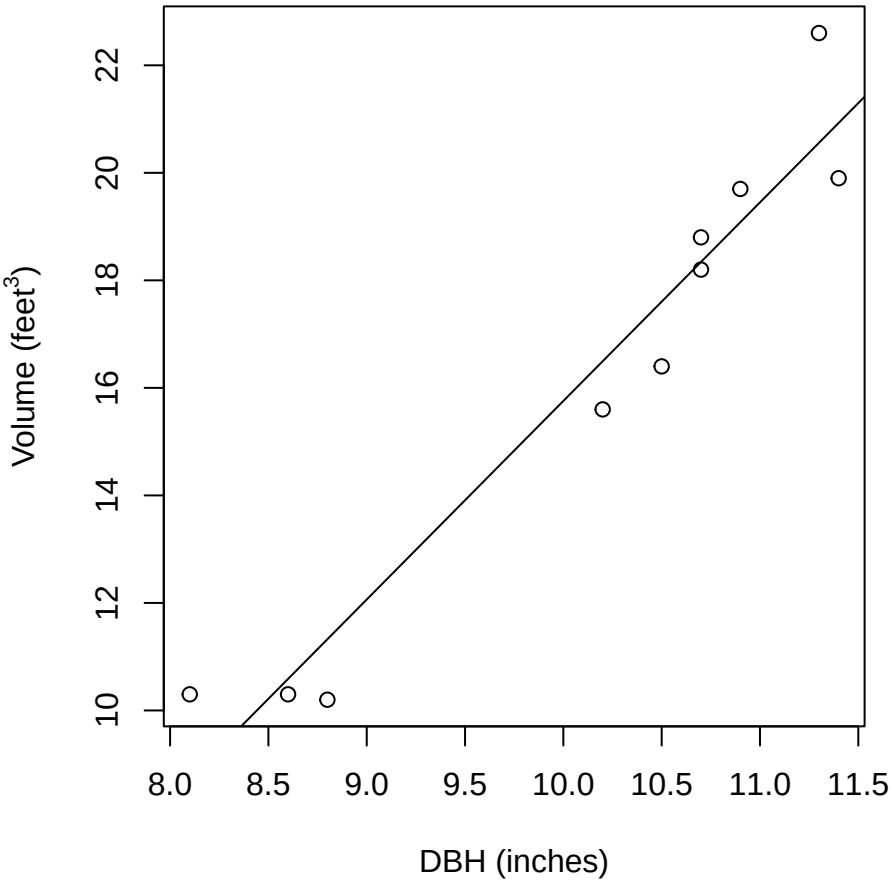


Figure 11.5. Plot of regression model from Ex. 1.

A simple linear regression model seem reasonable here.

We can easily obtain versions of Fig 11.3 and Fig 11.4 simultaneously, along with a plot that helps to identify outliers (Fig 11.6).

```
par(mfrow = c(2,2), mar=c(5,4,1.5,1.5))
plot(model)
```

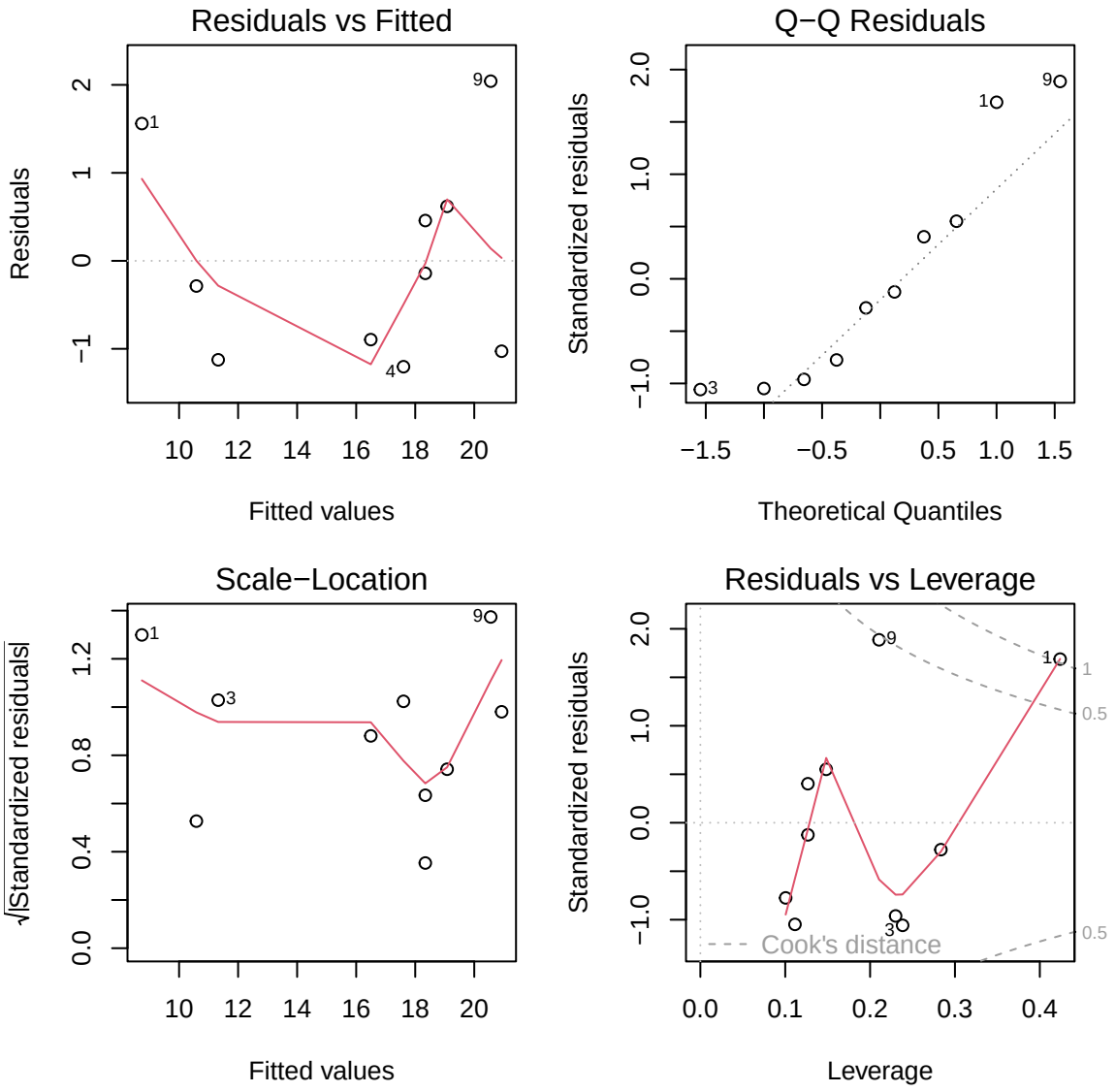


Figure 11.6. Default plots from `plot.lm` Solid red lines are smoother fits. The left-hand figures allow consideration of homoscedasticity; in fact, the top-left plot is equivalent to Fig 11.4. The top-right plot is identical to Fig 11.3. The bottom-right hand plot addresses outliers. The **leverage** of a particular point quantifies how unusual it is in explanatory variable space. For simple linear regression, oservations greater than $4/n$ can generally be considered outliers. Our cutoff for outliers is then $4/10 = 0.4$. One point (obs. 1) exceeds this high leverage cutoff. **Cook's distance** quantifies how much a model is affected by removing a point. If a point substantially affects a model it is said to be influential. The Cook's distances here have been converted into probabilities from an F-distribution. If a point has a probability > 0.5 it can be considered an outlier . Based on this criterion, observation one is an influential point.

■

Intervallic Estimators for Regression

Confidence intervals are calculable for all linear regression parameters, and for fitted and predicted values.

Confidence interval for β_1

We can calculate a $(1-\alpha)100\%$ confidence interval for β_1 using Eq. 11.4.

$$\hat{\beta}_1 \pm t_{(1-\frac{\alpha}{2},\text{df}=n-2)} \cdot \hat{\sigma}_{\hat{\beta}_1} \tag{11.4}$$

where

$$\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{MSE}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}, \tag{11.5}$$

and

$$MSE = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{n-2}. \tag{11.6}$$

Example 11.2

Revisiting the cherry tree volume example (Table 11.1) we have the following 95% confidence interval for β_1:

$$\hat{\beta}_1 = r\frac{S_Y}{S_X} = 0.9671469 \cdot \frac{4.51762}{1.183028} = 3.693236.$$

$$\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{MSE}{\sum_{i=1}^{n}(X_i - \bar{X})^2}} = \sqrt{\frac{1.483833}{12.596}} = 0.3432228.$$

Applying Eq. 11.4 we have:

$$\begin{aligned}
\hat{\beta}_1 &\pm t_{(1-\frac{\alpha}{2},\ \text{df}=n-2)} \cdot \hat{\sigma}_{\hat{\beta}_1} \\
3.693236 &\pm t_{(1-\frac{0.05}{2},\ \text{df}=8)} \cdot 0.3432228 \\
3.693236 &\pm 2.306004 \cdot 0.3432228 \\
3.693236 &\pm 0.7914733 \\
(3.693236 - 0.7914733,&\ 3.693236 + 0.7914733) \\
(2.901763,&\ 4.484709).
\end{aligned}$$

Thus, we are 95% confident that the true model slope lies in the interval (2.901763, 4.484709). Performing these operations "by hand" in **R** we have:

```
x <- volume[,2]; y <- volume[,3]

beta.hat1 <- coef(model)[2]
e.hat <- resid(model)
n <- length(e.hat)

MSE <- sum(e.hat^2)/(n-2)
sigma.hat.beta.hat <- sqrt(MSE/sum((x - mean(x))^2))

alpha <- 0.05 # for 95% CI
t.crit <- qt(1 - alpha/2, n- 2)
margin <- sigma.hat.beta.hat * t.crit
CI <- c(beta.hat1 - margin, beta.hat1 + margin)
CI

     DBH      DBH
2.901763 4.484709
```

The function `confint` will provide confidence intervals for both regression parameters, β_0 and β_1. Thus, we can quickly obtain the result above, using:

```
confint(model, level = 0.95)

                2.5 %     97.5 %
(Intercept) -29.234363 -13.116733
DBH           2.901763   4.484709
```

■

Confidence interval for $E(Y_h)$

A $(1-\alpha)100\%$ confidence interval for the hth true fitted value, $E(Y_h)$, is constructed using:

$$\hat{Y}_h \pm t_{(1-\frac{\alpha}{2},\ \mathrm{df}=n-2)} \cdot \hat{\sigma}_{\hat{Y}_h} \tag{11.7}$$

where

$$\hat{\sigma}_{\hat{Y}_h} = \sqrt{MSE\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]}, \tag{11.8}$$

Example 11.3

To calculate a confidence interval for a true fitted value, we have to first define X_h, the explanatory outcome corresponding to the fitted value of interest, Y_h. For the cherry tree

volume example, we will consider the explanatory outcome DBH = 8.1. This was the first observed X outcome in the `volume` dataset.

Applying Eq 11.8 we have:

$$\begin{aligned}
\hat{\sigma}_{\hat{Y}_h} &= \sqrt{MSE\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]} \\
&= \sqrt{1.483833\left[\frac{1}{10} + \frac{(8.1 - 10.12)^2}{12.596}\right]} \\
&= \sqrt{1.483833\left[\frac{1}{10} + 0.3239441\right]} \\
&= 0.7931344.
\end{aligned}$$

Applying Eq 11.7 we have:

$$\begin{aligned}
\hat{Y}_h &\pm t_{(1-\frac{\alpha}{2},\ \text{df}=n-2)} \cdot \hat{\sigma}_{\hat{Y}_h} \\
8.739663 &\pm t_{(1-\frac{0.05}{2},\ \text{df}=10-2)} \cdot 0.7931344 \\
8.739663 &\pm 2.306004 \cdot 0.7931344 \\
8.739663 &\pm 1.828971 \\
(8.739663 - 1.828971,\ &8.739663 + 1.828971) \\
(6.910692,\ &10.56863)
\end{aligned}$$

Thus, we are 95% confident that the true fitted value for the explanatory value $X_h = 8.1$ will be in the interval (6.910692, 10.56863).

