Chapter 1: Overview and Descriptive Statistics

Curtis Miller

2018-05-13

Introduction

This chapter is devoted to basic statistical ideas and statistical summaries. We start with describing what statistics is, does, and what it uses. Next we see graphical and tabular methods for describing distributions. The last two sections discuss measures of location and measures of spread, respectively.

Section 1: Populations, Samples, and Processes

Data is a collection of facts. A **population** is a group of interest. If we collect data for the entire population, we have conducted a **census**. Usually, though, we collect data for a subset of a population, called a **sample**. Our objective is to use the data in the sample to reach conclusions about the population as a whole.

In a sample we have **observations**, individual data points that consist of **variables**, or quantities/characteristics of interest. **Univariate** data records the value of only one variable for each observation. **Multivariate** data records the value of multiple variables for each observation. **Bivariate** data is a special case of multivariate data; there are two variables quantified.

Categorical variables take values from a finite number of possibilities. **Quantitative** variables, however, take numerical values.[1]

Modern statistics depends heavily on probability theory. **Probability** is the field of mathematics that describes the behavior of objects in the presence of uncertainty (which we refer to as randomness). The diagram below illustrates the relationship between probability and statistics with relation to samples and populations.

[1] This may be the simplest dichotomy of types of data. Stevens [1946] classifies data into **nominal**, **ordinal**, **interval**, and **ratio** types, the first two breaking up the "categorical" data type and the second two breaking up the "quantitative" data type. The data types allow for different operations to be defined for different data; ordinal data allows for order relations, interval for addition and subtraction, and ratio allows for division and multiplication.

How we define a population depends heavily on our problem. In **enumerative studies**, the population is a fixed, finite, tangible group that presently exists. In **analytic studies** the population may not presently exist.

Statistics depends crucially on how data is collected in survey-style, observational studies. If data is collected poorly, the results of analysis cannot be trusted.

Below are two approaches for collecting data *correctly*:

- In a **simple random sample (SRS)**, each member of the population of interest is eligible to be randomly selected to be included in the sample. The usual analogy is that each individual in the population is written on a piece of paper and put in a hat; then, slips of paper are randomly chosen in the hat and those individuals are chosen to be in the sample.[2] The statistical methods seen in this class are appropriate for simple random samples *only*.
- In **stratified sampling**, the population is divided into observable **strata**. A SRS is then selected from individuals in each strata.[3]

[2] For example, a candidate for public office may use the registered voter list to randomly select voters in the area the candidate will represent and ask them who they plan to vote for in the upcoming election.

[3] For example, in a national election, an equal number of voters are selected from each state to participate in a poll.

Convenience sampling selects individuals in a way that is not completely random (in the sense that not all individuals from the population are equally likely to be selected, and the procedure is not intentionally stratified). The results of convenience samples cannot be trusted. Statistical descriptions of error account only for error due to randomness, not due to bad sampling procedures.

Section 2: Pictorial and Tabular Methods in Descriptive Statistics

A **distribution** describes what values a variable takes and how frequently it takes them. This section describes techniques for visualizing distributions of univariate data. Visualization is an important first step in a statistical project, as it reveals patterns that are difficult to describe using numbers only, and could suggest what statistical procedures are appropriate.

In statistics, n usually denotes the **sample size**, or the number of observations in the dataset. To denote the values of the dataset's variable, we often use the notation $x_1, x_2, ..., x_n$, where x_i is the ith observation of the dataset. Unless otherwise stated this notation says nothing about the dataset's values. That is, the data is *not* assumed to be ordered.

Stem-and-Leaf Plot

The first visualization of data is a **stem-and-leaf plot**. This plot is constructed using the following steps:

1. Select the number of leading digits to be the **stem** values. The remaining digits are the **leaf** values.
2. Draw a vertical line and list the stem values to the left of this line, in order.
3. Record the leaf of each observation in the row corresponding to its stem value. (Computers often order the leaf values, but when done by hand this is not necessary.)
4. Somewhere in the display, indicate the units of the stem and leaves. (For example, the stems start at the tens place, and the leaves start at the ones place.)

