Chapter 1: Drawing Statistical Conclusions

This chapter focuses on the concept of the *Scope of Inference* which depends on the study design. Statistical inference, including hypothesis testing and constructing a confidence interval, requires knowing the probability distribution of the relevant statistic (i.e., the **sampling distribution**). The sampling distribution of a statistic depends on mechanisms under the control of the researcher: **random assignment** of treatments in an experiment and/or **random sampling** from a **population** of interest.

Scientific Method

Statistical procedures are part of the **Scientific Method** (steps 2-5 below) first espoused by Sir Francis Bacon (1561-1626), who wrote *to learn the secrets of nature involves collecting data and carrying out experiments.* The modern methodology:

1. Observe some phenomenon
2. State a hypothesis explaining the phenomenon
3. Collect data
4. Analyze the data
5. Test: do the data support the hypothesis?
6. Conclusion. If the test fails, go back to step 2.

If you encounter a "scientific claim" that you disagree with, scrutinize the steps of the scientific method used. *Statistics don't lie, but liars do statistics.* - Mark Twain.

Hypothesis Testing

The Six Steps in Hypothesis Testing can be inserted into steps 5 and 6 of the Scientific method. Let's number the six steps in hypothesis testing to emphasize this relationship:

5. Test: Do the data support the hypothesis explaining the phenomenon?

    5.1 State H_0 and H_a with respect to the parameter of interest.

    5.2 Check the assumptions necessary so that the test is valid.

    5.3 Compute the test statistic.

    5.4 Compute the p-value.

    5.5 Make a decision about H_0.

6. Draw a conclusion.

1.2 Statistical Inference and Study Design

Individuals or **Cases** or **Units** or **Subjects**: The objects from which data are collected. Individuals may be people, places, animals, things, or time periods.

Variable: Any characteristic of an individual that can be measured.

Population: The entire group of individuals that we want information about. For example: all grizzly bears in Yellowstone National Park; all G.E. light bulbs (made now and in the future); all tosses with a weighted die

Sample: A part of the population from which data is collected. For example: 22 tagged grizzly bears in Yellowstone National Park; 1 box of G.E. light bulbs; 100 tosses with a weighted die.

Inference: One collects data from a sample and uses the sample results to draw conclusions about the population. Inference is necessary whenever it is unrealistic to perform a **census** (i.e., data from the entire population of interest). Your book states that an inference is a conclusion that "patterns in the data" are present in some "broader context." In other words, an inference is a conclusion that the sample represents the population.

Statistical inference is an inference justified by a probability model.

Parameter: A numerical value calculated from a population of individuals.

Statistic: A numerical value calculated from a sample of individuals.

Some common population parameters and the statistics calculated from samples that estimate them:

Statistics	Parameters	Description
$\bar{x}$	μ	mean
s	σ	standard deviation
s^2	σ^2	variance
$\hat{\pi}$	π	proportion

Sampling Distribution: The value of a statistic varies from sample-to-sample. In other words, different samples will result in different values of a statistic. Since the value of a statistic varies from sample-to sample, it is a variable! Therefore, it has a distribution! The distribution of a statistic is called a **Sampling Distribution.** We will construct this sampling distribution for the case studies in section 1.1.

How to Construct a Sampling Distribution (conceptually - this cannot be done in practice):

- Take all possible samples of size n from the population.
- Compute the value of the statistic for each sample.
- Display the sampling distribution of the statistic as a table, graph, or equation.

1.2.1 Study Design: Observation and Experimentation

Observational Study: A study which observes individuals and measures variables, but does not attempt to influence the responses.

- An observational study on individuals from a random sample allows one to generalize conclusions about the sample to the population.
- An observational study cannot show **cause-and-effect** relationships because there is the possibility that the response is affected by some variable(s) other than the ones being measured. That is, **confounding variables** may be present. *It ain't what you don't know that gets you into trouble. It's what you know for sure that just ain't so.* - Mark Twain
- In **prospective** observational studies, investigators choose a sample and collect new data generated from that sample. That is, the investigators "look forward in time."
- In **retrospective** observational studies, investigators "look backwards in time" and use data that have already been collected. Retrospective studies are often criticized for having more confounding and bias compared to prospective studies.

