Chapter 1: Introduction

Juhyung Lee

Department of Statistics
University of Florida

Population and Sample

Statistics consists of methods for **designing** studies, **describing** data obtained for those studies, and making **inferences** based on those data to answer a statistical question of interest.

Definition (Subjects)

The entities that we measure in a study are called the **subjects** (e.g., people, schools, counties, etc.).

Definition (Population)

The **population** is the set of all the subjects of interest (e.g., all students taking STA 3032 this semester).

Definition (Sample)

A **sample** is the subset of the population for whom we have data, often randomly selected (e.g., students in my section of STA 3032).

The ultimate goal of most studies is to **learn about the population** using a sample (why sample, not census?).

Descriptive Statistics and Inferential Statistics

Description in Statistical Analyses

Descriptive statistics refers to methods for summarizing the collected data. The summaries usually consist of graphs and numbers such as averages and percentages.

Inference in Statistical Analyses

Inferential statistics refers to methods of making decisions or predictions about a population, based on data obtained from a sample of that population.

- Descriptive statistics are useful for both census and sample.
- Inferential statistics are used when data are available for a sample only, which is usually the case.
- In general, we *describe* the sample, and we make *inferences* about the population.

Descriptive Statistics and Inferential Statistics

Example (Polling Opinions on Handgun Control)

In a recent poll of 834 Florida residents, 54.0% of the sampled subjects said they favored controls over the sales of handguns. Here, the margin of error (later in the course) is 3.4%. This means we can predict with 95% confidence (later in the course) that the percentage of all adult Floridians favoring control over sales of handguns falls between 50.6% and 57.4%.

What are the subjects, population, and sample? What is the descriptive statistical analysis and what is the inferential statistical analysis?

Sample Statistics and Population Parameters

Definition (Parameter and Statistic)

A **parameter** is a numerical summary of the population. A **statistic** is a numerical summary of a sample taken from the population.

Example (Polling Opinions on Handgun Control)

- The (unknown) percentage of the population of all adult Florida residents favoring handgun control is a parameter.
- The percentage 54.0% of the sample favoring handgun control is a sample statistic (sample proportion specifically).

- We hope to learn about parameters so that we can better understand the population.
- The true parameter values are almost always unknown.
- We use sample statistics to estimate the parameter values.

Discrete and Continuous Data

Definition (Variable)

A **variable** is any characteristic observed in a study.

Definition (Categorical and Quantitative Variables)

A variable is called **categorical** if each observation belongs to one of a set of categories (e.g., gender, blood type, letter grade, etc.).
A variable is called **quantitative** if observations on it take numerical values that represent different magnitudes of the variable (e.g., number of siblings, number of typos in a book, weight, commuting time, etc.).

Definition (Discrete and Continuous Variables)

A quantitative variable is **discrete** if its possible values form a set of separate numbers, such as $0, 1, 2, 3, \ldots$ (e.g., number of siblings, number of typos in a book, etc.).
A quantitative variable is **continuous** if its possible values form an interval (e.g., weight, commuting time, etc.).

Measures of Location: Mean and Median

Measures of location provide some quantitative values of where the center, or some other location, of data is located.

Definition (Sample Mean)

Suppose that the observations in a sample are $x_1, x_2, \ldots, x_n$. The **sample mean**, denoted by $\overline{x}$, is

$$\overline{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

(i.e., a numerical average).

Definition (Sample Median)

Given that the observations in a sample are $x_1, x_2, \ldots, x_n$, arranged in **increasing order** of magnitude, the **sample median** is

$$\widetilde{x} = \begin{cases} x_{(n+1)/2}, & \text{if } n \text{ is odd,} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}), & \text{if } n \text{ is even,} \end{cases}$$

(i.e., the middle value).

Measures of Location: Mean and Median

Example (CO_2 Pollution)

Per capita CO_2 emissions (tons/person) for the nine largest countries in population size in 2011.

Nigeria	Bangladesh	Pakistan	India	Indonesia
0.3	0.4	0.8	1.4	1.8
Brazil	China	Russia	United States	
2.1	5.9	11.6	16.9	

$$\overline{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{0.3 + 0.4 + \cdots + 16.9}{9} = 4.6,$$

$$\widetilde{x} = x_{(9+1)/2} = x_5 = 1.8.$$

Measures of Location: Mean and Median

Example (CO_2 Pollution)

```
> x <- c(0.3, 0.4, 0.8, 1.4, 1.8, 2.1, 5.9, 11.6, 16.9)
>
> mean(x) # sample mean
[1] 4.577778
>
> median(x) # sample median
[1] 1.8
```

Measures of Location: Mean and Median

- The mean uses the numerical values of all the observations (informative, but sensitive to ouliers).
- The median uses only the ordering (less informative, but resistant to outliers).
- In practice, it is a good idea to report both the mean and the median.

