CHAPTER 1. OVERVIEW AND DESCRIPTIVE STATISTICS

The main outcome for this course is an introduction for how to formally think and reason about uncertainty and data. Without training, people tend to be very bad at such tasks.

To illustrate this point, try the following experiment:

Step (1). Imagine the process of tossing a fair coin 100 times, each time recording the outcome. For example,

H H T H T T H . . .

Step (2). This time, toss an actual fair coin 100 times, and record the outcome each time.

Step (3). Compare the 100 recorded outcomes of the imaginary coin tosses with the 100 actual tosses of a coin.
(i) What systematic differences do you observe in these two data sets?
(ii) What do the differences in these data suggest about the way we perceive uncertainty, versus reality?

After diving into that thought experiment about randomness, next think about how you would define the notion of "data". Write down your definition, and try to be as precise and exhaustive as possible. Keep your definition with you throughout the semester, and refer back to it often to think about how your thoughts about data evolve as your understanding develops.

SECTION 1.1. POPULATIONS, SAMPLES, AND PROCESSES

Whenever you think about data, the first notion to cross your mind should be the notion of a "population". In fact, without the notion of a population data would have no context nor meaning.

Population is a broadly defined concept, but has a very precise definition in the context of a particular data set and questions of interest. A population is an entity of interest. It could be the mental health status of citizens of a certain county, for example, with respect to dementia. Or the population could be the length of time between buses at a bus stop. A population may be fully or partially observable, or not directly observable.

Alternatively, data is any information used to learn about any feature of a population of interest. For example, information collected at mental health screenings, or randomly sampled recordings of the inter-arrival times at a bus stop. Data could come in numerical form or could be formatted as written text. For example, clinician notes about a patient at a hospital represent data containing information about numerous population features that may be of interest.

Once the theoretical population and features of interest have been defined a "sample" of data is collected to make inference on the population/feature. A "sample" is any subset of the population, and is only relevant when the full population is not available.

Note that the distinction between a sample and a population is extremely important, is often very subtle, and is entirely context dependent. To illustrate, consider the following example.

EXAMPLE. Suppose I need to learn the average GPA of college students in North Carolina. Then the population consists of all college students in North Carolina, and the feature of interest is the average GPA. In theory, I could learn this number if I had access to the GPA of every college student in North Carolina. However, this may not be practical, and instead I could estimate the average GPA by sampling some subset of North Carolina college students. For example, suppose I have access to the GPA of every NCSU undergrad. In that case, the collection of GPAs of all NCSU students is the sample.

Things to think about:

(i) Is this sample representative of the population? In what ways is it not representative? For instance, are there likely systematic differences in the GPAs of NCSU, UNC, and Duke students? What about about App State students? How about Wake Tech students? Does the population of "College students in North Carolina" include community college students or only students at 4-year institutions? Maybe the population needs to be defined more precisely for the question(s) of interest.

(ii) The questions in item (i) allude to possible sampling biases. Note that a bias exists only in the context of a question of interest, relating a population to a sample.

(iii) What if alternatively, the population is defined as "College students at NCSU"? In this case, my sample contains the entire population, and I can compute exactly the feature I care about with no uncertainty.

DEFINITION. A simple random sample is a sample from a population such that any particular subset, of pre-specified size, of the sample has the same chance of being selected.

Read Section 1.1 in Devore.

SECTION 1.2. PICTORIAL AND TABULAR METHODS IN DESCRIPTIVE STATISTICS

This section begins with the introduction of stem-and-leaf displays and dot plots. It is worth reading about them in the textbook, but the main graphical representation to which these plots relate is a histogram.

DEFINITION. A variable is a quantity with values that in some way characterizes the members or objects in a population.

For example, a variable

X = Number of conflict related deaths in a given nation/time
Then $X \in \{0, 1, 2, \dots\}$

Y = Distance to commute to/from school for NCSU students.
Then $Y \in [0, \infty)$

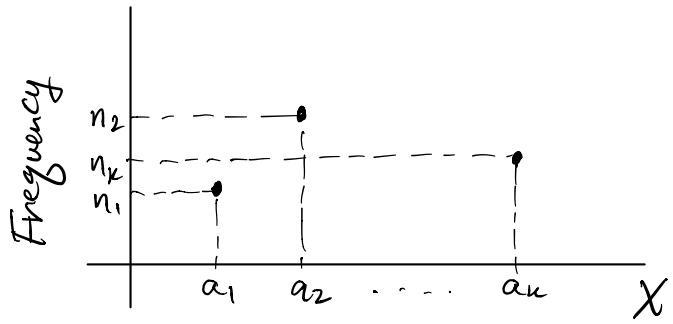
DEFINITION. A variable is discrete if it takes values only in a countable set. Alternatively, a variable is continuous if it takes values consisting of one or more entire intervals. A variable that is neither discrete nor continuous is called mixed.

Continuing with X and Y as in the example, first suppose that we have collected a sample of observations $x_1, x_2, \dots, x_n$. Assuming that these data have all come from the same population it is helpful to visualize them collectively to develop an intuition for what values of $X \in \{0, 1, 2, \dots\}$ are most common. There are many tools to visualize data to this end, but a histogram is typically among the most effective tools.

