Lesson 3
Association Between Variables

Outline of the Lesson

In Lesson 1 we described a survey administered the first day of class, with these twelve questions:

1. What is your gender (Male, Female)? _____
2. What is your class year? (Freshman, Sophomore, Junior, Senior, Other) _____
3. How many states have you visited (or lived in or even just driven through)?
4. Have you ever been a smoker (at least ½ pack a day)? (Yes/No) _____
5. How would you rate yourself politically? (1 = very liberal, 2 = liberal, 3 = slightly liberal, 4 = moderate, 5 = slightly conservative, 6 = conservative, 7 = very conservative)
6. What is your height (in inches)? _____
7. How many times a week (on average) do you read a daily newspaper? _____
8. What is your political affiliation? (D = Democrat, R = Republican, I = Independent) _____
9. Do you have a paying job during the school year at which you work on average at least 10 hours a week? (Yes/No) _____
10. How many minutes a day (on average) do you watch t.v.?_____
11. What is the distance (in miles) between your home and this campus? _____
12. Aside from class time, how many hours a week, on average, do you expect to spend studying and completing assignments for this course? _____

Based on this survey, we identified several categorical and numerical variables that could be used to describe the individuals in that class, as shown here (the numbers in parenthesis indicate the question number):

Categorical	Numerical
Gender (1)	States (3)
Class Yr (2)	Height (6)
Smoke (4)	Newspaper (7)
Politics (5)	TV (10)
Political Party (8)	Commute (11)
Job (9)	Course Time (12)

We commented that question #5 was really categorical, although the answers were written as numbers. But, since the numbers were on a scale from very liberal to very conservative, it does make sense to treat the variable as numerical for certain computations and analyses.

The following table contains the responses obtained from each of the students in that particular class. We will refer to this data from time to time throughout this lesson.

Gender	Class Yr	States	Smoke	Politics	Height	Newspaper	Political Party	Job	TV	Commute	Course Time
Male	Sophomore	31	Yes	2	72	2	D	Yes	120	0.5	1.5
Female	Sophomore	19	No	4	71	1	R	Yes	30	0	5
Male	Junior	16	No	2	70	1	D	No	60	0.3	7.5
Female	Junior	24	No		62	1	R	Yes	3	0.5	5
Female	Sophomore	43	No	7	63	4	R	Yes	120	1.5	8
Female	Junior	22	Yes	5	65	0	R	Yes	0	24	10
Male	Sophomore	17	No	2	60	3	D	No	90	1.5	3
Male	Sophomore	10	Yes	3	57	1	I	No	90	1.5	3
Female	Freshman	11	No	3	67	0	D	Yes	90	0	8
Female	Senior	16	No	5	60	0	R	Yes	120	0.5	3
Female	Freshman	8	No	2	66	0	D	Yes	180	6	7
Male	Sophomore	14	No	4	73	1	R	No	60	0	5
Female	Sophomore	33	Yes	4	60	0	R	Yes	60	1	6
Female	Sophomore	16	Yes	4	64.5	0	R	No	120	0.2	7
Female	Sophomore	13	Yes	6	68	2	R	No	60	1	5
Male	Sophomore	26	No	6	73	0	I	Yes	30	20	3
Female	Junior	11	No	6	64	0	R	Yes	60	0.5	5
Male	Sophomore	22	No	4	69	3	D	No	60	1	6
Female	Freshman	10	No	4	61	1	D	No	180	0.2	14
Male	Sophomore	13	Yes	1	72	0	D	No	120	150	3.5
Female	Freshman	21	No	5	64	2	I	Yes	30	110	7
Female	Freshman	13	No	4	63	0		No	60	0	2
Male	Senior	11	No	5	67	3	R	Yes	120	1	5
Female	Freshman	7	Yes	4	67	1	D	No	120	0.5	5
Male	Freshman	12	No	2	69	0	D	Yes	0	30	2
Male	Junior	12	No	3	74	1	R	No	120	0.5	2.5
Male	Sophomore	18	No	4	68	3	D	No	60	0	4
Male	Freshman	11	No	4	70	1.5	R	Yes	90	160	3.5
Male	Freshman	12	No	2	69	2	D	No	120	0.5	12.5
Male	Sophomore	9	Yes	6	66	3	R	No	120	1	3

In Lesson 2, we learned about ways to analyze a single variable. For categorical variables, the analysis concentrated on proportions (that is, percentages or probabilities), and we used pie charts and bar graphs to provide a graphical description of the data. For numerical variables we studied several graphical representations: dot plots, stem-and-leaf plots, histograms, box plots. We learned about the mean and the median as measures of center, and about the standard deviation and the interquartile range (along with quartiles and the five-number summary) as measures of spread.

At this point we change our emphasis slightly. Instead of studying a single variable, we will study two variables at once. This will allow us analyze questions such as the following:

- Have the women in that class visited more states than the men?
- Are the men in that class taller than the women?
- Are smokers in the class more likely to have a job than non-smokers?
- Are students in the class who have visited a lot of states more likely to be conservative?
- Do students in the class who spend more time watching TV spend less time studying?
- Are the juniors and seniors in the class more likely to be Democrats than the freshmen and sophomores?

Each of these questions relates to two variables from that survey.

Exercise 1[1]: For each question, identify by name the two variables the question relates to. Also, identify the two variables as N/N (both numerical), or C/C (both categorical), or C/N (one categorical and the other numerical).

a. Have the women in that class visited more states than the men?
b. Are the men in that class taller than the women?
c. Are smokers in the class more likely to have a job than non-smokers?
d. Are students in the class who have visited a lot of states more likely to be conservative?
e. Do students in the class who spend more time watching TV spend less time studying?
f. Are the juniors and seniors in the class more likely to be Democrats than the freshmen and sophomores?

There is a generic way to phrase each of these questions, once we have identified the variables involved. In fact, there are several ways to phrase the questions. For example, consider the first question: "Have the women in that class visited more states than the men?" Before we go any further, however, let's remove the built-in "bias" from our question, rewriting it as:

- Have the women in that class visited more states than the men, fewer states than the men, or about the same number?