Performing these operations "by hand" in **R** we have:

```
x.h <- x[1]
y.hat.h <- fitted(model)[1]# or: predict(model, newdata = data.frame(DBH = 8.3))

sigma.hat.Y.hat.h <- sqrt(MSE * (1/n + (x.h - mean(x))^2/sum((x - mean(x))^2)))

alpha <- 0.05 # for 95% CI
t.crit <- qt(1 - alpha/2, n- 2)
margin <- sigma.hat.Y.hat.h * t.crit
CI <- c(y.hat.h - margin, y.hat.h + margin)
CI

        1         1
 6.910692 10.568635
```

We can easily obtain confidence intervals for $E(Y_h)$ using the function `predict`. Here is the 95% confidence interval for $E(Y_h)$ for $X_h = 8.1$.

```
predict(model, newdata = data.frame(DBH = 8.1), interval = "confidence", level = 0.95)
```

```
       fit      lwr      upr
1 8.739663 6.910692 10.56863
```

Here are 95% confidence intervals for $E(Y_h)$, for all observed X outcomes.

```
predict(model, interval = "confidence", level = 0.95)
```

```
         fit       lwr      upr
1   8.739663  6.910692 10.56863
2  10.586281  9.090837 12.08173
3  11.324929  9.953600 12.69626
4  17.603430 16.665609 18.54125
5  18.342077 17.342186 19.34197
6  19.080724 17.998980 20.16247
7  16.495459 15.604920 17.38600
8  18.342077 17.342186 19.34197
9  20.558018 19.269107 21.84693
10 20.927342 19.579976 22.27471
```

■

Prediction interval for $Y_{h(new)}$

A **prediction interval** for $Y_{h(new)}$ (the hth true predicted value) represents not a range that the true fitted value will lie in at a certain level of confidence (this is the confidence interval for $E(Y_h)$), but the range that a new response, given X_h will fall into at a particular level of confidence. A prediction interval for $Y_{h(new)}$ is constructed using:

$$\hat{Y}_h \pm t_{(1-\frac{\alpha}{2},\ \text{df}=n-2)} \cdot \hat{\sigma}_{\hat{Y}_{h(new)}} \tag{11.9}$$

where

$$\hat{\sigma}_{\hat{Y}_{h(new)}} = \sqrt{MSE\left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]}. \tag{11.10}$$

Example 11.4

As with a confidence interval for $E(Y_h)$, to calculate a prediction interval for $Y_{h(new)}$, we must define X_h, the explanatory outcome corresponding to the fitted value of interest, Y_h. Using the cherry tree volume model, we will again consider the explanatory outcome DBH = 8.1 (the first observed X outcome in the `volume` dataset).

Applying Eq 11.10 we have:

$$\begin{aligned}
\hat{\sigma}_{\hat{Y}_{h(new)}} &= \sqrt{MSE\left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]} \\
&= \sqrt{1.483833\left[1 + \frac{1}{10} + \frac{(8.1 - 10.12)^2}{12.596}\right]} \\
&= \sqrt{1.483833\left[1 + \frac{1}{10} + 0.3239441\right]} \\
&= 1.45358
\end{aligned}$$

Applying Eq 11.9, we have:

$$\begin{aligned}
\hat{Y}_h &\pm t_{(1-\frac{\alpha}{2},\ \text{df}=n-2)} \cdot \hat{\sigma}_{\hat{Y}_{h(new)}} \\
8.739663 &\pm t_{(1-\frac{0.05}{2},\ \text{df}=10-2)} \cdot 1.45358 \\
8.739663 &\pm 2.306004 \cdot 1.45358 \\
8.739663 &\pm 3.351961 \\
(8.739663 - 3.351961,\ &8.739663 + 3.351961) \\
(5.387702,\ &12.09162)
\end{aligned}$$

Thus, we are 95% confident that a future observed response, $Y_{h(new)}$, given the explanatory value $X_h = 8.3$, will lie in the interval (5.387702, 12.09162).

Performing these operations "by hand" in **R** we have:

```
sigma.hat.Y.hat.h.new <-
  sqrt(MSE * (1 + 1/n + (x.h - mean(x))^2/sum((x - mean(x))^2)))

alpha <- 0.05 # for 95% CI
t.crit <- qt(1 - alpha/2, n- 2)
margin <- sigma.hat.Y.hat.h.new * t.crit
CI <- c(y.hat.h - margin, y.hat.h + margin)
CI
```

```
        1         1
 5.387702 12.091625
```

It is also straightforward to obtain prediction intervals for $Y_{h(new)}$ using the function `predict`. Here is the 95% prediction interval for $Y_{h(new)}$, for $X_h = 8.3$.

```
predict(model, newdata = data.frame(DBH = 8.3), interval = "prediction", level = 0.95)
```

```
       fit      lwr      upr
1 9.478311 6.198898 12.75772
```

Here are 95% prediction interval for $E(Y_{h(new)})$, for all observed X outcomes.

```
predict(model, interval = "prediction", level = 0.95)
```

```
Warning in predict.lm(model, interval = "prediction", level = 0.95): predictions
on current data refer to _future_ responses
```

```
         fit       lwr      upr
1   8.739663  5.387702 12.09163
2  10.586281  7.404008 13.76856
3  11.324929  8.199060 14.45080
4  17.603430 14.642008 20.56485
5  18.342077 15.360417 21.32374
6  19.080724 16.070627 22.09082
7  16.495459 13.548669 19.44225
8  18.342077 15.360417 21.32374
9  20.558018 17.467419 23.64862
10 20.927342 17.811912 24.04277
```

We can also use **R** to make a plot of the regression line, with confidence intervals for $E(Y_h)$ and the prediction intervals for $Y_{h(new)}$ overlaid (Fig. 11.7).

```
with(volume, plotCI.reg(DBH, volume, xlab = "DBH (inches)", CI.col = 1, PI.col =
1, ylab = expression(paste("Volume (", feet^3,")"))))

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
    -21.176        3.693
```

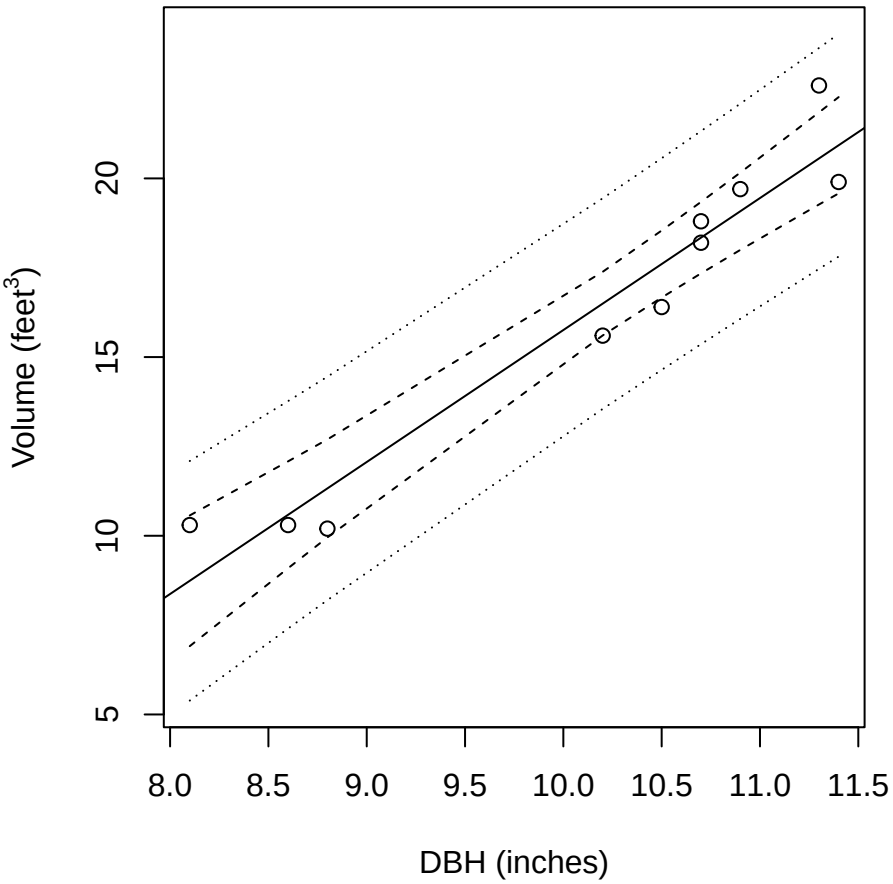


Figure 11.7. Regression line (solid line), 95% confidence intervals for $E(Y_h)$ (dashed lines) and 95% prediction intervals for $Y_{h(new)}$ (dotted lines) for the cherry tree example.

■

Assignment 11

Answer all questions in one MS Word document and submit to Moodle. At the beginning of the document include the assignment number, the date, your

name and section number.

Use complete sentences when appropriate, and make sure any tables, figures and computer output you include adhere to class standards (see Syllabus).

Model assumptions

- Open **R**
- Load the *asbio* package by typing `library(asbio)` or by going to **Packages > Load packages > *asbio***.
- Type `book.menu()` in the **R** console.

1. Go to the Ch. 9 pulldown menu. Click on **Regression mechanics**. Mac-users and others who wish to obtain the GUI directly can type: `see.regression.tck()`.

a) (2 pts) What do the Gaussian curves represent?

b) (1 pt) Do the the variances appear equal for the population curves that are shown?

c) (2 pts) What do the dots represent?

d) (2 pts) Why is the sample regression line (to the right) different from the population regression line (to the left)?

e) (4 pts) Make the true slope equal zero. Now the null hypothesis, $H_0 : \beta_1 = 0$, is true. Is it still possible to reject H_0 at $\alpha = 0.05$ using random sampling? Include an example figure in your homework. What is this outcome called?

2. Table 11.2 contains data comparing chicken weight gain (in grams) and a lysine diet additive (in grams). The data are also on Moodle.

Table 11.2. Chicken lysine diet additive and weight gain data.

Chicken	Weight gain (g)	Lysine (g)
1	14.7	0.09
2	17.8	0.14
3	19.6	0.18
4	18.4	0.15
5	20.5	0.16
6	21.1	0.23
7	17.2	0.11
8	18.7	0.19
9	20.2	0.23
10	16	0.13
11	17.8	0.17
12	19.4	0.21

a) (6 pts) Create a regression model of weight gain as a function of lysine.

- State the null and alternative hypotheses for the regression.
- Run the regression in **R** using `lm`. Use snapshots to show work.
- State your conclusions. Can we reject H_0 at $\alpha = 0.05$?

b) (4 pts) Check for independence of outcomes in the regression model. Comment with respect to the diagnostic results and attach graphs and computer output as necessary.

c) (6 pts) Check for normality of errors in the regression model using a normal quantile plot and the Shapiro Wilk test. Comment with respect to the diagnostic results. Attach graphs and computer output as necessary.

d) (4 pts) Check for homoscedasticity in the regression model using graph of residuals versus fits. Attach the graph and comment with respect to the diagnostic results.

e) (4 pts) Check for general linearity of the relationship between weight gain and lysine eaten by plotting weight gain as a function of lysine. Attach the graph and comment with respect to the diagnostic results.

3. (2 pts) Based on the results in **Q 2b-d**, are the results in **Q 2a** trustworthy? Why or why not?

Intervallic estimates

4. (6 pts) Calculate a 90% confidence interval for β_1

a) Calculate the confidence interval "by hand", using **R** to help. Include snapshots to show work.

b) Correctly summarize your results.

c) Check you results from **Q 4a** using `confint`. Include snapshots to show work.

5. (6 pts) Calculate a 95% confidence interval for $E(Y_h)$, when $X_h = 0.18$.

a) Calculate the confidence interval "by hand", using **R** to help. Include snapshots to show work.

b) Correctly summarize your results.

c) Check you results from **Q 5a** using `predict`. Include snapshots to show work.