Example (CO_2 Pollution)

Change 16.9 to 90 for United States:

$$0.3, 0.4, 0.8, 1.4, 1.8, 2.1, 5.9, 11.6, \mathbf{90}.$$

Now, $\overline{x}^* = 12.7 > 4.6 = \overline{x}$, but $\widetilde{x}^* = \widetilde{x} = 1.8$.

Example (Airplane Crashes)

One variable in a study measures how many airplane crashes a commercial airline company has had in the past year. Why would the mean likely be more useful than the median for summarizing the responses of 60 airline companies?

Measures of Location: Percentiles, Quantiles, and Quartiles

Definition (Percentile and Quantile)

The ***p*th percentile** is a value such that p percent of the observations fall below or at that value. The pth percentile is also called the **$p/100$ quantile**.

Definition (Quartiles)

- The **first quartile Q_1** is the 0.25 quantile (i.e., the 25th percentile).
- The **second quartile Q_2** is the 0.5 quantile (i.e., the 50th percentile).
- The **third quartile Q_3** is the 0.75 quantile (i.e., the 75th percentile).

- The quartiles split the data into four parts, each containing one quarter (25%) of the observations.
- Q_2 is the median.
- Q_1 is the median of the lower half of the observations (excluding the median itself if n is odd).
- Q_3 is the median of the upper half of the observations (excluding the median itself if n is odd).

Measures of Location: Percentiles, Quantiles, and Quartiles

Example (CO_2 Pollution)

Nigeria	Bangladesh	Pakistan	India	Indonesia
0.3	0.4	0.8	1.4	1.8
Brazil	China	Russia	United States	
2.1	5.9	11.6	16.9	

$$
\begin{aligned}
Q_2 &= \text{median}(0.3, 0.4, 0.8, 1.4, 1.8, 2.1, 5.9, 11.6, 16.9) = 1.8, \\
Q_1 &= \text{median}(0.3, 0.4, 0.8, 1.4) = \frac{0.4 + 0.8}{2} = 0.6, \\
Q_3 &= \text{median}(2.1, 5.9, 11.6, 16.9) = \frac{5.9 + 11.6}{2} = 8.75.
\end{aligned}
$$

```
> x <- c(0.3, 0.4, 0.8, 1.4, 1.8, 2.1, 5.9, 11.6, 16.9)
>
> quantile(x, c(0.25, 0.5, 0.75), type = 6) # quartiles
25%  50%  75%
0.60 1.80 8.75
```

Measures of Variability: Range, Variance, and Standard Deviation

Definition (Sample Range)

The **sample range** is the difference between the largest and the smallest observations.

Definition (Sample Variance and Standard Deviation)

The **sample variance**, denoted by s^2, is given by

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-1}.$$

The **sample standard deviation**, denoted by s, is the positive square root of s^2, that is,

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

Measures of Variability: Interquartile Range (IQR)

Definition (Interquartile Range)

The **interquartile range** is the distance between the third and first quartiles,

$$\mathrm{IQR} = Q_3 - Q_1.$$

Measures of Variability: Interquartile Range (IQR)

Example (CO_2 Pollution)

```
> x <- c(0.3, 0.4, 0.8, 1.4, 1.8, 2.1, 5.9, 11.6, 16.9)
>
> max(x) - min(x) # range
[1] 16.6
>
> var(x) # sample variance
[1] 34.60944
> sd(x) # sample standard deviation
[1] 5.882979
> sqrt(var(x)) # same as above
[1] 5.882979
>
> quantile(x, c(0.25, 0.75), type = 6) # Q1 and Q3
25%  75%
0.60 8.75
> IQR(x, type = 6) # IQR
[1] 8.15
```

Measures of Variability

- The range uses only the largest and smallest observations (sensitive to outliers, not used much).
- The standard deviation s uses all observations (sensitive to outliers).
- The IQR is not affected by outliers.
- The IQR is preferred over s when there are severe outliers.