Example 1

The following is a subset of Macdonell's data on height and finger length of criminals imprisoned in England and Wales [Macdonell, 1902]. Here I report only the (rounded) heights of the subset.[4]

```
height <- c(5.55, 5.30, 5.63, 5.30, 5.13,
            5.05, 5.38, 5.96, 5.21, 5.38)
```

Use this dataset to construct a stem-and-leaf plot.

[4] Throughout this course I will be including R code that answers the questions I ask. This is so you can see how to do these techniques in R. *You are not expected to understand any of the code at the start of the course!* I *do not* attempt to simplify the code to account for what you have learned so far in the lab. The more you see R code, though, the more familiar and less scary it will become, and I invite you to revisit these lectures at the end of the course and see how much you can understand. Additionally, I hope some of my code will stimulate your curiosity, including the more complicated code.

```r
stem(height, scale = 2)
```

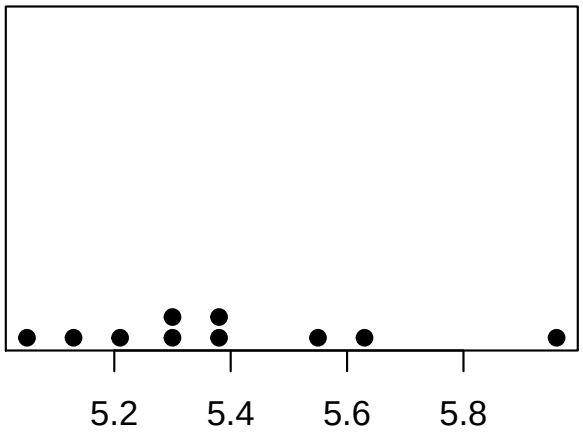
```
##
##   The decimal point is 1 digit(s) to the left of the |
##
##   50 | 5
##   51 | 3
##   52 | 1
##   53 | 0088
##   54 |
##   55 | 5
##   56 | 3
##   57 |
##   58 |
##   59 | 6
```

A **dotplot** represents each data point as a dot along a real number line, putting the point on the line according to its value. If two points would be almost overlapping, they would instead be stacked.

Example 2

Using the data in Example 1, create a dotplot.

```r
stripchart(height, method = "stack", pch = 19, offset = 0.5, at = 0)
```



A quantitative variable is **discrete** if its possible values are countable. It is **continuous** if possible values consist of entire intervals of the real number line (which could be the whole line, in principle).[5]

The **frequency** the value of a variable occurs is the number of times that value was seen in a dataset. For discrete variables it's reasonable to list the frequency of each observed value, but for continuous variables this is not reasonable. Instead, for continuous variables, we list the frequency of a **bin**, which is a range in which a datapoint could be. We would then count how many data points fell within that range.

The **relative frequency** is the frequency a value occured divided by the number of data points. (This is defined analogously for continuous variables.) That is:

[5] As a rule of thumb, discrete variables arise from counting, while continuous variables arise from measurements.

A **frequency distribution** is a tabulation of frequencies or relative frequencies.

Example 3

A statistically minded parent tracks the number of points scored by his daughter's little league soccer team during regular season. Below is the dataset.

```
soccer <- c(9, 6, 5, 5, 5, 6, 2, 8, 3, 4, 8, 1)
```

Construct a frequency distribution for this dataset.

```
table(soccer)
```

```
## soccer
## 1 2 3 4 5 6 8 9
## 1 1 1 1 3 2 2 1
```

When working with continuous data we need to construct bins when creating a frequency distribution, and list the frequency each bin occurs. How do we do this?

1. Decide on the number of bins. There are rules of thumb for doing this, such as choosing approximately $\sqrt{n}$ bins.[6]
2. Divide the segment of the number line where your data lies into that many equal-length bins.[7]
3. Depending on where each datapoint falls, assign it to a bin. If it falls on a border between bins, assign it to the bin on the right. (In other words, bins are right-inclusive.)
4. Construct a frequency distribution for the bins.