6. (6 pts) Calculate a 95% prediction interval for $Y_{h(new)}$, when $X_h = 0.18$.

a) Calculate the confidence interval "by hand", using **R** to help. Include snapshots to show work.

b) Correctly summarize your results.

c) Check you results from **Q 6a** using `predict`. Include snapshots to show work.

7. (4 pts) Create a plot of the confidence and prediction intervals for the population regression line using `plotCI.reg` from *asbio*. Insert the plot into your document. Summarize the figure.

Q1 11pts, Q2 20pts, Q3 2pts, Q4 6pts, Q5 6pts Q6 6pts, Q7 4pts. **Total pts: 54.**

12

ANOVA I

Lab 12 Topics

1. Introductory topics in ANOVA
    - One-way ANOVA model
    - Partitioning sums of squares
    - hypothesis testing

Introduction

Recall that with t-tests we were limited to testing the effect of a single categorical explanatory variable with no more than two categorical levels. For example, in Lab 7 we compared dissolved O_2 in water with respect to locations above and below a riverside community using a pooled variance t-test. **Analysis of Variance** (**ANOVA**) provides a way to quantify the effect of one or more categorical explanatory variables (**factors**), each with two or more categories (**factor levels**), on a quantitative response variable. Consider Table 12.1, which lists plant growth data with respect to two factors (Soil N and Grazing). Soil N has three factor levels (Hi N, Lo N, Control), as does grazing (Hi Grazing, Lo Grazing, Control). We could analyze these data with ANOVA, but not with a t-test.

Table 12.1. Data from a hypothetical plant growth experiment.

Plot	Plant height (cm)	Soil N	Grazing
1	15	Control	Hi Grazing
2	12	Control	Lo Grazing
3	13	Control	Control
4	16	Lo N	Lo Grazing
5	17	Lo N	Hi Grazing
6	19	Lo N	Control
7	23	Hi N	Lo Grazing
8	22	Hi N	Hi Grazing
9	24	Hi N	Control

One-way ANOVA Model

In this lab we will consider ANOVA frameworks with a single factor (although this factor may have many factor levels). These are called **one-way ANOVA** models. If factor levels are assigned to experimental units in a randomized fashion we call this a **completely randomized design** or **CRD**. Recall from Lab 1 that this sort of design strengthens causal inferences concerning the effect of X on Y. This is because the effect of confounding and lurking explanatory variables on the response is averaged out.

Like regression, ANOVA is a type of general linear model. That is, we assume that Y can be modeled as linear transformation of X, and the underlying model errors are normally distributed. In a one-way ANOVA we have the model:

$$Y_{ij} = \mu_i + \varepsilon_{ij} \tag{12.1}$$

where Y_{ij} represents the jth observation from the ith factor level, μ_i represents the ith factor level true mean and ε_{ij} is the jth error from the ith factor level. As with regression, we assume: $\varepsilon_{ij} \sim N(0, \sigma^2)$. The one-way ANOVA model can also be expressed in a slope intercept form:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \tag{12.2}$$

where μ is the **true grand mean** across all combined factor levels and α_i is the ith **effect size**, and is calculated as:

$$\alpha_i = \mu_i - \mu. \tag{12.3}$$

We will formally address ANOVA model assumptions in Lab 13.

Partitioning the Sums of Squares

Let r be the number of factors levels, $i = 1, 2, \ldots, r$, and let the number of observations for the ith factor level be n_i. Then, the total number of observations across all r factor levels, n, is:

$$n = \sum_{i=1}^{r} n_i \tag{12.4}$$

The sample mean for the ith factor level, $\bar{Y}_i$, estimates the true mean of the ith factor level, μ_i. It is calculated as:

$$\bar{Y}_i = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i} \tag{12.5}$$

The sample grand mean, $\bar{Y}$, estimates the true grand mean, μ.

$$\bar{Y} = \frac{\sum_{i=1}^{r} \sum_{j=1}^{n_i} Y_{ij}}{n} \tag{12.6}$$

The jth model residual from the ith factor level, $\hat{\varepsilon}_{ij}$, estimates the true error, ε_{ij}.

$$\hat{\varepsilon}_{ij} = Y_{ij} - \bar{Y}_i \tag{12.7}$$

We can partition the **total sum of squares** (***SSTO***) in the ANOVA model into two components: the **treatment sum of squares** (***SSTR***), and the **sum of squares error** (***SSE***). $SSTR$ quantifies the variability of the r factor level sample means with respect to the sample grand mean (Eq 12.8), whereas SSE quantifies the variability of individual observations around their respective factor level sample means (Eq 12.9). $SSTO$ quantifies the variability of individual observations with respect to the sample grand mean.

$$SSTR = \sum_{i=i}^{r} n_i (\bar{Y}_i - \bar{Y})^2. \tag{12.8}$$

$$SSE = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 = \sum_{i=1}^{r} \sum_{j=1}^{n_i} \hat{\varepsilon}_{ij}^2. \tag{12.9}$$

$$SSTO = SSTR + SSE = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2. \tag{12.10}$$

By dividing $SSTR$ and SSE by their respective degrees of freedom, we obtain the variance estimates $MSTR$ and MSE. MSE is an unbiased estimator for the error term variance σ^2 in the ANONA general linear model (see Eq 12.1 and 12.2).

$$MSTR = \frac{SSTR}{r-1}. \tag{12.11}$$

$$MSE = \frac{SSE}{n-r}. \tag{12.12}$$

Hypothesis Testing

The fundamental question we ask with ANOVA is: "do all factor level groups have the same population mean?" Thus, the hypotheses in a one-way ANOVA are:

$$\begin{aligned} \mathrm{H}_0 &: \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_r \\ \mathrm{H}_\mathrm{A} &: \text{At least one } \mu_i \text{ not equal to the others.} \end{aligned}$$

This is equivalent to:

$$\begin{aligned} \mathrm{H}_0 &: \alpha_1 = \alpha_2 = \alpha_3 = \ldots = \alpha_r = 0 \\ \mathrm{H}_\mathrm{A} &: \text{At least one } \alpha_i \text{ does not equal 0.} \end{aligned}$$

where α_i is the ith effect size (Eq. 12.3).

To consider these hypotheses, we compare the variability within groups, $MSTR$, to the variability between groups, MSE. It may seem strange that we compare the means of treatments using variances, but this is what ANOVA does. The ANOVA test statistic is computed as:

$$F^* = \frac{MSTR}{MSE}. \tag{12.13}$$

Recall that an F-test is used to test a null hypothesis of homoscedasticity for two normal populations (Lab 9). If H_0 is true and assumptions for the ANOVA model hold, then F^* will be a random outcome from an F-distribution with $r-1$ numerator degrees of freedom and $n-r$ denominator degrees of freedom. We calculate the P-value as $P(X \geqslant F^*)$ where $X \sim F(r-1, n-r)$.

If the variability between treatments, $MSTR$, is large, and the variability within treatments, MSE, is small, then the ANOVA test statistic, F^*, will also be large and provide appreciable evidence against the null hypothesis of no difference among treatment means. Terms discussed thus far are summarized with an ANOVA table (Table 12.2).

Table 12.2. Components of an ANOVA table.

Variation source	df	SS	MS	F^*
$Treatment$ (Among groups)	$r-1$	$SSTR = \sum_{i=1}^{r} n_i(\bar{Y}_i - \bar{Y})^2$	$MSTR = \frac{SSTR}{r-1}$	$\frac{MSTR}{MSE}$
$Error$ (Between groups)	$n-r$	$SSE = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$	$MSE = \frac{SSE}{n-r}$	

The means of the sampling distributions of $MSTR$ and MSE are:

$$E(MSTR) = \sigma^2 + \sum_{i=1}^{r} n_i \frac{\alpha_i^2}{r-1}$$

and

$$E(MSE) = \sigma^2$$

Thus, if H_0 is true, and all α_is $= 0$, then $E(MSTR) = E(MSE) = \sigma^2$, and on average F^* will equal 1.

Example 12.1

Bean beetles are a pest to bean crops. To determine the best refrigeration temperature to prevent loss of harvested crops to bean beetles (or freezer burn), an experiment was conducted. Random samples were taken from a large population of beetles, with each sample containing 25 female and 25 male beetles. Samples were placed in jars with the same amount of food. Four chambers were created with a particular level of temperature. To conform to the protocol of a completely randomized design, five jars were randomly assigned to each of four temperatures of particular interest. The response variable, counts of beetle eggs, was tabulated a week after placing the beetles in the chambers (Table 12.3) One jar in treatment 3 broke and was discarded from the experiment.

Table 12.3. Been beetle data for Example 1

Eggs	Treatment
11	1
17	1
16	1
14	1
15	1
12	2
10	2
15	2
19	2
11	2
23	3
20	3
18	3
17	3
27	4
33	4
22	4
26	4
28	4

We want to test if any of the treatments produce a different mean number of eggs. We have the following hypotheses:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$
$$H_A : \text{At least one } \mu_i \text{ not equal to the others.}$$

We will use $\alpha = 0.05$. We find:

$$\bar{Y}_1 = \frac{\sum_{j=1}^{n_1} Y_{1j}}{n_1} = \frac{73}{5} = 14.6, \qquad \bar{Y}_2 = \frac{\sum_{j=1}^{n_2} Y_{2j}}{n_2} = \frac{67}{5} = 13.4,$$
$$\bar{Y}_3 = \frac{\sum_{j=1}^{n_3} Y_{3j}}{n_3} = \frac{78}{4} = 19.5, \qquad \bar{Y}_4 = \frac{\sum_{j=1}^{n_4} Y_{4j}}{n_4} = \frac{136}{5} = 27.2$$

and

$$\bar{Y} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{n_i} Y_{ij}}{n} = \frac{354}{19} = 18.63.$$

$$\begin{aligned} SSTR =& \sum_{i=i}^{r} n_i(\bar{Y}_i - \bar{Y})^2 \\ =& 5(14.6 - 18.63)^2 + 5(13.4 - 18.63)^2 + 4(19.5 - 18.63)^2 + 5(27.2 - 18.63)^2 = 588.2211. \end{aligned}$$

$$\begin{aligned} SSE =& \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2 \\ =& (11 - 14.6)^2 + (17 - 14.6)^2 + (16 - 14.6)^2 + (14 - 14.6)^2 + (15 - 14.6)^2 + \\ & (12 - 13.4)^2 + (10 - 13.4)^2 + (15 - 13.4)^2 + (19 - 13.4)^2 + (11 - 13.4)^2 + \\ & (23 - 19.5)^2 + (20 - 19.5)^2 + (18 - 19.5)^2 + (17 - 19.5)^2 + \\ & (27 - 27.2)^2 + (33 - 27.2)^2 + (22 - 27.2)^2 + (26 - 27.2)^2 = 158.2. \end{aligned}$$

$SSTO = SSTR + SSE = 588.2211 + 158.2 = 746.4211.$

Thus, we have:

$$MSTR = \frac{SSTR}{r-1} = \frac{588.22}{3} = 196.0737.$$
$$MSE = \frac{SSE}{n-r} = \frac{588.22}{15} = 10.54667.$$
$$F^* = \frac{MSTR}{MSE} = \frac{196.07}{10.55} = 18.59106.$$

We can summarize our work with an ANOVA table (Table 12.4):

Table 12.4. ANOVA table for Example 1

Source of variation	*df*	*SS*	*MS*	*F**
Between treatments	$r - 1 = 3$	588.22	196.07	18.591
Error (within treatments)	$n - r = 15$	158.2	10.55	
Total	$n - 1 = 18$	746.42		

To calculate the P-value we find $P(X \geqslant F^*)$ when $X \sim F(r-1, n-r)$. We find:

```
pf(18.59106, 3, 15, lower.tail = F)

[1] 2.584959e-05
```

Because the P-value is less than α we reject null and conclude that there is at least at least one μ_i is not equal to the others.

