# Lecture 8

**Nancy Pfenning Stats 1000**

## Pitfalls of Sample Surveys

Examples discussed in the preceding lecture illustrate some of the most common problems in taking samples for surveys:

- Using the wrong sampling frame
- Not reaching the individuals selected
- Nonresponse or volunteer response
- Self-selected sample
- Convenience or haphazard sample

The best survey design uses a sampling frame that matches the population of interest. A probability sampling design is used to make the selection at random, rather than relying on volunteers or convenience. Every effort is made to reach the individuals selected, and follow-up efforts are made to attempt to contact non-respondents.

Once a representative sample has been correctly obtained, the respondents must be surveyed in such a way as to elicit honest and accurate responses. The following pitfalls should be avoided:

1. **Deliberate bias:** A man with a clipboard stopped me on DeSoto Street one day, asking, "Ma'am, do you smoke? No? Good, then would you sign this petition to keep smokers away from non-smokers like you in the workplace?" After I declined, he confronted a man: "Sir, do you smoke? Yes? Good, you can win a free pack of Kools by agreeing to sign this petition to provide for designated smoking areas in the workplace." Questions should not be worded in such a way as to influence the response.

   After a backcountry camping trip in Glacier National Park, I was asked to complete a survey by University of Idaho researchers on interactions between backcountry travelers and grizzly bears. One question read, *"Because people and cattle live practically everywhere in the United States, and grizzly bears only in Wyoming, Montana, and Alaska, I think Montana should forego some grazing when there is a conflict with a bear: (circle one) strongly agree/agree/neither agree nor disagree/disagree/strongly disagree."* This was a leading question which put pressure on respondents to agree, and should have been more neutrally worded.

2. **Unintentional bias:** USA Today reported on a survey of 102,263 randomly selected adults in 49 states: 87% of the people rated their health as "good" to "excellent". No wonder: the question read, "Is your health generally excellent, very good, good, fair, or poor?"! The percentage rating health as good to excellent would almost certainly have been lower if the question had listed the options as "excellent, good, adequate, fair, or poor".

3. **Desire to please:** People frequently claim to have voted in a past election, even if they hadn't, because they want to satisfy the interviewer with a substantial response. To avoid this bias, a Gallup Survey question reads, "In the election in November 19–, did things come up which kept you from voting, or did you happen to vote? For whom?"

4. **Asking the uninformed:** People may claim to know something because they are embarassed—or not given an opportunity—to admit they don't. Many pollsters do not include "don't know" as an option because such answers may seem to dilute the impact of a poll. Gallup is careful to include "don't know" or "not sure" for most of their questions.

5. **Unnecessary complexity:**

(a) A personality test asks, "Do you sometimes find that you have arguments with your family members and co-workers?" What if you only argue with one sibling? Questions should be kept simple.

(b) A Gallup Surveyor in a telephone surevey asked me the following question, to which I was to assign a level at which I agreed: "I don't go out of my way to purchase low-fat foods unless they are also low in calories."

6. **Ordering of questions:** Often you can put an idea in someone's head by asking a question. This then could influence the response on subsequent questions. For example, in February 1998, U.S. president Bill Clinton was under investigation for allegedly having had an extramarital affair. A Gallup Poll asked the following two questions:

   (a) "Do you think most presidents have or have not had extramarital affairs while they were president?" [59% responded "have had".]

   (b) " Would you describe Bill Clinton's faults as worse than most other presidents, or as no worse than most other presidents?" [75% responded "no worse".]

   Do you think the percentage responding "no worse" to the second question would have been higher or lower than 75% if the first question had been omitted?

   A recent poll in France discovered that 54% of respondents felt there were *not too many foreigners* in France. In a subsequent question, however, 51% felt there were *too many immigrants*. Did respondents really make such a distinction between foreigners and immigrants? Or perhaps, after magnanimously stating that the percentage of foreigners was not too high, some respondents subconsciously tempered their tolerance with a negative reaction in the subsequent question.

7. **Confidentiality and anonymity**: The 1995 National Survey of Adolescent Males showed different prevalence of less socially acceptable behaviors, depending on whether responses were on paper or by laptop computer—respondents would tend to feel more anonymous, and answer more honestly, when using a laptop.

Refer to article *Parents fear school survey could lead to trouble*. What would be the best way to phrase the first three survey questions in order to elicit honest responses about sensitive, illegal, or risky behavior, but at the same time not to suggest that these behaviors are being condoned?

## Be Sure You Understand What Was Measured

### Example

An article **Fudging the diaries** describes how psychiatrists asked 80 adults with chronic pain to make entries in either a paper diary or on a hand-held computer for three weeks. They were supposed to answer questions about their pain at 10 a.m., 4 p.m., and 8 p.m. each day. The paper diaries had photosensors that recorded when the binder was opened and closed. The researchers reported that "90% of the 40 people using the paper diaries claimed they followed instructions, but the photosensors showed actual compliance of 11%. The hand-held computers, by contrast, would not accept entries except at the designated times and boasted a 94% compliance rate."