[6] Actually, $n^{1/5}$ may work better.

[7] Some people consider bins of unequal length. When constructing a histogram, do not do this. It makes the histogram more difficult to read correctly.

Example 4

Using the data in Example 1, construct a frequency distribution.

```
length(height)  # The sample size
```

```
## [1] 10
```

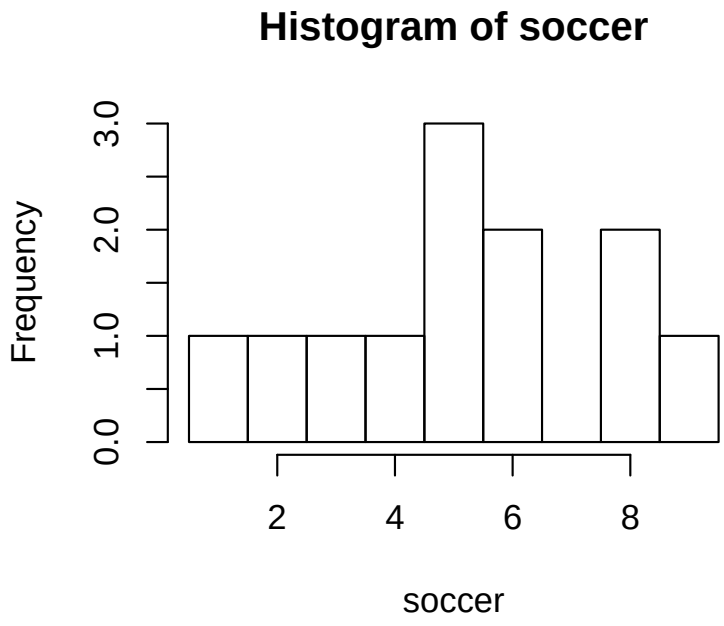
Once we have a frequency distribution, we can construct a **histogram**, a plot for visualizing the distribution of quantitative data. Do the following:

1. Draw a number line and mark the location of the bins. For discrete data, center the bins on the corresponding value.
2. For each class, draw a bar extending from the number line to either the frequency or relative frequency of the number/bin. Do this for each bin.

Example 5

Draw a histogram for the dataset in Example 3 (the soccer dataset).

```r
hist(soccer, breaks = min(soccer):max(soccer + 1) - 0.5)
```



Example 6

For the dataset in Example 1 (the height dataset), create a histogram.

```
hist(height)
```

When looking at plots visualizing distributions we are looking for certain qualities. We want to decide:

- Is the data **unimodal** (only one "peak")? Is it **bimodal** or **multimodal** (multiple "peaks")? Below are illustrations.

- Is the data **positively-skewed**? **Negatively skewed**? **Symmetric**? Below are illustrations.

- Are there **outliers**, points that are distant from the rest of the data?
- How spread out is the data?

A **bar plot** is a method for visualizing categorical (sometimes referred to as **qualitative**) data. To construct a bar plot:

1. List each possible value of the variable and how frequently each value is taken.
2. Draw a horizontal line and along that axis mark each possible value of the variable. The vertical axis will correspond to different possible frequencies.
3. Draw a bar for each category extending to the category's observed frequency.

Example 7

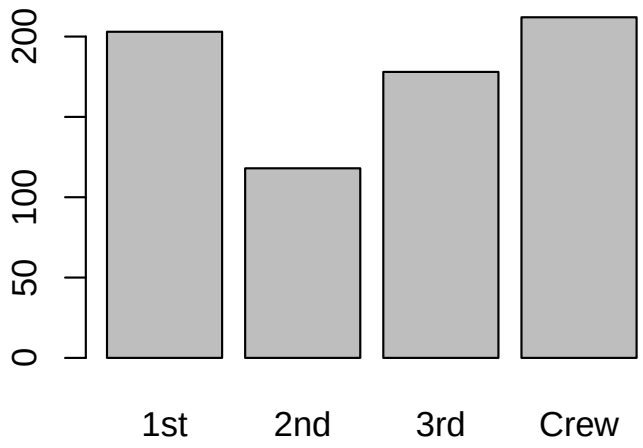
Below is a dataset showing the frequency of the class of passengers aboard the *Titanic* who survived her sinking.

```
(t_survive_class <- apply(Titanic[, , , 2], 1, sum))
```

```
##  1st  2nd  3rd Crew
##  203  118  178  212
```

