1 Overview and Descriptive Statistics

Chapter Preview

In this chapter, we'll introduce some of the basic concepts and definitions that are fundamental to the study of Probability and Statistics. In particular, sampling and sample statistics will be examined.

1.1: Populations, Samples, and Processes

Population vs. Sample

Def: A population is a well-defined complete collection of objects.

Def: A sample is a subset of the population.

Example: Population: all eligible voters in the U.S. / Sample:

Example: Population: all textbooks on Statistics / Sample:

Example: Population: all circuit boards / Sample:

Illustration of population and sample:

Methods of Collecting a Sample

There are many ways to collect a sample, many of which lead to biased results. An easy way to avoid bias is to do simple random sampling (SRS). In SRS, all members of the populations are equally likely to be chosen.

Example:

In addition to simple random sampling, there are more advanced techniques for selecting a sample. In stratified sampling, the population is broken down into sub-populations that are considered homogeneous. A sample is then taken from each of these sub-populations.

Example:

Care must be taken when selecting a sample. If a sample is either not randomly chosen or if the sample is not representative of the full population, our results may be biased.

Example:

While there are many other more complex techniques to sample, we'll generally assume that any sample collected in this class is a simple random sample.

Data can be either collected in an observational study or in an experiment.

In an observational study, data is collected in a way that doesn't directly interfere with how the data arise.

Example:

In a experiment, participants are separated into separate treatment groups and the differences between the groups are compared. A well-designed experiment can investigate the possibility of causation. Typically, experiments have a control, that is, participants receive an experimental unit that is constant or unchanged.

Example:

In addition to the sampling method, bias may still exist if participants in an experiment are aware of their treatment.

Example: 20 patients are being treated for depression, half are given an antidepressant and half are given a sugar pill (placebo). Suppose the patients know which pill they have been given. How might this cause the results to be biased?

Def: A blinded experiment is an experiment in which participants do not know which treatment they are receiving.

Def: A double-blinded experiment is an experiment in which the participants and the reseachers do not know which treatment is being administered during the experiment.

Types of Data

We will break data into two categories: **quantitative** and **qualitative**.

I. Quantitative data is numeric data or numbers and can broken into two further categories, **discrete** and **continuous**.

a) Def: Discrete data is quantitative data with a finite or countably infinite number of values.

Example:

b) Def: Continuous data is quantitative data with an uncountably infinite number of values or data taken from an interval.

Example:

II. Qualitative data refers to names, categories, or descriptions and can also be broken down into two further categories, **nominal** and **ordinal**.

a) Def: Nominal data is qualitative data with no natural ordering.

Example:

b) Def: Ordinal data is qualitative data with a natural ordering.

Example:

1.2: Pictorial and Tabular Methods in Descriptive Statistics

Notation: The size of our sample is typically denoted n.

Summarizing Qualitative Data

While we will mostly focus on quantitative data in this class, there are a few simple methods for describing qualitative data that you should be aware of.

For the following examples, let's collect some data on eye color from members of the class.

Data:

Bar Graph:

Proportions: $p = \frac{\#\text{ in category}}{\#\text{ total}}$

What proportion are brown?

Percentages: $P\% = p \cdot 100$

What percentage are blue?

Pictorial Methods for Quantitative Data

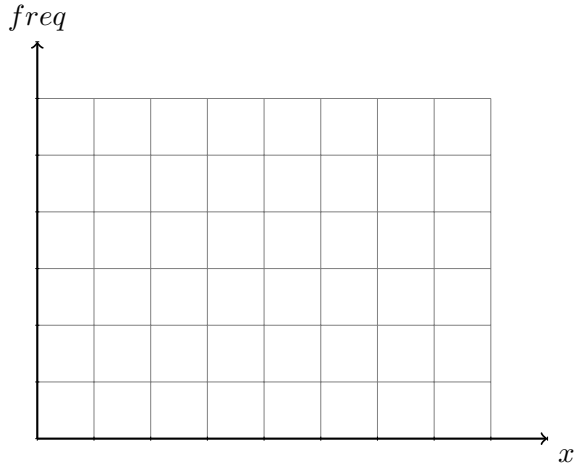
Histograms: One graphical method for visualizing quantitative data.

