Lecture Notes for Stat 1000 by Dr. Nancy Pfenning

Note: Survey/article exercises must be handed in to me in lecture along with the textbook homework problems by the due date. They must be your own individual work. Each is worth a maximum of 2 points. For problems involving survey variables, access the survey data via my website `nancyp+@pitt.edu www.pitt.edu/ nancyp/stat-1000/index.html` where there is a link to the most recent survey data at `surveymm-dd-yy.txt` followed by instructions for downloading into MINITAB. Be sure to choose different variables from those used in lecture examples. For problems involving your own newspaper articles or internet reports, you must hand in a copy of the article or report itself.

Lecture 1

Chapter 1: Statistics Success Stories and Cautionary Tales

Statistics is a collection of procedures and principles for gathering data and analyzing information in order to help people make decisions when faced with uncertainty. Some years ago, Rutger's football team managed to rack up seven turnovers from Pitt's team, but Pitt won 29-17. Pitt's coach Walt Harris remarked, "The scoreboard is what matters. Statistics are for proverbial losers." Pitt's team may have come out ahead in spite of the statistics involved, but studying statistics can actually put you ahead of the game in that it will help you understand the world around you much better than you would otherwise.

Consider the following quotation: "The country is hungry for information; everything of a statistical character, or even a statistical appearance, is taken up with an eagerness that is almost pathetic; the community have not yet learned to be half skeptical and critical enough in respect to such statements." If this was true back in the 1870's when spoken by General Francis A. Walker (superintendent of the 1870 census) how much more true is it today, when we are bombarded with information of a statistical nature in virtually all aspects of our lives?

Here are some examples of questions which can be answered with the help of statistics:

Example

Suppose we have the following data on how much money three students earned (in thousands of dollars) last year:

Name	Earned
Jessica	10
Nicole	0
Brian	2

If someone asked us to summarize the information, we could simply state that the students earned 10, 0, and 2 thousand dollars, respectively.

Now imagine data given for an entire class of 80 or 90 students. Could we look at the list of all of their earnings and discover an overall pattern? If so, could we identify any clear exceptions to this pattern? How could we briefly summarize the data, using just a few words or numbers? How were the data produced and measured? My data came from a **sample** of students attending class. Could we use information about their earnings to draw conclusions about the earnings of the **population** of all Pitt students? How reliable would those conclusions be?

First (Chapter 2), we perform **data analysis**, summarizing the data at hand with graphical displays or key numerical and verbal descriptions. Next (Chapters 3 and 4) we'll find out about good **data production**, via experiments, observational studies, or surveys. Then we look at **relationships**: between two quantitative variables in Chapter 5, between two categorical variables in Chapter 6. Next (Chapter 7), we'll establish enough groundwork in **probability** so that we can understand the behavior of **random variables** (Chapter 8). In Chapter 9 we focus on particular random variables, **sample mean** and **sample proportion**, in order to establish the theory needed to perform **statistical inference** in Chapters 11 through 16: given information from a random sample, we will draw conclusions about the entire population from which the

sample was obtained. All of this will be facilitated with **MINITAB**, an easy-to-use statistical package. [Altogether five recitations are to be held in the Stats Lab, including the first two.]

Example

In May, 2000, .56 of 1,012 respondents to an Associated Press survey supported gays' rights to inherit from their partners. "The AP poll of 1,012 people was taken May 17-21. Its error margin was plus or minus 3 percentage points, slightly larger for the split sample." What do those 3 percentage points mean, and how were they calculated? In fact, methods of statistics will eventually tell us (in Chapter 10) that we can be pretty sure that the percentage of *all* American adults who support gays' rights to inherit is within 3% of 56%. We call 3% the **margin of error**. A rough approximation for the margin of error can be found by taking 1 divided by the square root of the sample size:

$$\frac{1}{\sqrt{1012}} = \frac{1}{32} = .03 = 3\%$$

In general, we can be 95% sure that the population percentage comes within one margin of error of the sample percentage, as long as the sample has been chosen *at random* from the population. Assuming our 1012 American adults were sampled at random, we can be 95% sure that between 53% and 59% of all American adults support gays' rights to inherit. If about 500 each Democrats and Republicans were surveyed, then the margin of error within each group would be about $\frac{1}{\sqrt{500}} = \frac{1}{22} = .04 = 4\%$, which is why the article reports that the error margin is slightly larger for the split sample.

