MATH 1150 Chapter 2 Notation and Terminology

Categorical Data

The following is a dataset for 30 randomly selected adults in the U.S., showing the values of two categorical variables: whether or not the person has access to a smartphone (Y/N), and whether they rent their current residence, own their current residence, or would classify their living situation in a different way, such as living rent-free with parents or friends, living in a shelter, etc. (Rent/Own/Other).

Person	Smartphone?	Home Status
A	Y	Rent
B	Y	Rent
C	Y	Own
D	N	Rent
E	Y	Own
F	Y	Rent
G	N	Other
H	Y	Rent
I	N	Rent
J	Y	Own
K	Y	Own
L	Y	Rent
M	Y	Own
N	Y	Own
O	Y	Other
P	N	Rent
Q	N	Other
R	Y	Own
S	N	Rent
T	Y	Rent
U	Y	Other
V	Y	Rent
W	Y	Own
X	Y	Own
Y	N	Rent
Z	Y	Own
AA	Y	Rent
BB	Y	Rent
CC	N	Own
DD	N	Other

Since there are only a few possible values for each of our two categorical variables, and we do not seem to be interested in other details about the randomly selected individual cases, we can present this data in more efficient ways. For example, we can use the **frequency tables** shown below:

Smartphone?	**Frequency**
Yes	21
No	9

Home Status	**Frequency**
Rent	14
Own	11
Other	5

Moreover, if we wish to analyze the relationship between two quantitative variables, it is useful to present the data in the following way, known as a **two-way table**:

Home Status/Smartphone?	Yes	No	Total
Rent	9	5	14
Own	10	1	11
Other	2	3	5
Total	21	9	30

With categorical variables, information that we are frequently interested in is the **proportion** of individual cases that lie in a particular category. For example, suppose we were interested in the question "What proportion of adults in the U.S. have access to a smartphone?" For the answer to such a question, we use the notation

$$p = \frac{\text{number of individual cases in the population that lie in the specified category}}{\text{total number of individual cases in the population}}.$$

In practice, the proportion p for the full population is usually not available to us, so we instead try to approximate it using a sample. In this case, we use the notation

$$\hat{p} = \frac{\text{number of individual cases in the SAMPLE that lie in the specified category}}{\text{total number of individual cases in the SAMPLE}}.$$

For example, if our question is "What proportion of adults in the U.S. have access to a smartphone?", then we use p to denote the actual answer, and using our data we would write

$$\hat{p} = \frac{21}{30} = 0.7 = 70\%.$$

To clarify, any of those ways (fraction, reduced or not, decimal, percentage) of presenting the proportion are totally fine.

The idea is that, if the sample is good, $\hat{p}$ should be a good approximation for p. Another good use for two way tables is to address questions like the following, which investigate relationships between the two categorical variables:

What proportion of U.S. adults who own their home have access to a smartphone? From our sample data, that proportion is $10/11 \approx 90.9\%$. Note how this compares to the sample proportion of 70% of the full sample who have access to smartphones.

What proportion of U.S. adults who have access to a smartphone own their home? From our sample, that proportion is $10/21 \approx 47.6\%$. Note how this compares to the sample proportion of $11/30 \approx 36.7\%$ of the full sample who own their home.

It is important to note that the two questions above are different questions, with different answers. However, in this case, they both indicate a potential association between home ownership and access to a smartphone. Although, to refer back to some important chapter 1 terminology, this association is likely not causual, but rather due to confounding variables such as income or wealth.

Shape of Quantitative Data

The following is a frequency table showing the number of tattoos for 100 randomly selected adults.

Number of Tattoos	Frequency
0	41
1	28
2	10
3	11
4	8
9	1
23	1

As discussed in class, data from a frequency table can also be represented visually with a **bar chart** or **histogram**.

In this example, the small number of distinct responses would make a bar chart more effective, and the chart would show that the data is piled up on the left side, meaning the low end, of the range of possible values, with a tail going far to the right. We say that such data is **skewed to the right**.

Analogously, if the data was piled up on the right side, meaning the high end, of the range of values, with a tail going far to the left, we say the data is **skewed to the left**. An example we discussed was scores on an easy exam with a handful of unprepared students. To avoid mixing up the terminology (like I did), remember that the TAIL is the skew. If instead the bar chart or histogram appears that the data can roughly be "folded in half" onto itself, then we say the data is **symmetric**.

If, in addition to being symmetric, the bar chart or histogram has a peak in the center and falls consistently on either side, we say the data is **bell-shaped**. This shape is very common in naturally observed data.

Center of Quantitative Data

We discussed two ways of measuring the "center" of a set of numbers. Specifically, if we conduct a sample for a quantitative variable and collect n numbers $x_1, x_2, \ldots, x_n$, then we define the **mean** of those numbers to be

$$\overline{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Let that second equality serve as a reminder of what the Σ notation means, which is just to add everything up. If, rather than a sample, our numbers represent the FULL population of interest, we use the notation μ instead of $\overline{x}$. Just like with p and $\hat{p}$, the idea is that, if the sample is good, $\overline{x}$ should be a good approximation for μ.