We perform these operations "by hand" below using **R** to help. We first calculate the sums of squares:

```
beetle <- read.csv("beetle.csv")
y <- beetle[,1]; x <- factor(beetle[,2])
Ybari <- tapply(y, x, mean)
Ybari

   1    2    3    4
14.6 13.4 19.5 27.2

ni <- tapply(y, x, length)
ni

1 2 3 4
5 5 4 5

r <- nlevels(x)
n <- sum(ni)
Ybar <- mean(y)

SSTR <- sum(ni * (Ybari - Ybar)^2)
SSTR

[1] 588.2211

fits <- rep(Ybari, ni)
SSE <- sum((y - fits)^2)
SSE

[1] 158.2

SSTO <- SSTR + SSE
SSTO

[1] 746.4211
```

Here are the mean squares:

```
MSTR <- SSTR/(r - 1)
MSTR

[1] 196.0737

MSE <- SSE/(n - r)
MSE

[1] 10.54667
```

Here is the test statistic and P-value:

```
F.star <- MSTR/MSE
F.star

[1] 18.59106

pf(F.star, r - 1, n - r, lower.tail = F)

[1] 2.584961e-05
```

We could have easily obtained this result in just two lines of code:

```
model <- lm(y ~ x)
anova(model)

Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x          3 588.22 196.074  18.591 2.585e-05 ***
Residuals 15 158.20  10.547
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

■

Assignment 12

Answer all questions in one MS Word document and submit to Moodle. At the beginning of the document include the assignment number, the date, your name and section number.

Use complete sentences when appropriate, and make sure any tables, figures and computer output you include adhere to class standards (see Syllabus).

One-way ANOVA model

- Open **R**
- Load the *asbio* package by typing `library(asbio)` or by going to **Packages > Load packages > *asbio***.
- Type `book.menu()` in the **R** console.

1. Go to the Ch. 10 pulldown menu. Click on **ANOVA mechanics**. Mac-users and others who wish to obtain the GUI directly can type: `see.anova.tck()`.

a) (2 pts) What do you think the three normal distributions represent?

b) (2 pts) What do think the numbers (ones, twos, and threes) inside the distributions are?

c) (3 pts) The quantities α_1, α_2, and α_3 are called effect sizes. They are differences of the true factor level means, μ_i, and the true grand mean, μ. Given this, are the hypotheses below equal? Why?

$$\mathrm{H_0} : \mu_1 = \mu_2 = \mu_3$$
$$\mathrm{H_A} : \text{At least one } \mu_i \text{ not equal to the others.}$$

$$\mathrm{H_0} : \alpha_1 = \alpha_2 = \alpha_3 = 0$$
$$\mathrm{H_A} : \text{At least one } \alpha_i \text{ does not equal 0.}$$

d) (3 pts) σ^2 represents the true variance within the factor level populations. Alter the σ^2 slider. Does it appear as if the populations are assumed to have equal variances in ANOVA? Why?

e) (2 pts) Based on the GUI sliders, what are two ways to increase the size of F^* (and decrease P-values)?

f) (4 pts) Create a situation in which $\mathrm{H_0} : \mu_1 = \mu_2 = \mu_3$ is true. Resample from the populations repeatedly by clicking on the **Refresh** button.

- Is it still possible to reject $\mathrm{H_0}$ at $\alpha = 0.05$? Insert an example plot into your document.
- What is this called?

Application of the ANOVA model

2. A large laboratory has four types of devices to determine the pH of soil samples. The laboratory wants to determine whether there are differences in the true average readings given by these devices. The lab uses 24 soil samples with a known pH, and randomly assigns six of these samples to each of the four devices. The soil samples are tested and the difference in pH reading of the device from the true (known pH) is measured. These results are shown in Table 12.5. The data are also in Moodle under the name `pH.csv`.

Table 12.5. Data for Question 2.

difference from known pH	device
0.079	A
0.738	A
-1.045	A
-0.424	A
0.59	A
-1.244	A
0.006	B
-0.596	B
0.509	B
-0.082	B
-0.442	B
-1.877	B
-0.576	C
-1.54	C
-0.85	C
-0.163	C
-1.457	C
-0.284	C
-0.904	D
0.767	D
-1.239	D
0.017	D
0.056	D
0.142	D

a) (2 pts) Is this a completely randomized design? Why?

b) (2 pts) Correctly state the null and alternative hypotheses.

c) (5 pts) Calculate $\bar{Y}_i$ and n_i, (there will be four of each of these) and the sample grand mean, $\bar{Y}$. use **R** to help with calculations.

d) (3 pts) Calculate the sums of squares *SSTR*, *SSE* and *SSTO* "by hand" using **R** to help.

e) (2 pts) Calculate the mean squares *MSTR* and *MSE* "by hand" using **R** to help.

f) (4 pts) Create an ANOVA Table (refer to Table 12.2) to summarize your results from **d** and **e**. Insert the table into your document.

g) (2 pts) Calculate the *P*-value, using the function `pf` in **R**.

h) (2 pts) What are your conclusions? Do we reject H_0 at $\alpha = 0.05$?

i) (2 pts) Verify your results in **R** using the functions `lm` and `anova`. Insert the output into you document.

Q1 16pts, Q2 24pts. **Total pts: 40.**

13

ANOVA II

Lab 13 Topics

- Assumptions and diagnostics for ANOVA

Introduction

We learned last week that ANOVA is a type of general linear model. That is, we assume that Y can be modeled as linear transformation of X, and that the underlying model errors are normally distributed. In a one-way ANOVA we have the model:

$$Y_{ij} = \mu_i + \varepsilon_{ij} \tag{13.1}$$

where Y_{ij} represents the jth observation from the ith factor level, μ_i represents the ith factor level true mean and ε_{ij} is the jth error from the ith factor level. We assume: $\varepsilon_{ij} \sim N(0, \sigma^2)$. The one-way ANOVA model can also be expressed in a slope intercept form:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \tag{13.2}$$

where μ is the true grand mean across all combined factor levels and α_i is the ith effect size, and is calculated as:

$$\alpha_i = \mu_i - \mu. \tag{13.3}$$

There are three assumptions for ANOVA:

1. Observations are independent.
2. The error term distributions are normally distributed.
3. The error term distributions are homoscedastic.

Most of these assumptions concern the form of the error term distribution. Thus, we generally check model assumptions by by examining the model residuals. This is because the residuals estimate the true errors, ε_i. When model-checking, we remain attentive to the presence of outliers because estimators in general linear models are not resistant to outliers.

Example 13.1

Last week we used a bean beetle dataset to demonstrate one-way ANOVA. We tested the null hypothesis that all four refrigeration units tested would result in the same true mean number of bean beetle eggs.

```
beetle <- read.csv("beetle.csv")
model <- lm(Eggs ~ Treatment, data = beetle)
anova(model)

Analysis of Variance Table

Response: Eggs
          Df Sum Sq Mean Sq F value    Pr(>F)
Treatment  1 483.08  483.08  31.186 3.289e-05 ***
Residuals 17 263.34   15.49
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the P-value, we reject $\mathrm{H}_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ at $\alpha = 0.05$. However the P-value, and downstream decisions are only reliable if assumptions for the model are valid.

Independence

To check for temporal independence we can plot model residuals against the order that observations were taken. If independence is satisfied there should be no patterns to points in this plot. We obtain residuals using:

```
e.hat <- resid(model)
```

Fig 13.1 plots residuals against the assumed order of observations. We assume that the order of observations given in the dataset represents the actual order that they were recorded.

```
plot(1:length(e.hat), e.hat, xlab = "Assumed order of observations",
     ylab = "Residuals", type = "o")
```

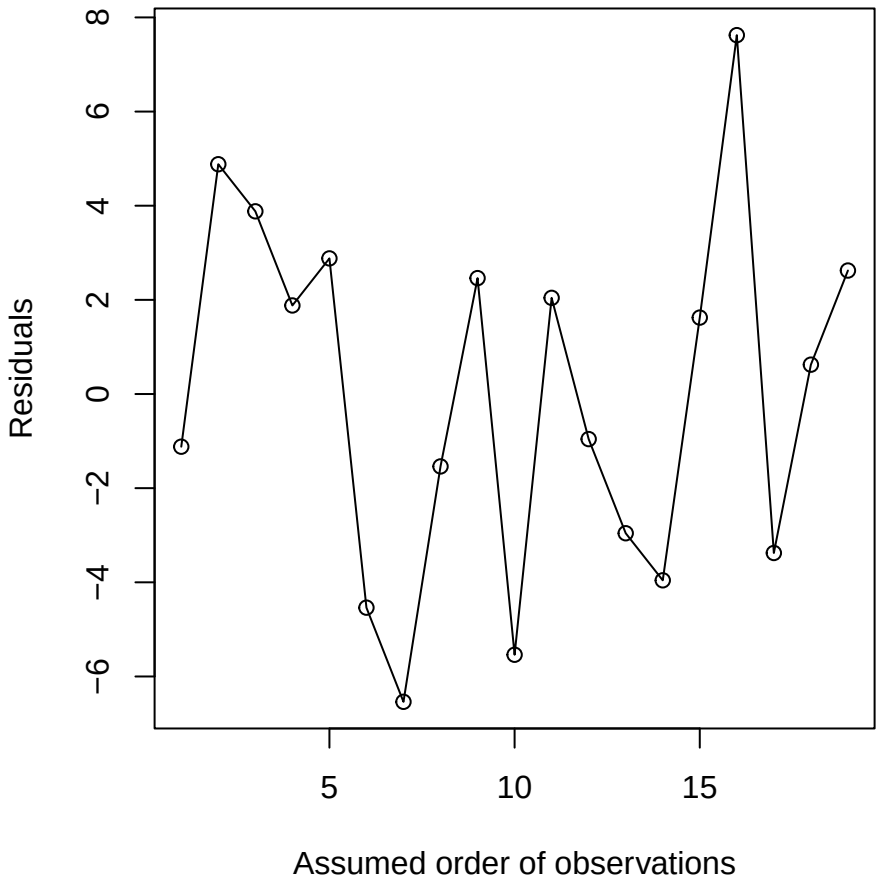


Figure 13.1. Residuals as a function of assumed order in Ex. 1.

There seems to be little pattern to points in the plot, supporting the assumption of independence.

Normality

We expect noise in ANOVA model have the distribution $N(0, \sigma^2)$ for each level in X. We can check this assumption with a normal quantile plot of residuals (Fig. 13.2), and the Shapiro-Wilk test for normality.

for normality.

```
qqnorm(e.hat, main = "")
qqline(e.hat)
```

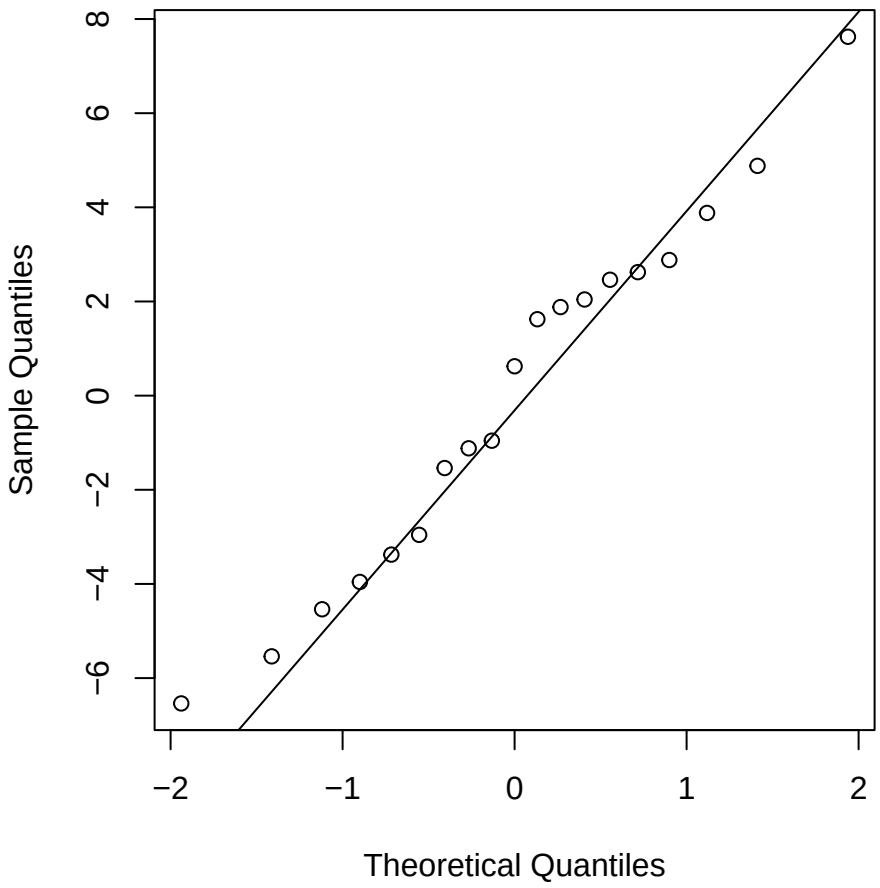


Figure 13.2. Normal quantile plot of residuals in Ex. 1.