Example

In a recent survey of 2500 American adults, 475 (that is, 19%) said they believed money could buy happiness. What does this tell us about how all American adults feel? Now the margin of error is

$$\frac{1}{\sqrt{2500}} = \frac{1}{50} = .02 = 2\%$$

Assuming our 2500 American adults were sampled at random, we can be 95% sure that between 17% and 21% of all American adults believe money can buy happiness.

Example

Larry Flynt, publisher of Hustler Magazine, spoke to an interviewer about the issue of exploitation: "Often some [women] on the fringe such as Gloria Steinem and that bunch see pornography as being exploitative and demeaning to women, but of the thousands of girls who have posed for my magazines, I've never had one who felt she had been exploited." Can we generalize from his sample to the population of all women? No: his was by no means a representative sample, and so it really tells us nothing about how women in general feel about pornography.

Example

A recent study found that men are twice as likely as women to be struck by lightning. Should men fear for their lives? No, because the **baseline risk** for women is only 1 in 10 million; thus the risk for men is still only 1 in 5 million.

Example

Why is it that many languages have no specific word for the color blue and do not distinguish between blue and green? Researchers from Ohio State University reviewed 203 languages from around the world and levels of ultraviolet B, which in high levels damage the eye to make it less able to distinguish between blue from green. In areas with low levels of UVB, languages tended to have a word for blue while areas with high levels tended not to.

Example

Psychologists and social scientists noted that children who grow up in families with fewer kids tend to have higher IQs. Should we conclude that parents can boost their kids' IQ scores by having fewer children? As with any observational study, it is difficult here to establish proof of a cause/effect relationship. In fact, researchers have reason to believe that causation goes more in the other direction: parents with higher IQs tend to have fewer children, and by heredity, those children tend to have higher IQs.

Example

Suppose we noticed that people who use stronger sunscreen tend to stay in the sun longer. Could we conclude that stronger sunscreen leads to more time in the sun? No; a **confounding variable** could be the person's inclination to seek or avoid time in the sun, which could also influence what type of sunscreen is used. In an **observational study**, where variables' values are observed as they naturally occur, it is common for a confounding variable to cloud the issue.

Example

In a study of 87 French and Swiss college students, researchers (randomly) gave half of them sunscreen with a protection factor of 10 and the other half with a factor of 30. The students, who weren't told which lotion they had received, went on summer vacation and recorded the amount of time they spent in the sun. Users of the stronger sunscreen spent 25% more time in the sun, mostly sunbathing, because they typically waited until their skin turned red before rushing to the shade. Can we conclude that in general, using a stronger sunscreen leads people to spend more time in the sun? Yes, because an **experiment** was performed whereby researchers imposed the type of sunscreen treatment at random. This controls for possible confounding variables such as we discussed for the observational study above, and lets us draw a conclusion about cause and effect.

Example

Students who take a formal SAT prep course score "significantly" better on the SATs. What does this mean? **Statistical significance** means that it would be unlikely to see such a difference in the sample if there were actually no difference in the population. Especially with a large sample (this study involved over 14,000 students), we may be able to produce *statistical* evidence of a difference that has little *practical* significance. In fact, research shows that coaching may only improve SAT scores by about 20 points out of the possible 1600.

Chapter 2: Turning Data Into Information

Example

Consider the following raw data, which have not yet been processed:

Name	Sex	Earned	Age	Year
Jessica	f	10	22.3	4
Nicole	f	0	19.4	2
Brian	m	2	19.8	2
...	...	...	...	...

Survey results from a class can be considered **sample data** if we think of these students as being a subset of the larger **population** of all Pitt students. A number that describes the sample is called a **statistic** (.19 was the proportion of sampled adults who believed money could buy happiness), whereas a number that describes the population is called a **parameter** (the unknown proportion of all adults who believe money could buy happiness).

If all the individuals in a sample or population were the same, then there would be nothing of interest to examine and statistics would be unnecessary. But (fortunately) characteristics do vary from one individual to the next, and so we call these characteristics **variables**. A variable may be **categorical**, like sex, or **quantitative**, like earnings or age. What about year? If I'd only permitted responses of 1, 2, 3, or 4, year could be treated as a quantitative variable. But since the response "other" was also possible, we must treat year as categorical. We can call it **ordinal** because the years do follow an order from lowest to highest, unlike major, for example, which cannot be ordered.

Lecture 2

The best way to handle statistical information depends to a large extent on the number and type of variables involved. Let's identify these for the previous examples:

Example

Earnings of students is a quantitative variable.

Example

19% of 2500 Americans said they believed money could buy happiness. Just one categorical variable (believing or not) is involved.