Create a bar plot for the frequency of each class's survival.

```
barplot(t_survive_class)
```



Section 3: Measures of Location

While visual summaries of data are nice, quantitative summaries are still important for describing datasets. We start with **measures of location**, which tell us where a dataset is located along the number line.

The first and most common measure of location for a sample is the **sample mean**[8], defined for a dataset $x_1, ..., x_n$ below:

[8] There is a physical interpretation of the mean; if you were to construct a dot plot of the data and made that plot a physical object, with a weight for each dot and the number line a teeter-totter, the mean would be the point where the teeter-totter balances.

The **sample proportion** for categorical data is defined below:

Example 8

What is the average number of points your daughter's soccer team scores? (Here's the dataset, as a reminder.)

```
soccer
```

```
##  [1] 9 6 5 5 5 6 2 8 3 4 8 1
```

```
mean(soccer)
```

```
## [1] 5.166667
```

Let's suppose that $r_1, r_2, ..., r_n$ is the *ordered* dataset corresponding to the dataset $x_1, ..., x_n$, so that $r_1 \leq r_2 \leq ... \leq r_n$. The **sample median**[9] is the number that splits this dataset in half. It is defined below:

[9] The physical/geometric interpretation of the median is obvious; when you arrange the data in order, it splits the data in half.

Example 9

Find the median of the first eleven games of your daughter's soccer team. (I have ordered the dataset for you below.)

```
sort(soccer[1:11])
```

```
##  [1] 2 3 4 5 5 5 6 6 8 8 9
```

```
median(soccer[1:11])
```

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

Let $\alpha \in [0,1]$. The $\alpha \times 100$**th percentile** is the number such that roughly $\alpha \times 100\%$ of the data in $r_1, ..., r_n$ lies to the left of that number. Perhaps the most common percentiles are the **quartiles**. The **first quartile** is the 25th percentile, and the **third quartile** is the 75th percentile. The **second quartile** is the median (the 50th percentile).[10]

Here is a procedure for finding quartiles:[11]

1. Find the median of the data $r_1, ..., r_n$.
2. Split the dataset into two datasets at the median. If n is odd, remove the datapoint corresponding to the median.[12]
3. The median of the lower dataset is the first quartile, and the median of the upper dataset is the third quartile.

---

[10] The 0th and 4th quartile are the minimum and maximum of the dataset. All quartiles together form the **five-number summary** of a dataset.

[11] Actually, this is a procedure for finding what your textbook refers to as **fourths**. The difference is negligible so I use the terms interchangeably.

[12] Not everyone does this, so software might give a different answer when computing medians. The difference is usually negligible.

Example 10

Find first and third quartiles for your daughter's first eleven soccer games.

Example 11

Find the 10th and 90th percentiles of the height data. (I have listed the data for you below, in order.)

```
sort(height)
```

```
##  [1] 5.05 5.13 5.21 5.30 5.30 5.38 5.38 5.55
##  [9] 5.63 5.96
```

```
quantile(soccer[1:11], c(.25, .75))
```

```
## 25% 75%
## 4.5 7.0
```

```
quantile(height, c(.1, .9))
```

```
##   10%   90%
## 5.122 5.663
```

The sample mean $\bar{x}$ is **sensitive** to outliers; that is, outliers in the dataset can have a profound effect on the sample mean. On the other hand, the sample median $\tilde{x}$ is **insensitive** to outliers, since outliers almost never alter the value of the sample median.

