Lecture 3

Nancy Pfenning Stats 1000

We learned last time how to construct a stemplot to display a single quantitative variable. A **back-to-back** stemplot is a useful display tool when we are interested in comparing the values of a single quantitative variable for two categorical groups.

Example

Let's use a back-to-back stemplot to compare earnings (in thousands of dollars) of 28 male and 51 female students. Since the earnings range from 0 to 22 thousand, we will split the stems 0, 1, and 2 five ways each. [Note that besides the quantitative variable "earnings", we are adding in a categorical variable "sex" that has two possible values, male and female.] Sharing the same stems, male earnings precede them right to left, while female earnings follow the stems left to right.

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

The center is clearly higher for the males (midpoint at 3) than for the females (midpoint at 2). Male earnings range from 0 to 22 thousand, whereas female earnings range from 0 to 15 thousand. However, the spreads appear comparable if we disregard the high outliers. Shapes are very skewed to the right, as is often the case with monetary variables such as earnings, costs of homes, etc. Both distributions have a single peak in the low thousands: 2 or 3 thousand dollars was a common amount for both sexes. Note that some of the stems have no leaves. These stems must **not** be omitted, otherwise we could not see outliers for what they are.

Since we see a tendency in this particular class for males to earn more than females, it is natural to wonder whether the same conclusion can be drawn about Pitt students, or all college students, in general. Our ultimate goal in this course is to go beyond the data at hand and draw conclusions about the larger population from which the data originated, a process called statistical inference. This requires careful development of needed theory over the course of the semester.

Up to now, we've mentioned "center" as simply the midpoint (median), and "spread" as the range. These only provide limited information from a couple of observations. Since center and spread are the most important features of a distribution, they should be defined carefully.

One measure of center is the **median**, or middle value. There is a single middle value for an odd number of observations. For an even number of observations, we take the median to be the average of the two middle values.

Example

The median earnings of the 28 male students is the average of the 14th and 15th, or $\frac{3+3}{2} = 3$ thousand dollars. The median earnings of the 51 female students is the 26th value, 2 thousand dollars. We can say that the typical male student earns 1 thousand dollars more than the typical female student.

Example

The median of 11 Math SAT scores is the 6th, or 592.

468 472 511 534 557 **592** 592 614 667 669 704

Another measure of center is the **mean**, or arithmetic average: just add up all the numbers and divide by how many there are. The mean of n observations $x_1, x_2, \cdots, x_n$ is denoted

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n}\Sigma x_i$$

Example

The mean earnings of 28 males is

$$\bar{x} = \frac{0 + \cdots + 22}{28} = 5.8$$

The mean earnings of 51 females is

$$\bar{x} = \frac{0 + \cdots + 15}{51} = 3.7$$

Example

The mean of 11 Math SAT scores is

$$\bar{x} = \frac{446 + \cdots + 682}{11} = 580$$

For fairly symmetric distributions, like the distribution of 11 SAT scores, mean and median are approximately the same. [median 592 vs. mean 580]. For a distribution that is skewed left or has low outliers, like age at death of all Americans, the mean tends to be less than the median. For a distribution that is skewed right or has high outliers, such as earnings of males or females, the mean tends to be greater than the median [males: median 3 vs. mean 5.8; females: median 2 vs. mean 3.7]. In general we prefer the mean as a measure of center because it includes information from all the observations. However, if a distribution has pronounced skewness or outliers, the median may be better because it is less affected by those few extreme values. For this reason, we call median a **resistant** measure of center.

If we use median as our measure of center, we can use **quartiles** to help describe the spread: they tell us where the middle half of the data values occur. The lower quartile (Q1) has one fourth of the data below it; it is the middle of the values below the median. The upper quartile (Q3) has three fourths of the data below it; it is the middle of the values above the median. For an odd number of values, we will exclude the median when finding the middles of the values below and above it. Software, or even other textbooks, may use a different algorithm and produce slightly different quartiles.