We run the Shapiro-Wilk test on the model residuals.

```
shapiro.test(e.hat)

Shapiro-Wilk normality test

data:  e.hat
W = 0.97327, p-value = 0.8394
```

The normal quantile plot and Shapiro-Wilk test support the assumption of normality.

Homoscedasticity

We assume errors have same the distribution, $N(0, \sigma^2)$, for each factor level in X,. Thus, we assume that the variance, σ^2 is constant among factor levels. To check for equal error

variances we can plot residuals as a function of fitted values (Fig 13.3). If the assumption of homoscedasticity is valid there should be no pattern to the points in the plot.

```
fits <- fitted(model)
plot(fits, e.hat, xlab = "Fitted values", ylab = "Residuals")
abline(h = 0, lty = 2)
```

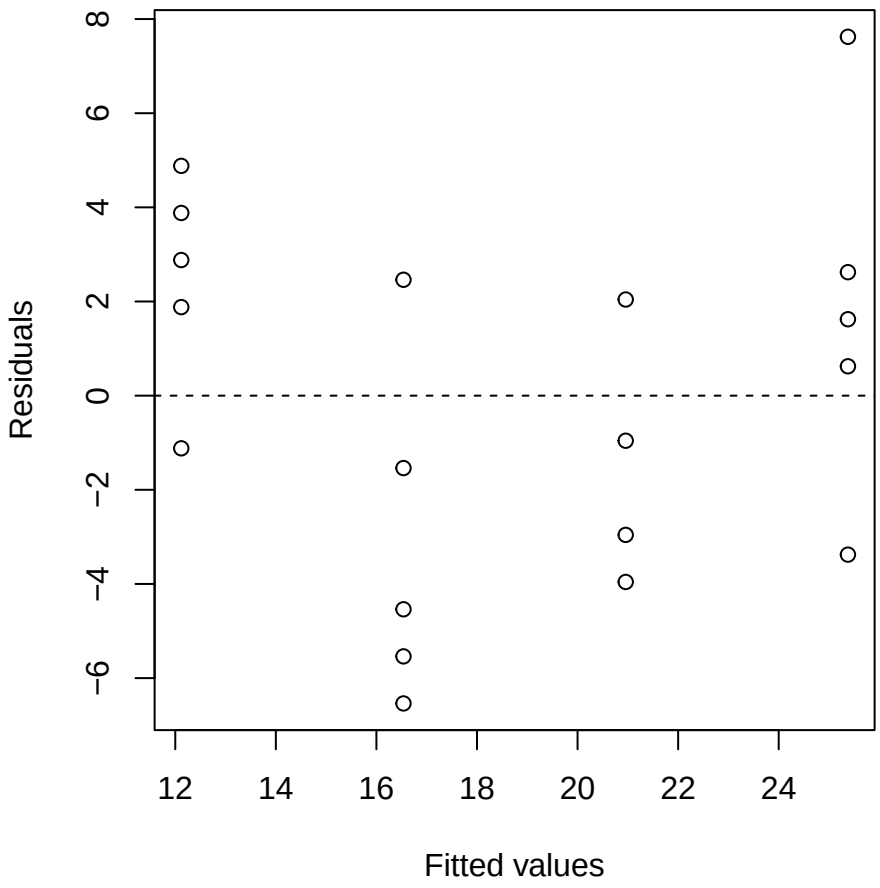


Figure 13.3. Plot of residuals as a function fits in Ex. 1.

We can also formally test for factor level homoscedasticity using the modified Levene's test.

```
library(asbio)
modlevene.test(e.hat, factor(beetle$Treatment))

Modified Levene's test of homogeneity of variances

df1 = 3,  df2 = 15,  F = 0.2417,  p-value = 0.86589
```

Based on homoscedasticity diagnostic plot and modified Levene's test, there appears to be little evidence of heteroscedasticity. All of our assumptions have been satisfied, indicating

that the ANOVA results for the beetle dataset are valid.

■

Assignment 13

Answer all questions in one MS Word document and submit to Moodle. At the beginning of the document include the assignment number, the date, your name and section number.

Use complete sentences when appropriate, and make sure any tables, figures and computer output you include adhere to class standards (see Syllabus).

1. (8 pts) An agricultural researcher wishes to test whether the potassium content of tree leaves vary from three different varieties of apple trees. Conduct an ANOVA of the potassium data shown in Table 13.1. . The data are also in Moodle under the name `K.csv`.

Table 13.1. Potassium content of tree leaves from three different varieties of apple trees (1, 2, and 3).

K-content	Variety
0.42	Var1
0.4	Var1
0.64	Var1
0.54	Var1
0.44	Var1
0.75	Var2
0.79	Var2
0.9	Var2
0.83	Var2
0.86	Var2
0.65	Var3
0.77	Var3
0.73	Var3
0.86	Var3
0.69	Var3

a) What are your null and alternative hypotheses?

b) Run the test in **R**, using `lm` and `anova`. Include snapshots to show your work.

c) Describe your results. Do you reject or fail to reject the null hypothesis defined in **(a)**?

d) Store the fitted values and the residuals as **R** objects. Include snapshots to show your work.

2. (4 pts) Check for independence of observations:

a) Attach a plot of residuals against against the order of observations given in the dataset.

b) Describe your results.

3. (7 pts) Check for normality of errors:

a) Create a normal probability plot.

b) What is the null hypothesis for the Shapiro-Wilk test?

c) Run the Shapiro-Wilk test using `shapiro.test`. Attach your results.

d) Interpret your results in **(a)** and **(c)**. Do you fail to reject the null hypothesis defined in **(b)**? What does this mean?

4. (7 pts) Check for homoscedasticity of errors:

a) Create a diagnostic plot for homoscedasticity by plotting residuals against fitted values.

b) What is the null hypothesis for the Levene's test?

c) Run the Levene's test using `modlevene.test` in *asbio*. Attach your results.

d) Interpret your results in **(a)** and **(c)**. Do you fail to reject the null hypothesis defined in **(b)**? What does this mean?

5. (8 pts) The one way ANOVA model is defined in Eq 13.2.

a) Why is this a linear model?

b) What is Y_{ij}?

c) What is α_i?

d) What do we mean when we say $\varepsilon_{ij} \sim N(0, \sigma^2)$. Be as detailed as possible.

Q1 8pts, Q2 4pts, Q3 7pts, Q4 7pts, Q5 8pts. **Total pts: 34.**

14

ANOVA III

Lab 14 Topics

1. Family-wise type I error
2. Simultaneous pairwise comparisons in the context of ANOVA
   - Fisher's LSD
   - Tukey's HSD

Introduction

Assume that we have conducted an ANOVA and have rejected $H_0 : \mu_1 = \mu_2 = \ldots = \mu_r$. While we conclude that at least one μ_i is not equal to at least one of the others, we do not know which particular factor levels should be considered different. In addition, we don't know if there is a trend among μ_is with some less than or greater than others. To acquire this information we will conduct what are called ***post hoc*** (Latin for "after this") tests.

Pairwise Comparisons

Pairwise comparisons are the most common type of *post hoc* test. Pairwise comparisons are generally used to test null hypotheses that all possible differences of factor level means, considered one pair of differences at a time, will equal zero. This can be stated summarily as:

$$H_0 : \mu_i = \mu_{i'}$$
$$H_A : \mu_i \neq \mu_{i'}$$

where $\mu_{i'}$ indicates a factor level true mean other than the ith factor level true mean.

Let r be the number of factor levels. There will always be $\binom{r}{2} = \frac{r^2-r}{2}$ possible pairwise comparisons.

As an example of a single pairwise test, suppose we have rejected the omnibus ANOVA null hypothesis, and we are interested in the true difference, D, between two particular factor level population means, say $\mu_1 - \mu_2$. To estimate D we use the difference of sample means, $\hat{D} = \bar{Y}_1 - \bar{Y}_2$. This difference will be an unbiased estimator for D.

Because in ANOVA we assume all of the factor level populations have the same variance, σ^2, and because MSE is an unbiased estimator of σ^2, the estimator for $\sigma^2_{\hat{D}}$ (i.e., the variance of the sampling distribution of $\hat{D}$), is:

$$\hat{\sigma}^2_{\hat{D}} = MSE\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \tag{14.1}$$

where $MSE = \sum_{i=1}^{n} \hat{\varepsilon}_i^2/(n-r)$ can be obtained directly from the the ANOVA model.

To test the null hypothesis $\mathrm{H}_0 : D = 0$, we calculate the test statistic:

$$t^* = \frac{\hat{D}}{\sqrt{\hat{\sigma}^2_{\hat{D}}}} = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\hat{\sigma}^2_{\hat{D}}}}. \tag{14.2}$$

We calculate the P-value as $2 \cdot P(T \geqslant |t^*|)$ where $T \sim t(n-r)$. Note Eq. 14.1 and 14.2 have the exact same format as a pooled variance t-test (Lab 7). The only difference here is that more than two factor levels will be used in computing MSE.

It is important to note that pairwise comparisons are not the only possible post *post hoc* comparisons. For instance, given a factor with three levels, level 1 could be compared to the combined (average) effect of levels 2 and 3.

It may be tempting to look at only the comparisons one is interested in after the data are gathered and summaries are examined. However, this is statistically incorrect and is known as **data snooping**, (a lack of independence in tests due to the cherry-picking results). Data snooping occurs because all possible comparisons are being made implicitly in the mind of the investigator as he or she reviews the data while deciding which tests would make his or her experiment look better. Thus, we need to either specify what contrasts we are interested in before we sample, or look at an entire *family* of comparisons (e.g., all possible pairwise tests) in our *post hoc* tests.

Family-wise Type I Error

The implementation of multiple *post hoc* tests presents a problem. Recall that a type I error occurs when a null hypothesis is rejected when it is actually true (Lab 6), and that the acceptable probability of type I error for a test is defined with the significance level, α. When multiple related null hypothesis tests are run, then multiple related type one errors (**false discoveries**) can occur. The **family-wise error rate** (**FWER**) is the probability of making one or more type I errors when performing multiple null hypothesis tests. Given m independent tests, this probability will be:

$$1 - (1-\alpha)^m \tag{14.3}$$

where α is the significance level to be used for each test. Thus, five independent tests, each using $\alpha = 0.05$, will have a family-wise α of $1-(1-\alpha)^5 = 0.226$. That is, the probability of falsely rejecting H_0 at least once across all five tests will be 0.226.

The probably of ballooning FWER due to multiple comparisons has resulted in the development of many different approaches. I will only consider two methods in the context of all possible pairwise comparisons here: Fisher's LSD and Tukey's HSD.