Example 12

Compute both the sample mean and the sample median when the value of your daughter's 12th soccer game is one of the following:

```
(outlier_game <- c(soccer[12], soccer[12] + 3, soccer[12] * 2, max(soccer) * 2))
```

```
## [1]  1  4  2 18
```

```r
# This loop will compute each of the requested values. I will display the result
# in a table, which is formed when the loop runs.
soccer_tab <- sapply(outlier_game, function(g) {
  dat <- c(soccer[1:11], g)
  return(c(g, median(dat), mean(dat)))
})
soccer_tab <- t(soccer_tab)  # Transpose matrix (I don't want this shape)
# Row/column naming
rownames(soccer_tab) <- 1:nrow(soccer_tab)
colnames(soccer_tab) <- c("Outlier Value", "Median", "Mean")
round(soccer_tab, digits = 2)
```

```
##   Outlier Value Median Mean
## 1             1    5.0 5.17
## 2             4    5.0 5.42
## 3             2    5.0 5.25
## 4            18    5.5 6.58
```

There is in fact a relationship between the mean and median depending on whether the data is negatively-skewed, positively-skewed, or symmetric, illustrated below:

The median is preferred for skewed data while the mean is preferred for symmetric data. (It is better behaved and has great analytic results.)

So far I've discussed only sample means and medians but *population* means, medians, and percentiles are also defined. They have similar properties to their sample analogues.

A compromise between the mean's sensitivity to outliers and the median's ignorance of nearly all of the dataset is the **trimmed mean**, which I denote by $\bar{x}_{\text{tr}(100\alpha)}$. The trimmed mean is the mean of the data when $100\alpha\%$ of the is removed from each end of the dataset.[13]

[13] It may not be possible to remove $100\alpha\%$ of the data *exactly*. You can approximate it with interpolation.

Example 13

Find $\bar{x}_{\text{tr}(10)}$ for the height data.

```r
mean(height, trim = 0.1)
```

```
## [1] 5.36
```

Section 4: Measures of Variability

Consider the following three datasets:

1	2	3
4	2	1
5	5	3
6	6	6
7	7	9
8	10	11

Construct dot plots for each dataset, then compute the mean and median of each dataset.

Now suppose each dataset represented waiting time (in minutes) for the red line train to arrive to take you home. Which dataset would you prefer to see? Why?

The above example illustrates that measures of center are insufficient for describing a dataset. We also want a **measure of variability**, which describes how "spread out" a dataset is.

How can we measure spread? This should be based on **deviations**. The deviation of data point i is $x_i - \bar{x}$.

Compute $\sum_{i=1}^{n}(x_i - \bar{x})$.

This result suggests we should measure variability with something else. The most common measure for variability is the **sample variance** and the **sample standard deviation**, defined below:

The sample standard deviation can roughly be interpreted as the "typical" deviation of a datapoint from the mean.[14]

There are population analogues to both of these quantities: the **population variance**, σ^2, and the **population standard deviation**, $\sigma = \sqrt{\sigma^2}$.

Ideally you should use software or a calculator to compute the sample variance, but in a pinch you can use this handy formula:

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

[14] The sample mean is sensitive to outliers. The sample standard deviation is *more* sensitive to outliers than the sample mean.

Example 14

Compute the sample variance and sample standard deviation of the soccer game scores (listed below, as a reminder).

```
soccer
```

```
##  [1] 9 6 5 5 5 6 2 8 3 4 8 1
```

```
length(soccer)
```

```
## [1] 12
```

```
summary(soccer)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.
##   1.000   3.750   5.000   5.167   6.500
##    Max.
##   9.000
```

```r
var(soccer)  # Sample variance
```

```
## [1] 5.969697
```

```r
sd(soccer)  # Sample standard deviation
```

```
## [1] 2.443296
```

Proposition 1. *Let $x_1, ..., x_n$ be a sample and c a constant. Then:*

1. *If $y_i = x_i + c$ for all i, $s_y^2 = s_x^2$*
2. *If $y_i = cx_i$ for all i, $s_y^2 = c^2 s_x^2$ and $s_y = |c|s_x$*

The **fourth spread** (also known as the **inter-quartile range (IQR)**) is the third quartile minus the first quartile; denote this with f_s. This is another measure of dispersion.