Example (CO_2 Pollution)

Change 16.9 to 90 for United States:

$$0.3, 0.4, 0.8, 1.4, 1.8, 2.1, 5.9, 11.6, \mathbf{90}.$$

Now,

$$\begin{aligned} \text{range}^* &= 89.7 > 16.6 = \text{range}, \\ s^* &= 29.2 > 5.9 = s, \\ \text{IQR}^* &= 8.15 = \text{IQR}. \end{aligned}$$

Histograms

Histogram

A histogram is a graph that uses bars to portray the frequencies or the relative frequencies of the possible outcomes for a quantitative variable.

Example (Car Battery Life)

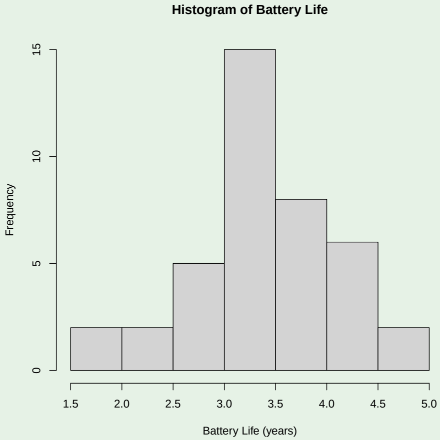
Life of 40 similar car batteries recorded to the nearest tenth of a year (see Table 1.4 of WMMY).

Interval	Frequency	Relative Frequency
$(1.5, 2.0]$	2	0.050
$(2.0, 2.5]$	2	0.050
$(2.5, 3.0]$	5	0.125
$(3.0, 3.5]$	15	0.375
$(3.5, 4.0]$	8	0.200
$(4.0, 4.5]$	6	0.150
$(4.5, 5.0]$	2	0.050

Histograms

Example (Car Battery Life)

```
> hist(x, main = "Histogram of Battery Life",
+      xlab = "Battery Life (years)")
```



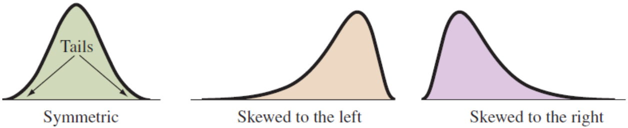
The Shape of a Distribution

Symmetric Distribution

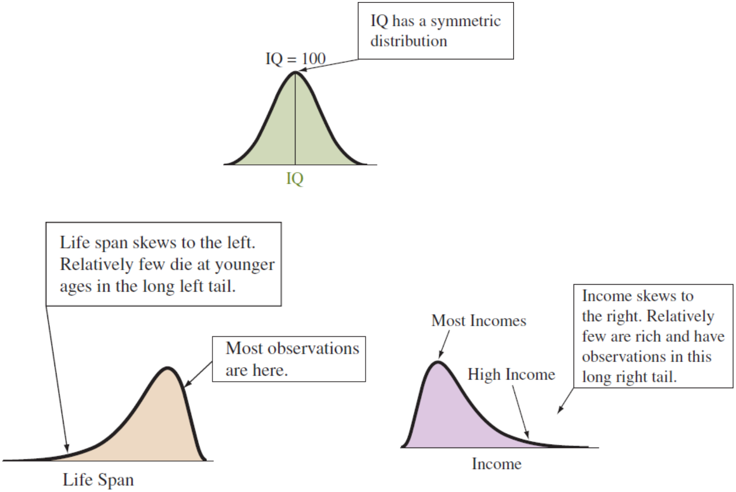
A distribution is **symmetric** if the side of the distribution below a central value is a mirror image of the side above that central value.

Skewed Distribution

A distribution is **skewed to the left** if the left tail is longer than the right tail.
A distribution is **skewed to the right** if the right tail is longer than the left tail.



The Shape of a Distribution



Detecting Potential Outliers

The $1.5 \times \text{IQR}$ Criterion for Identifying Potential Outliers

An observation is a potential outlier if it falls more than $1.5 \times \text{IQR}$ below Q_1 or more than $1.5 \times \text{IQR}$ above Q_3.

Example (Car Battery Life)

For the car battery life data,

$$Q_1 = 3.1, Q_3 = 3.85, \text{IQR} = 0.75 \text{ (check!)},$$

and

$$Q_1 - 1.5 \times \text{IQR} = 3.1 - 1.5 \times 0.75 = 1.975,$$
$$Q_3 + 1.5 \times \text{IQR} = 3.8 + 1.5 \times 0.75 = 4.975.$$

By the $1.5 \times \text{IQR}$ criterion, observations below 1.975 or above 4.975 are potential outliers. The observations 1.6 and 1.9 are the only potential outliers (check!).