What does it mean to have an actual compliance rate of 11%? Did only 11% of the subjects make all 63 entries correctly? (That is, three weeks times seven days a week times three times a day.) Is it counted as non-compliance if a subject fails to make 1 of the 63 entries correctly? Or do they mean that 11% of the 63 times 40 entries of paper diary-users were entered at the correct time? Is it counted as non-compliance if a subject made the entry at 11 a.m. instead of 10 a.m.? Clearly there are many ways to measure compliance, and the researchers could interpret it as they saw fit. Noteworthy is the fact that one of the two investigators is founder and chief scientist for a company that sells software for electronic diaries. Do we have reason to suspect that he may have classified subjects as non-compliant when another researcher would have construed the same behavior to be compliant?

## Hard-to-Define Concepts

**Example**

One of the first examples we discussed was a survey that found 19% of adult Americans to believe that money can buy happiness. What did they mean by happiness? How much of it? And for how long? According to Robert Frost, "Happiness makes up in height for what it lacks in length," suggesting that temporary pleasures, such as a nice (expensive) weekend in the Caribbean, may bring about happiness. On the other hand, Albert Camus has said, "But what is happiness except the simple harmony between a man and the life he leads?", which suggests that happiness is more of a long-term thing. How would you yourself define it? At any rate, using precise numbers to summarize imprecise concepts may still leave us in the dark as to what exactly is being measured.

## Open vs. closed questions

1. What is an open question?
2. What kind of question is this? (i) open (ii) closed

Responses to open questions are harder to analyze and summarize. But responses to closed questions may be influenced by the options given, as in our good-to- excellent health example.

**Example**

An article entitled **Taunts cut girls more than sticks or stones** reports: "Girls 8 to 17 are as concerned about emotional violence—teasing, gossip, and name-calling—as they are about physical violence, from street and date violence to car accidents and war, a new Girl Scout Research Institute survey has found... The institute surveyed 2,279 girls, only some of them Girl Scout members, in April, using a self-administered online questionnaire." An accompanying bar graph entitled **What scares girls most?** lists the following responses and percentages:

Being teased or made fun of: 32%

Being attacked with a weapon (such as a gun or knife): 28%

Being kidnapped: 26%

Being forced to do something sexual: 24%

Being gossiped about: 24%

Getting into a car accident: 21%

Getting a disease (such as AIDS or cancer): 21%

Being called names: 18%

Natural disasters (e.g. earthquakes, tornadoes, or floods): 18%

Terrorist attacks: 16%

War: 15%

The sampling frame may or may not be adequate to draw conclusions about all girls ages 8-17: the article reports that not all of the girls were scouts, but if Girl Scouts constituted a disproportionately high number in the sample, then it is risky to assume the sample is representative of all girls in that age group. (Perhaps girls who join organizations such as this are more concerned in general about what other girls think of them, or perhaps the level of teasing differs among girl scouts compared to other girls...) A web search located the official report at

`http://www.girlscoutsofscc.org/publications/downloads.html`

where it is stated that 56% of the respondents were never Girl Scouts and the remaining 44% were either currently or previously scouts. The other four potential problems are probably of little concern, because the survey was conducted by a very reputable polling agency (Harris). This should alleviate at least to some extent any suspicions of deliberate bias, but there may be some unintentional bias due to the ordering of questions (original survey was not to be accessed). Desire to please, asking the uninformed, and unnecessary complexity do not seem relevant, and since the survey was online, there should be little concern about confidentiality/anonymity.

The exact wording of the question is not mentioned; the girls may have been asked something like, "What concerns you the most...", but since the percentages add up to more than 100%, they were apparently permitted more than one response. We can assume the question was closed, listing all of the options above.

In my own opinion, the article's suggestion that girls are more afraid of taunts than they are of physical violence is a misinterpretation of the results. Taunts are the most immediate concern because they are the most common type of harm that would be encountered. I suspect the percentages would turn out much differently if the girls were asked, "Which of these would be most harmful to you, if they were to actually happen?"...

**Exercise:** Find an article or internet report about a sample survey. Tell if the variable(s) of interest is quantitative or categorical. Then tell how the individuals were selected and whether or not you believe they adequately represent the population of interest. Discuss whether any of the 5 common problems in the selection process (using the wrong sampling frame, etc.) apply, or if any of the 7 pitfalls in the surveying process (deliberate bias, etc.) apply. Were the questions open or closed?

# Lecture 9

# Chapter 5: Relationships Between Quantitative Variables

The most fascinating and important statistical problems are about relationships. In this chapter, we focus on relationships between two quantitative variables. Examples of such relationships already encountered in this course are the relationship between family size and IQ score, and the relationship between weight and time spent on the phone. The data generally consist of two columns of numbers, one for each of the two variables.

**Example**

Is there a relationship between Stat students' height and shoesize? Below is part of the data set I used:

| HEIGHT | SHOE |
|---|---|
| 67 | 7.5 |
| 73 | 11.0 |
| 70 | 10.5 |
| ... | ... |

In this particular example, choice of which variable is explanatory and which is response depends on how we intend to utilize our results, or may just be a matter of personal taste. Since my aim is to predict a person's shoesize given his or her height, I will take height to be the explanatory variable. On the other hand, an article called **350,000-year-old human footprints found** states: "The footprints' makers were short, just under 5 feet tall, based on the prints' size of less than 8 inches in length." This was a situation where researchers took foot size to be the explanatory variable and used it to predict the response, height.