Example

Let's find the quartiles for earnings of male and female students. For the 28 males, Q1 is the middle of the lower 14 values, that is, the average of the 7th and 8th: $\frac{2+2}{2} = 2$. Q3 is the middle of the upper 14 values, that is, the average of the 21st and 22nd: $\frac{8+10}{2} = 9$. For the 51 female earnings, Q1 is the middle of the lower 25 values, or 13th, which is 1. Q3 is the middle of the upper 25 values, or 39th, which is 5.

The **Five Number Summary** is a good way to describe a quantitative data set. It lists the minimum, Q1, median, Q3, and maximum.

Example

The Five Number Summary for male earnings is

$$0 \quad 2 \quad 3 \quad 9 \quad 22$$

The Five Number Summary for female earnings is

$$0 \quad 1 \quad 2 \quad 5 \quad 15$$

Note that only one quarter of the males earned 2 thousand or less, whereas one half of the females earned 2 thousand or less.

A **boxplot** lets us take in the information from the Five Number Summary visually.

1. The bottom whisker extends to the minimum.
2. The bottom of the box is at Q1.
3. There is a line through the box at the median.
4. The top of the box is at Q3.
5. The top whisker extends to the maximum.

For a modified boxplot, denote outliers with a "*" or other symbol, and extend whiskers to the minimum and maximum non-outliers.

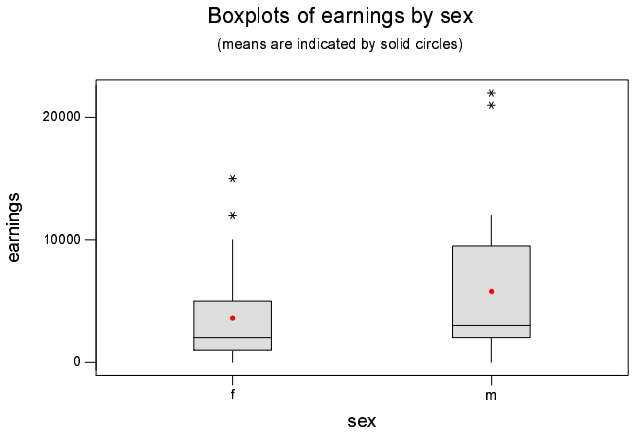
A simple criterion to identify outliers is based on the **interquartile range**, $IQR = Q3 - Q1$, which tells the range of the middle half of the data. Any value below $Q1 - 1.5 \times IQR$ will be considered a low outlier, and any value above $Q3 + 1.5 \times IQR$ will be considered a high outlier.

Example

We can draw **side-by-side boxplots** of the male and female students' earnings for a good visual comparison. The males' IQR is $9 - 2 = 7$, so low outliers would be below $2 - 1.5 \times 7 = -8.5$ (of course there are none) and high outliers would be above $9 + 1.5 \times 7 = 19.5$. Thus, the values 21 and 22 would be considered high outliers for the males. The females' IQR is $5 - 1 = 4$, so low outliers would be below $1 - 1.5 \times 4 = -5$ (there are none) and high outliers would be above $5 + 1.5 \times 4 = 11$. Thus, the values 12 and 15 would be considered high outliers for the females.

[Note: side-by-side boxplots have the advantage over back-to-back stemplots in that we can

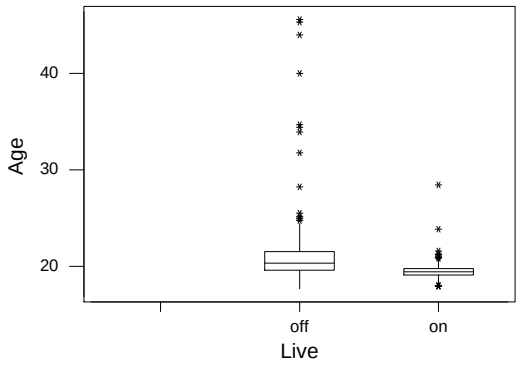
compare more than two distributions at a time.]



