Chapter 1 Descriptive Statistics

1.1 Populations, Samples, and Processes

Statistics provides methods to summarize and draw conclusions from data.

Typically, we focus on a particular population of interest.

Ex:
- all students in our class
- all classes in math department
- group of animals
- repeated tests on object

When we have information about all objects in the population, this is called a census.

Often a census is too difficult or expensive to obtain, so instead we get a sample, which is a subset of the population.

Ex:	Census	Sample
	all U of U students	100 U of U students
	population census	one neighborhood

The characteristics that change among the objects in the population are called variables.

Data with one variable is univariate, two variables is bivariate, and more than one variable is multivariate.

A variable can be numerical or categorical.

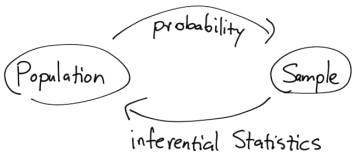
Ex:
- height of students — univariate, numerical
- hair color of students — univariate, categorical
- (height, hair color) — bivariate

Branches of Statistics:

Descriptive Statistics (Ch 1): Summarize and describe important features of the data. Includes calculating the mean, or drawing a histogram.

Inferential Statistics (Ch 6–9, 3080): Using the sample to make inferences or conclusions about the population. Includes confidence intervals and hypothesis tests.

Inferential Statistics is the main focus of 3070/3080. But first in Chapters 2-5 we cover Probability.



Ex:

Probability	Inferential Statistics
20 slices in loaf of bread, slice picked at random	We have a sample of slices, want to use that to understand the whole loaf
$\frac{2}{20}$ prob. of picking end piece at random	
 3 fair dice, roll them. prob. of rolling 10	roll the dice, use the samples to understand the dice, e.g. are they fair dice?

Enumerative vs Analytic Studies:

A study is enumerative if it focuses on a finite, unchanging, identifiable collection of people or objects which make up the population. It is possible to list all the members of the population, which is called a sampling frame.

Otherwise the study is analytic.

Ex:

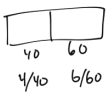
enumerative	analytic
voting in class/city	future voters testing new car model (some aren't built yet)
	prototype, initial testing
all lightbulbs produced yesterday	all lightbulbs produced

Methods for Collecting Data:

Simple Random Sample: Any subset/sample of size n is equally likely to be picked.

ex: draw names from hat, n of N students

Stratified Sample:  Split population into non-overlapping groups, and sample from each group.



ex: sample class by class

Convenience Sample: Sampling without randomness, done because it's easier

ex: sample at particular place / time

using our neighborhood, our class, our friends, ...

1.2 Visual Displays of Data:

Stem-and-Leaf Display: Say we have numerical data $x_1, x_2, \ldots, x_n$ with at least two digits. We can visualize the data with a stem-and-leaf plot.

We pick 1 or more leading digits to be the stem values, and the trailing digits become the leaves.

Example 1: Make a stem-and-leaf plot for the data:

04, 12, 15, 15, 19, 22, 23, 23, 26, 28, 29, 30, 31, 31, 31, 44, 53, 55

Stem	leaf
0	4
1	2 5 5 9
2	2 3 3 6 8 9
3	0 1 1 1
4	4
5	3 5

Sometimes to have more stems, we may repeat stem values as high and low.

Ex:

3L	0 2 3 4
3H	5 6 6 9 9 9
4L	0 2 3
4H	5 7 8 8
5L	1 1 4
5H	8 9

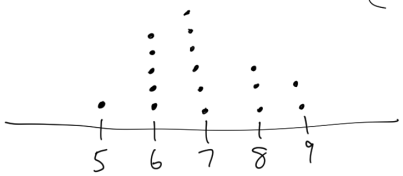
Can use stem( ) in R to create stem-and-leaf plots.

Dotplots: An alternative for numerical data is to use a dot plot. The scale is horizontal with dots placed vertically for each data point.

Example 2: Make a dot plot for the data:

5.1, 5.7, 5.9, 6.2, 6.3, 6.3, 6.7, 6.8, 7.0, 7.1, 7.1, 7.4, 7.9, 8.2, 8.4, 8.8, 9.1

(rounded to nearest whole number)



Dotplots become messy with large samples, and get too sparse with many different values in the data. So often we use a histogram instead.

Discrete vs Continuous Variables

A numerical variable is discrete if there are finitely many, or a countably infinite number, of possible values.

A numerical variable is continuous if its possible values consist of an entire interval of the real line.