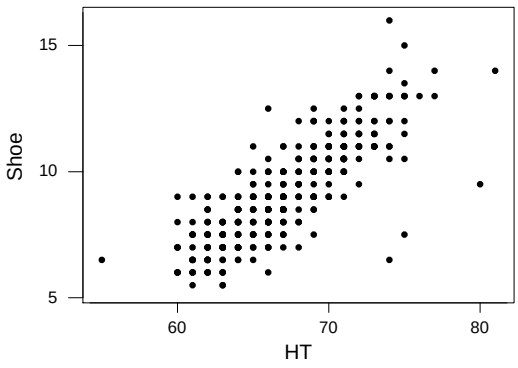
In many, or perhaps most, relationships, there is only one reasonable assignment of the roles of explanatory and response variables. For example, researchers have explored the question of whether or not taller men in general are more likely to receive higher salaries. Here height would be the explanatory variable, salary the response variable. [Note that vice-versa would not make sense.] In general, an explanatory variable attempts to explain the observed outcome. A response variable refers to the outcome of a study. Of course, relationships between variables are often much more complex than a simple cause/effect situation, for example, the link between obesity and low socio-economic status in women. [See article **There is a combination of reasons why so many poor are obese**.]

## Scatterplots

We begin analyzing relationships, just as we began for single variables, with a graphic display. A **scatterplot** is the most useful display technique for the relationship between two quantitative variables. We plot $y_i$, values of the response variable, along the vertical axis, corresponding to the $x_i$, values of the explanatory variable, along the horizontal axis for each individual i.

**Example**

To draw a scatterplot for our example, since height is the explanatory variable, we plot each height value horizontally, and plot the corresponding shoesize vertically.

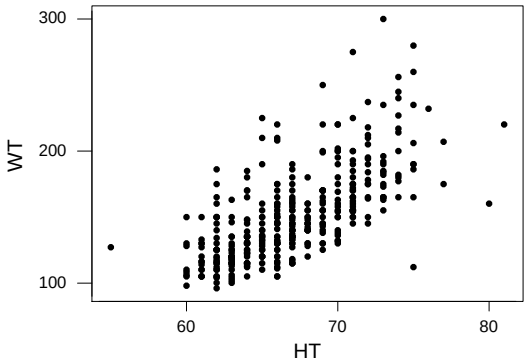



**Example**

Is there a relationship between Stat students' heights and weights? Explain how to draw the scatterplot.

| HEIGHT | WEIGHT |
|---|---|
| 67 | 150 |
| 73 | 165 |
| 70 | 163 |
| ... | ... |

Because height is more pre-determined than weight is, it should play the role of explanatory variable, and be plotted along the horizontal axis:

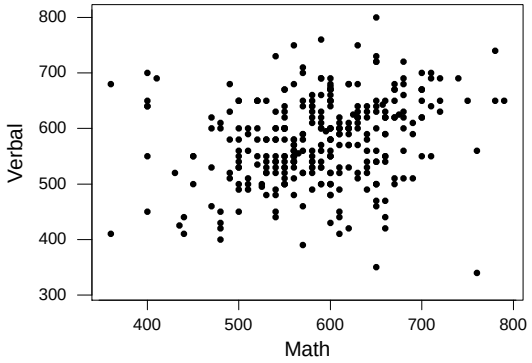



**Example**

Is there a relationship between Stat students' Math and Verbal SAT scores? Explain how to draw the scatterplot.

| MATH | VERBAL |
|---|---|
| 500 | 650 |
| 630 | 620 |
| 700 | 670 |
| 600 | 480 |
| 680 | 560 |
| 500 | 550 |
| ... | ... |

In this situation, neither variable has any more reason than the other to serve as explanatory variable, so the scatterplot could show either verbal vs. math or math vs. verbal.




To summarize a relationship between two quantitative variables, we should tell **direction**, **form**, and **strength**. We should also mention **outliers** if there are any present.

**Example**

Based on the scatterplot for heights and shoesizes, summarize the relationship.

1. **direction:** We see that below-average values of height tend to be accompanied by below-average values of shoesize, and the same for above-average values of height and shoesize. Thus, we have a **positive association**. If large values of one variable tend to occur with small values of the other (such as high IQs with small family sizes), and vice versa, we have a **negative association**.
2. **form:** This particular plot is clearly **linear**, not curved.
3. **strength:** Until we establish a universal measure of strength, it is difficult to characterize the strength of the relationship in a single scatterplot. For now, we will note that the cluster of points is tighter for this relationship than it is for the relationship between Verbal and Math SATs; in other words, this relationship is **stronger**.
4. **outliers:** There are a few points that stray somewhat from the bulk of the scatterplot points, representing people with unusual heights, shoesizes, or combinations of these.

Note: you may be concerned about lumping male and female students' heights and shoesizes or weights together: isn't gender often a confounding variable? It's not a bad idea to examine scatterplots of each relationship separately for males and females. It turns out that each gender group displays a similar positive relationship between height and shoesize, likewise between height and weight. Therefore, combining the two gender groups for these variables should not produce misleading results. For other variables—such as weight and time spent on the telephone—combining both gender groups would be a mistake.

## Least Squares Regression

Recall: In Chapter 2, we said the best idealized description of many *distributions of a single quantitative variable* is provided by the normal curve (for instance, the distribution of heights).