Clearly, earnings are higher for the males. Disregarding the outliers, overall spreads are comparable, but the middle half of the males' earnings has a considerably wider range than that for the females.

Example

Let's consider the values of the variable "age" compared for off and on campus students. I expect off campus students to be older overall (higher center), with more spread (because off campus there may be students beyond their twenties), and right skewness/high outliers for both groups, more so for off campus students.



Descriptive Statistics: Age by Live

Variable	Live	N	N*	Mean	Median	TrMean
Age	off	222	1	21.253	20.330	20.630
	on	222	0	19.488	19.420	19.424
	*	0	1	*	*	*

Variable	Live	StDev	SE Mean	Minimum	Maximum	Q1
Age	off	3.824	0.257	17.670	45.580	19.580
	on	0.901	0.060	17.920	28.420	19.080

```
                  *          *          *          *          *          *

Variable   Live             Q3
Age        off          21.520
           on           19.750
                  *          *
* NOTE * N missing = 2
```

Side-by-side boxplots confirm that center and spread are both greater for off-campus students, and there are many high outliers in both groups. Surprisingly, a low outlier appears for the on-campus students. Five Number Summary values are 17.67, 19.58, 20.33, 21.52, and 45.58 for off-campus, 17.92, 19.08, 19.42, 19.75, and 28.42 for on-campus. According to the 1.5*IQR Rule, boundaries for low and high outliers are 16.67 and 24.43 for off-campus students (none are below the lower bound; upper bound exceeded by many); 18.075 and 20.755 for on-campus students (there are a few below the lower bound, and many above the upper bound).

Exercise: Consider the values of one quantitative variable in our survey compared for two categorical groups. First, state your expectations about how the quantitative values would compare for the two groups. Then use MINITAB to get side-by-side boxplots and report the Five Number Summary for each. Tell how their centers, spreads, and shapes compare. Use the 1.5*IQR Rule to report the boundaries for low and high outliers in both groups, and tell whether there are any outliers according to the Rule.

Lecture 4

The median is an OK measure of center, especially in the case of skewness or outliers, but in general the mean is our preferred measure of center. The measure of spread to accompany the mean is the **standard deviation** s, or square root of the "average" squared deviation from the mean. s tells us how far the observations tend to be from their mean $\bar{x}$. If we solve first for the average squared deviation from the mean, or **variance** s^2, and then take its square root to find s, we can write the variance and standard deviation of n observations $x_1, x_2, \cdots, x_n$ as

$$s^2 = \frac{(x_1 - \bar{x})^2 + \cdots (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1}\Sigma(x_i - \bar{x})^2$$

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

[It is natural for students to wonder why we divide by n-1 instead of n. Ultimately, variance s^2 from a sample is used to estimate the variance of the entire population. It does a better job of estimating when we divide by $n-1$ instead of n.]

Example

468 472 511 534 557 592 592 614 667 669 704

It can be shown that the 11 Math SAT scores have mean $\bar{x} = 580$, standard deviation $s = 80$. How do we interpret these numbers? They are telling us that these students typically scored about 580, give or take about 80 points.

How do we calculate the standard deviation by hand? We must find square root of average squared deviation from the mean:

1. Find the mean: 580.
2. Find the deviations from the mean:

$$468 - 580 = -112, \quad 472 - 580 = -108, \cdots, 704 - 580 = 124$$

3. Find the squared deviations from the mean:

$$-112^2 = 12544, \quad -108^2 = 11664, \cdots, 124^2 = 15376$$

4. "Average" the squared deviations, dividing their sum by **the number of observations minus one**. This gives us the variance s^2.

$$s^2 = \frac{12,544 + 11,664 + \cdots + 15,376}{11 - 1} = \frac{63,924}{10} = 6392$$

5. Take the square root of the variance to find the standard deviation:

$$s = \sqrt{s^2} = \sqrt{6392} \approx 80$$

Example

For the male earnings, we can calculate $\bar{x} = 5.8$ thousand and $s = 5.7$ thousand. This tells us the "typical distance" of their earnings from their mean, 5.8, is 5.7 thousand dollars. Are they really that far away? If we exclude the outliers 21 and 22, $\bar{x} = 4.6$ and $s = 3.7$. The outliers had substantially inflated the value of the mean, and also the value of the standard deviation.