We could do the same for each of the other questions. Now, we are wondering if there is some kind of connection between a person's gender and their amount of traveling. If the men and the women tend to have visited about the same number of states, then there is no connection between the two variables. If, however, persons of one gender tend to have visited more states than persons of the other gender, then for the students in the particular class that was surveyed there definitely is a connection between a person's gender and the number of states they have visited.

Some synonyms for *connection* are *relationship*, or *association*, or *correlation*. Some statisticians and authors tend to use *association* when both variables are categorical and *correlation* when both are numerical. However, the various terms are essentially interchangeable.

Here is the question about gender and number of states, written using the *association* terminology. The word "association" could be replaced by any of the synonyms we just listed.

[1] Solutions to the exercises may be found at the end of the lesson.

- For the students in the class that was surveyed, is there an *association* between gender and the number of states visited?
    - A "no" answer implies that men and women have visited about the same number of states.
    - A "yes" answer" implies either that men have visited more states, or that women have visited more states.

There is one other way this question is frequently phrased:

- For the students in the class that was surveyed, does the number of states visited *depend on* the gender?
    - A "no" answer implies that men and women have visited about the same number of states, and we say that Gender and States are *independent*.
    - A "yes" answer" implies either that men have visited more states, or that women have visited more states.

Explanatory and response variables

When there *is* an association (or a suspected association) between two variables, it is sometimes but not always possible to identify one of the variables as somehow "explaining" the association. For example, if we study age at death and smoking status, one might believe that any association we uncover is in some sense explained by whether or not the person smokes. In this case we would identify the smoking status variable as the **explanatory variable**, and the age at death variable as the **response variable**.

Sometimes which variable we consider the explanatory variable may be reflected in the way the question is phrased. For example, consider the question, "Are smokers in the class more likely to have a job than non-smokers?" This implies that we group the students by whether or not they smoke, thus making smoking status (what we called "Smoke") the explanatory variable and employment status ("Job") the response variable. But if the question had been, "Are employed students more or less likely to smoke that those that are not employed?" we imply a grouping by employment status, making "Job" the explanatory variable and "Smoke" the response variable.

As you can see, the distinction can be subtle. This is especially true when both variables are categorical, or when both variables are numerical. We will discuss this further in the remainder of the lesson. Fortunately, for most situations the more important question is the neutral question, "Is there an association between the variables?"

3.1 – Association between a categorical variable and a numerical variable

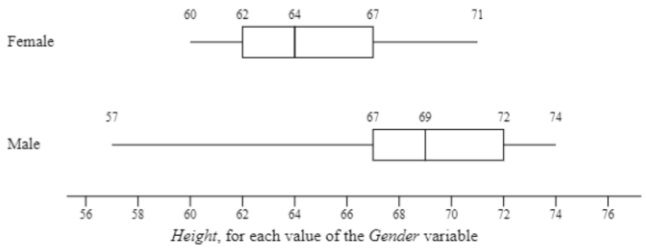
Note: In most cases involving one categorical variable and one numerical variable, the categorical variable will be the explanatory variable, and the numerical variable the response variable. We think of the categorical variable as dividing our subjects up into groups, and then we wonder whether the numerical variable results depend on which group you are in.

For example, in the question about gender and number of states visited, gender would be the explanatory variable and number of states the response variable.

Graphical analysis of association

An important tool in visualizing and analyzing categorical/numerical association is the **side by side box plot**. Thinking of the categorical variable as identifying two or more groups, we simply create a

box plot of the numerical variable for each group, graphing them on the same scale. For example, here is a side by side box plot of the heights from the first-day survey described above:



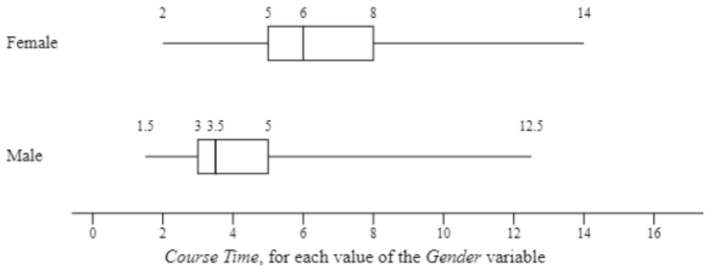
The graph makes it very clear that, for the individuals in that particular class, the males were in fact taller than the females. Using some of our terminology, the graph illustrates that, for these students:

- There is an association (connection, relationship, correlation) between gender and height.
- Height depends on gender.
- Gender and height are *not* independent.

Comment. The term *side by side* is used to describe this graph, even though you might be thinking that the two graphs are really "one-above-the-other." In some situations you may see the same graph using a vertical rather than a horizontal box layout, which would result in graphs that are physically side by side.

The same can be said for other "side by side" graphs – frequently they are laid out on the page one above the other rather than side by side.

Exercise 2: Based on this graph for the students in that same class, does there appear to be an association between *Gender* and *Course Time* (the amount of time the students expect to spend per week on the statistics class)?



In addition to side by side box plots, it is possible to use other graphical techniques. For example, here is a graph for the same question as in Exercise 2 but using **side by side histograms**.

Clearly the answers for the males tended to be smaller than those for the females. As one specific example, fully 60% of the males gave an answer less than 4; fewer than 15% of the female answers were in that range.

Numerical analysis of association

Along with graphical techniques, the most common numerical tool for analyzing association between a categorical and a numerical variable is to compare means. (Although the box plot is based on medians, when we do numerical comparisons we generally use means.) For this class, for example, the mean for the women in the class was 6.5 and that for the men was 4.3. Clearly the women's mean was larger than the men's. There are two ways we can numerically combine these numbers, and each is a reasonable approach for our analysis.

1. ***Subtraction***. We can subtract 6.5 – 4.3 = 2.2, and report, "The mean time the women plan to spend per week is 2.2 hours more than that for the men."
    a. Notice that since we subtracted "mean for women minus mean for men," a positive number indicates that women plan to spend more time than the men.
    b. If we had chosen to subtract "mean for men minus mean for women" we would have calculated 4.3 – 6.5 = –2.2, and the negative number would indicate that the men plan to spend less time than the women.