The algorithm for constructing a histogram for discrete data $x_1, \dots, x_n$ is as follows.

(1) Identify the set of unique values observed in the set $\{x_1, \dots, x_n\}$. Denote the k unique values as $a_1, \dots, a_k$.
(2) Count the number of times that each value $a_1, \dots, a_k$ occurs in the set $\{x_1, \dots, x_n\}$. Denote these frequencies by $n_1, \dots, n_k$.

(3) Plot the values $n_1, \dots, n_k$ against $a_1, \dots, a_k$



Alternatively, for step (3) you may want to plot the values $n_1/n, \dots, n_k/n$ versus $a_1, \dots, a_k$. The values n_i/n are called relative frequencies since

$$\sum_{i=1}^{k} \frac{n_i}{n} = \frac{n_1}{n} + \cdots + \frac{n_k}{n} = 1.$$

Note that it must be the case that

$$\sum_{i=1}^{k} n_i = n.$$

Note that for discrete data the histogram plot can be presented as points, lines, or bars. It is better to NOT use bars, however, to reflect that the data cannot take values continuously in an interval.

The algorithm for constructing a histogram for continuous data $y_1, \dots, y_n$ is as follows.

(1) Divide the interval $\left[\min_{1 \le j \le n}\{y_j\}, \max_{1 \le j \le n}\{y_j\}\right]$ into a partition of m bins of equal length. Denote the edges of the bins as

$$b_0 \le \min\{y_j\}, b_1, \dots, b_m \ge \max\{y_j\}$$

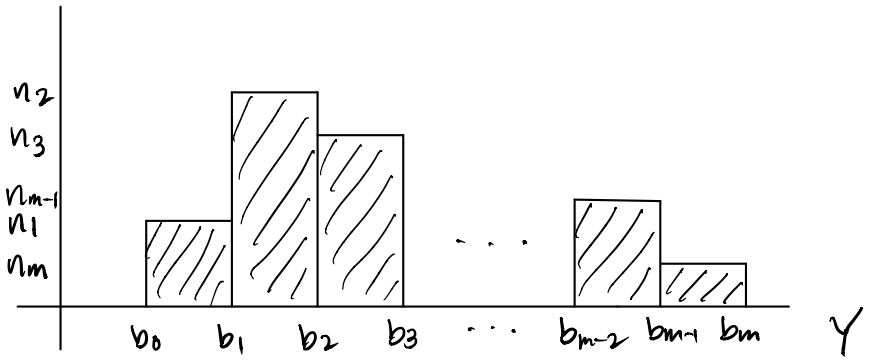
where $b_1 - b_0 = b_2 - b_1 = \cdots = b_m - b_{m-1}$.

(2) Construct the frequencies $n_1, \dots, n_m$ as

$$n_i = \left|\{j : y_j \in [b_{i-1}, b_i)\}\right|$$

for $i \in \{1, \dots, m\}$.

(3) Plot m rectangles with respective heights $n_1, \dots, n_m$ (or $n_1/n, \dots, n_m/n$) and width edges $b_0, b_1, \dots, b_m$.



Note that for continuous data the bars are touching because the variable Y could have taken any value in the intervals/bins $[b_{i-1}, b_i)$.

We also could have constructed the bins on $(b_{i-1}, b_i]$.

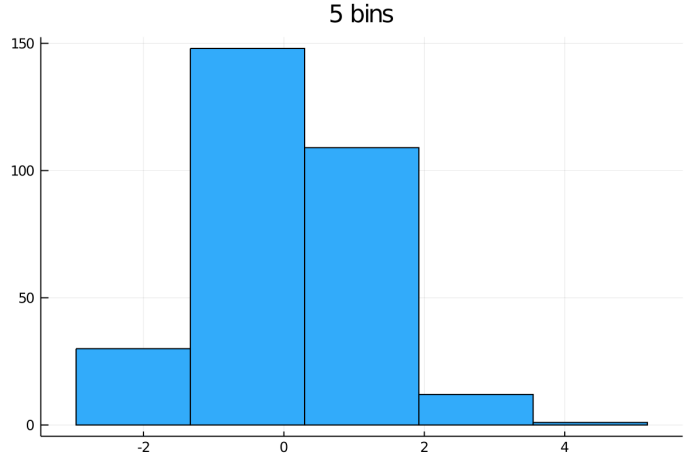
The bins all have to the same width so that the n_i values are comparable. However, what is it about the construction of the histogram for continuous data that has NOT been fully specified?