Example: 14 students were randomly sampled and were asked how many credits they are taking this semester.

Data: 12, 14, 15, 12, 9, 18, 22, 15, 16, 14, 14, 12, 3, 17

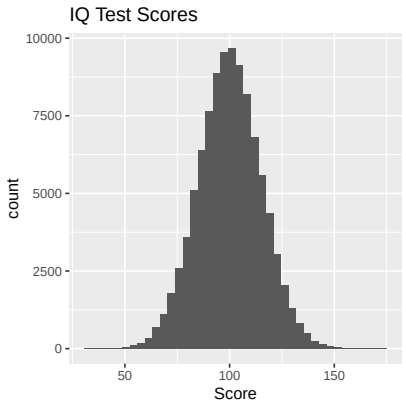
Class	Frequency	Relative Frequency	Percentage Frequency
$0 \leq x < 3$			
$3 \leq x < 6$			
$6 \leq x < 9$			
$9 \leq x < 12$			
$12 \leq x < 15$			
$15 \leq x < 18$			
$18 \leq x < 21$			
$21 \leq x < 24$			

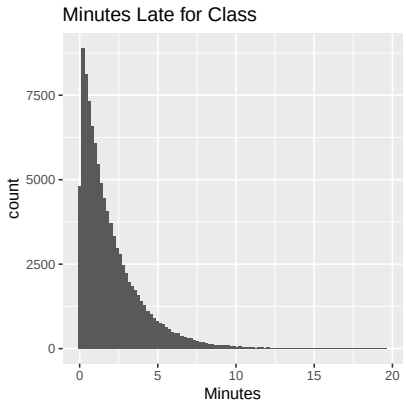
Histogram:



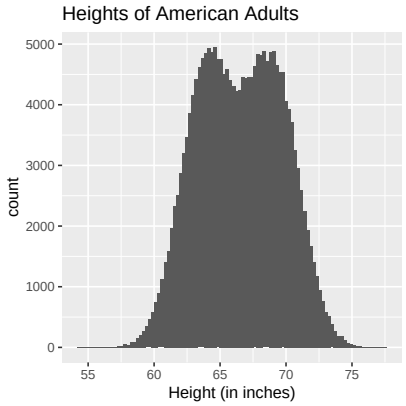
Histogram Shapes

1. Unimodal

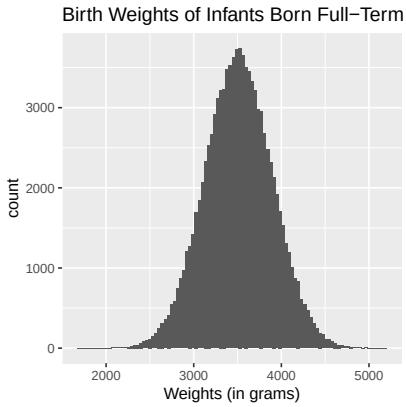




2. Bimodal

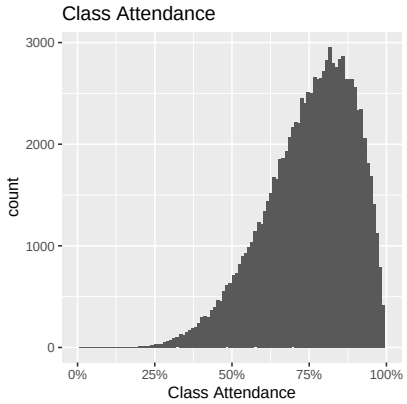


3. Bell-shaped

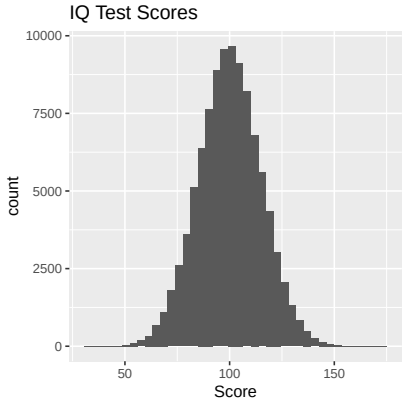


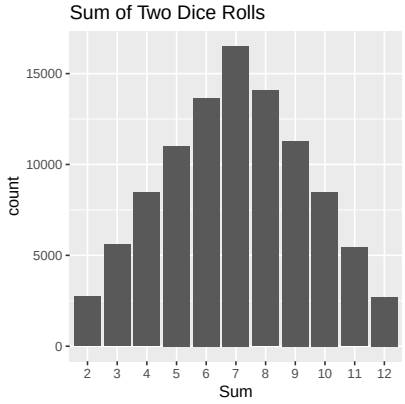
4. Positively (right) skewed - tail is on the right