Ex:

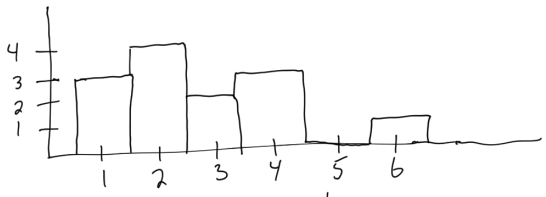
Discrete	Continuous
Year (2020, 2021, ...)	time (8.43 mins)
# of students in class	speed of car (0-60 mph)
	length (2.32 m)

Histograms:

To make a histogram with discrete data, put the possible values on the horizontal scale and above each value draw a rectangle whose height is the frequency of that value.

Example 3: Make a histogram for the data:

1, 1, 1, 2, 2, 2, 2, 3, 3, 4, 4, 4, 6



We can use hist( ) to make a histogram in R.

Note we can also use the relative frequencies to make the histogram.

Frequency: number of times the value appears in the data

Relative Frequency: $\frac{\text{number of times the value appears in data}}{\text{number of observations in data}}$

How do we make a histogram for continuous data?

We need to divide the interval of possible values into classes (bins).

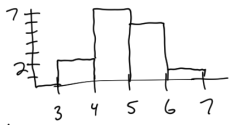
There are no specific rules about how to do this. A good rule of thumb is to use $\sqrt{\text{number of observations}}$ as the approximate number of classes/bins.

Example 4: Make a histogram for this data:

3.6, 3.9, 4.1, 4.2, 4.4, 4.6, 4.7, 4.7, 4.8, 5.1, 5.2, 5.3, 5.4, 5.8, 6.4

15 observations, $\sqrt{15} \approx 4$ classes/bins

3.0 – 3.99
4.0 – 4.99
5.0 – 5.99
6.0 – 6.99



$6.4 - 3.6 = 2.8$

$\frac{2.8}{4} = 0.7$

3.6 – 4.3
4.31 – 5.0
5.01 – 5.7
5.71 – 6.4

different histograms depending on classes!

many other possibilities

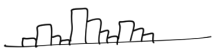
Histogram Shapes:

Modality: How many peaks does the histogram have?

unimodal: one peak

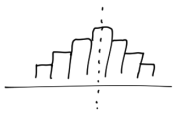
bimodal: two peaks

multimodal: 3 or more peaks



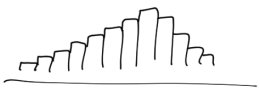
When a histogram is unimodal, we may also wonder if it is symmetric or skewed.

Symmetric: left half is a mirror image of the right half



Positively Skewed: the right tail is stretched out compared to the left.
(skewed right)

Negatively Skewed: the left tail is stretched out compared to the right.
(skewed left)



How the histogram looks can vary significantly depending on how the classes/bins are chosen.

1.3 Measures of Location:

Now we want to look at numerical summaries of the data.

For now, say we have numerical data

$x_1, x_2, \ldots, x_n$

say "x bar"

The Mean: The sample mean $\bar{x}$ is the arithmetic average:

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

Example 1: Find the sample mean of the data:

3, 6, 8, 9, 9

$$\bar{x} = \frac{3+6+8+9+9}{5} = \boxed{7}$$

R command is mean( ).

We can also calculate the population mean which is the average over the entire population. We usually write the population mean as μ.

mu

The mean is susceptible to outliers:

3, 6, 8, 9, 9 has mean 7

3, 6, 8, 9, 359 has mean 77

The Median

The median $\tilde{x}$ of $x_1, x_2, \ldots, x_n$ is the middle value when the observations are ordered from smallest to largest.

When n is odd, there is a unique middle value:

2, 4, 6, 7, 8, 9, 12

median $\tilde{x} = 7$

When n is even, there are two middle values and the median is the average of the two middle values.

3, 6, 8, 10, 11, 14

median $\tilde{x} = \frac{8+10}{2} = 9$

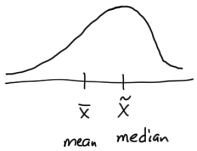
In R, we can use median( ).

Unlike the mean, the median is not susceptible to outliers.

We can find the population median, which we will write as $\tilde{\mu}$.

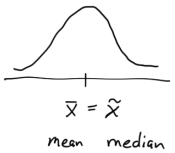
How do the mean and median compare if the data is:

Negatively Skewed

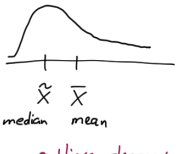


outliers drag mean left

Symmetric



Positively Skewed



outliers dragged mean right

Quartiles:

The median divides the data in 2 halves. We can also divide the data into 4 parts. These values are called quartiles.

