

Lectures 7-9: Measures of Central Tendency, Dispersion, and Assumptions

Lecture 7: Descriptive Statistics: Measures of Central Tendency and Cross-tabulation

Lecture 8: Measures of Dispersion

Lecture 9: Assumptions for Measures of Central Tendency/ Measures of Dispersion

Descriptive Statistics: Measures of Central Tendency and Cross-tabulation

Introduction

Measures of central tendency (MCT) allow us to summarize a whole group of numbers using a single number. We are able to say “on average” this variable looks like this. It is better than handing them raw data (IN PERSON: $n=3000$ like in kid prison study). Pretend I had 100 people in my study and I handed you all of the data for gender-- if I just handed you a bunch of numbers that makes it hard for you to figure out how to describe the ‘average’ gender in my study, or average age, or average income, etc.

There are three types of MCT (averages) in statistics: mean, median, and mode.

Once again, any average in statistics is just a “single number used to describe a group of numbers.”

Here are two fake “variables” and their data we will use for this lecture.

Both variables are ratio level variables. One is “age in years.” The other is “amount of \$ in their wallet.” The sample size in each case is $n=9$. The far left hand column x_1, x_2, x_3 etc are the individual people. It is just to show you that there are 9 people (each data point is often referred to as “ x ” in statistics).

Fake Variables *age* and *\$ in wallet* (n=9)

data point	age	\$ in wallet
X1	19	100
X2	20	110
X3	22	120
X4	22	120
X5	24	130
X6	24	130
X7	24	130
X8	26	180
X9	75	2000
SUM or $\sum x$	256	3020
MEAN	28.4	335.555
MODE	24	130
MEDIAN	24	130
SKEWNESS	2.9	3.0

Mode : most frequently occurring score

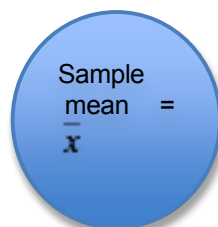
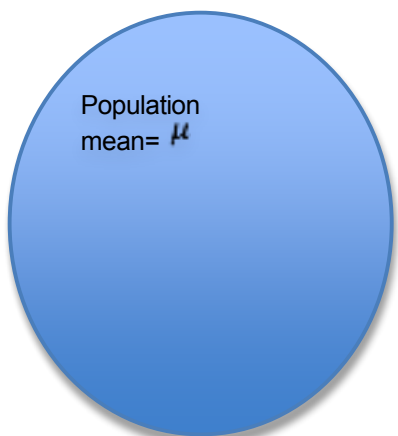
The mode is the most frequently occurring score. Sometimes there is no mode, two modes, or more than 2 modes. For the two variables above there is only one mode.

For the variable *age* above, the age 24 occurs the most (three times) and it is the mode.

For the variable *\$ in wallet*, \$130 occurs the most (three times) and it is the mode.

Mean: nothing more than arithmetic average!

The mean is the MCT we will use the most in this class. To compute the mean you simply add up all of the scores and then divide by the number of scores you have. In “statistics speak” we have two formulas for the mean. Recall when we refer to any thing that describes a population we call it a “population parameter” and it will have a different symbol than a “sample statistic” (recall something that describes a sample it is called a “statistic.”)



Formulas for the mean

$$\mu = \frac{\sum x}{N} \quad \text{This is the formula for the "population mean"}$$

$$\bar{x} = \frac{\sum x}{n} \quad \text{This is the formula for the "sample mean"}$$

Don't panic about those formulas!

Before you panic, both of these formulas tell you to do the same thing. They simply have different symbols for when you have sample data or population data. Don't panic. Each of these formulas for the mean tell you to do the same thing:

Add up all the x's ($\sum x$) and divide by the number of data points you have.

Recall each data point is an x. There are nine data points or x's in the two examples below. That means N or n =9.