Example 15

Compute the fourth spread for the soccer game scores.

---

f_s can be used for outlier detection. We may call an observation that is further than $1.5f_s$ from its nearest quartile a **mild outlier**, and an observation that is more than $3f_s$ away from the nearest quartile an **extreme outlier**.

Example 16

Use the fourth spread to detect outliers in soccer game scores. What is the minimum score needed for a data point to be a mild outlier? Extreme outlier?

A **boxplot** is a plot visualizing a dataset. A boxplot is created in the following way:[15]

1. Compute the minimum, maximum, median, first and third quartiles for the dataset.
2. On a number line, draw a box with one end at the first quartile and the other at the third quartile.
3. Within the box, draw a line at the median.
4. Extend a line from one end of the box to the minimum and a line from the other end to the maximum. (These are called **whiskers**.)

Boxplots give both a sense of location and a sense of spread. They're especially useful when placed side-by-side; they then are called **comparative boxplots**.

[15] Often software will not extend the whiskers of box plots to the extrema of samples, instead ending at the largest value that is *not* an outlier. The outliers are then denoted with dots. R, for example, does this by default. While this is more informative it's more difficult to do by hand. The instructions provided here are good enough when not using software.

Example 17

The following dataset contains the tooth growth for guinea pigs given vitamin C via orange juice at three different dosage levels.

```
suppressPackageStartupMessages(library(dplyr))  # Provides %>% operator
```

```
## Warning: package 'dplyr' was built under R
## version 3.4.3
```

```
OJ <- ToothGrowth %>% filter(supp == "OJ") %>% select(len, dose) %>% unstack %>%
  lapply(sort) %>% as.data.frame
```

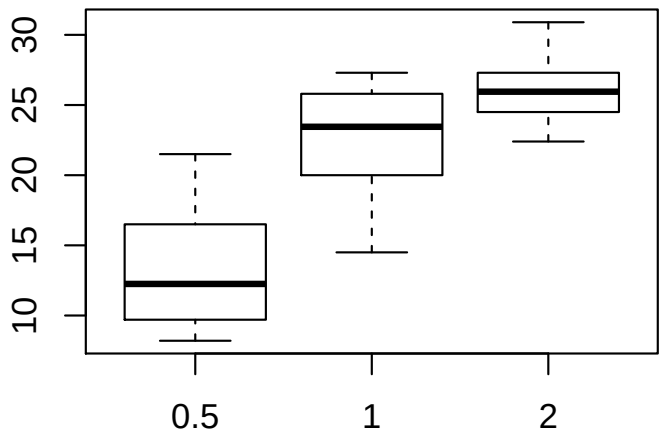
```
## Warning: package 'bindrcpp' was built under R
## version 3.4.3
```

```
names(OJ) <- c(0.5, 1, 2)
OJ
```

```
##     0.5    1    2
## 1   8.2 14.5 22.4
## 2   9.4 19.7 23.0
## 3   9.7 20.0 24.5
## 4   9.7 21.2 24.8
## 5  10.0 23.3 25.5
## 6  14.5 23.6 26.4
## 7  15.2 25.2 26.4
## 8  16.5 25.8 27.3
## 9  17.6 26.4 29.4
## 10 21.5 27.3 30.9
```

Construct a comparative box plot for the lengths. Compare.

```r
boxplot(len ~ dose, data = ToothGrowth %>% filter(supp == "OJ"))
```



References

W. R. Macdonell. On criminal anthropometry and the identification of criminals. *Biometrika*, 1(2):177–227, 1902. ISSN 00063444. URL http://www.jstor.org/stable/2331487.

S. S. Stevens. On the Theory of Scales of Measurement. *Science*, 103: 677–680, June 1946. DOI: 10.1126/science.103.2684.677.