Ex: 2 3 5 5 6 7 9 10 11 12 15

- 5: 1st quartile, Q1
- 7: median, 2nd quartile, Q2
- 11: 3rd quartile, Q3

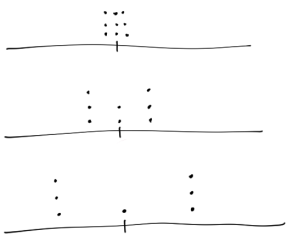
Sample Proportion and Population Proportion

When the data is categorical, we can look at the relative frequency of a particular value occurring. This is called the sample proportion or when looking at the whole population, the population proportion.

1.4 Measures of Variability

Different data sets can have the same measures of center but with different variability, how spread out the data is around the center.

Ex:



Range: The range is the difference between the largest and smallest sample values. This is a simple way to measure the variability, but the downside is it's very susceptible to outliers.

Instead, we usually look at deviations from the mean:

$x_1 - \bar{x}$
$x_2 - \bar{x}$
$\vdots$
$x_n - \bar{x}$

$2 - 3 = -1$
$3 - 3 = 0$
$4 - 3 = 1$

average deviation 0

We want to combine this information into a single number.

Look at average of the deviations, but we don't want the positives and negatives to cancel out. One idea is to take absolute value, another idea is to square the data. To make calculations easier, we'll square the deviations, and then take the average.

Sample Variance and Standard Deviation:

The sample variance s^2 is

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

The sample standard deviation is $s = \sqrt{s^2}$.

s is always nonnegative and has the same units as the data.
It represents a typical (standard) deviation from the mean.

Example 1: Find the sample standard deviation for the data:

2, 3, 4, 7 $\qquad \bar{x} = \frac{2+3+4+7}{4} = \frac{16}{4} = 4$

deviations	squared deviations
$2 - 4 = -2$	$(-2)^2 = 4$
$3 - 4 = -1$	$(-1)^2 = 1$
$4 - 4 = 0$	$0^2 = 0$
$7 - 4 = 3$	$3^2 = 9$

$s^2 = \frac{1}{3}(4+1+0+9)$

$= \boxed{\frac{14}{3}}$

$\boxed{s = \sqrt{\frac{14}{3}}}$

In R, we can use sd( ) and var( ) to find the sample standard deviation and variance.

Why n-1 in the denominator? (Revisit in Chapter 6)

The population variance and population standard deviation are written σ^2 and σ. (sigma σ)

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

(μ: average of whole population)

Since we typically don't know μ, in the sample variance we use an approximation, $\bar{x}$.

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

($\bar{x}$: average sample)

We'll revisit this later, but the idea is approximating μ with $\bar{x}$, we lose a degree of freedom.

Alternative Formula for s^2 and s: (We'll skip this since usually we use technology)

We will mostly utilize technology to compute s and s^2. But here are alternative formulas which can be easier to use by hand.

$$s^2 = \frac{S_{xx}}{n-1} \quad \text{where} \quad S_{xx} = \sum(x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}.$$

Pf is in the book, but we'll skip it for now.

Scaling and Shifting Effects:

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

Say data $x_1, x_2, \ldots, x_n$ has sample variance s^2.

What is the sample variance of $x_1+c, x_2+c, \ldots, x_n+c$? ($c$ is constant)

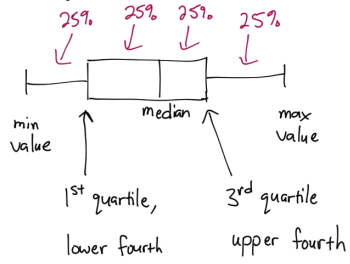
It stays the same!

What is the sample variance of $cx_1, cx_2, \ldots, cx_n$?

sample var: $c^2 s^2$

sample sd: $\sqrt{c^2 s^2} = |c| \cdot s$

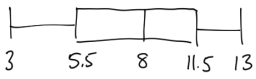
Boxplots: We can create a visual representation of the data with the **quartiles**, or **fourths**.



The **fourth spread**, or **IQR**, is upper fourth − lower fourth.

Example 2: Draw the boxplot and find the fourth spread f_s for the data: 3, 4, 7, 8, 8, 10, 11, 12, 13

lower fourth 5.5 | median 8 | upper fourth 11.5



$f_s = 11.5 - 5.5 = \boxed{6}$

In R, we can use boxplot( ).