So add up all the x's and then divide by the number of x's (n or N).

n= the number of data point in the sample

N= the number of data points in the population.

data point	age	\$ in wallet
X1	19	100
X2	20	110
X3	22	120
X4	22	120
X5	24	130
X6	24	130
X7	24	130
X8	26	180
X9	75	2000
SUM or $\sum x$	256	3020
MEAN	28.4	335.56
SKEWNESS	2.9	3.0

Let's pretend the data for "age" and "\$ in wallet" are sample data so we would use the sample formula:

$$\text{Age } \bar{x} = \frac{\sum x}{n} = \frac{256}{9} = 28.4$$

$$\text{\$ in wallet } \bar{x} = \frac{\sum x}{n} = \frac{3020}{9} = 335.56$$

As we will see both later in this lecture and in the lecture on “misuses of statistics” the mean is thrown off or distorted by extreme scores. But as a preview, look at each of those means and then the data set. In both cases, there is an extreme value (in the x9 place) and you can see that the extreme score “inflates” the mean. (By the way extremely small scores would “deflate” the mean.) For *age* 28.4 is a really bad “single number used to describe all the numbers” and the same thing can be said for *\$ in wallet*: \$335.56 does a very poor job of representing all of the numbers in the sample. These are examples of the mean being negatively influenced by extreme scores. Below is an example of a data set without extreme scores.

Example of mean without extreme scores

data point	age	\$ in wallet
X1	19	100
X2	20	110
X3	22	120
X4	22	120
X5	24	130
X6	24	130
X7	24	130
X8	26	180
Σx	181	1020
mean	22.6	127.5
skewness	-0.3	1.7

Now, above is just the same data but I eliminated the x9 data point. You will notice that the mean for age 22.6 and \$ in wallet \$127.50 each do a pretty good job of describing all of the numbers in the data set. Remember, the mean (any average) is just a single number used to describe a group of numbers. In this case each mean does a pretty good job of this, as there are no extreme scores. You do not have to be a rocket scientist or mathematician to see this. Look at the group of numbers and then look at the mean. Do it. In the example w/ extreme scores, the mean does a poor job of describing all of the numbers. In the example without the extreme scores, the mean does a pretty fair job.

Median: refers to position when data is put in an array.

The median refers to the number in the middle position when the data is placed in an array. An array is a fancy mathematical term that means put the numbers in order from the smallest to the biggest. Note that the variables in the tables above are in an array –smallest to biggest. So place the numbers in order smallest to biggest (an array) and the median is the middle number.

Fake Variables age and \$ in wallet (n=9)

data point	age	\$ in wallet
X1	19	100
X2	20	110
X3	22	120
X4	22	120
X5	24	130
X6	24	130
X7	24	130
X8	26	180
X9	75	2000
Σx	256	3020
mean	28.4	335.56
median	24	130
mode	24	130
skewness	2.9	3.0

The median refers to a position if you think about it. It is the middle score. It is easiest when there is an odd number of data points in your data set: then there are an equal number of data points above the median and an equal number below the median. But the median can be computed using the following formula regardless of whether or not there is an odd or even number of data points.

To compute the median: $(n+1)/2$

For both *age* and *\$ in wallet* $n=9$. So:

$(n+1)/2 = (9+1)/2 = 10/2 = 5$. So count down to the 5th number. Note this refers to position in the array, not the actual number! For the variable \$ in wallet the 5th number is \$130, so that is the median. For the variable age, the 5th number in the array is 24.

Once again, the median formula computes the position of the median in the data array, not the actual number!

How do I compute the median if there is an “even” # of data points?

If there are an even number of data points you do the same thing, but you take the average of the two numbers on either side of the median position. Below is a modified version of the data set we’ve been using in this lecture (I removed the x9 data point)

Fake Variables *age* and *\$ in wallet* (n=9)

data point	age	\$ in wallet
X1	19	100
X2	20	110
X3	22	120
X4	22	120
X5	24	130
X6	24	130
X7	24	130
X8	26	180
Σx	181	1020
mean	22.6	127.5
median	23	125
mode	24	130
skewness	-0.3	1.7

$$(n+1)/2 = (8 + 1)/2 = 9/2 = 4.5$$

So we look at the numbers in the 4th and 5th place and take a mean of the two. [For your information this is what SPSS does.]