Fisher's LSD Procedure

Fisher's **least significant difference** (**LSD**) method fixes the probability of a false rejection of H_0 for each single pair of means being compared. It does not, however, control the overall probability of false rejection of H_0 for comparisons of all other pairs of means. The LSD procedure is essentially a series of pooled variance t-tests, using the ANOVA MSE as the test statistic pooled variance. In fact, the only family-wise adjustment used in the LSD method is the requirement of a rejection of the omnibus ANOVA H_0 hypothesis.

The method, developed by Fisher (1936), has the following steps:

1. We first need to reject $H_0 : \mu_1 = \mu_2 = \ldots = \mu_r$. If we can't do this, we can go no further in factor level comparisons. This caveat is often referred to as Fisher's "protected" LSD.

2. We define the least significant difference to be the observed difference between sample means necessary to reject $H_0 : \mu_i = \mu_{i'}$ as:

$$LSD_{i,i'} = t_{1-(\alpha/2),n-r} \cdot \sqrt{\hat{\sigma}^2_{\hat{D}}} \tag{14.4}$$

   where $\hat{\sigma}^2_{\hat{D}}$ is given in Eq 14.1.

3. Compare all pairs of means.

4. If $|\bar{Y}_i - \bar{Y}_{i'}| \geqslant LSD_{i,i'}$ for a particular comparison, then we reject $H_0 : \mu_i = \mu_{i'}$.

Example 14.1

Humans are born with rudimentary reflexes for walking, but these largely disappear by the age of eight weeks due to disuse. Accordingly, walking movements must be relearned by an infant following a significant passage of time, through a process of trial and error. Zelazo *et al.* (1972) performed a series of experiments to determine whether certain exercises could allow infants to maintain their walking reflexes, and allow them to walk at an earlier age. Study subjects were 24 white male infants from middle class families. Infants were randomly assigned to one of four groups immediately following birth.

- Active exercise (AE): parents were taught and were told to apply exercises that would strengthen the walking reflexes of their infant.

- Passive exercise (PE): parents were taught and told to apply exercises unrelated to walking.
- Test-only (TO): investigators did not specify any exercise, but visited and tested the walking reflexes of infants in weeks 1 through 8. This treatment was established to account for the potential effect of the walking reflex tests themselves and thus served as a control group.
- Control (C): no exercises were specified, and infants were only tested at weeks one and eight.

The data are in *asbio* in a dataframe called `baby.walk`.

```
library(asbio)
data(baby.walk)
```

Running the ANOVA model we have:

```
model <- lm(date ~ treatment, data = baby.walk)
anova(model)

Analysis of Variance Table

Response: date
          Df Sum Sq Mean Sq F value  Pr(>F)
treatment  3 16.602  5.5340  3.5676 0.03482 *
Residuals 18 27.921  1.5512
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Because we reject $H_0 : \mu_{PE} = \mu_{AE} = \mu_{TO} = \mu_C$ we can proceed to Fisher's LSD pairwise comparisons.

Here are the factor level means, Y_i and factor level sample sizes, n_i, the number of factor levels, r, and the total sample size, n.

```
means <- tapply(baby.walk[,1], baby.walk[,2], mean)
ns <- tapply(baby.walk[,1], baby.walk[,2], length)

means

      AE        C       PE       TO
10.12500 12.35000 10.65000 11.70833

ns

AE  C PE TO
 6  5  5  6

r <- nlevels(baby.walk[,2])
r

[1] 4

n <- sum(ns)
n

[1] 22
```

MSE can be obtained from the linear model:

```
MSE <- sum(resid(model)^2)/(n - r)
MSE

[1] 1.551157
```

We have $(r^2 - r)/2 = (16 - 4)/2 = 6$ possible pairwise comparisons. If we had an equal number of samples for each treatment then we would only have to calculate LSD once. However, there are two different sample sizes. We have $n = 6$ for the AE and TO groups and $n = 5$ for the C and PE groups. Defining $\alpha = 0.05$, we have:

$$\begin{aligned} LSD_{AE,TO} &= t_{1-(\alpha/2),n-r} \cdot \sqrt{MSE\left(\frac{1}{n_{AE}} + \frac{1}{n_{TO}}\right)} \\ &= t_{0.975,18} \cdot \sqrt{1.551157 \cdot \left(\frac{2}{6}\right)} \\ &= 2.100922 \cdot \sqrt{1.551157 \cdot \left(\frac{2}{6}\right)} \\ &= 1.510717 \end{aligned}$$

We obtain $t_{1-(\alpha/2),n-r}$ using the t inverse CDF.

```
alpha = 0.05
qt(1 - alpha/2, n - r)

[1] 2.100922
```

$$\begin{aligned} LSD_{C,PE} &= t_{1-(\alpha/2),n-r} \cdot \sqrt{MSE \cdot \left(\frac{1}{n_C} + \frac{1}{n_{PE}}\right)} \\ &= t_{0.975,18} \cdot \sqrt{1.551157 \cdot \left(\frac{2}{5}\right)} \\ &= 2.100922 \cdot \sqrt{1.551157 \cdot \left(\frac{2}{5}\right)} \\ &= 1.654908 \end{aligned}$$

For all other comparisons, sample sizes are 5 and 6 for the two groups. Thus,

$$\begin{aligned} LSD_{AE,C} = LSD_{AE,PE} = LSD_{C,TO} = LSD_{PE,TO} &= 2.100922 \cdot \sqrt{1.551157 \cdot \left(\frac{1}{6} + \frac{1}{5}\right)} \\ &= 2.100922 \cdot \sqrt{1.551157 \cdot \left(\frac{11}{30}\right)} \\ &= 1.584454 \end{aligned}$$

Pairwise mean differences can be quickly computed for all pairs, using:

```
mean.diff <- round(abs(outer(means, means, "-")),3)
mean.diff[upper.tri(mean.diff)] <- ""
data.frame(mean.diff)

      AE     C    PE TO
AE     0
C  2.225     0
PE 0.525   1.7     0
TO 1.583 0.642 1.058  0
```

1. For AE and C, we have $|\bar{Y}_{AE} - \bar{Y}_C| = 2.225$ and $LSD_{AE,C} = 1.584454$. Because $|\bar{Y}_{AE} - \bar{Y}_C| > LSD_{AE,C}$, **we reject** $H_0 : \mu_{AE} = \mu_C$.
2. For AE and PE, we have $|\bar{Y}_{AE} - \bar{Y}_{PE}| = 0.525$ and $LSD_{AE,C} = 1.584454$. Because $|\bar{Y}_{AE} - \bar{Y}_{PE}| < LSD_{AE,PE}$, **we fail to reject** $H_0 : \mu_{AE} = \mu_{PE}$.
3. For C and PE, we have $|\bar{Y}_C - \bar{Y}_{PE}| = 1.7$ and $LSD_{C,PE} = 1.654908$. Because $|\bar{Y}_C - \bar{Y}_{PE}| > LSD_{C,PE}$, **we reject** $H_0 : \mu_C = \mu_{PE}$.

4. For AE and TO, we have $|\bar{Y}_{AE} - \bar{Y}_{TO}| = 1.583$ and $LSD_{AE,TO} = 1.510717$. Because $|\bar{Y}_{AE} - \bar{Y}_{TO}| > LSD_{AE,TO}$, **we reject** $H_0 : \mu_{AE} = \mu_{TO}$.

5. For C and TO, we have $|\bar{Y}_{C} - \bar{Y}_{TO}| = 0.642$ and $LSD_{C,TO} = 1.584454$. Because $|\bar{Y}_{C} - \bar{Y}_{TO}| < LSD_{PE,TO}$, **we fail to reject** $H_0 : \mu_{C} = \mu_{TO}$.

6. For PE and TO, we have $|\bar{Y}_{PE} - \bar{Y}_{TO}| = 1.058$ and $LSD_{PE,TO} = 1.584454$. Because $|\bar{Y}_{PE} - \bar{Y}_{TO}| < LSD_{PE,TO}$, **we fail to reject** $H_0 : \mu_{PE} = \mu_{TO}$.

We can quickly obtain these results using the function `pairw.anova` from *asbio*.

```
LSD <- pairw.anova(baby.walk[,1], baby.walk[,2], method = "lsd")
LSD

95% LSD confidence intervals

              LSD     Diff    Lower    Upper  Decision Adj. p-value
muAE-muC  1.58443   -2.225 -3.80943 -0.64057 Reject H0      0.00856
muAE-muPE 1.58443   -0.525 -2.10943  1.05943    FTR H0      0.49523
muC-muPE  1.65489      1.7  0.04511  3.35489 Reject H0      0.04467
muAE-muTO  1.5107 -1.58333 -3.09403 -0.07264 Reject H0      0.04095
muC-muTO  1.58443  0.64167 -0.94277   2.2261    FTR H0      0.40604
muPE-muTO 1.58443 -1.05833 -2.64277   0.5261    FTR H0      0.17754
```

The function also provides confidence intervals for the true difference, D. Note that in cases when 0 is in the interval we fail to reject $H_0 : D = 0$. The function also allows graphical summarization of the comparisons (Fig 14.1, 14.2).

```
plot(LSD, type = 1, ylab = "Onset of walking (days)")

Bars are means.  Errors are SEs.

The population means of factor levels with the same letter are not
significantly different at alpha = 0.05 using the Fisher LSD method.
```

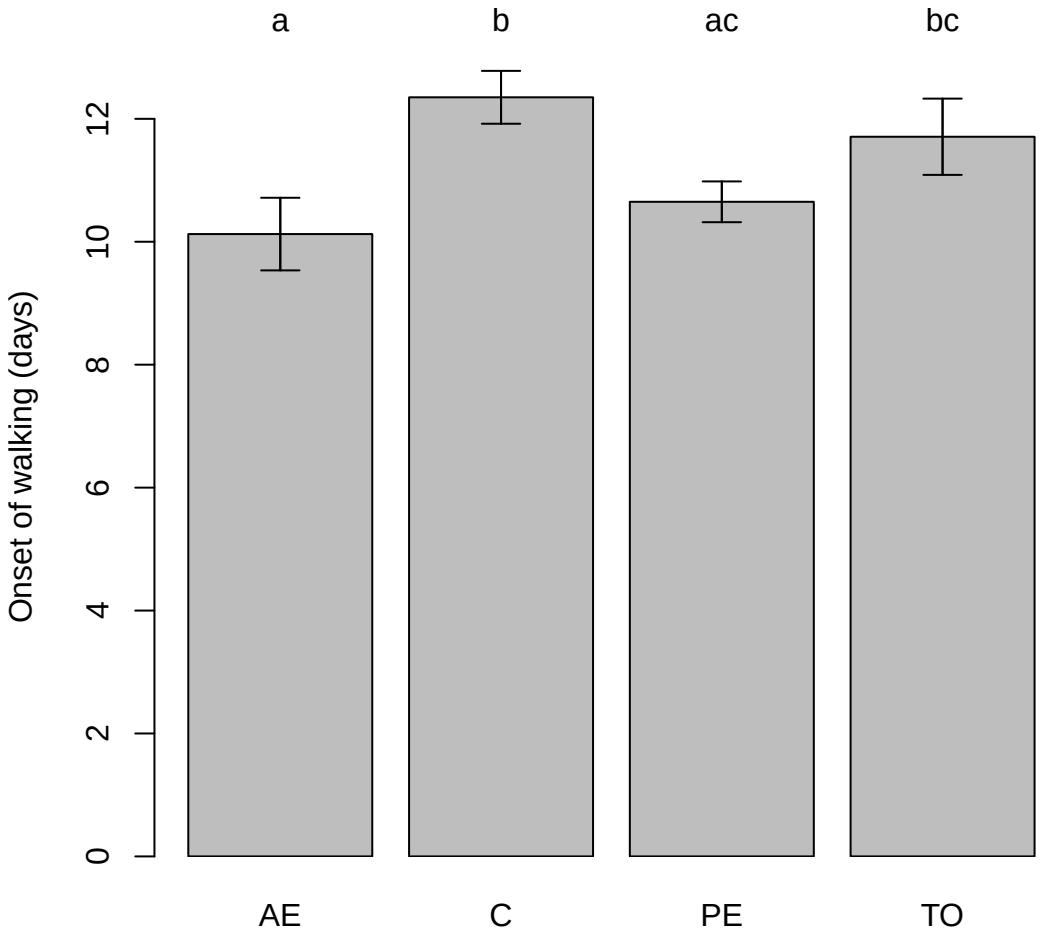


Figure 14.1. The population means of factor levels with the same letter are not significantly different at alpha = 0.05 using the Fisher LSD method. Bars are means. Errors are SEs.