The other notion of the "center" for quantitative data is the **median**. When a list of numbers is written in increasing order, the median is the midpoint, the number that splits the list in half. If the length of the list is odd, then the median is the actual member of the list right in the middle, while if the length of the list is even, the median is the average of the two members of the list in the middle.

For example, if the heights of seven adult males are sampled, yielding the list

$$67, 73, 68, 68, 71, 75, 69,$$

then the mean for the sample is

$$\overline{x} = \frac{67 + 73 + 68 + 68 + 71 + 75 + 69}{7} \approx 70.143,$$

while the median is 69, since when the list is put in increasing order, there are three numbers below 69 and three numbers below it. However, if we had a sample with the weights of six men

$$162, 170, 194, 226, 188, 379$$

then the mean is $1319/6 \approx 219.83$, while the median is 191 (even though there is no 191 on the list), since when the list is put in increasing order, the two numbers in the middle are 188 and 194.

Observing the difference between the mean and the median is a way of measuring how much the data for a quantitative variable is skewed. If the data is symmetric, they will be about the same. If the data is skewed right, the mean will be higher (like in the weight example above, or our income example in class). If the date is skewed left, the median will be higher.

Spread of Quantitative Data

Recall our example from class, where we considered the high temperatures on January 26 in two cities for five randomly selected years since 1900:

City A: 37, 39, 36, 41, 27 City B: 55, 19, 37, 16, 53

For both of these lists, the mean is 36 and the median is 37, but there is a clear difference: the data for City A is very consistent while the data for City B is much more volatile.

To measure this "spread" of quantitative data, if we collect n numbers $x_1, x_2, \ldots, x_n$, we define the **sample variance** by

$$\frac{\sum(x_i - \overline{x})^2}{n-1},$$

and the **standard deviation** is $s = \sqrt{\text{sample variance}}$. In the example, the sample variance for City A is

$$\frac{(37-36)^2 + (39-36)^2 + (36-36)^2 + (41-36)^2 + (27-36)^2}{4} = 29,$$

so the standard deviation is $s = \sqrt{29} \approx 5.385\ldots$

If we are considering a full population instead of a sample, we use σ instead of s to denote the standard deviation. Once again, the idea is that, if the sample is good, s is a good approximation for σ.

95% Rule: If the data for a quantitative variable is bell-shaped, then roughly 95% of the individual cases should lie within two standard deviations of the mean. In our example, this would say that for a randomly selected year, there is about a 95% chance that the high temperature on January 26 will fall between $36 - 2(5.385\ldots) \approx 25.23$ and $36 + 2(5.385\ldots) \approx 46.77\ldots$

The **z-score** for a value, x, of a quantitative variable, defined by

$$z = \frac{x - \overline{x}}{s} \quad \left(\text{or } \frac{x-\mu}{\sigma} \text{ if you have the full population data}\right),$$

measures how many standard deviations x is above or below the mean. This is a unitless measurement of the "extremity" of a single data point. A positive z-score means x is above the mean, while a negative z-score means x is below the mean.

The 95% can then be rephrased as saying that roughly 95% of the values of a bell-shaped quantitative variable should have z-score between -2 and 2.

In our example, if the high temperature in City A on January 26, 1971 was 49 degrees, that would be a z-score of

$$\frac{49-36}{5.385\ldots} \approx 2.414\ldots,$$

which indicates a pretty extreme value.

In a list of quantitative data, an **outlier** is a number that is notably distinct from the rest of the list. As we saw in class with our annual income example, outliers can have a major impact on the mean of the list, but a much smaller impact on the median. Analogously, outliers have a major impact on the standard deviation, so it is useful to have alternative notions of spread that are resistant to outliers, like the median is.

We define the P^{th} percentile for a list of quantitative data to be, in principle, the smallest number that is greater than $P\%$ of the data. In particular, we focus on:

$$\textbf{first quartile: } Q_1 = 25^{\text{th}} = \text{ median of the "lower half"}$$

and

$$\textbf{third quartile: } Q_3 = 75^{\text{th}} = \text{ median of the "upper half".}$$

We define the **Inner Quartile Range**, or IQR, by IQR$= Q_3 - Q_1$, which is a good single number measurement of spread that is resistant to outliers.

In our example, the median for both cities temperature data is 37, so for City A, the "lower half" of the list is $27, 36$ and the "upper half" is $39, 41$, hence $Q_1 = 31.5$ and $Q_3 = 40$, and IQR$= 8.5$. Similarly, for City B, $Q_1 = 17.5$ and $Q_3 = 54$, so IQR$= 36.5$.