2. ***Division*.** We can divide instead of subtracting, again in either order with a different interpretation depending on the order.
    a. If we calculate "mean for women divided by mean for men" the calculation yields 6.5/4.3 = 1.51. Because the ratio is larger than 1, we know the women plan to spend more time.
        i. This result might be reported as "The women plan to spend about one and a half times as long studying for the class per week as the men."
        ii. Another way to describe this is, "The women's anticipated study time per week is 50% higher than that for the men." The 50% comes from the .51 portion of the 1.51 calculated result. (More precisely, we should say 51%, but this is so close to 50% that it might be rounded to 50%.)
    b. On the other hand, we can calculate "mean for men divided by mean for women," yielding 4.3/6.5 = 0.66. A number lower than 1 indicates that the men plan to spend less time.

i. This might be reported as "Men plan to spend about 2/3 as much time as women do for this course."
ii. Another rendition might be, "Men's anticipated study times are only 66% as high as the anticipated times for women."

Here is another example:

Example. Researchers randomly assign mice to two groups, and withhold food from the first group for three days. The average time to complete a maze is 302.17 seconds for this group. For the group that is fed normally, the average time is 267.63 seconds. Calculate the ratio of the means, rounding to two places, and use the results to fill in the blanks in these statements. (Group 1 is the group who had food withheld.)

On average, the first group's time was _____ % higher than the second group's time. Put another way, the second group's time was _____ % as large as that for the first group.

Solution: mean1/mean2 = 302.17/267.63 = 1.13. Notice that if the means had been the same the ratio would have been 1. So the .13 part of the answer represents the excess time for the first group. As a percent, .13 is 13%.

mean2/mean1 = 267.63/302.17 = 0.89, and 0.89 is the same as 89%.

On average, the first group's time was __13__ % higher than the second group's time. Put another way, the second group's time was __89__ % as large as that for the first group.

The app at the following link provides practice calculating and interpreting the ratio of two means.

Ratio of means

Exercise 3: A recent study reported that employed women averaged 16 hours of housework per week, and employed men averaged 11 hour per week. For the people in the study, compare the mean amount of time spent on housework for men and women. Use both subtraction and division. Give a sentence reporting each result "in the context of the problem," similar to how it might be reported in the media. Also, report your conclusions using words such as *association, depends on, independent.*

3.2 – Association between Two Categorical Variables

One of the questions we asked in the introduction was this: "Are smokers in the class more likely to have a job than non-smokers?" In the terminology we have been using, we would rephrase this as, "Is there an association between smoking and having a job for this group of students?" Notice that in this case both variables are categorical variables – in fact, they are very simple "yes/no" questions with only two categories for each variable.

Note: When both variables are categorical, it may not be obvious which should be treated as the explanatory variable and which as the response variable. In general, we think of the explanatory variable as splitting the subjects into groups, where for each group we will analyze the occurrence of the response variable answers.

For our example, we will use Smoke as the explanatory variable and Job as the response variable, although the other choice is equally defensible for this example.

Contingency tables and conditional proportions

Earlier in this lesson we reproduced the data from the particular class of students that participated in the survey. However, to begin our analysis we need to summarize that data. A very useful format is that of a *contingency table*. We make each possible value of the explanatory variable a row of the table, and each possible value of the response variable a column of the table, as illustrated here:

	Have a job?		
Smoke?	Yes	No	Totals
Yes			
No			
Totals			

The table has a place for listing how many smokers have a job, how many do not, how many non-smokers have a job, and how many do not. Notice that in addition to the counts, we will enter the total for the "Smoke/yes" row, the "Smoke/no" row, the "Job/yes" column, and the "Job/no" column. Finally, there is a place for the sum of the "Totals" row, which should match with the sum of the "Totals" column; the result should be the number of people who participated in the survey.

Then we examine the data, counting how many individuals fall into each category in the table. Here is the resulting contingency table, with the counts and totals filled in.

	Have a job?		
Smoke?	Yes	No	Totals
Yes	3	6	9
No	12	9	21
Totals	15	15	30

Exercise 4: Create a contingency table for the Gender and Political Party variables in the data, with Gender as the explanatory variable. (Just ignore the person who didn't answer the Political Party question.)

In Lesson 1 you calculated a variety of different types of proportions. Using the contingency table to organize the various counts concerning two categorical variables makes it relatively easy to answer questions of the type posed in Lesson 1. This is especially true if you have calculated the totals for each value of the explanatory variable (the row totals) and the totals for each value of the response variable (the column totals).

Some questions you might ask are questions about the entire group (in this case, the entire class that responded to that survey).

Example: Answer these questions for the data presented in the contingency table above.

a. What proportion of the students in the class have a job?

Solution: There are 30 students, and 15 have a job. So, $\frac{15}{30} = 0.5 = 50\%$

b. What proportion are smokers?
Solution: There are 30 students, of whom 9 are smokers, so $\frac{9}{30} = 0.3 = 30\%$. Notice that this question is the same as the question, "What proportion of the students in the class are smokers?" The phrase "of the students in the class" is implied.

c. What proportion are non-smokers with a job?
Solution: This is still a question about the entire class, so the denominator is still 30. For the numerator, the students we are counting are those who answered "no" to the question about smoking and "yes" to the question about having a job. The contingency table show that count is 12, so 12 of the 30 students are non-smokers with a job. The proportion is $\frac{12}{30} = 0.4 = 40\%$

It is also possible to consider proportion questions that relate not to the entire class, but only to some smaller group of students in the class. These proportions are referred to as *conditional proportions* (there is a condition which describes the group for which the question is being asked). Here again is the contingency table for the entire class:

	Have a job?		
Smoke?	Yes	No	Totals
Yes	3	6	9
No	12	9	21
Totals	15	15	30

Example: Answer these questions for the data presented in the contingency table above.

a. What proportion of the smokers have a job?
Solution: This is a question concerning only the smokers in the class, whose counts appear in the first row of the table. As we can see from the table, there are 3 smokers who have a job and 6 who do not (a total of 9 smokers). Of the 9 smokers, 3 have a job, so the proportion is $\frac{3}{9} = 0.3333 = 33.33\%$.

b. What proportion of the non-smokers have a job?
Solution: This time we look only at the non-smokers (row 2 of the table). Of the 21 non-smokers, 12 answered yes to the question about a job. The proportion is $\frac{12}{21} = 0.5714 = 57.14\%$.

c. What proportion of the students with a job are non-smokers?
Solution: This is a question about the students with a job, whose counts are in the first column of the table. There are 15 of these students, and 12 of them are non-smokers, so the proportion is $\frac{12}{15} = 0.8 = 80\%$.