Example

Men are twice as likely as women to be hit by lightning. Here we consider two categorical variables: gender and whether or not a person is hit by lightning. When the relationship between two variables is being considered, it helps to decide which, if any, plays the role of **explanatory** variable and which is the **response** variable. In this case gender would be the explanatory variable, and being hit by lightning or not is the response.

Example

Languages tend not to have a specific word for the color blue if they are spoken in countries with high levels of UVB. Low or high levels of UVB is the explanatory variable, and having a word for the color blue or not is the response; both are treated as categorical variables.

Example

Do children in smaller families have higher IQs? We are interested in two quantitative variables: family size and IQ score. Researchers originally thought of family size as being the explanatory variable, but then they realized that it is closer to being the response, explained by IQ of parents and heredity.

Example

Do people stay in the sun longer if they use a stronger sunscreen? Type of sunscreen is the explanatory variable, and it's categorical. Time in the sun is the response, and it's quantitative.

Example

Students who take SAT prep courses score "significantly" better. SAT score is a quantitative variable. If we compared scores for two groups, those who did and did not take a prep course, then we'd be introducing an additional categorical variable.

Example

A December 2003 New York Times article stated: "It's not your imagination; it really is taking longer to get there. Scheduled travel time between many major cities by air, rail and bus all increased from 1995 to last year, according to the Transportation Department's Bureau of Transportation Statistics. The bureau studied 261 city-pair markets and found that in 68 percent of them scheduled air travel time increased for direct service. The scheduled trip took longer by train in 61 percent of those city pairs and by bus in 52 percent." The variables involved are (change in) scheduled travel time and mode of transportation. As reported, travel time is not summarized quantitatively; rather, it was recorded whether or not the time increased from 1995 to 2002—a categorical variable. Mode of transportation is a categorical variable allowing for three possibilties. Whether travel time increased is summarized with percentages, and these are cited for the various modes of transportation. There is not much emphasis on making comparisons for plane vs. train vs. bus, but if we wanted to make an assignment of explanatory/response, mode of transportation would be explanatory and whether time increased would be response.

Exercise: Hand in an article or report about a statistical study; tell what variable or variables are involved and whether they are quantitative or categorical. If there are two variables, tell which is explanatory and which is response.

Chapter 2: Turning Data Into Information

Now we'll begin to learn to how to display and summarize data depending on the number and type of variables involved.

There are various ways to display and summarize data, depending on whether there are quantitative or categorical variables (or some combination) involved. We summarize categorical variables by recording the count, or (usually preferable) the percent or proportion in the category of interest.

Example

In the 1311 crimes in Maryland from 1978 to 1999 where the defendants were eligible for the death penalty, 690 involved black victims. There is one categorical variable involved here, namely race of the victim. The percentage involving black victims was $\frac{690}{1311} = 53\%$ and so the percentage involving white victims was 47%. This information could be displayed with a piechart (a 53% slice for black victims and a 47% slice for white) or a bar graph (bars of height 53% for black victims, 47% for white). Either way, we see the percentages of victims in the two races are comparable, both close to 50%.

If two categorical variables are involved, we can record counts in the various category combinations with a **two-way table**. Once we've determined which is the explanatory variable, we can compare percentages in the response category of interest for each of the explanatory groups.

Example

Here is a two-way table for the Maryland crimes, now classified not only according to the race of the victim, but also as to whether or not the defendant was given the death penalty:

	Death Penalty	No Death Penalty	Total
Black Victim	15	675	690
White Victim	61	560	621
Total	76	1235	1311

Now there are two categorical variables involved. It only makes sense to take the victim's race to be the explanatory variable, and whether or not the death penalty was imposed would be the response. (Why would the reverse be nonsensical?) Thus, we would compare percentages

sentenced to death for the case of black victim $\frac{15}{690} = .022$ vs. white victim $\frac{61}{621} = .098$. To display the relationship between two categorical variables we list the possible explanatory values along a horizontal axis, and represent percentages in the various response categories with bars of the appropriate height: bars of heights 2.2% and 97.8% showing the death or no death rate when the victim was black, next to bars of heights 9.8% and 90.2% showing the rates when the victim was white. The fact that the death penalty rate was more than four times higher in the case of white victims leads us to suspect that the victim's race plays a role in sentencing. In order to convince someone that this difference cannot be explained away by attributing it to chance, a statistical procedure called the **chi square test** is needed. This will be presented in Chapter 15, after the necessary theory has been developed, and we will indeed show that there is strong statistical evidence of discrimination. Interestingly, the race of the defendant did not appear to impact sentencing.