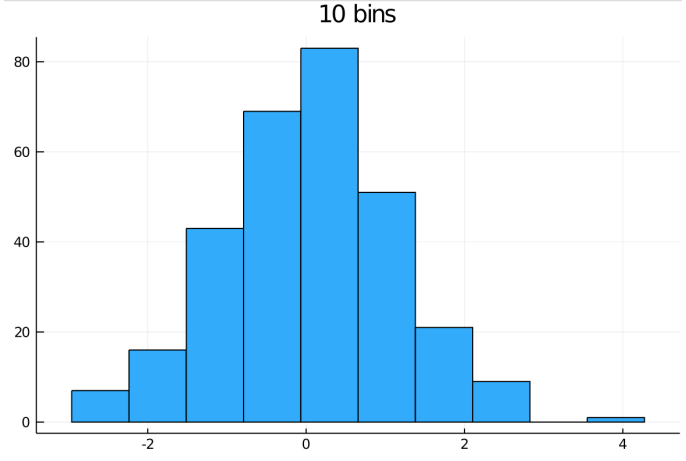
→ How to determine the correct number of bins, m?

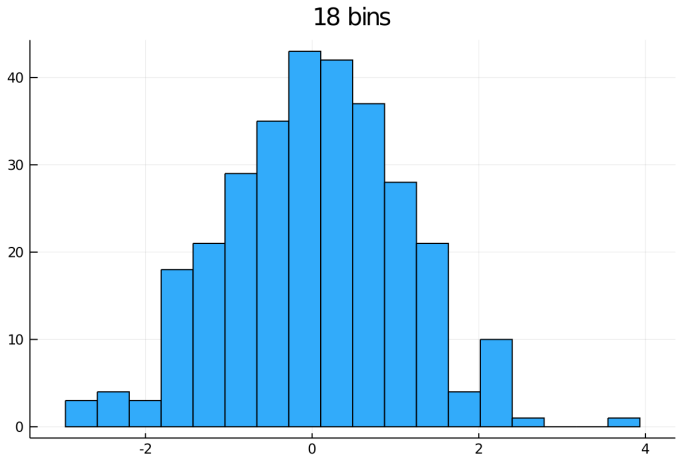
(i) There is no correct number of bins to choose
(ii) Academic papers have been written to investigate an optimal choice of number of bins.
(iii) A good choice is related to the sample size, n.
(iv) The representation of a histogram is highly sensitive to the number of bins.
(v) A reasonable choice is $\sqrt{n}$.

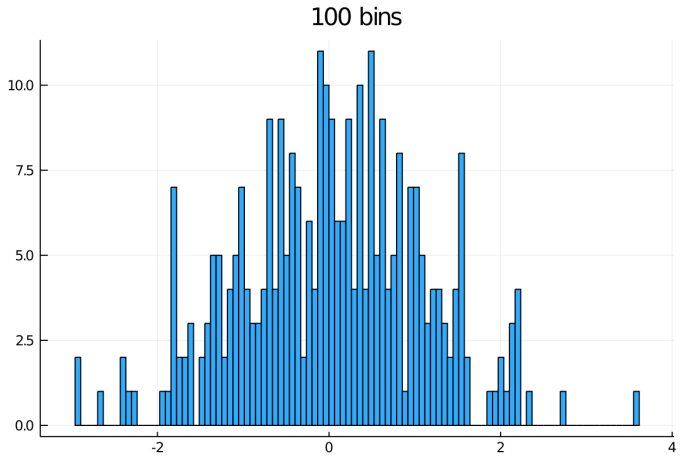
To illustrate these points, consider the following data and corresponding histograms. I have used a random number generator in the computer programming language Julia to generate $n = 300$ data points from a continuous random variable (rounded to two decimal places).