For a data set with outliers, caution should be used in describing its center and spread with mean and standard deviation, because their values can be severely affected by just one or a few extreme observations. For such data sets, **resistant measures** like median and quartiles should be used. They are hardly affected at all if one or a few extreme observations are included or not. The distribution of SAT scores was fairly symmetric and outlier-free, so standard deviation should provide an adequate measure of spread. The earnings have high outliers and right skewness, and so would be best summarized with the Five Number Summary.

In many cases, an outlier occurs because of faulty recording of data. A student may report her height in inches as "52" when she means "5 feet 2 inches". Or I may mistype a height of 62 inches as "662". In these instances, the outliers should be corrected or eliminated.

2.7: Bell-Shaped Distributions of Numbers

Some quantitative variables, such as SAT scores, have a distribution with a symmetric, single-peaked shape. Such a shape occurs naturally in all sorts of contexts.

Example

Suppose I constructed a histogram for heights of 50 female students, using classes of width 2 inches. Then the total area taken up by the histogram's rectangles would be $100\% * 2\text{inches} = 200\%\text{inches}$. Let's divide the percentages by 2 inches, so that our vertical scale is now "percent per inch" and the total area will be 100%. Then the *area* of any block tells the percentage of females in that height range. For example, if the median height is 65 inches, then the area under the histogram to the left of 65 would be 50%.

In general, if the vertical scale of *any* histogram is adjusted so that the total area of all rectangles together is 1 or 100%, then the area of the rectangles over any interval tells us the proportion or percentage of observations which fall in that interval. Imagine making more and more observations on a continuous variable, like height of female college students in this class, at Pitt, in the U.S., ... and recording their values to the utmost accuracy. Then the profile of our histogram would be smoothed out. Idealized, we would have a smooth curve.

A **density curve** is an idealized representation of a distribution, where the area under the curve between any two values gives the proportion or percentage of observations which fall between those values. By this

construction, the total area under a density curve must be 1 or 100%. Whereas a frequency histogram displays sample data values, a density curve displays the behavior of a continuous quantitative variable for an entire population. We denoted the mean of sample data as $\bar{x}$ and the standard deviation s. Now, we denote the mean of a density curve with the Greek letter μ (called "mu") and the standard deviation with the Greek letter σ (called "sigma").

The density curve for heights of males or females in a certain age group, like many naturally occuring density curves, follows a symmetric bell-shape called **normal**. Besides providing a good model for many actual data sets (eg. heights, weights, test scores, measurement errors, etc.), the normal curve also approximates typical long-run random behavior (eg. dicerolls, coinflips). Most importantly, it approximates the shape of the distribution of sample mean or sample proportion when large enough random samples are taken from a quantitative or categorical population whose shape is not necessarily normal.

The normal curve is symmetric about its mean μ, indicating that it is just as likely for the variable to take a value below its mean as above. It is single-peaked and bell-shaped, with tapering ends, showing that values closest to the mean are most likely, and values further from the mean are increasingly less common. The total area, as for any density curve, must be 100% (or 1).

In fact, the mean μ and standard deviation σ tell us *everything* about a normal distribution, but it is easiest to begin by specifying three useful landmarks on the normal curve.

Empirical Rule

For any normal curve, approximately

68% of the values fall within 1 standard deviation of the mean;
95% of the values fall within 2 standard deviations of the mean;
99.7% of the values fall within 3 standard deviations of the mean.

Example

The distribution of verbal SAT scores in a certain population is normal with mean $\mu = 500$, standard deviation $\sigma = 100$. What does the Empirical Rule tell us about the distribution of scores?