The Five-Number Summary

The Five-Number Summary

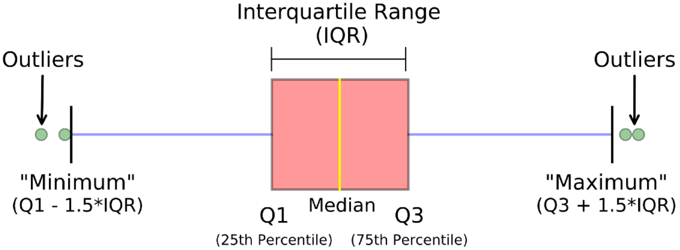
The five-number summary of a dataset is the minimum value, Q_1, Q_2 (i.e., median), Q_3, and the maximum value.

The five-number summary is the basis of a graphical display called the boxplot.

The Boxplot

Constructing a Boxplot

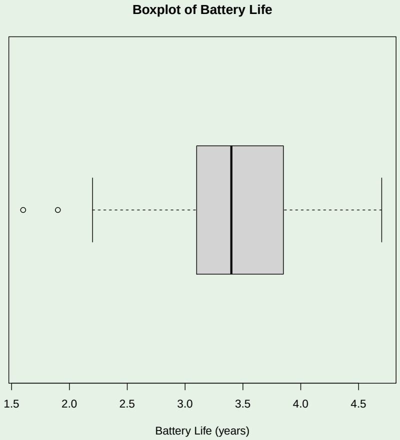
- A box goes from Q_1 to Q_3.
- A line is drawn inside the box at the median.
- A line goes from the lower end of the box to the smallest observation that is not a potential outlier.
- A separate line goes from the upper end of the box to the largest observation that is not a potential outlier.
- The potential outliers are shown separately.



The Boxplot

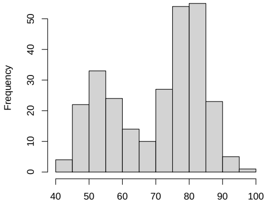
Example (Car Battery Life)

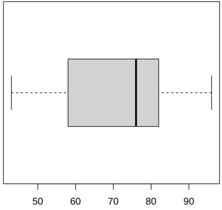
```
> boxplot(x, horizontal = TRUE, main = "Boxplot of Battery Life",
+         xlab = "Battery Life (years)")
```



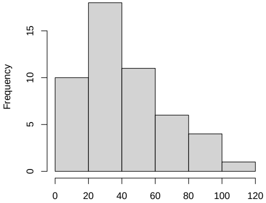
The Boxplot Compared with the Histogram

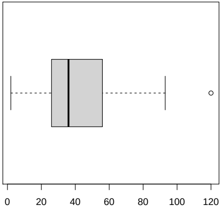
- A boxplot does not portray certain features of a distribution, such as distinct mounds and possible gaps, as clearly as does a histogram.





- A boxplot does indicate skew and is useful for identifying potential outliers.

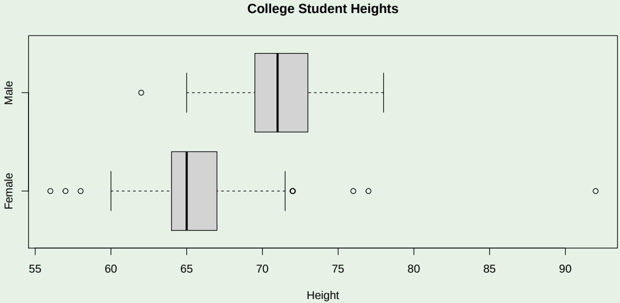




Side-by-Side Boxplots Help to Compare Groups

- Boxplots are also useful for graphical comparisons of distributions.

Example (College Student Heights)



- Both distributions are approximately symmetric.
- Although the centers differ, the variability of the middle 50% of the distribution is similar.
- The upper 75% of the male heights are higher than the lower 75% of female heights.

Response Variables and Explanatory Variables

Definition (Response Variable and Explanatory Variable)

The **response variable** is the outcome variable on which comparisons are made (e.g., survival status, GPA, etc.).
When the **explanatory variable is categorical**, it defines the groups to be compared with respect to values for the response variable (e.g., smoking status).
When the **explanatory variable is quantitative**, it defines the change in different numerical values to be compared with respect to values for the response variable (e.g., number of hours a week spent studying).

Scatterplot

Scatterplot

A **scatterplot** is a graphical display for two quantitative variables using the horizontal (x) axis for the explanatory variable x and the vertical (y) axis for the response variable y. The values of x and y for a subject are represented by a point relative to the two axes. The observations for the n subjects are n points on the scatterplot.

Example (Positive Association and Negative Association)