CAUTION: You need to pay attention when you read the question. Here are three questions we have answered which are quite similar in appearance but which have totally different meanings:

- What proportion are non-smokers with a job? This is a question about the entire class; it could have been phrased as, "What proportion *of the students in the class* are non-smokers with a job."
- What proportion of the non-smokers have a job? This is a question about the non-smokers, whose counts are contained in the second row of the table.
- What proportion of the students with a job are non-smokers? This is a question about the students who have a job, whose counts are contained in the first column of the table.

Exercise 5: Use the contingency table from Exercise 4, reproduced here, to calculate these proportions. (Again we totally ignore the one person who didn't answer both questions on the survey.)

	Party			
Gender	D	R	I	Totals
Male	8	5	2	15
Female	4	9	1	14
Totals	12	14	3	29

a. The proportion of males who are Republican.
b. The proportion who are male Democrats.
c. The proportion of the females who are Democrats.
d. The proportion of the Democrats who are female.

The app at the following link provides practice calculating proportions, including conditional proportions, for data presented in a contingency table.

Proportions for contingency tables

Connections to probability

Here again is the contingency table relating the "smoking" variable to the "has a job" variable:

	Have a job?		
Smoke?	Yes	No	Totals
Yes	3	6	9
No	12	9	21
Totals	15	15	30

We can see that 50% of that entire class of students had a job (15 out of 30). Just to keep you always aware of the connection, we want to point out once again that we can think of this in terms of probability: The probability that a randomly selected individual from that class has a job is 50% or 0.5 (15 divided by 30). In symbols, P(Job) = 0.5.

The contingency table also contains conditional proportions. For example, 3 of the 9 smokers (33.33%) have a job. This can also be thought of in terms of probability, and the resulting probability is a "conditional probability." Here are a couple of ways to phrase the question; both have the same meaning although you may find one or the other way of saying it more meaningful:

- If you randomly select one of the smokers from the class, what is the probability that person has a job? Since there are only 9 smokers to select from, and 3 of them have a job, the answer is 3/9 = 0.3333 or 33.33%.
- What is the probability that a randomly selected person has a job, given that they are a smoker? The phrase "given that they are a smoker" indicates that the person selected is one of those 9 smokers, 3 of whom have a job, so the answer is again 3/9 = 0.3333 or 33.33%.

- The notation for this is P(Job | Smoker) where the vertical bar is read as "given" or "given that they are." The important thing to realize is that the "| Smoker" part of the notation tells us we are looking *only at the 9 people who are smokers.*

Note: It is also possible to pose conditional probability questions about the columns of the table. For example:

a. If you select one of the people without a job at random, what is the probability they are not a smoker? There are 15 people without a job; 9 of them are non-smokers; so the probability is 9/15 = 60%.
b. Using the word "given" we could ask this as: What is the probability that a person in the class is not a smoker, given that they do not have a job.
c. The notation for this would be P(Non-smoker | No job)

Exercise 6: Use the contingency table from Exercise 4, reproduced here, to calculate these probabilities. (Again we totally ignore the one person who didn't answer both questions on the survey.)

	Party			
Gender	D	R	I	Totals
Male	8	5	2	15
Female	4	9	1	14
Totals	12	14	3	29

a. The probability that a randomly selected student is a male Democrat.
b. The probability that a randomly selected male is an Independent.
c. The probability that a randomly selected Independent is male.
d. The probability that a randomly selected individual is male, given that they are Republican.
e. P(Republican | male)

The app at the following link provides practice calculating probabilities, including conditional probabilities, for data presented in a contingency table.

Probabilities for contingency tables

Using proportions to analyze association

The proportions and conditional proportions we have been calculating give us a first step to aid in analyzing the possible association between two categorical variables. The idea is that we look at the proportion of the entire group falling in each of the response variable's categories. We also look at the corresponding conditional proportions for each row of the table – for each group identified by the explanatory variable. We can enter these proportions along with the counts, as illustrated in this version of our "smoking/has a job" contingency table:

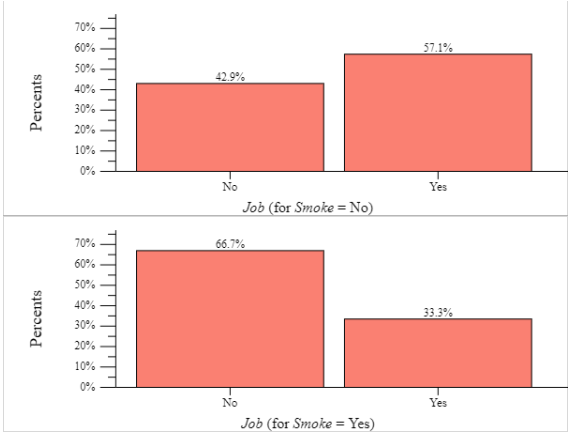
	Have a job?		
Smoke?	Yes	No	Totals
Yes	3 *33.33%*	6 *66.67%*	9
No	12 *57.14%*	9 *42.86%*	21
Totals	15 *50%*	15 *50%*	30

For the smokers, 3 of the 9 smokers have a job, which is 33.33%. On the other hand, 12 of the 21 non-smokers have a job, which is 57.14%. For the group as a whole, 15 of the 30 students (50%) have a job.