Now, the best idealized description of many *relationships between two quantitative variables* (for instance, heights and weights in certain age groups) is provided by a straight line. This line has been called a **regression line** because the first such studies in the mid 1800's by Sir Francis Galton involved children of tall parents whose heights were shown to "regress" to mediocrity.

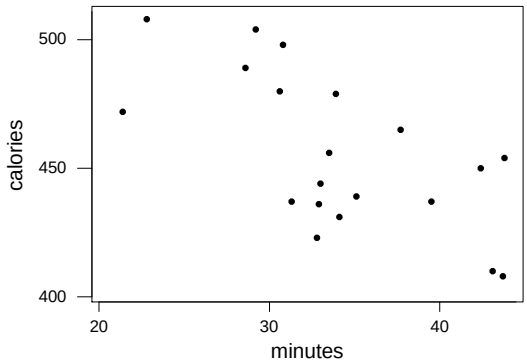
We want to use a straight line to describe how the response variable $y$ tends to change as the explanatory variable $x$ changes. To regress, we fit a line $\hat{y} = b_0 + b_1 x$ to the data which comes "as close as possible" to the observations.

### Example

Is how long children remain at the lunch table (in minutes) related to how much they eat (in calories)? Twenty toddlers were observed over several months at a nursery school, and for each child the average time in minutes and average calories consumed were recorded:

| Time | Calories |
|---|---|
| 21.4 | 472 |
| 30.8 | 498 |
| 37.7 | 465 |
| 33.5 | 456 |
| 32.8 | 423 |
| 39.5 | 437 |
| 22.8 | 508 |
| 34.1 | 431 |
| 33.9 | 479 |
| 43.8 | 454 |
| 42.4 | 450 |
| 43.1 | 410 |
| 29.2 | 504 |
| 31.3 | 437 |
| 28.6 | 489 |
| 32.9 | 436 |
| 30.6 | 480 |
| 35.1 | 439 |
| 33.0 | 444 |
| 43.7 | 408 |

We should always begin with a scatterplot and a report of apparent direction, form, and strength.




The relationship appears negative and moderately strong. The fact that it shows a linear, rather than curved, pattern suggests that a straight line is an appropriate summary. For this example, the regression line $\hat{y} = b_0 + b_1 x$ expresses that

$$\text{calories [idealized]} = b_0 + b_1 * \text{minutes}$$

If we picture a line through the points on the scatterplot graph, $b_0$ is the line's intercept, or idealized y-value for $x = 0$. $b_1$ is the slope, or idealized change in $y$ for every unit change in $x$.

Here are some questions to think about:

1. Is there only one "best" line?
2. If so, how do we find it?
3. Assuming we have fitted a line to the data, what can it tell us?

We'll begin by answering the last question, which will provide insight into the first two.

We can use the regression line $\hat{y} = b_0 + b_1 x$ to **predict** a value $\hat{y}$ for any given value $x$.

The "best" line should make the best predictions: the observed $y$-values should stray as little as possible from the line. The vertical distances, or **residuals**, of the points in our scatterplot from the fitted line are errors in predicting $y$. As a group, we want to make them as small as possible. If we simply required the sum of all residuals to be zero, we could have a badly fitting line, but large positive residuals cancelling out large negative residuals. Thus, we choose the **least squares method**: minimize the sum of squared residuals. [Note: this method is also preferable to minimizing absolute values because the least squares method favors two residuals of size 1 to one of size 0 and one of size 2, since $1^2 + 1^2 = 2 < 2^2 + 0^2 = 4$.]

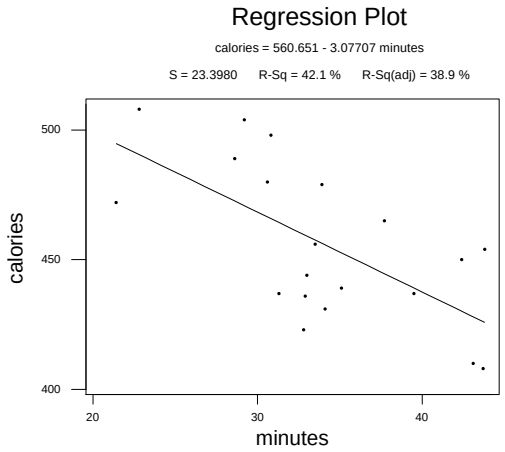
This turns out to be a calculus problem with a unique solution, assuring us that there is in fact one "best" line $\hat{y} = b_0 + b_1 x$, with values for $b_0$ and $b_1$ calculated from the data ($n$ pairs of values $(x_i, y_i)$). Such calculations are quite messy, even for small data sets, and are better left to the computer. For the regression of calories on time, MINITAB produces the following output:

```
Regression Analysis: calories versus minutes

The regression equation is
calories = 561 - 3.08 minutes

Predictor        Coef     SE Coef          T        P
Constant       560.65       29.37      19.09    0.000
minutes       -3.0771      0.8498      -3.62    0.002

S = 23.40       R-Sq = 42.1%     R-Sq(adj) = 38.9%

Pearson correlation of minutes and calories = -0.649
P-Value = 0.002
```

To interpret the regression equation, we can say that the actual prediction made by the best-fitting line, if a child sat at the lunch table for 0 minutes, is 561 calories. This may be the "correct" prediction, but for practical purposes it is nonsense, of course, because no-one can eat lunch in zero minutes. Since 0 minutes is outside the range of times observed, we should not use the line to predict calories for such a time value. Using the regression equation to make predictions for x-values that are beyond the range of those used to fit the line is called **extrapolation**, a regression pitfall that can lead to nonsensical results. As for the slope of the line being -3.08, it tells us that for every additional minute at the lunch table, we predict calorie consumption to go down by about 3. [Note: the "P-Value" of .002 provides a vital piece of information, which we will understand later in the course when we learn how to perform statistical inference in the form of hypothesis tests.]