QUESTION: Prospective or Retrospective Observational Study?

1. A study that follows marijuana users in Colorado for 5 years.
2. A study of illegal immigrant activity last year in Arizona.

Experiment: A study in which treatment(s) are deliberately imposed on individuals in order to observe their response. This is called a **randomized trial** in clinical settings.

- An experiment in which the treatments are randomly assigned to individuals can provide evidence for a **cause-and-effect** relationship. Furthermore, if the individuals are from a random sample, then one can generalize conclusions from the experiment to the population.

To recognize the difference between an Observational Study and an Experiment, ask yourself, "Was there a treatment imposed on the individuals?" In an experiment, the researcher determines (randomly) which individuals receive which treatment. In an observational study, the individuals have already self-chosen their groups.

QUESTION: Observational Study or Experiment?

1. A study of the birth weight of babies and the mother's level of coffee consumption.
2. A study of lab mice whose spinal cords have been severed.
3. A study of gender versus salary.
4. A study of grizzly bear attacks.
5. A study of the number of 1's rolled on a weighted die.

Confounding Variable: A variable that is related to the response variable and to the explanatory variable in such a way that makes it impossible to distinguish the effects of the confounding variable on the response from the effects of the explanatory variable on the response.

EXAMPLES:

- In a study of gender differences in salary, it was found that female nurses (in a certain hospital) have higher salaries, on average, than do male nurses. It also was found that female nurses have a greater number of years of experience than do male nurses. **Years of experience** is a confounding variable. It may be that the data give no clue as to whether the salary difference is due to gender discrimination or due to years of experience.
- In a study investigating the association between the occurrence of low birth weight babies and the mother's level of coffee consumption, it was found that an increase in the mother's coffee consumption is associated with an increase in the risk of having a low birth weight baby. It also was found that moms who smoke also consume large amounts of coffee and moms who do not smoke consume no or small amounts of coffee. **Smoking** is a confounding variable. Are the low birth weights due to the smoking or the coffee? CAN'T TELL!

IMPORTANT POINTS to remember:

- Time spent thinking and planning PRIOR to data collection is time well spent. You should know what statistical analysis you are going to use BEFORE collecting your data. A well thought out design can make statistical analysis and its interpretation easier.
- A statistical analysis cannot salvage a poorly designed study. Consulting statisticians encounter this problem all too frequently:

  To call in the statistician after the study is done may be no more than asking her to perform a post-mortem examination: she may be able to say what the study died of. – R. A. Fisher.

1.2.2 Sampling

Sampling Plans: methods of selecting individuals from a population. We are interested in sampling plans such that results from the sample can be used to make conclusions about the population.

Biased Samples: Bias occurs when the sample tends to differ from the population in a systematic way. When this happens, results from the sample can not be used to make conclusions about the population of interest.

1. **Convenience Sample** - An "easily available" sample of individuals which was convenient for the researcher to collect. Individuals in the convenience sample may systematically differ from the population and therefore may not represent the entire population.
2. **Voluntary Response Sample** - A sample of individuals who volunteer or are *self selected* to participate. Individuals who volunteer may systematically differ from the population and therefore may not represent the entire population.

EXAMPLES:

- Phone surveys exclude (1) households without a phone, (2)individuals who do not pick up their phone, and (3) individuals who refuse to participate and hang up
- Call-in polls on TV exclude (1) individuals without a TV, (2) individuals not watching the program, and (3) individuals who do not care to participate

Random Sampling: A sample of individuals who have been chosen randomly from the population. Random samples tend to represent the population from which they are chosen since randomization does not systematically favor some individuals in the population over others.

Since random samples are representative of the population of interest, then inference is valid. In other words, results from a random sample can be generalized to make conclusions about the population.

Simple Random Sample (SRS) - Each possible sample of size n has an equal chance of being selected from the population.

How to Select a SRS:

- Put slips of paper in a hat, mix well, then choose n slips.
- Use a computer:
  1. Create a **sampling frame**, a numbered list of all individuals in the population.
  2. Use a **random number generator** to select individuals from the list.

IMPORTANT POINTS to remember:

- A biased sample is a biased sample, regardless of its size! Collecting more data in a biased fashion will not correct the problem.
- A biased sample still contains information about a population, but this population is not the one that a researcher is interested in! Information can still be gleaned from biased samples, but one must be wary of the interpretation.