The first and third quartiles, together with the median and the smallest and largest value on the list, form what we call the **five number summary**. In increasing order, the five number summary is

$$\text{minimum}, Q_1, \text{median}, Q_3, \text{maximum}.$$

Boxplots and Side-by-Side Graphs

A **boxplot** is a visual representation of the five number summary, together with an identification of outliers. Specifically, we make a numerical scale, appropriate for the data, on a horizontal axis and draw:

- A box stretching from Q_1 to Q_3
- A line that divides the box drawn at the median.
- A line from each quartile to the most extreme data value that is NOT an outlier
- Label each outlier with something like a dot or asterisk

Check out section 2.4 of your textbook for some good examples and illustrations of boxplots. While our general definition of an outlier is quite vague, for the purpose of boxplots we use something more concrete. Specifically, for boxplots, an outlier is defined as anything less than $Q_1 - 1.5\text{IQR}$ or greater than $Q_3 + 1.5\text{IQR}$.

If we want to investigate an association between one categorical variable and one quantitative variable, we can use a **side-by-side graph**, in which we divide the full collection of data into groups according to the values of the categorical variable, then make a boxplot or bar chart for each group using the quantitative variable, and finally put the boxplots or bar charts side by side. For example, if we wanted to observe an association between gender and height amongst 1000 randomly selected U.S. adults, we could divide the 1000 cases into males and females, make a boxplot from the heights of the men, make a boxplot for the heights of the women, and then put the boxplots side by side. For examples of side-by-side graphs, see the end of Section 2.4 in the textbook.

Scatterplots and Correlation

A **scatterplot** is a way to visually investigate the relationship between two quantitative variables. Specifically, we plot for each case an ordered pair (x, y) in a plane by plotting the value of one quantitative variable on the x-axis, and the other on the y-axis.

One particular kind of relationship between quantitative variables that a scatterplot may reveal is a linear relationship. In other words, the scatterplot may roughly resemble a straight line. **Correlation** is a measurement of the strength and direction of the linear association between two quantitative variables, which in the case of a sample we denote by r, and in the case of a full population we denote by ρ. For now we'll use r for our discussion.

There is a formula for correlation, but we won't emphasize it here. Rather, we focus on certain key properties:

- r is always between -1 and 1
- the sign of r ($+$ or $-$) indicates the direction of linear association: $+$ is $\nearrow$, $-$ is $\searrow$
- r values close to 1 or -1 indicate strong linear association, while r values close to 0 indicate weak or no linear association

See Section 2.5 in the text for lots of good pictures and examples for correlation.

Linear Regression

While correlation measures how closely a scatterplot resembles a straight line, a natural follow-up question to ask is exactly WHICH line the scatterplot resembles. The process for determining the best choice of line is called **linear regression**, and the line itself is called the **regression line**. Like any line, the regression line is given be a unique slope-intercept form equation, which we write as

$$\hat{y} = a + bx.$$

The convention when doing linear regression is to put what you are thinking of as the explanatory variable on the x-axis and what you are thinking of as the response variable on the y-axis. The slope of the regression line, b, represents how much we expect the response variable to increase if the explanatory variable goes up by 1.

We use $\hat{y}$ for the y-coordinates on the line, which are the "predicted values" of the response variable, while we reserve plain y for the y-coordinates on the scatterplot, which are the actual values of the response variable.

Like with correlation, we will not be doing linear regression by hand, but you should be comfortable doing it on your calculator. What your calculator is doing to find the regression line is to minimize the sum of the squares of the **residuals**, which are the differences

$$\text{residual} = y - \hat{y}$$

between the predicted and observed values of the response variable.

Here is a table with some of the formulas and notation we have discussed:

Term	**Formula**	**Population Notation**	**Sample Notation**
Proportion	$\frac{\#\text{ of cases in the specified category}}{\text{total }\#\text{ of cases}}$	p	$\hat{p}$
Mean	$\frac{\sum x_i}{n}$	μ	$\overline{x}$
Standard Deviation	$\sqrt{\frac{\sum(x_i-\overline{x})^2}{n-1}}$	σ	s
z-score	$\frac{x-\overline{x}}{s}$ or $\frac{x-\mu}{\sigma}$	z	z
median	midpoint of list in increasing order		
First Quartile	median of "lower half"	Q_1	Q_1
Third Quartile	median of "upper half"	Q_3	Q_3
Inner-quartile range	$Q_3 - Q_1$	IQR	IQR
Outliers for Boxplots	Less than $Q_1 - 1.5IQR$ or greater than $Q_3 + 1.5IQR$		
Correlation	See Note Below	ρ	r

As noted on the previous page, there is a formula for correlation but we won't use it in class, as we won't be computing correlation by hand. Rather, you should be comfortable with the properties listed on the previous page, and finding correlation with your calculator.