Based on these results, for the students in this class we would say that there *is* an association between smoking and having a job. Having a job *depends on* whether or not they smoke. The students who smoke are much less likely to have a job than those who do not smoke. The variables are *not* independent.

Graphical analysis of association

Just as for the case of categorical/numerical association, graphs can make it easier to see categorical/categorical association. A useful tool is **side by side bar graphs**, as illustrated here for our example:



The graph shows visually that non-smokers in that class of students are more likely to have a job than smokers.

Note: The bar chart must be created using proportions or percents rather than raw counts in order for the comparisons to be meaningful.

Numerical analysis of association

Just as for the numerical analysis of categorical/numerical association, there are two useful ways to compare the conditional proportions for the categorical/categorical association. We can subtract the conditional proportions, or we can divide them. We illustrate each approach for our example.

1. **Subtraction.** Concentrating on the "yes column" of our table, we subtract: 57.14% - 33.33% = 23.81% = 0.2381. In this group, the proportion of non-smokers with a job is 0.2381 larger than the proportion of smokers.

2. **Division**. Since 57.14% / 33.33% = 1.71, we can say that non-smokers are almost twice as likely to have a job as smokers; they are 71% more likely to have a job.
   Or, reversing the division, 33.33%/57.14% = 0.58, so smokers are only 58% as likely to have a job as non-smokers.

Comment. Suppose group 1 has 0.70 = 70% proportion, and group 2 has 0.20 = 20%. When we subtract, we obtain 70% - 20% = 50% = 0.50. Should we say that group 1 proportion is 50% larger than group 2. The answer is no, since the terminology "50% larger" is widely used with a different meaning. To be 50% larger than 0.20 means 0.10 larger, or 0.30. In fact, 0.70 is more than 3 times as large as 0.20. When we use subtraction, we should report our answers using the decimal difference, not the percentage form of the difference.

Exercise 7: Here is a contingency table summarizing answers in a small town to a survey with two questions: "Do you have children in the family?" and "Is paying bills a major concern in the family?"

	Bills a major concern?		
Have children?	Yes	No	Totals
Yes	510	340	850
No	473	437	910
Totals	983	777	1760

Do the following using the data:

a. Using "do you have children" as the explanatory variable, add the conditional proportions for each value of the explanatory variable; the sum of the rows should be 100%.
b. Create side by side bar charts.
c. Use subtraction and division to compare the proportions answering "yes" to the question about bills, for the two groups.
d. In that small town, is there an association between having children and viewing bills as a major concern?

Exercise 8: Consider the survey of the previous exercise. In another town, the results were somewhat different, as shown here:

	Bills a major concern?		
Have children?	Yes	No	Totals
Yes	492	328	820
No	459	306	765
Totals	951	634	1585

Do the following using the data:

a. Using "do you have children" as the explanatory variable, add the conditional proportions to the table; the sum of the rows should be 100%.
b. In that small town, is there an association between having children and viewing bills as a major concern?

If you worked exercise 8, you may have noticed the following. The percentage answering "yes" to the question about bills was the same for both groups (60%). As a result, the percentage for the entire town was also 60%. This is a pattern which always holds:

> *If the percentages for each possible value of the explanatory variable are the same, the percentages for the entire set of data will be the same also. Put another way, if there is no association between the variables, the percentages for each group will match the percentages for the entire set of data.*

This pattern is also true for situations where the variables have multiple values, as illustrated in the following exercise.

Exercise 9: Suppose a survey question has three possible answers (call them choice (a), choice (b), and choice (c)). Of the 1200 males surveyed, 52% choose (a), 37% (b), and the rest (c). For the 900 females, the percentages are the same. Fill in the counts in this contingency table. Then use the results to calculate the percentages for the entire set of data.

	(a)	(b)	(c)	Totals
Males				1200
Females				900
Totals				2100

3.3 – Association between Two Numerical Variables

The methods used in the previous two sections are quite similar. The graphical analysis utilized *side by side* graphs (bar charts, histograms, boxplots, etc.), and the numerical analysis involved either subtracting or dividing the summary information for the two groups (means for numerical variables, proportions for categorical). When both variables are numerical, the methods used are significantly different.

You are already quite familiar with associations between numerical variables, from your experience with high-school algebra. For example, for a person who earns $10 an hour at his or her job, there is an association between the hours worked and the amount earned. When that person works more hours, that person will receive more pay. In fact, unless overtime is involved, there is a quite simple algebraic formula that spells out the association: Pay = 10 * Hours. If you graph this equation, the result is a straight line, so we say the association is linear.

With variables of the type we are considering in this course, the association (if any) is almost never this straightforward. When we graph the data, it usually does not give a perfect linear graph. However, we will be looking for two things similar to the Pay/Hours situation: 1) does there appear to be an *approximate* straight-line relationship between the variables; and 2) if so, what algebraic equation can we use to describe this relationship?

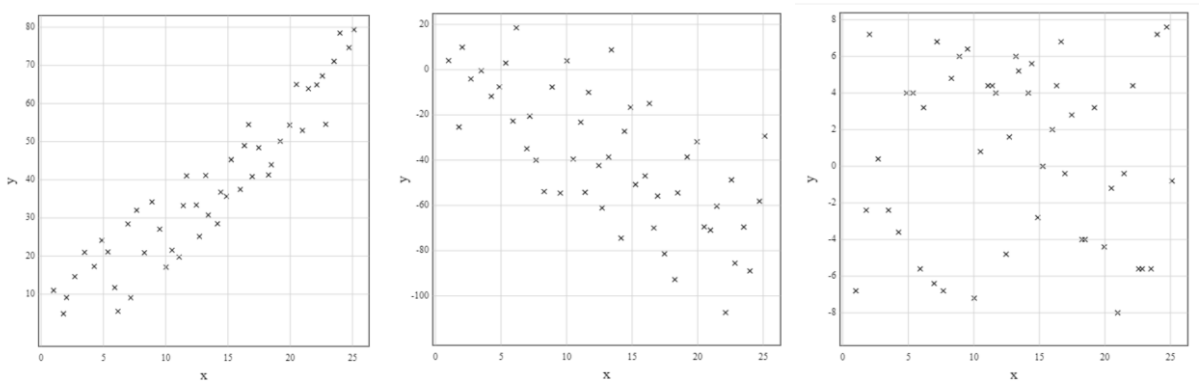
Note: When both variables are numerical, once again it may not be obvious which should be treated as the explanatory variable and which as the response variable. In general, we think of the response variable as the one we are studying, with the explanatory variable "explaining" the results we see in the response variable. However, this does not necessarily imply a cause/effect relationship.