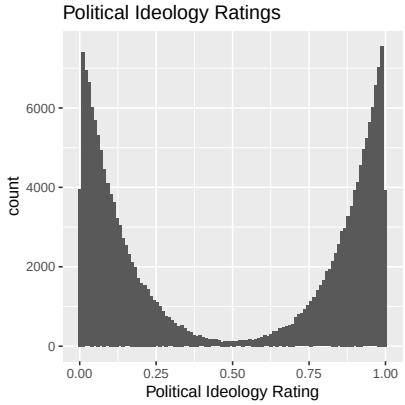
5. Negatively (left) skewed - tail is on the left



6. Symmetric







1.3: Measures of Location

For the following examples, we will use the following data set. The heights of five students are measured to be (in cm): 183, 165, 165, 175, 187.

Notation: $\sum_{i=1}^{n} x_i = x_1 + x_2 + ... + x_n$

Example: What is $\sum x_i$, the sum of all heights in the sample?

There are many ways to measure the "center" of our data set. We will look at mean, median, and mode.

Mean

Def: The sample mean (or sample average) $\bar{x}$ is given by: $\boxed{\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}}$.

Example: What is the sample mean of the original heights data set, $\bar{x}$?

Median

Def: The sample median $\tilde{x}$ is obtained by first ordering the n observations from smallest to largest and then calculated by the following:

$$\tilde{x} = \begin{cases} \text{the single middle value if } n \text{ is odd,} & \tilde{x} = x_{(n+1)/2} \\ \text{the average of the two middle values if } n \text{ is even,} & \tilde{x} = (x_{n/2} + x_{n/2+1})/2 \end{cases}$$

where x_i is the i^{th} smallest value in the data set.

Example: What is the sample median of the heights data set, $\tilde{x}$?

Example: If we measure an additional person at 174cm, what is the new median of the data set?

Mode

Def: The sample mode is the most commonly occurring data point.

Example: What is the sample mode of the heights data set?

1.4: Measures of Variability

There are a few different ways to describe the variability of a data set, including range, variance, standard deviation, and interquartile range.

Range

Def: The range of a data set is the difference between the largest and the smallest values.

Example: What is the range of the heights data set?

Variance and Standard Deviation

Typically, the most common method for measuring the spread of a data set is variance or standard deviation.

Def: The sample variance, denoted by s^2, is given by: $\boxed{s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}}$

Def: The sample standard deviation, denoted by s, is given by: $\boxed{s = \sqrt{s^2} = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}}}$

Example: Using the heights data set, calculate the sample variance and standard deviation.

i	1	2	3	4	5	Σ
x_i						
$x_i - \bar{x}$						
$(x_i - \bar{x})^2$						

Z-score

Def: The z-score measures the number of standard deviations an observation is from the mean: $\boxed{z_i = \frac{x_i - \bar{x}}{s}}$

Percentiles

Def: A percentile is a measure of relative standing. The p^{th} percentile is the number where at least p% of the data values are less than or equal to this number.

Special percentiles:
1. 25^{th} percentile = 1^{st} quartile = Q_1
2. 50^{th} percentile = 2^{nd} quartile = Q_2 = median
3. 75^{th} percentile = 3^{rd} quartile = Q_3

Computing Quartiles

1. Order the list of data values from smallest to largest
2. Compute the median. This is Q_2.
3. Split the data in half at the median. If there are an odd number of data values, the median is included in neither half.
4. Q_1 is computed by finding the median of the lower half.
4. Q_3 is computed by finding the median of the upper half.

Example: Calculate Q_1 and Q_3 of the heights data set.

Def: The interquartile range (IQR) is given by: $Q_3 - Q_1$.

Example: Calculate IQR for the heights data set.

Def: An outlier is a data point that lies far outside of the normal range of the rest of the data.