  EXAMPLE:
  - Drug trials using human volunteers
  - Studies on animals which have been specifically bred for experiments

1.2.3 Determining the Scope of Inference

This semester, you will be writing a *Scope of Inference* statements in every statistical report.

To determine the appropriate scope of inference of a study, there are TWO questions to answer.

1. Were the individuals randomly assigned to groups? In other words, was the study a randomized experiment or an observational study?
2. Were the individuals units randomly selected from some larger population? In other words, are the data from a random sample?

After answering the TWO questions, use this table to help you write the Scope of Inference.

	Randomized Experiment	Observational Study
Random Sampling	Cause-and-Effect in population of interest	Association in population of interest
Non-random Sampling	Cause-and-Effect in restricted population	Association in restricted population

Table 1: Scope of inference depending on the sampling plan and study design. Compare with Display 1.5 in your text.

- Write the *Scope of Inference* **in the context of the specific problem** rather than in vague statistical generalities.
- It may be challenging to identify the "**restricted population**" to which inference can be made. Many times, the restricted population is merely the sample itself.
- Cause-and-effect relationships may still be drawn from **randomized experiments without random sampling**. There is just no justification that the relationship extends to a larger population.
- An **observational study with no random sampling** is especially worrisome, since inferring to a larger population is often the goal. However, it is not always possible (or practical) to obtain a random sample from the larger population of interest. It is common to *pretend* that the non-random sample is representative of the population, but this should be accompanied by strong justification and even then the potential for bias cannot be ruled out. The researcher must decide whether statistical inference based on assumed (but technically incorrect) models is better than no statistical inference at all. The assumptions and justification should be clearly reported with the work.
- When designing an experiment/study, the **ideal scenario** is to:

  1. Randomly select units from population of interest
  2. Randomly assign units to comparison (study) groups

  Unfortunately, in many fields of science, these types of studies are rare and sometimes impossible to conduct.

Chance Models and Statistical Inference

- Statistical analysis relates the available data to some broader context through uncertainty using chance (or probability) models.
- The probability models are associated with the chance mechanisms used to *select units from a population* (random sampling) OR to *assign units to groups* (randomized experiment).

- The probability models allow the researcher to calculate statistical measures of uncertainty to accompany inferential conclusions. This is the fundamental goal of statistics. If you do not want a measure of uncertainty (i.e., an error bar) you do not need to do statistics.

1.3 Measuring Uncertainty in Randomized Experiments

IDEA: If there is no treatment effect then the labels indicating group membership are meaningless and could have been randomly assigned. A method of evaluating a treatment is to **visualize hypothetical replications of a study** under different random group assignments.

Case study: 1.1.1 and 1.3.1 Motivation and Creativity - A Randomized experiment

Do grading systems promote creativity in students? Do ranking systems and incentive awards increase productivity among employees? Do rewards and praise stimulate children to learn? See p. 2 of your text for the experiment reported in Amabile, T., 1985, Motivation and Creativity: Effects of Motivational Orientation on Creative Writers, *Journal of Personality and Social Psychology*, 48(2), 393-399.

A common model of this study is write the response Y^* of a **writer in the intrinsic** group as

$$Y^* = \mu^* + \epsilon$$

where μ^* is the true mean score for the population of all intrinsic writers and ϵ is the error. The response Y of a **writer in the extrinsic group** is

$$Y = \mu + \varepsilon$$

where μ is the true mean score for the population of all extrinsic writers and ε is the error.

Your book considers **an additive treatment effect model**: Let Y denote the score of an individual in the extrinsically motivated group. This same subject would receive a score of

$$Y^* = Y + \delta$$

if she had been in the intrinsically motivated group. The treatment effect δ is the unknown parameter of interest,

$$\delta = \mu^* - \mu.$$

Note that this model assumes the same treatment effect for each individual. This is a key *assumption* of the model.