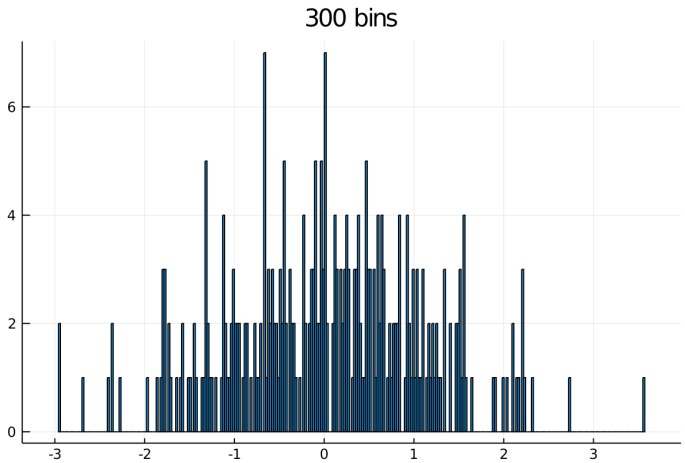
[-0.03, 3.55, 0.48, -0.88, -1.74, -1.02, 1.52, 0.24, 0.09, 0.15, 0.77, -0.22, -0.34, -1.58, -0.46, 0.03, 0.02, -0.22, 1.56, 0.51, 0.28, -1.29, -0.57, 0.15, -0.21, -2.94, -1.78, 2.22, -0.12, 1.52, -1.43, 1.21, 0.75, -0.27, -0.09, -1.08, -1.65, -0.48, 0.02, 0.47, 0.44, -0.61, -1.0, -0.45, 1.91, 0.96, 1.64, -2.96, 0.23, 1.34, 0.65, -1.14, 0.92, 0.01, -0.06, -2.28, -1.1, 0.5, -0.2, -1.82, 0.52, 0.28, 0.18, -1.31, 1.53, -1.78, 2.23, -1.73, 0.18, -0.67, 1.23, -0.03, 0.6, -0.85, -0.67, -0.94, 0.31, -2.37, -0.04, -0.47, -0.04, -1.79, -1.27, 0.51, 0.22, 0.42, 1.26, -1.87, 0.94, -0.12, -1.0, 0.2, 2.74, -1.37, -1.61, -1.03, 1.02, 0.8, -0.05, 0.04, 1.06, -1.51, -0.03, 1.15, -0.67, 0.33, 0.95, 2.1, -0.16, -0.71, 0.63, 0.58, -1.8, 0.67, 1.55, 0.74, 0.5, -0.58, 0.09, 1.11, 1.88, -1.3, -0.61, 0.36, -0.1, -0.62, 1.52, -0.31, -0.78, 0.56, 0.59, 1.49, 0.67, -0.42, 0.26, 2.14, -0.53, 2.22, 0.79, 0.13, 0.65, 0.4, -0.95, 1.46, -0.45, -0.87, -0.38, -1.34, -0.59, 0.92, -1.32, 0.36, 1.03, -0.07, 0.81, 0.83, 0.15, 1.22, -1.06, -2.4, -1.21, -0.42, -1.13, -0.85, 0.25, 0.5, -1.97, 0.33, -0.38, 0.12, -0.96, 1.34, -0.22, -0.09, 0.13, -1.78, 0.01, 1.11, -0.13, -0.73, -1.02, 0.38, 2.33, -1.31, -0.45, -0.54, 0.56, -0.22, 0.37, -0.71, 1.41, 0.7, 2.21, -1.11, 1.0, -0.09, 1.5, 0.65, 0.82, -0.44, -0.65, 0.02, 0.99, 0.66, 0.26, -0.01, 1.47, 0.34, 0.48, 1.29, -0.49, 0.6, 0.61, -0.08, 1.07, -1.32, -0.35, -1.02, 1.55, -1.5, -0.78, 0.55, -0.14, -0.12, 1.11, 0.12, 0.73, 1.19, -1.1, 0.39, -0.96, -0.75, -0.15, 0.83, -0.1, 0.9, -0.91, -1.13, -0.48, -0.51, 0.83, -2.36, -0.33, -1.24, -1.04, 0.41, -0.56, 0.19, -1.71, 1.0, 2.16, 0.93, -0.67, 0.01, -0.58, 0.93, -1.44, 0.28, 0.26, -0.66, -1.58, 0.65, 0.78, 0.84, 1.58, -0.01, -0.82, 1.16, -0.67, -0.02, 2.03, -0.35, 1.28, 1.99, -1.79, 0.38, 2.11, 1.41, 0.38, -0.37, -0.44, -1.12, 1.35, -0.17, 1.25, 1.56, 0.63, 0.01, -1.44, 0.2, 1.03, 0.48, 0.47, 1.03, -0.59, -2.68, -0.67, -1.32, -0.55, 1.16]











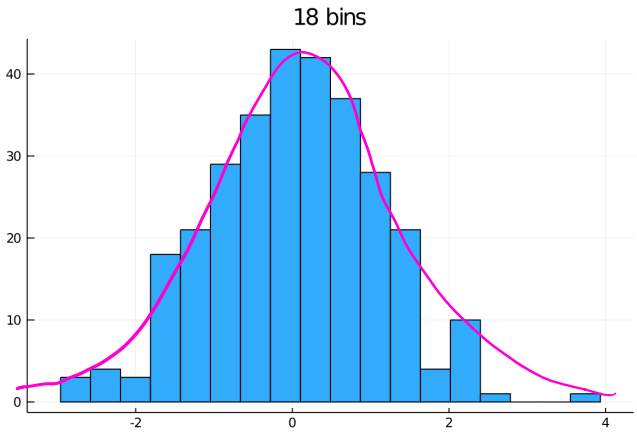
The figures show five different histograms of the same data, each using a different number of bins. Note that

$$\sqrt{n} = \sqrt{300} \approx 18,$$

and observe how different the picture of the data looks for different numbers of bins.

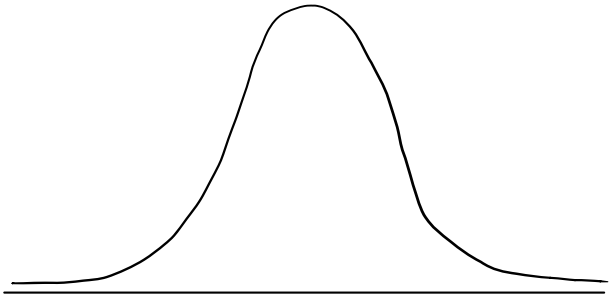
Aside from this topic of simple visual representations of univariate data, histograms are an approximation to a deeper and intrinsic feature of a random variable, the probability density (or mass) function. These functions will be an important topic that we will focus on in coming chapters.