To plot the line, we could solve for two points and connect them, for example $(30, 468)$ and $(40, 437.8)$.

MINITAB produces a fitted line plot upon request.




# Lecture 10

# Chapter 5: Relationships Between Quantitative Variables

Last time we learned to fit a line $\hat{y} = b_0 + b_1 x$ coming "as close as possible" to the scatterplot points by minimizing the sum of squared residuals $y_i - \hat{y}_i$. This least squares regression line can be used to predict a response $\hat{y}$ for any explanatory value $x$.

**Example**

MINITAB produced the regression line for Stat students' heights and shoesizes:

```
The regression equation is
Shoe = - 17.2 + 0.392 HT
```

The line predicts the shoesize of a 70-inch-tall student to be $-17.2 + .392(70) = 10.24$. In fact, a 70-inch-tall student actually wore size 10.5 shoes, so the prediction error, or residual, in this case is $y_i - \hat{y}_i = 10.5 - 10.24 = +.26$. Our prediction wasn't bad!

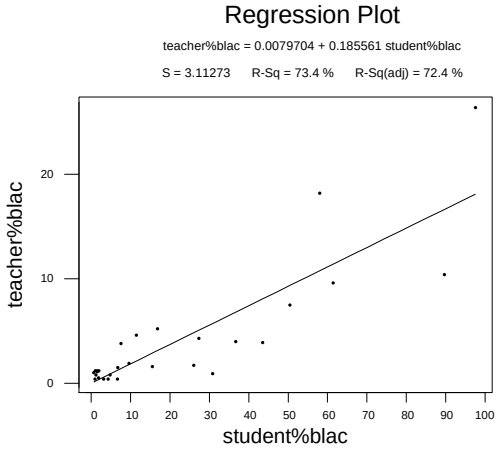
**Example**

Of the 43 school districts in Allegheny County, 28 employed black teachers during the year 2001-2002. Do there tend to be more black teachers in schools that have more black students? Here are the results of a MINITAB regression, where the data consisted of values for percentage of black students (explanatory variable) and percentage of black teachers (response) in each of those 28 districts:

```
Regression Analysis: teacher% versus student%

The regression equation is
teacher% = 0.008 + 0.186 student%

S = 3.11273      R-Sq = 73.4 %      R-Sq(adj) = 72.4 %
```

The regression line predicting percentage of black teachers based on percentage of black students is provided by MINITAB . We can predict the percentage of black teachers to be $.008+.186(1) = .194$ when the percentage of black students is 1, and $.008 + .186(58) = 10.796$ when the percentage of

black students is 58. Looking at the scatterplot, do our predictions tend to be better for low or high percentages of black students? Low, because the scatterplot cluster is tighter in the low $x$-value range, looser in the high $x$-value range. A plot of residuals would show a fan-shaped pattern. In particular, the school district with 1 percent black students had 1.2 percent black teachers, so this residual (prediction error) is just $1.2 - .194 = 1.006$. The school district with 58 percent black students had 18.2 percent black teachers, so this residual is $18.2 - 10.796 = 7.404$. The district with 90 percent black students would be predicted to have $.008 + .186(90) = 16.748$ percent black teachers, but it only had 10.4 percent. Now the prediction error is $10.4 - 16.748 = -6.348$. Note that points above the regression line produce positive residuals and points below the line produce negative residuals.

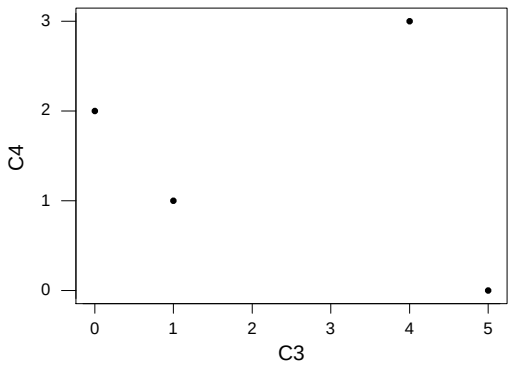



## Roles of $x$ and $y$ in Regression

The choice of which is to be the explanatory variable and which the response variable does not affect $r$, but it does affect the regression line.

**Example**

Picture a scatterplot consisting of points (0,2), (1,1), (4,3), (5,0):




Regressing $y$ on $x$, we are minimizing the sum of squared *vertical* distances of observations from the fitted line, whereas regressing $x$ on $y$ we are minimizing the sum of squared *horizontal* distances. These result in two very different regression lines. [Note that $r^2$ is unaffected by our choice of regressing $y$ on $x$ or $x$ on $y$; in either case, $r^2 = 4.7\%$, so $r = .22$.]

```
 MTB > regress c4 1 c3

 The regression equation is
C4 = 1.79 - 0.118 C3

 s = 1.543       R-sq = 4.7%      R-sq(adj) = 0.0%

 MTB > regress c3 1 c4

 The regression equation is
 C3 = 3.10 - 0.40 C4

 s = 2.846       R-sq = 4.7%      R-sq(adj) = 0.0%
```

Putting C4 in terms of C3, we have $C4 = 7.75 - 2.5C3$. Thus, the first line is almost horizontal (slope is -.118) whereas the second line has a sharp downward slope of -2.5.