SCIENTIFIC QUESTION OF INTEREST: **Is there an effect of the instrinsic treatment over the extrinsic treatment in the population of all experienced writers?**

5.1 Null and Alternative Hypotheses: Translate question of interest hypotheses about the parameter δ in the additive treatment effects model:

5.2 Check assumptions: Because random assignment was used in this study, we will proceed with a randomization test. In other words, the probability model that we will assume for the test statistic is a randomization distribution. Random assignment also helps ensure that the overall response of the group is due to the intrinsic or extrinsic treatment and not some confounding variable such as gender or writing experience or age.

5.3 A test statistic measures the plausibility of H_a relative to H_0.

- How big (or small) is the test statistic as compared to what might have happened under a different randomization if there were no effect of the treatment?
- For the additive treatment effects model: A suitable test statistic is the difference in the sample averages of the scores.

  Let $\overline{Y}_1$ denote the average of the extrinsically motivated group and let $\overline{Y}_2$ denote the average of the intrinsically motivated group. If there is no treatment effect then the difference $\overline{Y}_2 - \overline{Y}_1$ should be close to 0. We observed

  $$\overline{Y}_2 - \overline{Y}_1 = 4.14$$

  Is this "close to" or "far from" 0?

5.4 To calculate the p-value, we could perform a t-test. Instead, we consider the **Randomization distribution** of the test statistic: all test statistic values for every possible outcome of randomly assigning the units to intrinsic and extrinsic treatment groups.

The p-value in the context of a randomized experiment is the probability that the randomization alone leads to a test statistic as extreme or more extreme than the observed one.

Ways to obtain a p-value:

1. EXACT: Enumerate all possible regroupings of the data (create the randomization distribution) and find the proportion of these that produce a test statistic as extreme or more extreme than the observed one.
2. The total number of possibilities gets unmanageably large as the sample size increases. Hence usually we APPROXIMATE the randomization distribution by simulating a large number of randomizations and finding the proportion of these that produce a test statistic as extreme or more extreme than the observed one.
3. MOST COMMON: Approximate the randomization distribution with a mathematical curve based on assumptions about the distribution of the response and the form of the test statistic. Eg. If we assume that the data are normally distributed and that the additive model is correct, then the normal curve provides a good approximation of the randomization distribution for the difference between averages ($\hat{\delta}$) (Chapter 2).

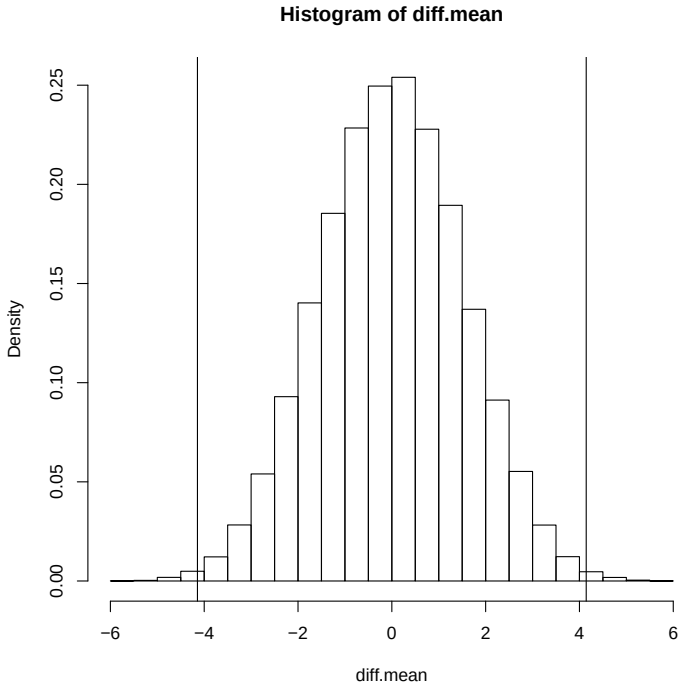
Here is the R code that approximates the randomization distribution of $\bar{Y}_2 - \bar{Y}_1$ in order to find the p-value for testing the hypotheses in 5.1 for the creativity study:

```
require(Sleuth3)
poetry.dat<-case0101
names(poetry.dat)
[1] "Score"     "Treatment"

# How many possible randomizations are there?
choose(47,23)
[1] 1.612380e+13 # a reasonably large number

tapply(Score,Treatment,mean)
Extrinsic Intrinsic
 15.73913  19.88333
# observed difference is 19.8833 - 17.73913 = 4.144203

# Draw 100000 randomizations
diff.mean<-numeric(100000) # storage vector

# generate 100000 random assignments and calculate difference in means
for(i in 1:100000)
    {
    grp<-sample(Treatment,47,replace=F)
    diff.mean[i]<- mean(Score[grp=="Intrinsic"])- mean(Score[grp=="Extrinsic"])
    }

# Graph the approximate randomization distribution
hist(diff.mean,prob=T) # a density histogram
abline(v=4.144203) # puts a vertical line at the observed difference
abline(v=-4.144203)

# Get the two-sided p-value
sum(abs(diff.mean)>=4.144203)/100000  # two-sided p-value
[1] 0.0055
```