The pink line is a representation of the (unnormalized) probability density function for these data, as estimated by the histogram. The larger the sample size, n, the better

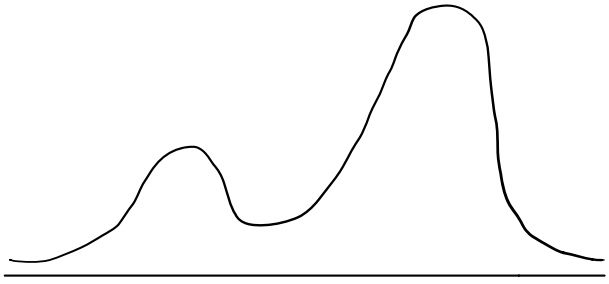


approximation to the true (or theorized) density function, from the histogram.

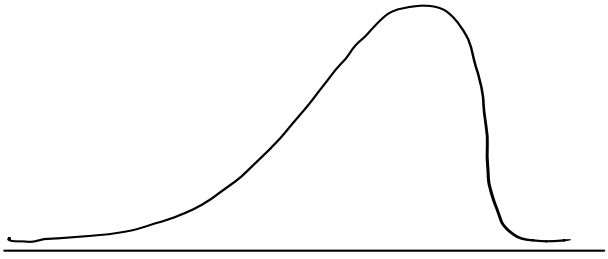
There are a variety of characteristics used to describe a given histogram. Most basically we describe the modality and the skew of the histogram.



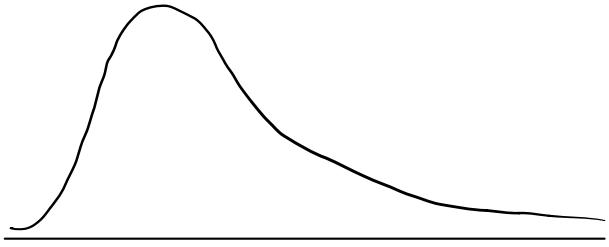
"Unimodal"
"symmetric"



"Bimodal" or "Multimodal"



"Negative Skew"



"Positive skew"

Read Section 1.2 in Devore.

SECTION 1.3. MEASURES OF LOCATION

In this section we develop more precise numerical summaries of data to describe the central tendency of a distribution of data. Namely, this includes the mean and the median.

Most people educated in the United States are familiar with the notion of a mean as an average of a collection of numbers. This concept is taught in primary school, and this is where we will begin our discussion.

DEFINITION. Suppose $x_1, \ldots, x_n$ is a collection of numbers. The arithmetic mean is defined as $\bar{x}_n := \frac{1}{n}\sum_{i=1}^{n} x_i$.

Note that $\bar{x}_n$ is called the arithmetic mean as opposed to the geometric mean or the harmonic mean. Further, we often call it the sample mean. Why?

→ This is where you should think a bit deeper. What is the point of computing or even defining $\bar{x}_n$?

→ We are trying to measure something about the broader population from which these data were generated/sampled.

→ This is why the notion of a population (and sample) are very important.

→ When you interpret an average you should ask whether the average is being used to reflect the unknown population mean, or is simply being used to describe a static collection of numbers.

We also have the notation μ used to refer directly to the population mean. Greek letters are commonly used in the field of statistics to denote population features. Much more will be said about population features in future chapters.

The next thing to consider is why the sample mean is thought of as a natural measure of location. In fact, this idea was understood as very strange a few hundred years ago.

→ First, for using a sample mean to describe the center of a set of data, $\bar{x}_n$ may not even represent an observed data value. For example, suppose $n=2$, $x_1 = 4$, $x_2 = 5$. Then

$$\bar{x}_2 = 4.5 \notin \{x_1, x_2\}$$

→ The sample mean is a feature of a set of data at the group level, not at the individual level.

→ Mathematical results for large sample theory shows that under reasonable conditions

$$\bar{X}_n \longrightarrow \mu \quad \text{as} \quad n \longrightarrow \infty$$

Before discussing a few other properties of the sample mean, I will introduce another measure of location. This will give helpful context for making better sense of the properties.

DEFINITION. For a sample of observed data $x_1, \ldots, x_n$ the sample median is any number M_n that satisfies

(a) $\sum_{i=1}^{n} 1\{x_i < M_n\} = \sum_{i=1}^{n} 1\{x_i > M_n\} = \frac{n}{2}$, if n is even.

(b) $M_n = x_j$ such that $\sum_{i=1}^{n} 1\{x_i < x_j\} = \sum_{i=1}^{n} 1\{x_i > x_j\} = \frac{n-1}{2}$, if n is odd.

where $1\{\cdot\}$ is the indicator function (1 if $\{\cdot\}$ true, 0 else).

As was the case with the mean, the sample median is the sample analogue of the population median. We will likely discuss the notion of the population median in coming chapters.

Do not confuse the definition of the sample median with the algorithm that you have surely learned for finding the sample median, in previous courses. The basic algorithm is as follows.