Example

In the first lecture, we considered data for how much money a large group of students earned (in thousands of dollars) the year before:

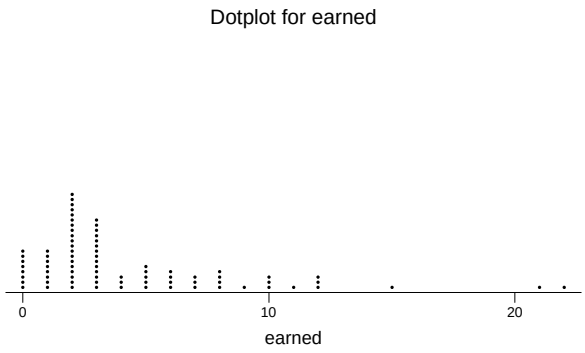
Name	Earned
Jessica	10
Nicole	0
Brian	2
...	...

To see the pattern of variation of a quantitative variable like earnings of class members, some common display tools are dotplots, stemplots, histograms, and boxplots. A good display will help us to summarize a distribution by reporting its **center**, **spread**, and **shape**. Until we learn more precise measures, we will mention the midpoint for center and the range (lowest to highest) for spread. As for shape, we will focus on whether the distribution is balanced (symmetric) or lopsided (skewed left or right), whether it has one peak or more, and whether outliers are present.

One very useful and straightforward display is the rather self-explanatory **dotplot**. A dotplot's horizontal axis corresponds to the full range of possible values; each occurrence of a value is marked with a dot in the appropriate horizontal position, and for multiple occurences the dots stack up vertically.

Example

Here is a dotplot for amounts earned (in thousands of dollars) by 79 students:



Another way to display the distribution of a quantitative data set is with a **histogram**. It is important to note that histograms differ from bar graphs in that they represent frequencies by *area*, not *height*. To construct a histogram, we

1. Divide the range of data into classes of equal width. [In this case, height and area correspond. However, it is possible to use classes of unequal width, in which case it is important to represent frequencies with *area*, not height.]
2. Count the number or percentage of observations in each class.
3. Draw the histogram, using the horizontal axis for the range of data values and the vertical axis for counts or percents.

Example

Construct a histogram for earnings of 79 students:

0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1

2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2

3 3 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 5 5 5 5 5

6 6 6 6 7 7 7 8 8 8 8 9 10 10 10 11 12 12 12 15 21 22

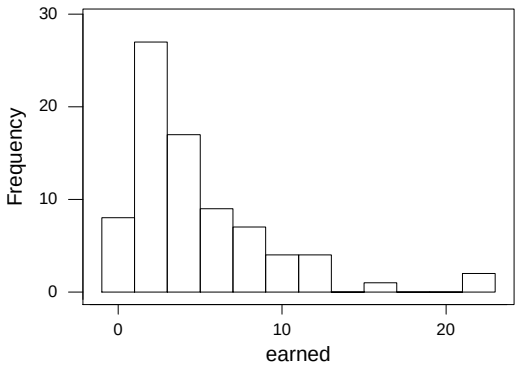
1. Since the earnings range from 0 to 22 thousand dollars, I could maybe use 5 classes of width 5, or maybe more classes of a narrower width.
2. A table helps to record the number of students with earnings in each class interval.

Class	Count	Percent
0 to 5	52	66%
5 to 10	17	22%
10 to 15	7	9%
15 to 20	1	1%
20 to 25	2	3%

3. Our horizontal axis, labeled "earnings", extends from 0 to 25. The vertical axis could be labeled "count" or "percent". If labeled with counts, there would be rectangles of height 52, 17, 7, 1, and 2. If labeled with percents, we could either divide each count by the total 79 and have rectangles of height 66, 22, 9, 1, and 3, or we could adjust the scale to represent "percent per thousand", and have rectangles of heights $\frac{66}{5}$, $\frac{22}{5}$, $\frac{9}{5}$, $\frac{1}{5}$, and $\frac{3}{5}$, resulting in a total area of 100%.

The distribution is **centered** in the 0 to 5 range, if we consider where the midpoint of all the values would be. Values are **spread** from 0 to 22, for a **range** of 22. The **shape** is extremely right-skewed with possible **outliers** in the twenties.

MINITAB opted to construct a histogram for the same data using 12 intervals of width 2:



The advantage of the histogram is that it is easily constructed for a large data set like earnings of a large group of students. For a quick, specific display of a relatively small data set, we can use a **stemplot**, consisting of a vertical list of stems, after each of which follows a horizontal list of one-digit leaves.