If we are able to develop an algebraic equation to describe the relationship, the explanatory variable is the one we will "plug in" to the equation to calculate the value of the response variable. In algebra you called this the independent variable and the variable you calculated the dependent variable, but we will stick with the explanatory/response terminology in this course.

Graphical analysis of association

The graph we use when both variables are numerical is called a *scatterplot*. It is very similar to what you did in high-school algebra. We make the explanatory variable the *x*-axis, with the response variable the *y*-axis, and we simply plot points based on our data. This can be done by hand, but we also discuss the use of technology in section 3.4 of this lesson.

Once we have the points plotted, we look for *form*. In particular, does there appear to be any straight-line pattern in the points? If so, we ask about *direction* and *strength*: is the line increasing from left to right or decreasing; and how close to being on a straight line are the points? For example, consider the following three scatterplots. We answer the questions for each plot.



Form. Does there appear to be any straight-line pattern in the points? The answer seems to be yes for the first two graphs, and probably no for the third.

Direction. Is the line increasing from left to right or decreasing? The first graph has an increasing direction, the second a decreasing direction.

Strength. How close to being on a straight line are the points? In the first graph, the points are "reasonably" close to being on a straight line, less so for the second.

Numerical analysis of association (the correlation coefficient *r*)

There is a calculation, whose details are beyond the scope of this lesson, which yields a number that indicates the *strength* of the linear association in a scatterplot[2]. **You are not responsible for the details of the calculation, but you are responsible for interpreting the result. Refer to section 3.4 of this lesson for information on using technology to calculate this number.**

The number is referred to as the *correlation coefficient*, or more simply as the *correlation*, and it is denoted using the letter *r*. It has these properties:

Value	Meaning
+1	Data is on a perfect straight line with a positive slope
−1	Data is on a perfect straight line with a negative slope
0	Data shows no linear association at all (a "blob" or perhaps some pattern that is not linear).
Between 0 and +1	Data shows some linear association, **increasing** from left to right. The closer *r* is to +1 the stronger the linearity.
Between 0 and −1	Data shows some linear association, **decreasing** from left to right. The closer *r* is to −1 the stronger the linearity.

In describing our data, the closer the correlation coefficient is to either plus or minus one, the stronger the (linear) association between the two variables.

To give you some sense of how the correlation coefficient indicates the strength, here are the correlation coefficients for the three graphs considered above.

Graph 1. $r = 0.9315$. The value is positive because the data is increasing from left to right, and the value 0.9315 is close to 1, capturing a numerical measure of our description of the data as being "reasonably" close to linear.

---

[2] For the record, the formula is

$$r = \frac{1}{n-1}\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$$

where x represents the explanatory variable values, y the response variable values, $\bar{x}$ and $\bar{y}$ are the means of the two variables, and s_x and s_y are the standard deviations of the two variables.

Graph 2. $r = -0.7008$. The value is negative because the data is decreasing from left to right. As we might have expected, the value is not as close to –1 as that for graph 1 was close to +1; intuitively we judged that the strength for the second plot was less than that for the first plot.

Graph 3. $r = -0.025$. As expected, the correlation is very close to 0 – there is very little linearity in this set of data (the data appears, loosely speaking, to be "all over the place").

Numerical analysis of association (the regression line)

If our graphical analysis suggests a linear association, we would like to have a formula for the corresponding straight line. Unless the data lies on a perfect straight line, our formula will be only an approximation, but it will still be useful as a tool for describing the association. Given a value for the explanatory variable, we can plug it into the formula to obtain a prediction for the response variable. We would like to choose the line that is "best" in some sense – we don't want a line that gives lousy predictions!

The method we choose for calculating our line is called the *least squares* method. Here is a very brief explanation. For each data point in the scatterplot, calculate the difference between the y value of the point and the prediction provided by the line, $y_{observed} - y_{predicted}$. Square these, then calculate the average of the squared values[3]. This is called the Mean Square Error (MSE), given as

$$MSE = \frac{\Sigma(y_{observed} - y_{predicted})^2}{n-2}$$

The line that makes this as small as possible will have reduced the overall errors in the predictions as much as possible, and that is the line we use. It is called the *linear regression line* or more simply just the *regression line*. The regression line will have the general form $\hat{y} = a + bx$, where a is the y-intercept of the line and b is the slope.

Comment on notation. It is customary to write this line using the notation $\hat{y}$ (read as *y* hat) rather than just *y* as a reminder that the line may be used to *predict* the response variable value for a given value of the explanatory variable. Computer software or calculators may or may not include the "hat" notation in their output, and it is typically not used when we write the equation using variable names for the explanatory and response variables instead of the generic variable names *x* and *y*.

We omit the details for the calculation of the slope and y-intercept for this line[4]. Fortunately, you are not responsible for these calculations. Once again, refer to section 3.4 of this lesson for details on using technology to do the calculations. The key thing to remember is this: Given a value for the explanatory variable, this line can be used to predict the corresponding value for the response variable.