For age: $22 + 24 / 2 = 46 / 2 = 23$. So the median of age is 23.

For \$ in wallet: $120 + 130 / 2 = 250 / 2 = 125$ So the median of money in wallet is \$125.

Assumptions for Measures of Central Tendency

The tables below do the best job of summarizing “when” to use the which measure of central tendency. Take a look at them and the look below where I will do my best to explain what’s going on in writing.

<u>measure</u>	<u>when used</u>	<u>advantages and disadvantages</u>
Mean	at least Interval or ratio & no extremes "no extremes" $= -2 < \text{skewness} < +2$	incorporates all the data, use in other statistics, influenced by extremes
median	at least ordinal data	good when have extreme values
Mode	any-- only choice for nominal	quick to get

Another Table

Type of Variable		What to Use
NOMINAL		only choice is MODE
ORDINAL		MEDIAN OR MODE
INTERVAL/RATIO	extremes in data skewness < -2 or skewness $> +2$	MEDIAN (or mode) not mean!!!
	no extremes in data ($-2 < \text{skewness} < +2$)	MEAN, MEDIAN, OR MODE probably should use mean

*By the way an explanation of “skewness” is below under the heading Use “**skewness**” to decide whether there are “**extremes**” in data (to eliminate the mean)*

First classify the variable Nominal Ordinal Interval or Ratio.

Nominal Variables

If you have a nominal variable you have one choice and once choice only: the MODE. Period, end of story. The median is out because nominal variables have no order to their numbers and the mean is out because they are not real numbers (the mean requires you to “do math” on the numbers and you need real numbers if you are going to add and divide).

How come you cannot use the median on a nominal variable? To use the median, you are referring to position so you need AT LEAST an ordinal variable (order!) to be able to refer to “position.” You need an order, a number that at least makes sense on that dimension. Consider a variable GENDER 1=male and 2=female. It is nominal. The numbers make NO SENSE. 2 or female is not “more” gender than 1 or male! So you can put the numbers in an array (in order), but the order is meaningless. Ordinal variables have an order that means something: 1 is less than 2 etc.

How come you cannot use the mean for a nominal variable? Well, because when you compute the mean you “do math” on the numbers – you add and divide and you need “real” numbers to do that. Nominal variables assign numbers but the numbers are meaningless so the computing the (ahem) mean would be meaningless.

Ordinal Variables

If you have an ordinal level variable you have two choices: MEDIAN or MODE, BUT YOU MAY NOT USE the MEAN. Ordinal variables are not “real numbers” so you cannot do math on them and compute a mean.

Think about this story about how a bunch of PhD’s evaluate teaching. It happens not only on this campus but on many campuses across the nation.

They compare the mean response to five item attitude scales which are ordinal variables by the way. That is material t-test of means [lecture 17-18]. Can you take the mean of an ordinal variable? You can but the number is meaningless.

So there are a bunch of questions in the course evaluations like this:

The professor does a good job of explaining complex ideas in plain English

1 = strongly disagree 2= strongly agree 3= neutral 4= agree 5= strongly disagree.

Our course evaluation reports compare our class’ mean score to the campus mean. Asking does your mean score differ from the mean campus score. **The mean score of an ordinal variable is [ahem] a MEANINGLESS number!** You can pretend it is not meaningless, but that is pretending. To make matters worse, some professors and administrators make career-altering decisions about other professors based upon this. Let me make it even more ironic [or infuriating] for you.

Some of us statistically literate people changed this as the mean score is [ahem] meaningless. Many colleagues were upset because they “liked the mean” [even though it was explained to them the mean was – ahem—a meaningless number]. So evaluations of teaching were done using a different, statistically valid, measure. My colleagues did not understand the new measure, or did not like it, or thought the mean made their teaching look good, so they voted to change it back to using a mean. This makes me sad.

So [some] teachers and administrators at this campus judge the teaching of their colleagues based upon a number that is, ahem, meaningless. In addition teaching evaluations are gathered from a **unrepresentative** sample as some may choose to fill out the course evaluation and other will not. It just so happens that those motivated to fill them out, tend to be students who are unhappy with your teaching.