Typically, the regression line is constructed as a tool for prediction, and so the variable to be predicted is regressed as a response to the other variable, which is explanatory. Thus, we'd use a different line if we were predicting height from shoesize, not vice versa.

Just as apparent spread in a single distribution was affected by our choice of stems for a stemplot or classes for a histogram, the apparent strength of a relationship as displayed in a scatterplot may be affected by our choice of scale. Standard deviation gave us a well-defined measure of spread; now we need a well-defined measure of the strength of a relationship.

## Correlation

Recall: We said that the best summary for many relationships between two quantitative variables is the **least squares regression line**:

$$\hat{y} = b_0 + b_1 x$$

Unless the scatterplot points fall exactly along a straight line, our predictions are not going to be perfect. How good or bad our predictions tend to be depends on how tightly clustered or loosely scattered the points are; in other words, it depends of the strength of the relationship.

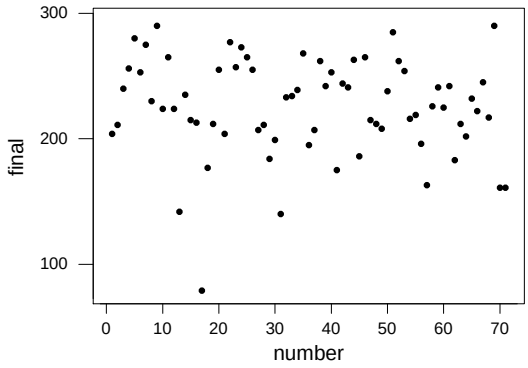
We can measure the exact strength of the *linear* association between two quantitative variables, regardless of their unit of measurement, by finding the average product of standardized $x$ and $y$ values, called the **correlation**, denoted

$$r = \frac{1}{n-1}\Sigma(\frac{x_i - \bar{x}}{s_x})(\frac{y_i - \bar{y}}{s_y})$$

Rather than devoting time and energy to performing tedious calculations with this formula, we will leave the calculations to MINITAB and concentrate instead on the most important properties of the correlation $r$:

- The correlation is always a number between -1 and +1.
- $r$ is close to $\pm 0$ for weak relationships (eg. verbal and math SATs has $r = .25$); $r$ is close to -1 for strong negative relationships; $r$ is close to +1 for strong positive relationships (eg. shoesize vs. height).
- The correlation equals -1 for a perfect negative relationship (a person's age vs. birthyear) and +1 for a perfect postitive relationship (a person's weight in kilogram vs. weight in pounds). Such relationships are called **deterministic** because the value of one variable completely determines the value of the other.
- The correlation is greater than zero for a positive association (eg. shoesize vs. height; weight vs. height) and less than zero for a negative association (eg. nursery school children's calories consumed vs. time at the lunch table; child's IQ vs. family size).
- The correlation equals zero when knowing the value of $x$ tells us nothing about the value of $y$, and so the best line to fit the data would simply be a horizontal line at the average $y$-value. [As students handed in their final exams at the end of spring semester 2003, I recorded the chronological number for each (first was 1, last was 71) and then plotted each student's exam score vs. this number. The plot showed completely random scatter, and the correlation was just about zero: time order for when

each exam was handed in told me nothing about what the score was going to be.]




- Correlation measures the strength of the linear association between two *quantitative* variables. (A Time magazine article on Gulf War syndrome stated there was little correlation between ill soldiers and where they served during the war. They should have said "association" not "correlation".)
- Correlation is independent of unit of measurement (eg. we could measure height in centimeters instead of inches, and/or weight in kilograms instead of pounds, and the correlation between height and weight would remain the same; this is because it is the average product of *standardized* $x$ and $y$ values.) Imagine adjusting the scale on the horizontal and/or vertical axes by relabeling the units: this in itself would have no impact on the tightness or looseness of the scatterplot cluster.
- Correlation measures the strength of a *linear* relationship only. There may be a strong curved relationship for which the correlation is calculated to be close to zero.
- The roles of $x$ and $y$ are interchangeable (as can be seen from the mechanics of the formula above), so the choice of explanatory/response variables does not affect the correlation. Imagine flipping the graph so that $x$ becomes $y$ and $y$ becomes $x$: this would have no impact on the tightness or looseness of the scatterplot cluster.

We have demonstrated that the roles of $x$ and $y$ impact the equation of the regression line but not the value of the correlation. However, correlation—specifically, squared correlation—has a special meaning in the context of regression. We can express $r^2$ as the ratio of the variation of the regression line from the mean to the total variation, which includes variation from the mean as well as variation from residuals. Thus, $r^2$ is the fraction of the variation in $y$ that is explained by least squares regression on $x$. If $r^2$ is close to 1, then $y$ doesn't vary much about the regression line (it just varies about its mean), and so the regression line does a good job of accounting for the variation in $y$. If $r^2$ is close to 0, the observed $y_i$ vary considerably from predicted $\hat{y}_i$, and the line isn't explaining much about the behavior of $y$.