---

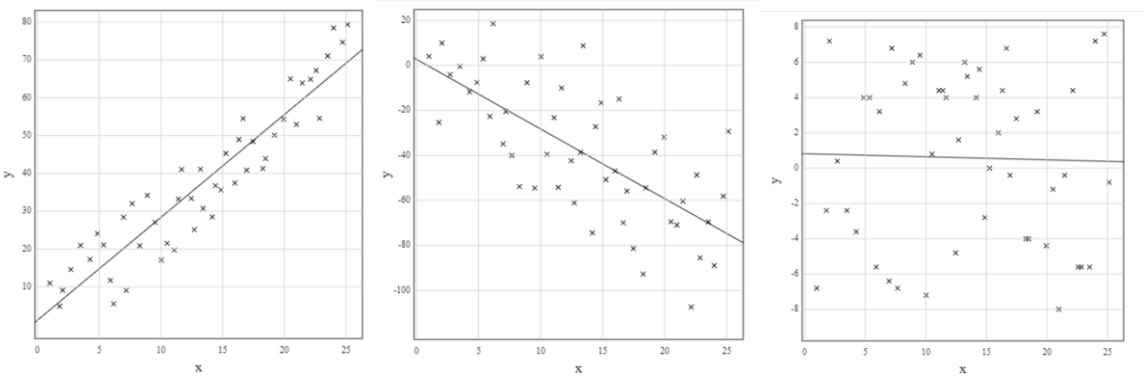
[3] For technical reasons, in calculating this average we divide by *n*-2 rather than dividing by *n*, the number of points. The reason for this is somewhat similar to the reason we divide by *n*-1 in the calculation of standard deviations. (See Lesson 2.)

[4] For the record, the formula for the regression line is $\hat{y} = a + bx$ where the slope b is calculated as $b = r\frac{s_y}{s_x}$ and the y intercept a is calculated as $a = \bar{y} - b\bar{x}$.

Example. For the three plots discussed earlier, the regression lines are, respectively:

1. $\hat{y} = 1.196 + 2.7182x$. As expected, the slope is positive.
2. $\hat{y} = 2.6427 - 3.0953x$. As expected, the slope is negative.
3. $\hat{y} = 0.8225 - 0.0174x$. Even though the plot showed little or no linearity, which was quantified by the value of *r* being very near 0, we can calculate the regression line. However, since the data was not linear in nature, this line is of no use in predicting values for the response variable!

Here are the three plots with the regression line plotted on the same set of axes as the scatterplot:



Example. Is there a correlation between the weight of a car and the gas mileage of that car? In a study of 25 randomly chosen vehicles, a researcher obtained the following regression line:

$$\text{mileage} = 46.51 - 0.0072 * \text{weight}$$

where mileage is in miles per gallon and weight is in pounds.

a. Predict the gas mileage for a vehicle that weighs 5010 pounds (round to 2 places).
Solution. Mileage = $46.51 - 0.0072(5010) = 10.44$ miles per gallon

b. Interpret the slope of the equation.

Solution. In general, the slope of a linear equation is the ratio is $\frac{change\ in\ y}{change\ in\ x}$. For this equation, this indicates $\frac{change\ in\ mileage}{change\ in\ weight}$. The slope is –0.0072; one way to write this as a ratio is $\frac{-0.0072}{1}$, representing a change in weight of 1 pound and a change in gas mileage of –0.0072. A positive change in weight is an increase of the weight, and a negative change in gas mileage represents a decrease for the gas mileage. So we could write:

For every increase of 1 pound for the vehicle weight, the gas mileage decreases by 0.0072 miles per gallon.

Note: Perhaps more meaningfully, we could write the slope as $\frac{-0.0072}{1} = -\frac{0.72}{100}$, by multiplying top and bottom of the fraction by 100. This would lead us to write: ***For every increase of 100 pounds for the vehicle weight, the gas mileage decreases by 0.72 miles per gallon.***

The app at the following link provides practice using a regression line for prediction, and interpreting the slope of the line.

Regression line calculations

Interpretation of r^2

One way to verbalize the significance of the correlation coefficient is actually based on the value of r^2 rather than *r* itself. To understand this interpretation, consider two different ways to predict the value of the explanatory variable for a particular data item:

1. We could always predict the mean $\bar{y}$ of the response variable, no matter what value the explanatory variable has.
2. We could use the regression line to make our prediction.

It turns out that r^2 measures how much better it is to use the regression line. For example, if the explanatory variable is Math SAT and the response variable Verbal SAT, and if r^2 for our data is equal to 0.72, we have this interpretation:

> Using the regression line rather than just using the mean verbal SAT score for the data will improve our predictions overall by 72%.

Here is another way to verbalize the significance of r^2, again using the SAT example to illustrate the idea. If you look only at the values of the response variable (verbal SAT) there is a great deal of variability in the data. Some of this variability is "explained" by the explanatory variable, in the sense that those with higher math SAT scores tend to have higher verbal SAT scores. But because the data points do not lie on a perfect straight line, for any given math SAT score there will be a variety of different verbal SAT scores – there will be variability that is not explained by the regression line's upward slope. The value of r^2 answers the question, "What percent of observed variability in the response variable is explained/predicted by the regression line? For our example:

> The regression line explains/predicts 72% of the variability in verbal SAT scores.

Example. For the three scatterplots we have been considering, these are the values for r^2:

1. 0.8676; the regression line explains/predicts a significant portion of the variability we see for the *y* values in this plot.
2. 0.4911. About half the variability is explained by the regression line; for each value of *x* in the plot, the observed *y* value can be close to the predicted value, but it can also be somewhat far away.
3. 0.0006. The regression line has almost nothing to do with the variability we see in the *y* values for this plot.

3.4 – Using Technology to Analyze Numerical/Numerical Association

This section provides information on using the author's online calculator to create scatterplots, and to calculate the correlation coefficient and the regression line[5]. Once you have read this material, you should practice using it. The app at the following link provides a tool for practicing.

Regression calculations

[5] As you learned in Lesson 2, in addition to the calculator described here for "ad hoc" calculations, the author has provided another calculator for use with data files. See Section 3.6 of this lesson for additional information on using that calculator.