```
plot(LSD, type = 2, ylab = "Onset of walking (days)")
```

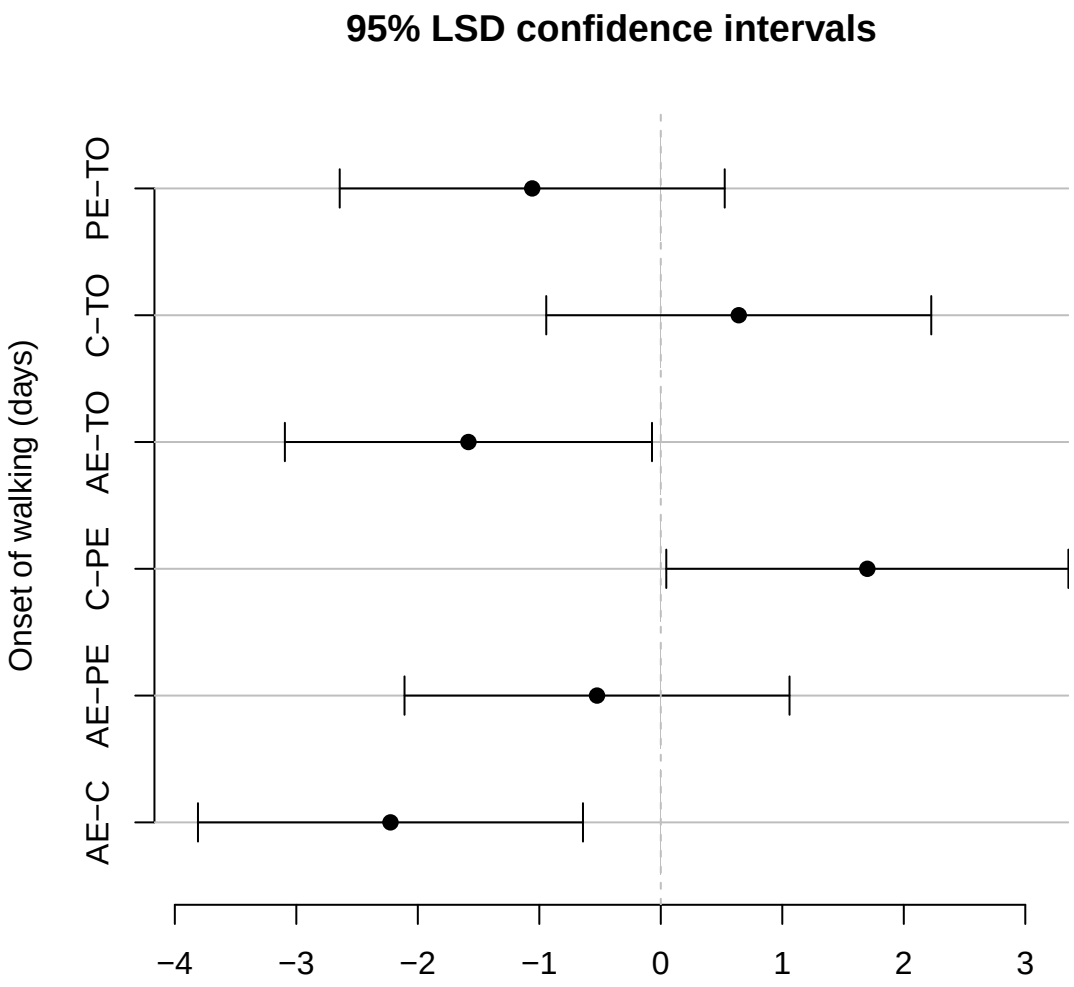


Figure 14.2. LSD 95% confidence intervals for the true difference, D.

■

Tukey's HSD Procedure

Tukey's honest significant difference (**HSD**) method (Tukey, 1949) explicitly controls FWER for the family of all possible pairwise tests. Thus, if we are using 100 pairwise comparisons there will be at most only a 5% probability of the overall family of tests being in error (i.e. that one or more of the 100 pairwise comparisons incorrectly reject H_0). In addition, Tukey's procedure provides narrower confidence intervals than two other popular

methods that control FWER (Scheffé and Bonferroni) when all possible pairwise comparisons are being considered.

When sample sizes are equal, then the Tukey HSD FWER is exactly α. When sample sizes are not equal, the FWER is less than α. Thus, the procedure is conservative when sample sizes are not equal. As before, rejection of the omnibus ANOVA null hypothesis should occur before proceeding with Tukey HSD comparisons.

Under the Tukey HSD approach we reject $\mathrm{H}_0 : \mu_i = \mu_{i'}$ if $|q^*| \geqslant q(1-\alpha, r, n-r)$, where:

$$q^* = \frac{\sqrt{2}\hat{D}}{\sqrt{\hat{\sigma}^2_{\hat{D}}}} \tag{14.5}$$

and q indicates the inverse CDF for the **studentized range distribution**; a distribution derived by Tukey (1949). The studentized range distribution has three parameters, the lower-tailed probability $1-\alpha$, r = the number of factor levels being compared, and the degrees of freedom error in the ANOVA model, i.e., $n-r$. We calculate $\hat{\sigma}^2_{\hat{D}}$ using Eq. 14.1.

We can also calculate Tukey's $(1-\alpha)100\%$ confidence intervals for D using:

$$\hat{D} \pm T \cdot \hat{\sigma}_{\hat{D}} \tag{14.6}$$

where

$$T = \frac{1}{\sqrt{2}} q(1-\alpha, r, n-r) \tag{14.7}$$

Example 14.2

We will reuse the `baby.walk` data of Exercise 1 to demonstrate Tukey's HSD method. The value for $q(1-\alpha, r, n-r)$ is 3.996978.

```
alpha = 0.05
q <- qtukey(1 - alpha, r, n - r)
q

[1] 3.996978
```

1. For AE and C, we have

$$\begin{aligned} q^* &= \frac{\sqrt{2} \cdot \hat{D}}{\sqrt{MSE(11/30)}} \\ &= \frac{\sqrt{2} \cdot 2.225}{0.754} = 4.220. \end{aligned}$$

Because $q^* > q$, **we reject** $\mathrm{H}_0 : \mu_{AE} = \mu_C$.

2. For AE and PE, we have

$$q^* = \frac{\sqrt{2} \cdot \hat{D}}{\sqrt{MSE(11/30)}} = \frac{\sqrt{2} \cdot 0.525}{0.754} = 0.985.$$

   Because $q^* < q$, **we fail to reject** $\mathrm{H}_0 : \mu_{AE} = \mu_{PE}$.

3. For C and PE, we have

$$q^* = \frac{\sqrt{2} \cdot \hat{D}}{\sqrt{MSE(2/5)}} = \frac{\sqrt{2} \cdot 1.7}{0.787} = 3.054.$$

   Because $q^* < q$, **we fail to reject** $\mathrm{H}_0 : \mu_C = \mu_{PE}$.

4. For AE and TO, we have

$$q^* = \frac{\sqrt{2} \cdot \hat{D}}{\sqrt{MSE(2/6)}} = \frac{\sqrt{2} \cdot 1.583}{0.719} = 1.261.$$

   Because $q^* < q$, **we fail to reject** $\mathrm{H}_0 : \mu_{AE} = \mu_{TO}$.

5. For C and TO, we have

$$q^* = \frac{\sqrt{2} \cdot \hat{D}}{\sqrt{MSE(11/30)}} = \frac{\sqrt{2} \cdot 0.642}{0.754} = 3.189.$$

   Because $q^* < q$, **we fail to reject** $\mathrm{H}_0 : \mu_C = \mu_{TO}$.

6. For PE and TO, we have

$$q^* = \frac{\sqrt{2} \cdot \hat{D}}{\sqrt{MSE(11/30)}} = \frac{\sqrt{2} \cdot 1.058}{0.754} = 2.003.$$

   Because $q^* < q$, **we fail to reject** $\mathrm{H}_0 : \mu_{PE} = \mu_{TO}$.

Note that Tukey's method gives a more conservative (and safe, with respect to family-wise type I error) interpretation of the data compared to LSD. LSD found two additional significant pairwise differences.

One again, we can quickly use `pairw.anova` to get these results. Graphical summaries are shown in Fig 14.3 and 14.4.

```
tukey <- pairw.anova(baby.walk[,1], baby.walk[,2])
tukey

95% Tukey-Kramer confidence intervals

              Diff    Lower    Upper  Decision Adj. p-value
muAE-muC    -2.225 -4.35648 -0.09352 Reject H0     0.038997
muAE-muPE   -0.525 -2.65648  1.60648    FTR H0     0.897224
muC-muPE       1.7 -0.52625  3.92625    FTR H0     0.172932
muAE-muTO -1.58333 -3.61562  0.44895    FTR H0     0.160457
muC-muTO   0.64167 -1.48981  2.77314    FTR H0     0.829542
muPE-muTO -1.05833 -3.18981  1.07314    FTR H0     0.513366
```

```
plot(tukey, type = 1, ylab = "Onset of walking (days)")

Bars are means.  Errors are SEs.

The population means of factor levels with the same letter are not
significantly different at alpha = 0.05 using the Tukey HSD method.
```

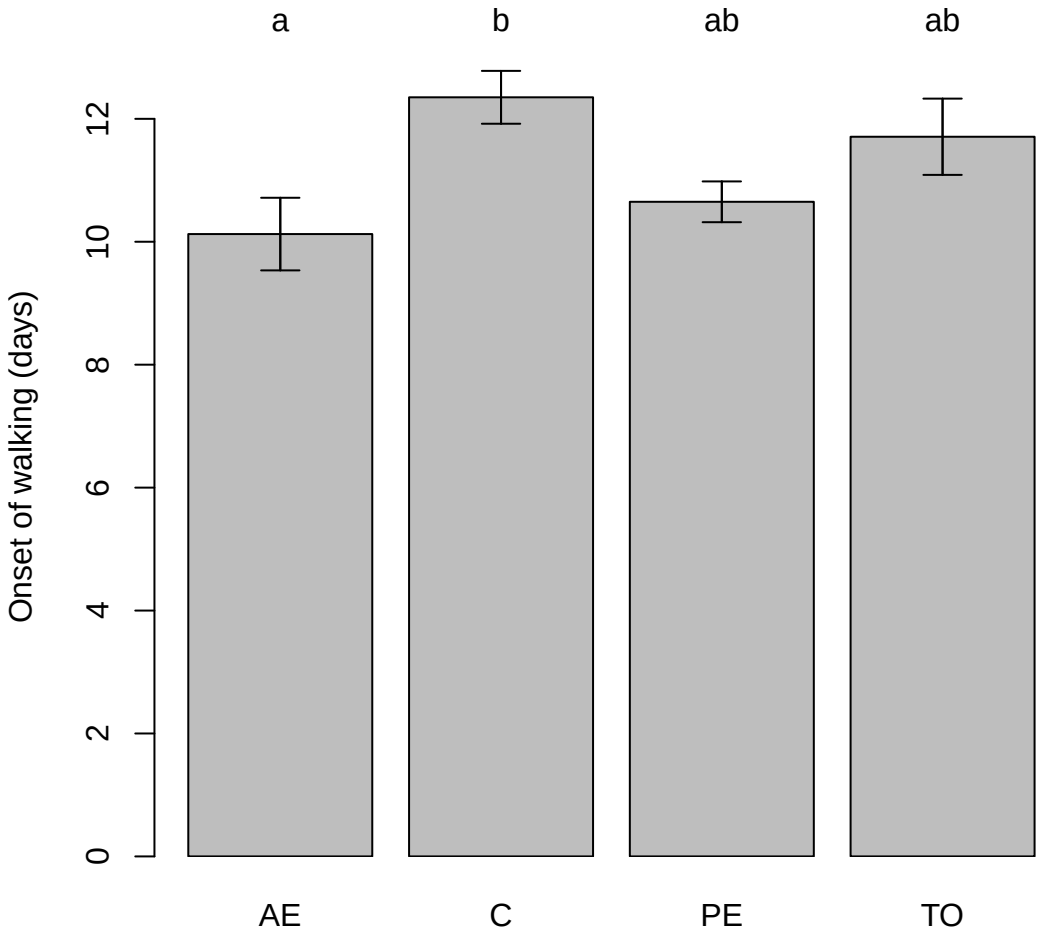


Figure 14.3. The population means of factor levels with the same letter are not significantly different at alpha = 0.05 using the Tukey HSD method. Bars are means. Errors are SEs.