**Example**

How much of the variation in shoesize can be explained by least squares regression on height? The output below shows $r^2 = 67.4\%$.

```
The regression equation is
Shoe = - 17.2 + 0.392 HT
369 cases used 2 cases contain missing values
Predictor        Coef     SE Coef          T        P
```

```
Constant     -17.1592      0.9568     -17.93    0.000
HT            0.39235     0.01424      27.55    0.000
S = 1.115       R-Sq = 67.4%     R-Sq(adj) = 67.3%
```

**Example**

How much of the variation in weight can be explained by least squares regression on height? For this relationship, $r^2 = 45.6\%$.: Height tells us more about someone's shoesize than it does about his or her weight.

```
The regression equation is
WT = - 232 + 5.72 HT
368 cases used 3 cases contain missing values
Predictor        Coef     SE Coef          T        P
Constant      -232.13       21.94     -10.58    0.000
HT             5.7189      0.3265      17.52    0.000
S = 25.59       R-Sq = 45.6%     R-Sq(adj) = 45.5%
```

Note: in MINITAB, the regression output automatically includes $r^2$. If we want the correlation $r$, we can take the square root of $r^2$, giving it the correct sign + or -, or we can specifically request correlation from the Basic Statistics menu.

**Example**

```
The regression equation is
calories = 561 - 3.08 minutes
Predictor        Coef     SE Coef          T        P
Constant       560.65       29.37      19.09    0.000
minutes       -3.0771      0.8498      -3.62    0.002
S = 23.40       R-Sq = 42.1%     R-Sq(adj) = 38.9%
Pearson correlation of minutes and calories = -0.649
```

For the regression of calories on minutes, $r^2$ was 42.1%. Because the relationship is negative, $r = -\sqrt{.421} = -.649$.

## Why the Answers May Not Make Sense

- **Outliers** are points lying far from the fitted line, producing large residuals. They may or may not have much influence on the regression line, depending on the configuration of points, but they should be looked into, in case for some reason they don't really belong with the data. Correlation can be affected by outliers. One class of mine had a very short son of a very tall father. The correlation for heights of sons vs. fathers was close to zero, but jumped to .4 when that one data pair was omitted. **Influential observations** are often stray points far removed from the others in a horizontal direction. Whether or not such an observation is included greatly affects the position of the regression line. The short son of a very tall father had a great deal of impact on the regression line because the fathers' height was indeed far in a horizontal direction. A very short son of a medium-tall father would appear as an outlier and have a large residual, but wouldn't have as much impact on the regression line.
- **Combining groups inappropriately** in a regression situation can produce misleading results.

**Example**

In Chapter 2 we discussed recording students' weights and how much time they spent on the phone in a given day: students who weighed more tended to spend less time on the phone. The resulting regression line would have a negative slope, but in fact for males and females treated separately, phone time does not go down as weight goes up.

**Example**

Shoe size vs. reading scores for elementary school children in all grades K through 12 taken together would have a high value of $r$.

Both of the preceding examples involve a confounding variable. If such a variable is not taken into account, we may find a substantial correlation between two variables even though changes in one don't necessarily *cause* changes in the other. In the case of phone time vs. weight, *gender* would be the confounding variable; for shoe size vs. reading scores, *age* would be the confounding variable. The way to control for such confounding variables is to analyze groups of similar individuals separately.

- If the data is **curvilinear**, then a straight line can make poor predictions. For example, a scatterplot of height vs. age for years zero to thirty would exhibit a curved shape, leveling off around the late teens. Summarizing such a relationship with a straight line is inappropriate, and would make outlandish predictions for elderly people (extrapolation). Often a regression analysis includes plots of residuals vs. the $x$-values. If the relationship itself is curved, then the residual plot would also be curved, confirming that a straight line is not a good summary. In general, if the residual plot displays some pattern, it indicates that a straight line is not providing an adequate description of the relationship between $x$ and $y$. If on the other hand we see no pattern in the residuals, then we have support for the assumption of linearity.

## Correlation Does Not Prove Causation

Confounding variables are one reason why evidence of a linear relationship between two quantitative variables does not necessarily mean that one of them is causing changes in the other.

**Example**

The article **There is a combination of reasons why so many poor are obese** mentions several studies over the past three decades which have determined that for women, low socio-economic status and obesity are definitely "linked". Can we conclude that obesity causes low socio-economic status in women? There are in fact several possible explanations for a relationship between two variables.

1. The explanatory variable does cause changes in the response: perhaps obesity in women leads to lower socio-economic status because of discrimination by employers.
2. Confounding variables are tied in with the explanatory variable, making it unclear whether the explanatory variable by itself is responsible for changes in the response: perhaps women in certain racial groups are more likely to be overweight, and this is where the discrimination comes in.
3. Other variables cause changes in both the explanatory and response variables, but neither of them influences the other: perhaps living in a certain region of the United States predisposes some women to nutritional habits that lead to obesity, and being in such regions also reduces their socio-economic level.
4. The response variable causes changes in explanatory values: perhaps being in a lower socio-economic bracket results in obesity via poorer nutrition options.