Step 1. Order the observed data $x_1, x_2, \ldots, x_n$ in increasing order

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)} \quad \text{— these are called order statistics.}$$

Step 2a. If n is odd, then $M_n = x_{\left(\frac{n+1}{2}\right)}$

Step 2b. If n is even, then any number $M_n \in \left(x_{\left(\frac{n}{2}\right)}, x_{\left(\frac{n}{2}+1\right)}\right)$ suffices.

Note that in the textbook it is recommended to choose the average value,

$$M_n = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2},$$

however, there is no particularly strong reason to do so. For example, suppose

$$\{x_1, \ldots, x_6\} = \{61, 45, 39, 18, 14, 33\}$$

Then $$\{x_{(1)}, \ldots, x_{(6)}\} = \{14, 18, 33, 39, 45, 61\}$$

So the median $M_n \in (33, 39)$. We could call $M_n = \frac{33+39}{2} = 36$, but is 36 any more representative of the "middle" than any other number in the interval $(33, 39)$, based on the definition of the median?

Suppose we observe another data point, x_7. Then

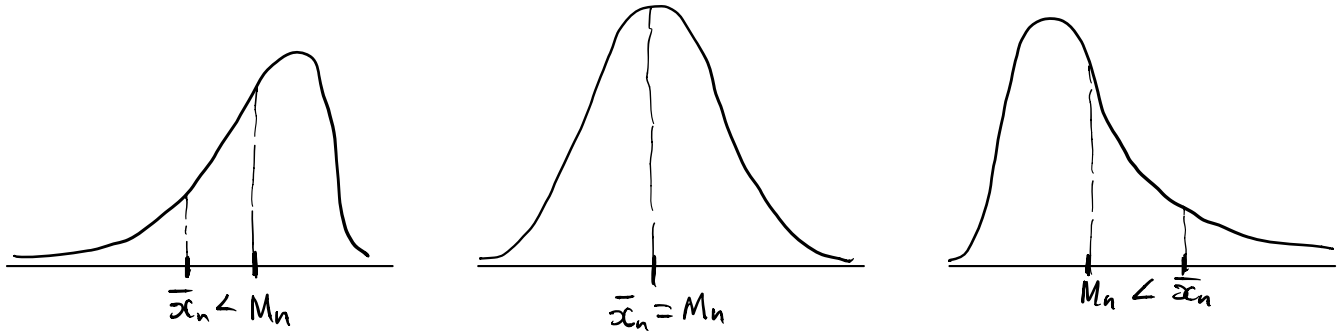
(a) If $x_7 < 33$, $M_n = 33$
(b) If $x_7 > 39$, $M_n = 39$
(c) If $x_7 \in [33, 39]$, $M_n = x_7$

None of these cases relate to 36, unless it happens that $x_7 = 36$.

Now consider a few distinguishing properties of the mean and median.

(1) The median is said to be "robust" to extreme values or outliers, while the mean is not.
EXAMPLE. Consider the set of numbers $\{1, 2, 3\}$. Then $\bar{x}_3 = 2 = M_3$. Observe that if the next observation is -100, then the data set $\{-100, 1, 2, 3\}$ has $M_4 \in (1, 2)$ which is not much different from $M_3 = 2$, but $\bar{x}_4 = -23.5 \ll 2 = \bar{x}_3$.

(2) The relative positions of the mean and median allude to the symmetry or skewness of the data. That is,



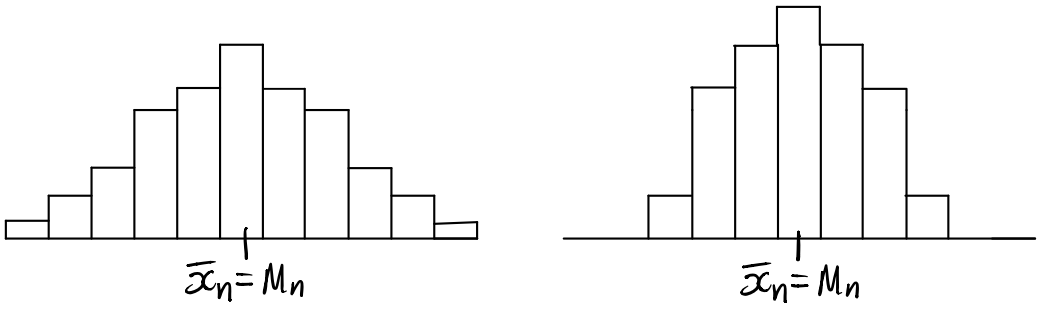
Negative skew | Symmetric | Positive skew

Other measures of location such as a trimmed mean are briefly discussed in Section 1.3.

Read Section 1.3 in Devore.

SECTION 1.4. MEASURES OF VARIABILITY

Measures of location are a useful first metric to consider when describing a sample of data, but the certainly do not give a full picture. For example,



These are two histograms representing data that has the same mean and median (i.e., location), but different variability.

Perhaps the simplest measure of variability is the range, $x_{(n)} - x_{(1)}$, and is useful for giving a sense of the magnitude and scale of the observed values.