So let’s review: use the mean of an ordinal variable and use a sample that is neither representative nor random. You can do the calculations on any data you wish in statistics but it does not logically follow that the numbers created have any meaning. **Now when you fill out these questions on teaching evaluations you should still answer the questions honestly, because your professors will [tragically] be evaluated for promotion and tenure based upon the mean class response by some [but not all] of the professors and administrators at this campus.**

Interval/Ratio Variables

You can use any of the three “averages” with an interval or ratio level variable. However, as you saw above, if there are extremes in the data, the mean is distorted. So if there are extremes in the data the mean is a poor choice and you should use the median or the mode. Below are the data from top of this lecture; look at the data for each of the variables “age” and “\$ in your wallet” and then look at the mean for each.

Example of mean with extreme scores

data point	age	\$ in wallet
X1	19	100
X2	20	110
X3	22	120
X4	22	120
X5	24	130
X6	24	130
X7	24	130
X8	26	180
X9	75	2000
SUM or Σx	256	3020
MEAN	28.4	335.56
SKEWNESS	2.9	3.0

In both cases, there is an extreme value (in the x9 place) and you can see that that “inflates” the mean. For *age* 28.4 is a really bad “single number used to describe all the numbers” and the same thing can be said for *\$ in wallet*: \$335.56 does a very poor job of representing all of the numbers in the sample.

So, the mean is distorted by extreme scores. This is why when you see “average house price” in the Honolulu Advertiser (or in the media) they use the “median house price.” Think about it. I live in Makaha. Most homes in Makaha are not very fancy but what if a few homes *right on the beach* sold? Well homes right on the beach are MUCH more expensive than homes just across the street. So extremes influence the mean negatively, they distort the mean. Thus, smart media sources and realtors use the “median.”

Use “skewness” to decide whether there are “extremes” in data (to eliminate the mean)

One way to decide whether or not there are extremes in the data set is to just look at the numbers.

That was relatively easy when we had a small number of data points, like our examples with $n=9$.

However, when we have a “real” data set with a large “ n ” or sample size, this is difficult if not impossible.

So an objective way is to look at a number called “skewness.” A computer program like Microsoft Excel or SPSS will “spit out” the skewness for you. I don’t know what the formula is and I do not want to know. (Actually I’ve seen it and it is so complex as to be confusing to me, but I understand what “skewed” data is.)

Here is a rule to use in this class. It is not a “universal rule” in statistics or anything, but it is an objective measure that you can use and that we will use in this class (for testing purposes). You will have to understand the number line to understand this rule.

Here is the rule: if the skewness is BETWEEN -2 and +2, there are NOT extremes in the data. If the skewness is less than -2 or greater than +2, there ARE extremes in the data. When there are extremes in the data do not choose the mean.

-4 -3 -2 -1 0 +1 +2 +3 +4

Above is a “color coded” number line to help you visualize the rule. The green section means there are NOT extremes and the red section means there are extremes and do not use the mean.

-2 < skewness < +2 NO extremes mean is okay

skewness < -2 or skewness > +2 EXTREMES do NOT use mean

When we look at our data set from above with extreme values in the 9th position we see that both age and \$ in wallet had skewness values above 2.0, so we would say there were extremes in the data and not use the mean. In this case we would use the median or the mode, which in this case are identical.

Example of mean *with extreme scores*

data point	age	\$ in wallet
X1	19	100
X2	20	110
X3	22	120
X4	22	120
X5	24	130
X6	24	130
X7	24	130
X8	26	180
X9	75	2000
Σx	256	3020
mean	28.4	335.56
median	24	130
mode	24	130
skewness	2.9	3.0

In the example where we removed the extreme data points from the x9 position, note the skewness for each variable changes and is now between -2 and +2 and is “okay.” We can use the mean as, according to the skewness, there are no extremes in the data.