The best way to verify *causation* is to perform an experiment: impose changes on the explanatory variable $x$, holding possible lurking variables constant so that changes in the response $y$ due to $x$ only are observed. Can we conduct an experiment to prove causation in either direction? To prove low socio-economic status causes obesity, we would have to impose changes to improve socio-economic status, and see if women's weight went down. This would be expensive and impractical. To prove obesity causes low socio-economic status, we could recruit participants in a weight-loss program, but the fact that the women must volunteer may predispose them to the improvement of their response (socio-economic status). In many situations, an experiment

is unethical, or impractical, or subject to volunteer bias. Then we would have to rely on results of an **observational study** to determine if one of these variables actually causes changes in the other. A list of criteria to establish causation in such a study is given:

1. There is a reasonable explanation for cause and effect.
2. The connection occurs under varying circumstances.
3. Potential confounding variables are ruled out.

Most researchers agree that these criteria have been satisfied to an overwhelming degree in the case of the question of whether or not more cigarettes smoked lead to higher chance of lung cancer.

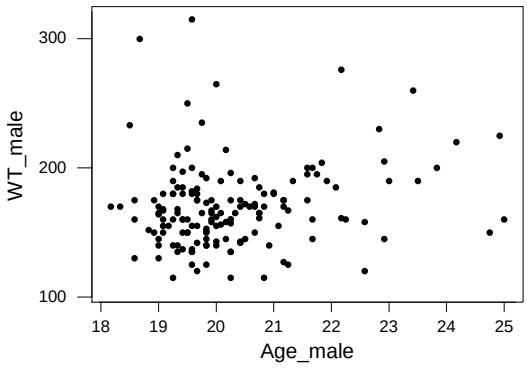
**Correlation based on averages** may be misleading: averaging out the $y$ values leads to a more linear plot, overstating the strength of the relationship.

**Example**

Mean age vs. year for students in year 1, 2, 3, or 4 at Pitt has a very high correlation, but age vs. year has a lower correlation because including all the many individuals results in much more scatter in the plot.

**Example**

To review analysis of relationships between two quantitative variables, I decided to look at students' ages and weights. Since gender plays such an important role in weight, and perhaps also in the relationship between age and weight, I unstacked the data in age and weight by gender, and then concentrated on male students. In order to eliminate the most pronounced age outliers, I sorted weights and ages in descending age order and deleted weights and ages for the handful of males who were over 25. Then I produced a scatterplot, graphing weight (response) vs. age (explanatory), since age may play a role in a man's weight, but not vice versa:




The plot's scatter is very loose but the points do seem to allign somewhat along a line of positive slope; at any rate, they do not seem curved. Thus, the relationship appears to be positive, linear, and weak. Since most males are in their late teens or early twenties, and of moderate weights, but a few are unusually old or unusually heavy, there are outliers on the high end of both variables, but none seem too extreme. I then proceeded with the regression analysis:

```
The regression equation is
WT_male = 82.2 + 4.33 Age_male

158 cases used 6 cases contain missing values

Predictor        Coef     SE Coef          T        P
Constant        82.18       39.55       2.08    0.039
Age_male        4.335       1.944       2.23    0.027

S = 32.16       R-Sq = 3.1%      R-Sq(adj) = 2.5%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1        5142        5142      4.97    0.027
Residual Error   156      161372        1034
Total            157      166514

Unusual Observations
Obs   Age_male    WT_male         Fit      SE Fit    Residual    St Resid
  6       25.0     160.00      190.56        9.50      -30.56       -0.99 X
  7       24.9     225.00      190.21        9.35       34.79        1.13 X
  8       24.8     150.00      189.48        9.03      -39.48       -1.28 X
  9       24.2     220.00      186.96        7.95       33.04        1.06 X
 10       23.8     200.00      185.49        7.33       14.51        0.46 X
 11       23.5     190.00      184.06        6.73        5.94        0.19 X
 12       23.4     260.00      183.71        6.59       76.29        2.42RX
 22       22.2     276.00      178.29        4.45       97.71        3.07R
 78       20.0     265.00      168.88        2.62       96.12        3.00R
100       19.8     235.00      167.80        2.77       67.20        2.10R
108       19.6     315.00      167.06        2.91      147.94        4.62R
120       19.5     250.00      166.72        2.99       83.28        2.60R
158       18.7     300.00      163.12        4.07      136.88        4.29R
162       18.5     233.00      162.38        4.33       70.62        2.22R

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.

Correlations: WT_male, Age_male

Pearson correlation of WT_male and Age_male = 0.176
P-Value = 0.027
```

The correlation of +.176 confirms that the relationship is positive and weak. The importance of the "P-value" will become clear in future chapters on statistical inference. Its size will tell us whether or not there is statistical evidence of analogous trends in the larger population from which our sample originated.

MINITAB identifies quite a few influential observations (marked X) and outliers (marked R to indicate large residuals) and one student aged 23.4 years, at 260 pounds, who was marked both R and X. The regression equation is included in the output. The slope of 4.33 predicts an additional 4.33 pounds for every additional year in age, which does seem plausible for young male adults.

**Exercise:** Pick two quantitative variables from our survey, decide on roles of explanatory and response, and tell what you expect to see in terms of their relationship. Use MINITAB to explore the relationship

between them: start by assessing the scatterplot. Be sure to mention direction, form, strength, and outliers. Your summary should tell the value of the correlation $r$ and the equation of the regression line, if the form appeared linear. Summarize your findings in context.