```
plot(tukey, type = 2, ylab = "Onset of walking (days)")
```

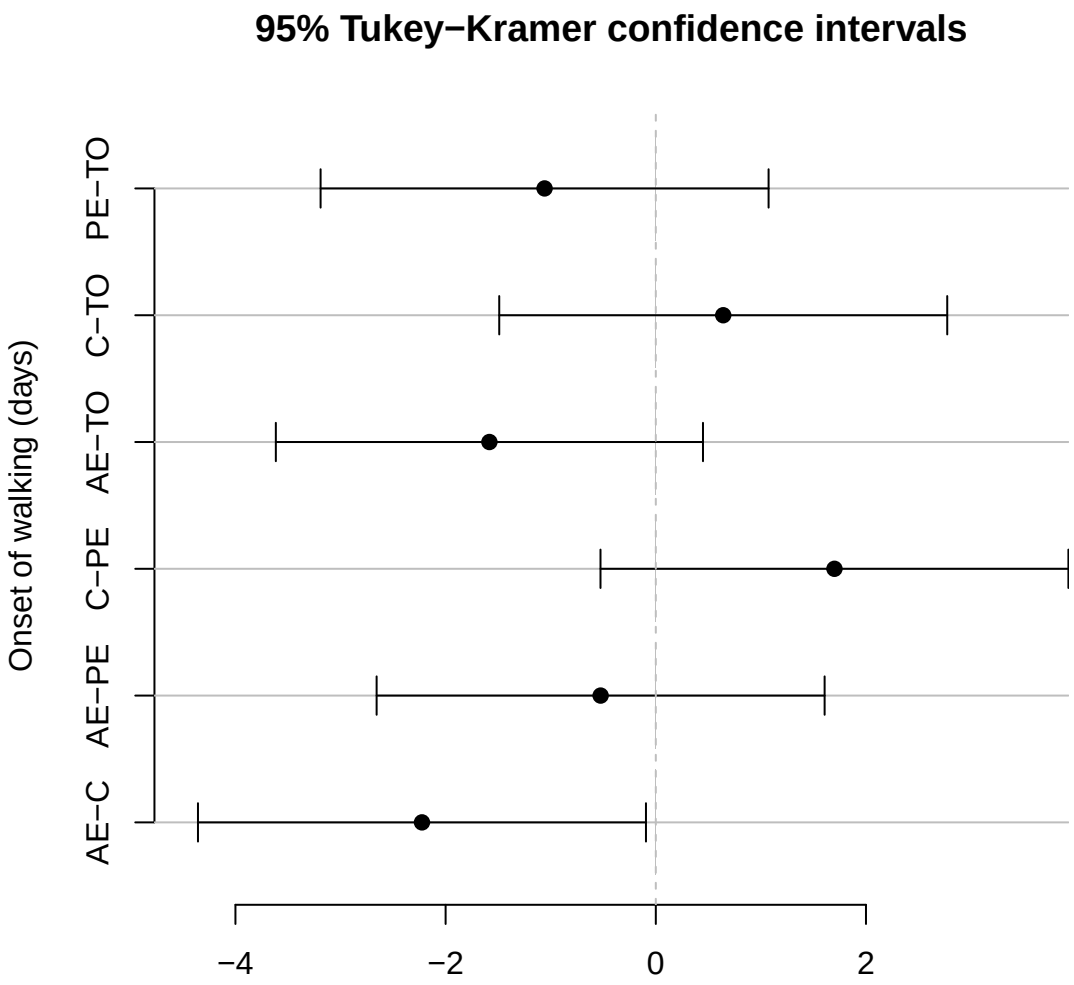


Figure 14.4. Tukey HSD 95% confidence intervals for the true difference, D.

■

Assignment 14

Answer all questions in one MS Word document and submit to Moodle. At the beginning of the document include the assignment number, the date, your name and section number.

Use complete sentences when appropriate, and make sure any tables, figures and computer output you include adhere to class standards (see Syllabus).

1. (2 pts) Define Family-wise Error Rate (FWER)

2. (2 pts) Define data snooping

3. (2 pts) Define D

4. (2 pts) Define $\hat{D}$

5. (9 pts) In Lab 13 we examined horticultural data concerning the potassium content of tree leaves from three different varieties of apple trees (1, 2, and 3). The data are in the Moodle data directory as `K.csv`. Re-test that varieties of apple trees differ in phosphorous. Use $\alpha = 0.05$.

a) What are your ANOVA hypotheses?

b) Run the ANOVA in **R**, attach the results.

c) Can we proceed with *post hoc* tests? Why?

d) How many pairwise comparisons among factor levels are possible?

6. (15 pts) For the analysis in Q. 5 run all possible pairwise tests with Fisher's LSD method "by hand" using **R** to help. Use $\alpha = 0.05$.

a) What does the null hypothesis $H_0 : \mu_i = \mu_{i'}$ mean?

b) What is calculated value of MSE from the ANOVA?

c) What is the calculated value of LSD (see Eq. 14.4)? You will only need to calculate LSD once because the design is balanced.

d) To summarize your analysis, fill out a facsimile of the Table below in your homework and discuss the results.

Factor levels under comparison	$\lvert\hat{D}\rvert = \lvert\bar{Y}_i - \bar{Y}_{i'}\rvert$	$LSD_{i,i'}$	Decision for $H_0 : \mu_i = \mu_{i'}$ Reject H_0 if $\lvert\hat{D}\rvert \geqslant LSD$ FTR H_0 if $\lvert\hat{D}\rvert < LSD$

e) Verify your LSD comparisons using `pairw.anova` from *asbio*. Include snapshots.

f) Use a `pairw.anova` object to create a graphical summary of your LSD comparisons. Use `type = 1` when plotting. Attach the graph and briefly describe it.

7. (15 pts) Run all possible pairwise tests using Tukey's confidence intervals "by hand" using **R** to help. Use $\alpha = 0.05$.

a) Calculate T (see Eq. 14.7). You will need to use the function `qtukey` to find q.

b) Calculate $\hat{\sigma}_{\hat{D}}$. You will only have to calculate this once because the design is balanced.

c) To summarize your analysis, fill out a facsimile of the Table below in your homework and discuss the results.

Factor levels under comparison	$\hat{D} = \bar{Y}_i - \bar{Y}_{i'}$	Confidence intervals $\hat{D} \pm T \cdot \hat{\sigma}_{\hat{D}}$	Decision for $H_0 : \mu_i = \mu_{i'}$ Reject H_0 if 0 is not in interval FTR H_0 if 0 is in interval

d) Verify your Tukey HSD comparisons using `pairw.anova` from *asbio*. Include snapshots.

e) Use a `pairw.anova` object to create a graphical summary of your Tukey comparisons. Use `type = 2` when plotting. Attach the graph and briefly describe it.

8. (2 pts) Which method, Fisher's LSD or Tukey's HSD, appears to be more conservative with respect to FWER? Hint: confidence interval methods with narrower intervals for the same level of confidence are *less* conservative.

Q1 2pts, Q2 2pts, Q3 2pts, Q4 2pts, Q5 9pts, Q6 15pts, Q7 15pts, Q8 2pts. **Total pts: 49.**

Index of Terms

Index of R Operators and Functions

Bibliography

Aho, K.A. (2014) *Foundational and Applied Statistics for Biologists Using R*. CRC Press.

Boik, R.J. (1987) The Fisher-Pitman permutation test: A non-robust alternative to the normal theory F test when variances are heterogeneous. *British Journal of Mathematical and Statistical Psychology*, **40**, 26–42.

Bradford, M.M. (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical Biochemistry*, **72**, 248–254.

Crowley, P.H. (1992) Resampling methods for computation-intensive data analysis in ecology and evolution. *Annual Review of Ecology and Systematics*, **23**, 405–447.

Dobzhansky, T. (1950) Evolution in the tropics. *American Scientist*, **38**, 209–221.

Fisher, R.A. (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, 179–188.

Gibbs, H.L. & Grant, P.R. (1987) Adult survivorship in Darwin's ground finch (Geospiza) populations in a variable environment. *The Journal of Animal Ecology*, pp. 797–813.

Hansen, S.G., Ford, J.C., Lewis, M.S., Ventura, A.B., Hughes, C.M., Coyne-Johnson, L., Whizin, N., Oswald, K., Shoemaker, R., Swanson, T. *et al.* (2011) Profound early control of highly pathogenic SIV by an effector memory t-cell vaccine. *Nature*, **473**, 523–527.

Hubbell, S.P. & Johnson, L.K. (1978) Comparative foraging behavior of six stingless bee species exploiting a standardized resource. *Ecology*, **59**, 1123–1136.

Manly, B.F. (2006) *Randomization, Bootstrap and Monte Carlo Methods in Biology*, volume 70. CRC press.

Murakami, H., Ogawara, H., Morita, K., Saitoh, T., Matsushima, T., Tamura, J., Sawamura, M., Karasawa, M., Miyawaki, S., Schimano, S. *et al.* (1997) Serum beta-2-microglobulin in patients with multiple myeloma treated with alpha interferon. *Journal of Medicine*, **28**, 311–318.

Ott, R.L., Longnecker, M. & Ott, L. (2004) *A First Course in Statistical Methods*. Thomson-Brooks/Cole.

Pitman, E.J. (1949) Notes on non-parametric statistical inference. Technical report, North Carolina State University. Dept. of Statistics.

Quinn, G.P. & Keough, M.J. (2002) *Experimental Design and Data Analysis for Biologists*. Cambridge university press.

Rakison, D.H. (2009) Does women's greater fear of snakes and spiders originate in infancy? *Evolution and Human Behavior*, **30**, 438–444.

Rao, P.M., Rhea, J.T., Novelline, R.A., Mostafavi, A.A. & McCabe, C.J. (1998) Effect of computed tomography of the appendix on treatment of patients and use of hospital resources. *New England Journal of Medicine*, **338**, 141–146.

Society for the Prevention of Cruelty to Animals (SPCA) (2020) Pet overpopulation. https://spcala.com/pet-library/general-articles/pet-overpopulation/. [Online; accessed 30-Dec-2020].

Stanford Blood Center (2020) Blood types: What's your type? https://stanfordbloodcenter.org/donate-blood/blood-donation-facts/blood-types/. [Online; accessed 29-Dec-2020].

Tukey, J.W. (1949) Comparing individual means in the analysis of variance. *Biometrics*, pp. 99–114.

Xie, Y., Mueller, C., Yu, L. & Zhu, W. (2018) *animation: A Gallery of Animations in Statistics and Utilities to Create Animations*. R package version 2.6.

Zelazo, P.R., Zelazo, N.A. & Kolb, S. (1972) "Walking" in the newborn. *Science*, **176**, 314–315.

Temporary page!

LaTeX was unable to guess the total number of pages correctly. As there was some unprocessed data that should have been added to the final page this extra page has been added to receive it.

If you rerun the document (without altering it) this surplus page will go away, because LaTeX now knows how many pages to expect for this document.