68% fall within 100 of 500, i.e. between 400 and 600.

95% fall within $2 * 100$ of 500, i.e. within 200 of 500, between 300 and 700.

99.7% fall within $3 * 100$ of 500, i.e. within 300 of 500, between 200 and 800.

Standardizing Normal Values

A way of assessing a particular value of any normal distribution is to identify how many standard deviations below or above the mean it is. We do this by finding its **standardized score**, or **z-score**:

$$z = \frac{\text{observed value} - \text{mean}}{\text{standard deviation}}$$

The z-score will be positive if the value is above the mean, negative if it is below the mean.

Example

Say a variable is normal with mean 50, standard deviation 10. If it takes the value 70, what is its standardized value? The standardized value of 70 is $\frac{70-50}{10} = +2$. For a variable with mean 50, standard deviation 10, the value 70 is 2 standard deviations above the mean.

Example

Say a variable is normal with mean 3.6, standard deviation .2. What is the standardized value of 3.3?

$$z = \frac{3.3 - 3.6}{.2} = \frac{-.3}{.2} = -1.5$$

In other words, 3.3 is 1.5 standard deviations below the mean.

Example

The distribution of heights of young women in the U.S. is normal with mean 65, standard deviation 2.7. The distribution of heights of young men in the U.S. is normal with mean 69, standard deviation 3.

Who is taller relative to other members of their sex—Jane at 71 inches or Joe at 75 inches?

Jane's standardized height is $\frac{71-65}{2.7} = 2.22$. Joe's standardized height is $\frac{75-69}{3} = 2.00$. Jane is taller, for a woman, than Joe is, for a man.

Lecture 5

Chapter 3: Gathering Useful Data

In the early part of this course (Chapters 2, 5, and 6), we use **descriptive statistics** to summarize the data at hand, that is, we summarize the **sample**. Our ultimate goal in the course (Chapters 10 to 16) is to use **inferential statistics** to draw conclusions about the larger **population** from which our sample originated. Such inferences can only be made if the sample data are truly representative of the population with regard to the question of interest.

Example

I could use heights of female students in this class to draw conclusions about heights of all college females. But I could not use SAT scores of class members to draw conclusions about SAT scores of all college students, because Pitt Stat 200 students would not be representative of all college students with regard to SAT scores.

Example

Larry Flynt's sample of girls who had posed for his magazines would not be representative of the general population of women with regard to whether or not they see pornography as being exploitative.

The simplest way—in theory, at least—to guarantee that the sample truly represents the population is to take a **simple random sample**, where every possible group of a given size has the same chance of being selected. This is sampling **at random** and **without replacement**. In practice, samples are chosen in a wide variety of ways. Common sense is often the best guideline in assessing whether or not a sample truly represents the population.

Statistical data are gathered via two basic types of research studies: observational studies and experiments. In an observational study, researchers note values of the variables of interest as they naturally occur. In an experiment, researchers impose a treatment, manipulating the explanatory variable so they can see the effect on the response variable of interest.

If researchers find out what type of sunscreen people use, and how much time they spend in the sun, then they are conducting an observational study. If researchers provide people with one or the other type of sunscreen and then find out how much time is spent in the sun, they are conducting an experiment. A **randomized experiment** is one in which the treatments are assigned at random, a good way to control for possible confounding variables, those which are tied in with the explanatory variable and may affect

the response of interest. If researchers had let the **subjects** (participants in an experiment) choose which sunscreen to use, their inclination to spend—or avoid—time in the sun could influence their choice, and also impact how much time they sunbathed. Another word for a confounding variable is a **lurking variable**: it is lurking in the background, clouding the issue of interest.

Example

I recorded students' weights and how much time they spent on the phone in a given day (two quantitative variables), and found that students who weighed more tended to spend less time on the phone. Does phone time really go down as weight goes up? A possible confounding variable would be gender: we know males tend to weigh more than females, and perhaps they tend to spend less time on the phone. In order to control for a confounding variable, study similar groups separately—that is, look at the relationship between weight and phone time for women, then for men. As if by magic, the relationship vanishes: in fact, weight has nothing to do with time spent on the phone.