In addition to calculating a p-value, we can also get a confidence interval (CI) from this randomization distribution. We cannot use it directly because it is centered on the null value of 0. But if we add the observed difference of 4.14 to each of the values the histogram is moved 4.14 units to the right and is now centered on the observed value. An approximate 95% confidence interval can be easily generated by picking off the $2.5th$ and 97.5 quantiles.

```
quantile(diff.mean + 4.14, p=c(0.025,0.975))
   2.5%    97.5%
1.166087 7.100689
```

5.5 Make a decision about H_0. At a significance level of $\alpha = 0.05$, because the p-value $= 0.0055 < \alpha$, then we reject H_0 in favor of H_a. Equivalently, because the 95% CI for δ does not contain the value $\delta = 0$ (under H_0), then we reject H_0 in favor of H_a.



6. Conclusion. In the reports that you will submit this semester, you will split the conclusion out into two sections:

   Summary of Statistical Findings: This experiment provides evidence that receiving the "intrinsic" rather than the "extrinsic" questionnaire caused students **in this study** to score higher on poem creativity **on the average** (p-value = 0.005). The **estimated increase** in score attributed to the "intrinsic" questionnaire is 4.1 points (95% CI: 1.3 to 7.0 points) on a 0-40 point scale.

   Scope of Inference: Subjects were not randomly selected from the population of all creative writers, although the subjects were randomized to the two treatment groups. Hence the data suggest that the mean difference in creativity scores in these 47 subjects was **caused** by the difference in the intrinsic and extrinsic motivational questionnaires. **Extending this conclusion to any larger population** is problematic due to the non-random sampling.

1.4 Measuring Uncertainty in Observational Studies

- In an observational study, there is no chance mechanism for group assignment. The investigator has no control over such assignment.
- But the investigator does have control over who is selected for study. We now use a chance mechanism for **sample selection** (via random sampling). This again allows us to formulate a probability model by connecting the distribution of the test statistic to the chance mechanism.

- IDEA: **Visualize hypothetical replications of a study** under different randomly selected samples.

1.4.1 A simulation model for random sampling

- Use a chance mechanism to randomly select units from one or more populations.
- This approach requires that we know a lot about the population(s) including how to sample.
- For comparing two populations:

  1. Select a sample of n_1 units from population 1 (with mean μ_1 and standard deviation σ_1) such that all subsets of size n_1 have the same chance of selection.
  2. Select a sample of n_2 units in the same manner from population 2 (with mean μ_2 and standard deviation σ_2) *independent* of those from population 1. (the sample drawn from population 1 should not influence how the sample from population 2 is drawn.)
  3. Calculate a statistic from the sample from population 1 and also from the sample from population 2 (e.g., $\bar{Y}_1$ and $\bar{Y}_2$ if we want to estimate μ_1 and μ_2).
  4. Repeat steps 1-3 many times.
  5. Construct an approximation to the sampling distribution of the statistic that compares the two populations (e.g., $\bar{Y}_2 - \bar{Y}_1$ if we want to estimate $\mu_2 - \mu_1$). Confidence intervals and p-values are based on this aproximate sampling distribution.