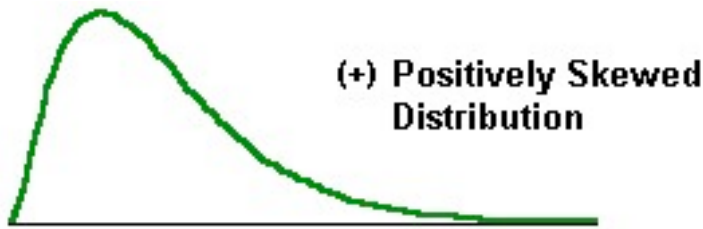
So it does not take a rocket scientist to see that the mean is now a pretty good average, or a single number used to represent a group of numbers. For age 22.6 does a pretty good of representing the average age of the group as a whole and so does \$127.50 for \$ in wallet.

Example of mean without extreme scores

data point	age	\$ in wallet
X1	19	100
X2	20	110
X3	22	120
X4	22	120
X5	24	130
X6	24	130
X7	24	130
X8	26	180
Σx	181	1020
mean	22.6	127.5
skewness	-0.3	1.7

skewness one last time

To help you visualize skewness, below is a figure of skewed distributions: w/ “negatively” skewed data there are some “small” extremes and w/ positively skewed data there are some “large” extremes.



Measures of Dispersion

Introduction

As we saw when we talked about misuses of statistics, sometimes a measure of central tendency, such as a mean, can be misleading. This is why when you really want to describe a variable you need not only a measure of central tendency, but you also need a measure of dispersion.

A measure of dispersion is a number that describes “How much the data varies.” Are the data points closely grouped together or spread out? Do the values tend to cluster in certain areas (i.e. skewed distributions)? These are the sorts of questions measures of dispersion answer.

One more time: a measure of dispersion is a single number that describes how “spread out” or how “smashed together” a group of numbers are. If the data is “really spread out” the number will be large and if the numbers are really close together, the number will be small.

So for example in a kindergarten classroom, the ages will be closely grouped together, as by definition kids start kindergarten at ages 4 and 5. In a college classroom, however, we would expect the ages of students in the classroom to vary or be more diverse or be more dispersed. At UHWO we tend to get quite few older “retuning” students and so the ages in our classrooms vary quite a bit. In fact we would expect the age in UHWO classrooms to be more dispersed than say UH Manoa which tends to have almost exclusively 18-22 year olds.

PROMPT: IS THAT CLEAR?

Computing Measures of Dispersion

Range

The highest value or data point minus the lowest value. For the data below the range=6.

data point	data (x)
x1	1
x2	3
x3	5
x4	7

Range = highest number – smallest number. $7-1=6$

Absolute Deviation

We compare all of the values to some sort of measure of central tendency (usually the mean). So we can see how the data as a whole differs or deviates from the mean.

data point	data (x)	mean	x - mean
x1	1	4	-3
x2	3	4	-1
x3	5	4	1
x4	7	4	3
Sum	16		0
mean	4		0

You can see the problem here. We come up w/ zero.

One way to “erase” the 0 is to square the negative numbers.

data point	data (x)	mean	x - mean	(x-mean) ²
x1	1	4	-3	9
x2	3	4	-1	1
x3	5	4	1	1
x4	7	4	3	9
Sum	16		0	20
Mean	4		0	5

Variance

The example above is the variance. Therefore the variance (5) is nothing more than ***“the average of the sum of the squared deviations from the mean.”***

The problem with the Variance:

Since we squared the values to get rid of the negative signs we have changed the data. How do we “undo” a square? THE SQUARE ROOT!

Standard Deviation

“the square root of the average of the sum of the squared deviations from the mean.”

The square root of 5 = 2.236

Formulas for standard deviation and variance

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N} = \text{population variance}$$

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1} = \text{sample variance}$$

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}} = \text{population standard deviation}$$

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} = \text{sample standard deviation}$$

What’s up with the “n-1” in the sample formulas?

It is to adjust for sampling error when the sample size (n) is small. Recall you would have more sampling error with a small sample that you would with a bigger sample. Common sense tells you

that I hope. If you had a population of 1,000 people would you rather have a sample size of 10 (n=10) or a sample size of 100 (n=100)?