Designing Experiments

Example

Does sugar cause hyperactivity in children? How can this be tested?

In order to prevent confounding variables from clouding the issue, researchers could conduct an experiment. The subjects would most likely be volunteers. There are two variables of interest: sugar intake is the explanatory variable and activity level the response. Each of these has the potential of being handled as quantitative or categorical. To keep things simple, let's take both of them to be categorical: sugar intake is low or high, activity level is normal or hyper.

The critical stage in which to employ randomization in an experiment is during the assignment of treatments: some children should be randomly assigned to higher levels of sugar. Should they be given Count Chocula cereal for breakfast while their counterparts are given half a grapefruit? No: the **treatment group** receiving sugar should be compared to a **control group** which is treated identically in all other respects. Thus, the children should be provided with the same diet, except that one group has foods sweetened with sugar, the other with an artificial sweetener. [When the treatment is a drug, the control is a **placebo** pill. Researchers know that subjects often respond to the *idea* of being treated, and are careful to prevent this from confounding their results.] Because many people have heard that sugar causes hyperactivity, if subjects knew they were given additional sugar, they might alter their behavior. Thus, the subjects should be **blind**, that is, not know whether they're given sugar or an artifical sweetener. [Is this really possible?] What if researchers know whether or not a child received sugar when it comes time to assess activity level? Since this evaluation could be rather subjective, it is important that the researcher not be aware of which treatment a subject received. In other words, the experiment should ideally be **double-blind**.

Activity habits and levels vary greatly from family to family, depending on region, socio-economic status, etc. One way to control for all of these influences would be to use a **matched pairs** design: select two siblings from each family and randomly assign one to the sugar diet, the other to an artificial sweetener.

Note: the most common matched pairs design is one in which the same individual is evaluated for both treatment and control, such as in a before-and-after study. If the order of treatments could play a role in the response, then order should be randomized. For example, if we want to see if people prefer Pepsi or Coke, each individual can taste both drinks (blind, of course), but which is tasted first should be determined at random, say by a coin flip.

Besides matched pairs designs, another way of controlling for outside variables is **blocking**, that is, dividing the units first into groups that are similar in a way that may play a role in their responses, then randomize assignment to treatments within these blocks. If age is important for

activity level, divide the children first into younger, medium, and older groups. If gender may be important, divide first into males and females.

Example

A recent newspaper article entitled **Heights fears ease with pill** reports: "In a small study released yesterday, a drug already on the market for tuberculosis helped people who were terrified of heights get over that fear with only two therapy sessions, instead of the usual seven or eight. The study, led by Michael Davis, a professor of psychiatry and behavioral sciences at the Emory University School of Medicine, was described at a session about unlearning fears at the Society for Neuroscience meeting. Davis based his work on research that had found the transmission of a certain protein to a brain receptor was critical to overcoming fear. He found the TB drug, D-cycloserine, aids transmission of the protein."

This was an experiment, because clearly the drug was administered to study participants by the researchers (as opposed to observing differences in height fears for people who do and do not happen to take that particular drug). The explanatory variable of interest is whether or not the TB drug is taken (a categorical variable). The response variable records the effectiveness of therapy by counting how many sessions are needed for patients to overcome their fear of heights—thus, it is a quantitative variable. The subjects were apparently a small group of people who were terrified of heights. The treatment was the TB drug D-cycloserine, and we can assume that researchers randomly assigned the drug or a placebo to patients in order to make the comparison. We can also assume the study was double-blind, because a reputable researcher would not compromise his study by allowing the experimenter effect to enter in.

Exercise: Find an article or report about an experiment. Tell what the variables of interest are, whether they are quantitative or categorical, and which is explanatory and response. Describe the subjects, treatments, whether or not the study was blind, etc.