  EXAMPLE: We will generate a set of 30 random observations from a normal distribution with a mean of 0 and a standard deviation of 2 and another set of 30 observations from a normal distribution with a mean of 3 and a standard deviation of 5.

```
diff.mean<-numeric(100000) # storage vector

for(i in 1:100000)
{
    # rnorm is the R function that samples from a normal distribution
    set1 = rnorm(30, mean=1, sd=2)  # RS1 from Population 1
    set2 = rnorm(30, mean=1.5, sd=5)  # RS2 from Population 2

    mean1 = mean(set1) # Statistic from RS1
    mean2 = mean(set2) # Statistic from RS2
    diff.mean[i]=mean2 - mean1
}

# Measure of center of sampling distribution
mean(diff.mean)

## [1] 0.5000842
```

```
# Measure of spread of sampling distribution
sd(diff.mean)
```

```
## [1] 0.9842521
```

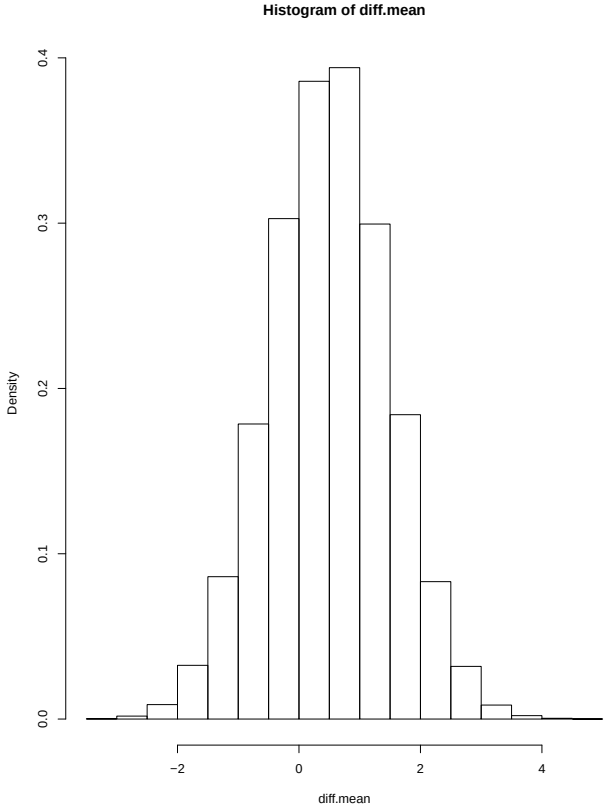
```
# Theoretical results:
# True mean of sampling distribution
1.5 - 1
```

```
## [1] 0.5
```

```
# True SD of sampling distribution
sqrt((2/sqrt(30))^2 + (5/sqrt(30))^2)
```

```
## [1] 0.9831921
```

```
# Graph the approximate sampling distribution
hist(diff.mean,prob=T) # a density histogram
```



1.4.2 Non-random sampling - A permutation distribution

Case study: 1.1.2 and 1.4.2 Sex Discrimination in Employment - an Observational Study

Did a bank discriminantly pay higher starting salaries to men than to women? See p. 4 of your text: $\bar{Y}_{\text{men}} - \bar{Y}_{\text{women}} = \$818$.

- Is it an experiment or observational study?

- Were people selected via random sampling?

- Based on our answers where does a chance mechanism come into the picture?

- Inference is based on a *pretend* probability model. We can *pretend* that the employer assigned the set of starting salaries to employees at random. We then compare the observed salary difference between males and females to what we would expect if the salaries were handed out at random.
  - Hypothesis tests are conducted (i.e., p-values are calculated) and confidence intervals are found just as we did in section 1.3 for randomized experiments. Because we are only pretending that this is an appropriate model, the sampling distribution of the test statistic in the observational study scenario is called a **Permutation distribution** (not a randomization distribution).
  - Draw a picture of the permutation distribution for the sex discrimination study.

  - How is a permutation distribution different from a randomization distribution?

  - What is the proper scope of inference for the sex discrimination case study?

1.5.1 Graphical Methods

Below is an example that shows you how to produce the following graphs in R:

- Histograms
- Stem-and-Leaf Diagrams
- Box plots
- Scatter plots

EXAMPLE: We will generate a set of 30 random observations from a normal distribution with a mean of 0 and a standard deviation of 2 and another set of 30 observations from a normal distribution with a mean of 3 and a standard deviation of 2.