More commonly, it is useful to consider deviations in the observed values x_i from some measure of location. This allows us to interpret the x_i relative to a meaningful and common point of reference. But how to measure deviations?

Start by considering $\bar{x}_n$ as the location point of reference. Then we can compute the deviations

$$x_1 - \bar{x}_n, \quad x_2 - \bar{x}_n, \quad \ldots, \quad x_n - \bar{x}_n$$

Reducing these deviations to a single value by summing them together gives

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x}_n) &= \left(\sum_{i=1}^{n} x_i\right) - n\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \\ &= \left(\sum_{i=1}^{n} x_i\right) - \left(\sum_{i=1}^{n} x_i\right) \\ &= 0\end{aligned}$$

This is because positive deviations cancel with negative deviations. Instead, consider first squaring the deviations so that they are all positive. This leads to what is referred to as the sample variance.

DEFINITION. The sample variance, denoted by S_n^2, is given by

$$S_n^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2.$$

The sample standard deviation is defined as $S_n = \sqrt{S_n^2}$.

Note that the sample variance is a natural analogue to the sample mean. The standard deviation is more interpretable because it is on the same scale as the data, whereas the variance is units squared.

The reason for using $\frac{1}{n-1}$ instead of $\frac{1}{n}$ in the expression for S_n^2 is so that it is an "unbiased estimator" of the population variance, σ^2. For a population of finitely many members, say N,

$$\sigma^2 := \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

A useful formula for computing S_n^2 is the following.

$$\begin{aligned}(n-1)\,S_n^2 &= \sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \\ &= \sum_{1}^{n}(x_i^2 - 2x_i\bar{x}_n + \bar{x}_n^2) \\ &= \left(\sum_{1}^{n} x_i^2\right) - 2\bar{x}_n\sum_{1}^{n} x_i + n\bar{x}_n^2 \\ &= \left(\sum_{1}^{n} x_i^2\right) - 2n\bar{x}_n^2 + n\bar{x}_n^2 \\ &= \left(\sum_{1}^{n} x_i^2\right) - n\bar{x}_n^2 \\ &= \left(\sum_{1}^{n} x_i^2\right) - \frac{1}{n}\left(\sum_{1}^{n} x_i\right)^2\end{aligned}$$

Next, consider two properties of the sample variance.

(1) S_n^2 is location invariant. That is, the sample of data $x_1, \dots, x_n$ has the same sample variance as the transformed values

$$y_1 := x_1 + c \;,\; y_2 := x_2 + c \;,\; \dots,\; y_n := x_n + c,$$

for any constant c.

Proof
$$\begin{aligned}\sum_1^n (y_i - \bar{y}_n)^2 &= \sum_1^n \left(x_i + c - \left[\tfrac{1}{n}\sum_1^n (x_j + c)\right]\right)^2 \\ &= \sum_1^n \left(x_i + c - \left(\tfrac{1}{n}\sum_1^n x_j\right) - \tfrac{1}{n}\sum_1^n c\right)^2 \\ &= \sum_1^n (x_i + c - \bar{x}_n - c)^2 \\ &= \sum_1^n (x_i - \bar{x})^2\end{aligned}$$

Hence, $\frac{1}{n-1}\sum_1^n (y_i - \bar{y}_n)^2 = \frac{1}{n-1}\sum_1^n (x_i - \bar{x})^2$ #

(2) If $y_1 = cx_1$, $y_2 = cx_2$, $\dots$, $y_n = cx_n$ for any constant c, then the sample variance of the values $y_1, \dots, y_n$ is the sample variance of the $x_1, \dots, x_n$ scaled by c^2.

Proof.
$$\begin{aligned}\sum_1^n (y_i - \bar{y}_n)^2 &= \sum_1^n \left(cx_i - \left[\tfrac{1}{n}\sum_1^n cx_j\right]\right)^2 \\ &= \sum_1^n c^2 \left(x_i - \tfrac{1}{n}\sum_1^n x_j\right)^2 \\ &= c^2 \sum_1^n (x_i - \bar{x}_n)^2\end{aligned}$$

So $\frac{1}{n-1}\sum_1^n (y_i - \bar{y}_n)^2 = c^2 \cdot \frac{1}{n-1}\sum_1^n (x_i - \bar{x})^2$

And $\sqrt{\frac{1}{n-1}\sum_1^n (y_i - \bar{y}_n)^2} = |c| \sqrt{\frac{1}{n-1}\sum_1^n (x_i - \bar{x})^2}$ #