So the n-1 adjusts mathematically for sampling error when sample size is small. as n grows it has less and less of an effect on the fraction. Try it. Divide any number -- literally any number -- by 9 (n=10) then 8 (n=9). Then divide by 99 (n=100) and then by 98 (n=99). Notice the difference from the two sets of numbers. The difference will be smaller for the two small samples than the two big samples.

$1/8 = 0.125$ $1/9 = 0.1111111111$ difference = 0.013888889

$1/98 = 0.0102$ $1/99 = 0.0101$ difference = 0.000103072

Example 1: Tightly Grouped Data means small SD

Note, in the three examples below we are using the sample formula for standard deviation, so we are assuming the data comes from a sample. So each of the formulas use the n-1 in the bottom part of the fraction.

This is a tightly grouped data set. Notice how the mean, median, and mode are all about the same and the SD is very small. 'the square root of the average of the sum of the squared deviations from the mean' is = to .83666 years. So another way of looking at this # we have for this SD is, "on average" each of the data points differ from the mean about .84 years.

data pt	age	mean	x-mean	(x-mean) ²
x1	24	25.2	-1.2	1.44
x2	25	25.2	-0.2	0.04
x3	25	25.2	-0.2	0.04
x4	26	25.2	0.8	0.64
x5	26	25.2	0.8	0.64
Sum	126			2.8
Mean	25.2		Ex/n-1	0.7
Sd	0.83666	years		

Example 2: an extreme value makes the data less tightly grouped and thus has a higher SD

Now look at this example where one extreme value replaces data point x5:

data pt	age	mean	x-mean	(x-mean) ²
x1	24	30	-6	36
x2	25	30	-5	25
x3	25	30	-5	25
x4	26	30	-4	16
x5	50	30	20	400
Sum	150			502
Mean	30		Ex/n-1	125.5
Sd	11.2026	Years		

Notice how it changes the mean considerably. And notice now that the data is more dispersed, the SD gets much bigger. In this example, 'the square root of the average of the sum of the squared deviations from the mean' is = to 11.2 years. So another way of looking at this # we have for this SD is, "on average" each of the data points differ from the mean about 11.2 years.

Example 3: no variation in data? SD=0

data pt	age	mean	x-mean	(x-mean) ²
x1	24	24	0	0
x2	24	24	0	0
x3	24	24	0	0
x4	24	24	0	0
x5	24	24	0	0
Sum	120			0
Mean	24		$\Sigma x/n-1$	0
sd=0				

Now look what happens when there is zero variation in the data. Since all of the data points are exactly equal to the mean, when you compare them (x-mean) it equals 0. You square and add it all up it is zero. The SD=0. So the SD="on average how much to all of these data points vary from the mean?" The answer to that for this group of data is "it does not vary at all." Or "on average there is no difference from the mean." Thus the SD=0 indicating that the data points show ZERO average variation from the mean.

Assumptions for Measures of Dispersion

The assumptions for MCT are much simpler. The most important assumption of which to be aware is the following: both the standard deviation and the variance are based upon the mean, so for either to make sense, the variable MUST be at least interval or ratio. You can compute the standard deviation/variance for nominal and ordinal variables, but they make no sense whatsoever.

I would not worry too much too much if there are "extremes" in the data as with the mean. The standard deviation/variance show "how spread out" the data are, so they will illustrate the extremes, even though technically speaking if the mean was a poor measure of central tendency, the standard deviation/variance will be somewhat poor measures of dispersion as they are both based upon the mean!

So if you have a nominal or ordinal level variable, use the range as your measure of dispersion. The range can be used for all but not very useful – just quick and dirty. But it is the only one that can be used for Nominal or Ordinal data!

Other ways to “summarize” data: cross-tabulation and controls

The following things are not “measures of central tendency” or averages, but they are ways to summarize or describe data. **Right now, the stuff below here is not on the exams for this course.**

Cross Tabulation

Another way to “describe data” is to perform or compute what are called “cross-tabulations” of two variables. Essentially you “compare” two variables that should be related in some way. ***It is best to “compute” cross tabulations using a computer!***

Example 1: Opinion and Gender

Many students do projects where they want to compare two groups on some measure. For example, maybe you think that surfers and non-surfers would have different opinions on the importance of beach access or you think that Democrats and Republicans would have different satisfactions levels with our current President. The point is that you may think that two groups will be different on some second variable or measure.