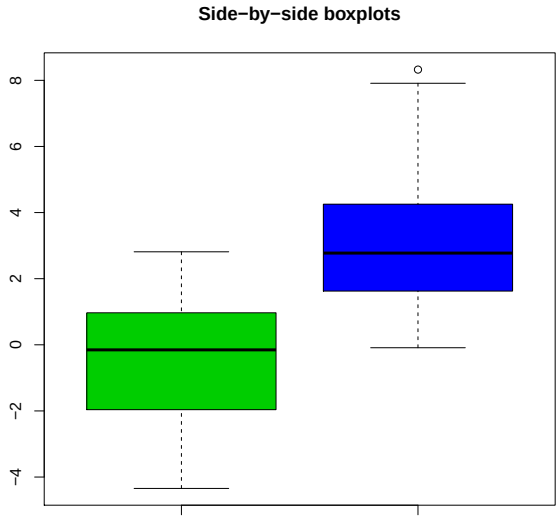
```
# rnorm is the R function that samples from a normal distribution
set1 = rnorm(30, mean=0, sd=2)
set2 = rnorm(30, mean=3, sd=2)
```

1. Stem-and-leaf plot of the first set of 30 observations.

   ```
   stem(set1)
   ##
   ##  The decimal point is at the |
   ##
   ##  -4 | 3
   ##  -3 | 91
   ##  -2 | 66520
   ##  -1 | 653
   ##  -0 | 75521111
   ##   0 | 668
   ##   1 | 03367
   ##   2 | 148
   ```

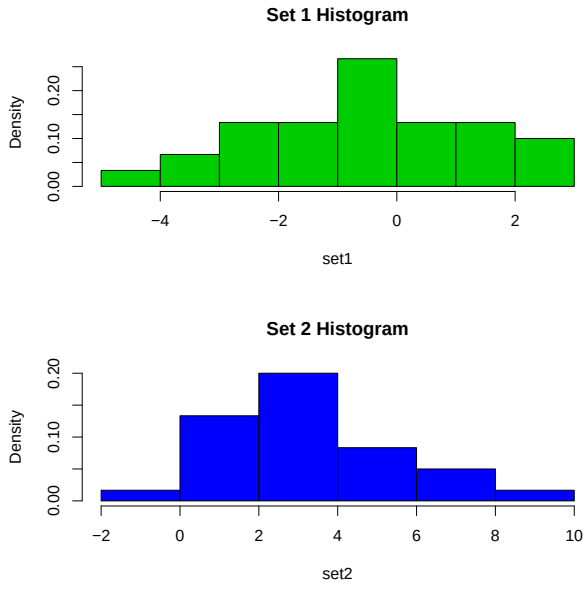
2. Make side-by-side boxplots of the two sets of observations. Label the parts of the boxplot by hand.

   ```
   boxplot(set1,set2, col=c(3,4), main="Side-by-side boxplots")
   # side-by-side boxplots
   # col determines colors
   # main specifies a main title or label for the plot
   # if you want to see all the options and get details
   # of how boxplot works type ?boxplot
   ```



3. Make histograms of your two sets of data.

```
par(mfrow=c(2,1))
# splits plot area into 2 rows and 1 column, i.e 2
# plots in the same panel.
hist(set1, col=3, nclass=10, prob=T, main="Set 1 Histogram")
hist(set2, col=4, nclass=5, prob=T, main="Set 2 Histogram")
# nclass specifies the number of class intervals
# prob=T (TRUE) tells R to create a density
# histogram, i.e. the total area under the histogram is equal to 1.
```



Look at the distribution of the two data sets you simulated carefully. They may not look normally distributed even though the samples came from a normal distribution.

4. Put a histogram and a box plot together on the same plot for the first set of data.

```
dev.new() #  - opens a new plot window instead of writing over the previous one
par(mfrow=c(2,1))
hist(set1, col=3, nclass=6, prob=T, main="Set 1 Histogram")
boxplot(set1, col=3, horizontal = TRUE)
# horizontal=TRUE tells R to create a horizontal boxplot
```

5. Make a scatter plot with the first set of data on the x-axis and the second set on the y-axis.

```
#Scatterplot
dev.new()
plot(set1, set2,pch=18,col=2,main="Scatter plot", xlab="Set 1 data",
     ylab="Set 2 data")
# xlab and ylab specify x and y axis labels.
# pch is for print character and tells R to use little diamonds
# the default print character is an open circle
```