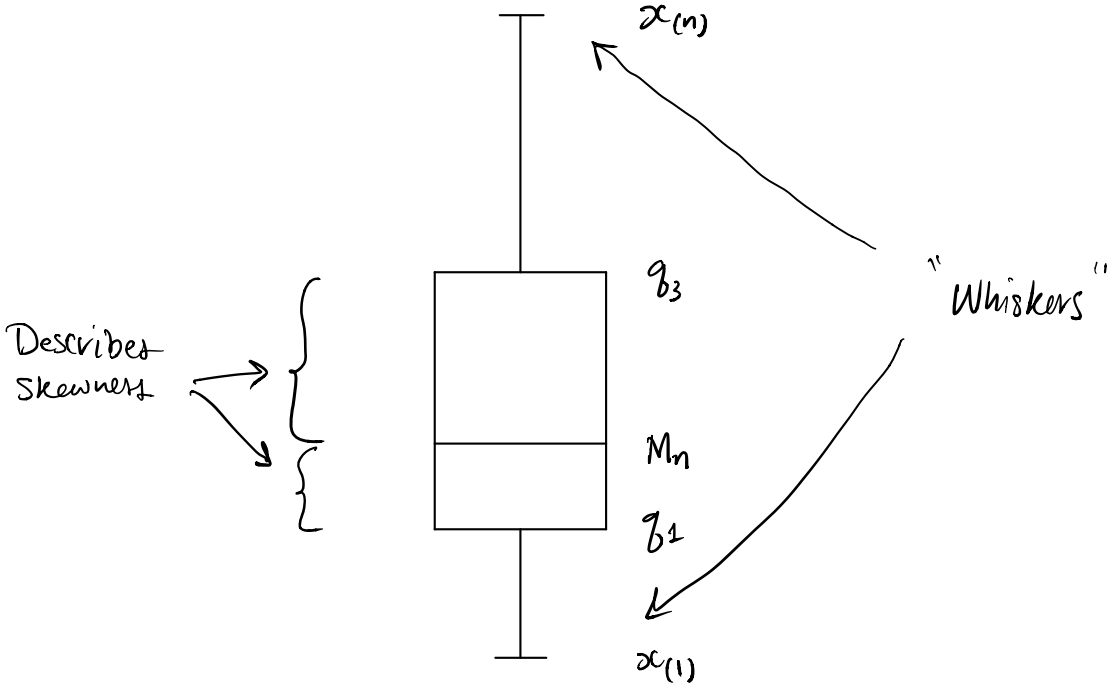
The last topic of chapter 1 is boxplots. These are super useful for presenting a summary of data, and are inherently robust to extreme values in the data. A boxplot is most commonly constructed from the five number summary,

$$x_{(1)} \;,\; q_1 \;,\; M_n \;,\; q_3 \;,\; x_{(n)}$$

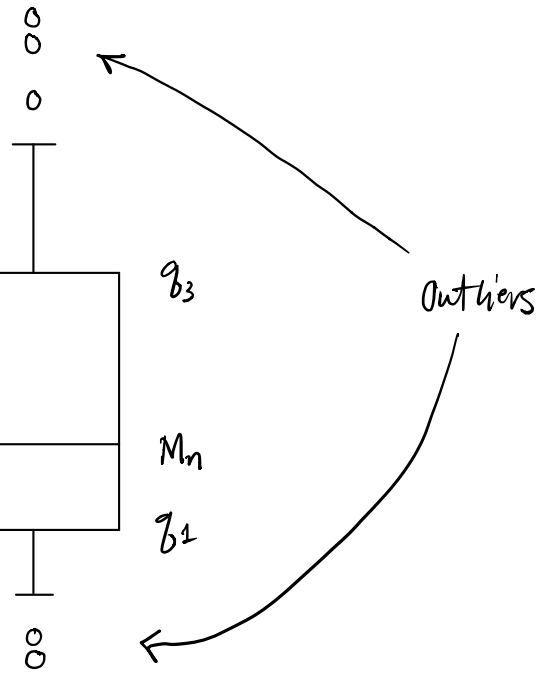
where q_1 and q_3 are the 25th and 75th quantiles, respectively. Precisely, q_1 is the median of the lower half of the ordered set of data values, where the lower half are all values less than the median, M_n. Similarly, q_3 is the median of the upper half of the observed data values.

Commonly, an outlier is defined as any data value that exceeds $1.5(q_3 - q_1)$ distance from q_1 and q_3. The quantity $q_3 - q_1$ is commonly called the interquartile range.

Box plots are often presented in the form



Alternatively, the whiskers can be shortened to extend only to the smallest and largest observed values that are not considered outliers. That may look like



Read Section 1.4 in Devore.

CHAPTER 2. PROBABILITY

This chapter begins our rigorous treatment of studying and quantifying uncertainty.

SECTION 2.1. SAMPLE SPACES AND EVENTS.

DEFINITION. An experiment is any activity or process whose outcome is subject to uncertainty.

We will use the term "experiment" loosely throughout this semester. It can refer to a controlled scientific investigation such as clinical trials, or to a process as simple as tossing a coin.

DEFINITION. The sample space, denoted by S, is the set of all possible outcomes of an experiment.

For example, tossing a coin has two possible outcomes, H or T. So

$$S = \{H, T\}$$

If instead the experiment consisted of tossing two coins, then

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

DEFINITION. An event is any subset of outcomes contained in the sample space S.

For example, if the experiment consists of tossing two coins, then $A = \{HH, TH\}$ is an event since $A \subseteq S$.