Rule of thumb: Data points smaller than $Q_1 - 1.5 \cdot IQR$ or larger than $Q_3 + 1.5 \cdot IQR$ are outliers.

Boxplots

Def: The five number summary is another way to describe a data set.
The five numbers are: minimum, Q_1, median, Q_3, maximum .

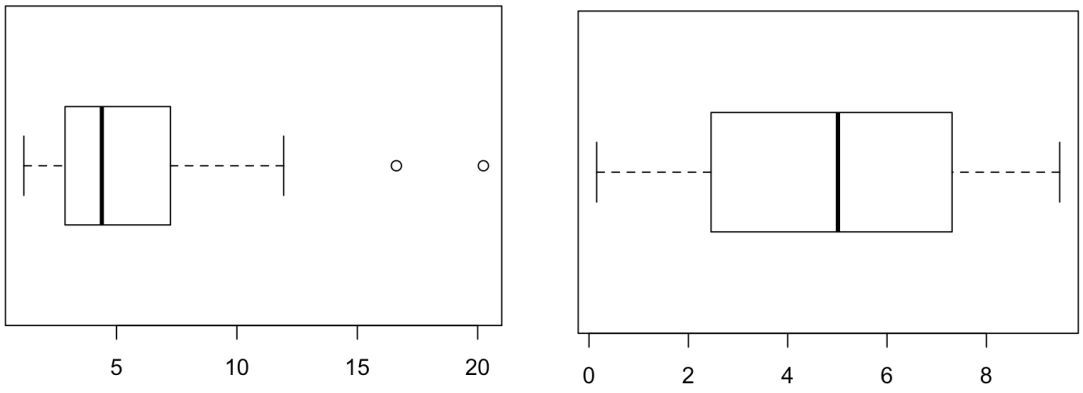
Example: What is the five number summary for the heights data set?

Once a five number summary has been created, it's easy to create a boxplot which is another graphical representation of the data set.

Note: There are numerous ways to create a boxplot, but we'll be following the standard of the textbook.

Drawing a boxplot

1. Determine if there are any outliers using the $Q_1 - 1.5IQR$, $Q_3 + 1.5IQR$ standard. Label these on a boxplot as circles.
2. Draw vertical lines at the following points:
   (a) Smallest data point that is not an outlier.
   (b) Largest data point that is not an outlier.
   (c) Q_1, median, Q_3
3. Make a box using the middle three lines.
4. Connect the box with the outer, non-outlying points.



Example: Draw a boxplot for the heights data set.

Notation for Population Parameters and Sample Statistics

Some care must be given when reporting a mean, variance, or standard deviation, as it is more important to distinguish between population and sample.

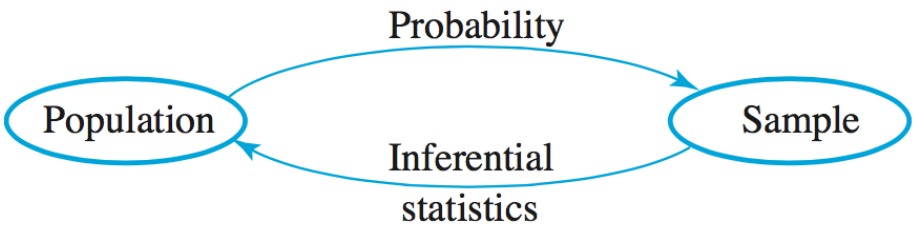
	Median	Mean	Variance	Standard Deviation
Population	$\tilde{\mu}$	μ	σ^2	σ
Sample	$\tilde{x}$	$\bar{x}$	s^2	s

Descriptive Statistics vs. Inferential Statistics

After a sample is taken, there are a variety of ways to describe the sample. This is the realm of Descriptive Statistics. We can describe our sample with numbers (statistics) or pictorially.

Later in the semester, we will examine Inferential Statistics. In Inferential Statistics, our goal is to use characteristics of a sample to say something about the full population (see Inferential Statistics).

Connection between Probability and Statistics



If we know the characteristics of a population, the study of Probability will help us determine the likelihood that a sample will have certain characteristics.

Likewise, Statistics can potentially help us use the characteristics of a sample to say something about the population.