Let’s pretend you wanted to know if gender affected the amount of domestic “house” work done by heterosexual married couples. The classical hypothesis would be that, despite the gains of “the women’s movement” women still tend to do most of the “housework.” By the way we are looking at “aggregate data – so we can say things about groups of people but not specific individuals. For example, I do far more of the cooking than my wife. It does not follow that this is true for US couples in general. For example, Professor Chinen assures me that scientific evidence shows that women do more “domestic chores” than men do. So, this is true for a group of people, but not necessarily for individuals.

Let’s also pretend you have two variables: gender and opinion on housework. (NOTE: for simplicity we are measuring “opinion about who does the most housework” rather than actual housework.)

Table 1: Gender

Women	10
Men	10
Total	20

Table 2: “I do most of the housework”

Agree	10
Disagree	10
Total	20

Looking at the tables above separately we have no idea as to whether or not there is a relationship between gender and the amount of housework. What we do is combine the variables into one table.

Table 3: Crosstabulation of Gender and Housework

Most work?	Women	Men
agree	8	5
disagree	2	5
total	10	10

Looking at the tables above we can see there **appears to be a relationship** between gender and opinion about who does the most housework. Specifically 8 out of 10 women agree they do more housework compared to only 5 out of 10 men.

Example 2: Religion and Birth Control

Perhaps we do a study to investigate the relationship between religious preference and use of birth control and come up with the following numbers. First we have two variables: Religion – Catholic or Protestant and Use Birth Control – yes or no. *We would expect Catholic to use birth control less as it is official church doctrine that Catholics are not supposed to use birth control.*

Table 4: Religion

Catholic	190
Protestant	160
Total	350

Table 5: Use of Birth Control

No	180
Yes	170
Total	350

Looking at the tables above separately we have no idea as to whether or not there is a relationship between being Catholic and not using birth control. What we do is combine the variables “use birth control” and “religion” into one table or cross tabulation.

Table 6: Crosstabulation of Use of Birth Control by Religion

Use BC?	Protestant n=160	Catholic n=190
no	75	105
yes	85	85
total	160	190

Table 7: Percentages of those not using Birth Control by Religion

	Protestant	Catholic
Don't Use BC Rate in %	47%	55%
	75 of 160	105 of 190

Looking at the two tables above, we might conclude that there is a relationship between being Catholic and not using birth control. However we have not “controlled for education.” See below.

Using a statistical control

For example, to help reduce the risk of nonspuriousness, statistical controls can be used. For example, perhaps there is a third or “hidden” variable that is making it seem as if Catholics use birth control more than Protestants – when in reality that is not true! For example what if we “controlled” for educational level? Could it be that more educated women are more likely to use birth control regardless of religion? We can “statistically control” for education to see if that is the case.

statistical control = "one variable is held constant so the relationship between two or more other variables can be assessed without the influence of variation in the control variable" (Schutt: 164)

Using statistical controls can help reduce of spuriousness by holding one variable constant. But when we control for educational level we come up with the following numbers (see tables 3 and 4)

Table 8 Crosstabulation of Use of Birth Control by Religion (Controlling for Education Level)

Use BC?	High Ed		Low Ed	
	Protestant n=100	Catholic N=50	Protestant n=50	Catholic n=150
No	40	20	30	90
Yes	60	30	20	60

Table 9 Percentages of those not using Birth Control by Religion (Controlling for Education Level)

Use BC?	High Ed		Low Ed	
	Protestant n=100	Catholic n=50	Protestant n=50	Catholic n=150
No	40% 40 of 100	40% 20 of 50	60% 30 of 50	60% 90 of 150
Yes	60	30	20	60

We can see that educational level may play a very significant role in whether or not one uses birth control as opposed to religious preference. It appears now that religion does not play a very great role in whether or not one uses birth control. We have reduced the risk of spuriousness by controlling for education level.