Example

Use a stemplot to display the Math SAT scores of eleven students:

511 592 704 667 468 592 614 472 534 669 557

Sorting the data helps us to keep organized:

468 472 511 534 557 592 592 614 667 669 704

If we used the hundreds and tens digits as stems and the ones as leaves, we would have 25 stems (46, 47, ..., 70) with only 11 leaves, not a very useful display. Instead, we will simply truncate the ones digits, then use hundreds for stems and tens for leaves:

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

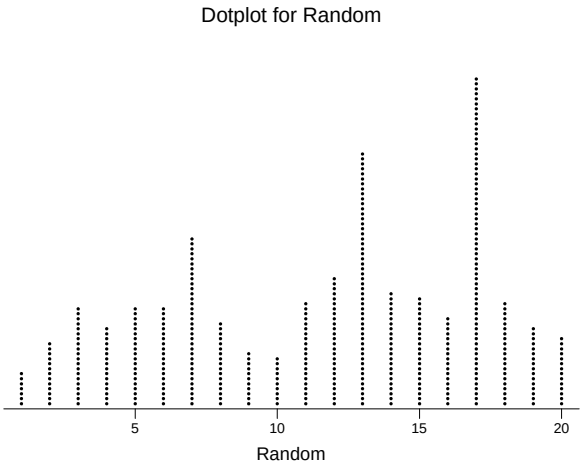
But this stemplot has rather few stems, which may prohibit us from getting a feel for the shape of the distribution. Besides truncating digits, another option in constructing a stemplot is to *split* the stems two, five, or ten ways. Splitting two ways [first stem gets leaves 0-4, second stem gets leaves 5-9] would result in this plot:

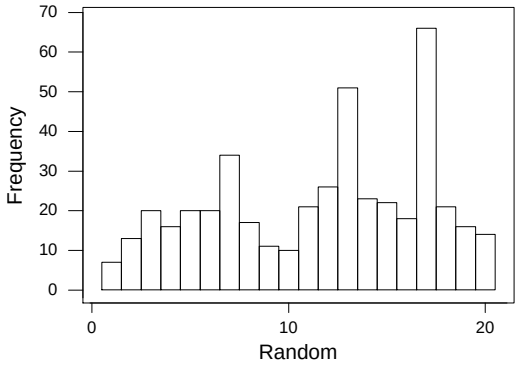
4	6	7	
5	1	3	
5	5	9	9
6	1		
6	6	6	
7	0		

We see the distribution to be **centered** at 592 (not 59!), with a **spread** of values ranging from the 4 hundreds to the 7 hundreds (specifically, from 468 to 704), and with a more or less single-peaked **shape** (called unimodal) which is fairly symmetric. There are no apparent outliers.

Example

Below are dotplot, histogram, and stemplot displaying results of the survey question which asked several hundred students to pick a number at random between 1 and 20, followed by MINITAB's descriptive statistics.






```
Stem-and-leaf of Random    N  = 446
Leaf Unit = 1.0

    7    0 1111111
   40    0 22222222222223333333333333333333
   76    0 4444444444444445555555555555555555
  130    0 66666666666666666667777777777777777777777777777777
  158    0 88888888888888888999999999
  189    1 00000000011111111111111111111
  (77)   1 22222222222222222222222223333333333333333333333333333333333333+
  180    1 444444444444444444444455555555555555555555
  135    1 666666666666666667777777777777777777777777777777777777777777+
   51    1 88888888888888888888999999999999999
   14    2 00000000000000
```

```
Descriptive Statistics: Random

Variable            N       Mean     Median     TrMean      StDev    SE Mean
Random            446     11.614     13.000     11.714      5.283      0.250

Variable       Minimum    Maximum         Q1         Q3
Random           1.000     20.000      7.000     17.000
```

The median is 13; selections range from 1 to 20, and the shape exhibits peaks at certain "popular" numbers like 17, troughs at numbers like 1, 10, or 20 that students tend to avoid. There was a tendency for students to pick numbers greater than 10, which results in some left skewness. This data set should not—and does not—contain outliers. The stemplot is not a good display in this case because it does not handle the large sample size very well (additional three's and seven's are replaced by plus signs) and doesn't let us focus easily on one number at a time.

Exercise: Pick a quantitative variable from those in the survey. Use MINITAB to display the variable's values with all three graphs discussed: a dotplot, a histogram, and a stemplot. Report the median for center, range for spread, and describe the shape. Be sure to mention if there are outliers.