For our explanation, we will be working with a very small set of points:

$$(4,100), (17,165), (12,137), (23,180), (45,320).$$

Scatterplot

Open the provided calculator using the following link, then choose menu option *Graphs* and submenu option *Scatterplot.*

Statistical calculator

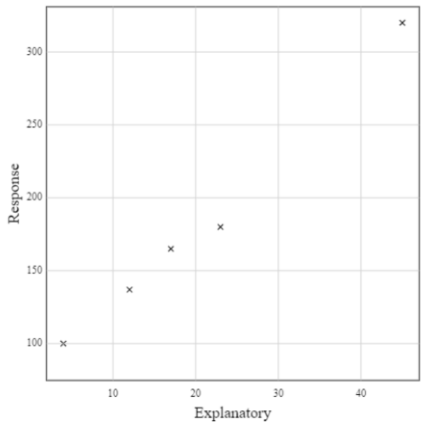
You will obtain a data entry screen for entering the explanatory (x) values and the corresponding response (y) values:

Explanatory Size: 3	Response Size: 3

Since our data consists of 5 points, the first step is to use the drop-down list to change the size of the Explanatory list to 5; the Response list will also automatically increase to size 5. Enter the data as shown here:

Explanatory Size: 5	Response Size: 5
4	**100**
17	**165**
12	**137**
23	**180**
45	**320**

When you press the *Computations* button, you will obtain this scatterplot:



Notice that this data displays fairly strong increasing linearity.

Correlation and Regression

We will use the same data. Since it is a very small dataset, you may choose to simply reenter the data. However, as we did in Lesson 2, you can instead choose *Modify data*, then *Save to file* to obtain a text file similar to this obtained by the author:

```
data-2024-0207-1536 created by "Scatterplot" application
Explanatory,4,17,12,23,45
Response,100,165,137,180,320
```

Then *Exit (return to menu)* and choose option *Descriptive statistics*. You will see two submenu options, *Correlation* and *Correlation with linear regression*. If you do not want to obtain the regression line, you may choose the former. We will choose the latter to obtain both the correlation information and the corresponding regression line. After entering the data manually or by loading the text file we just created, press *Computations* to obtain the following:

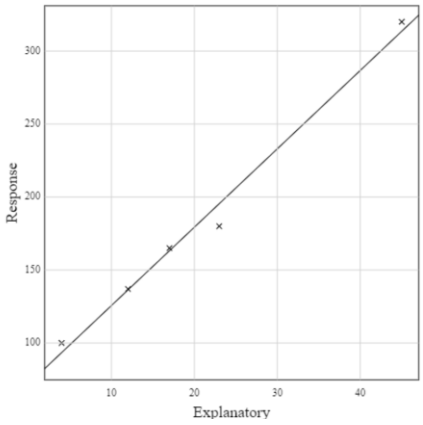
☐ Display scatterplot

r : 0.9942

r^2 : 0.9884

Regression line : $\hat{y} = 71.9603 + 5.3683x$

As expected, r is very close to +1. By checking the *Display scatterplot* checkbox, we can obtain the following scatterplot with the regression line included. Notice that the points are all quite close to the line; that is, the regression line does a very good job of predicting the value of the response variable for the given explanatory variable values.



3.5 – Association is not Causation, and Other Warnings

In this section we will mention and briefly describe some cautions that must be observed when examining the association between two variables. Your instructor may refer you to more detailed information on this topic, perhaps in a textbook or in publicly available materials.

Association is not causation. This may be the most important caveat, and it applies to all three types of association we have examined: categorical/categorical, categorical/numerical, and numerical/numerical. It may be stated a bit more carefully as, "Association between an explanatory variable and a response

variable *does not necessarily imply* that the explanatory variable differences *cause* the response variable differences."

For example, for the class of students we have been using in various examples in these lessons, we found an association between smoking and having a job. Specifically, for the smokers, 3 of the 9 smokers have a job, which is 33.33%. On the other hand, 12 of the 21 non-smokers have a job, which is 57.14%. We concluded that, for that particular class, the students who smoke are much less likely to have a job than those who do not smoke. If I was trying to guess whether a student has a job, knowing whether or not they smoke might inform my guess.

So, does smoking *cause* a person to be less likely to have a job? That is not at all clear; in fact, one might guess that a smoker would be more likely to get a job, to help pay for what is an expensive habit. Perhaps, on the other hand, we have the explanatory and response variables backwards. Here is the contingency table with having a job as the explanatory variable:

	Smoke?		
Have a job?	Yes	No	Totals
Yes	3 20%	12 80%	15
No	6 40%	9 60%	15
Totals	9 30%	21 70%	30

Those with no job are twice as likely to be smokers. So, does not having a job cause one to smoke? Perhaps, but one might be more likely to believe the opposite – that not having a job would imply not being able to afford smoking.

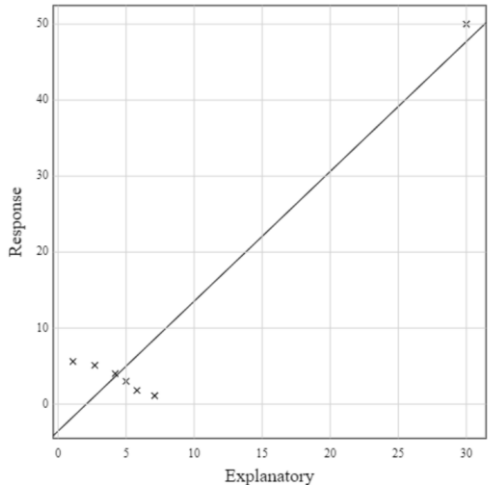
There is another possibility that might be involved. Perhaps there is a **lurking variable,** a third variable that explains what is happening for both the variables being studied. Here is a classical example: for students in the public schools of the USA, there is a strong positive correlation between reading ability (as measured on standardized reading comprehension tests) and shoe size. So does practicing reading make your feet grow? Does growth in the foot send special hormones to the brain that trigger easier reading? Or, perhaps more likely, there is a lurking variable, the age of the student. Younger children have smaller feet and have not yet learned to read as well as older students.

For our example, it might be interesting to consider that personal wealth might be a lurking variable. Students who are less well off financially are less likely to be able to afford to smoke, and are more likely to need a job to afford their education. (Note: We do not know if this is what is happening with these students; we only raise it as a possibility.)

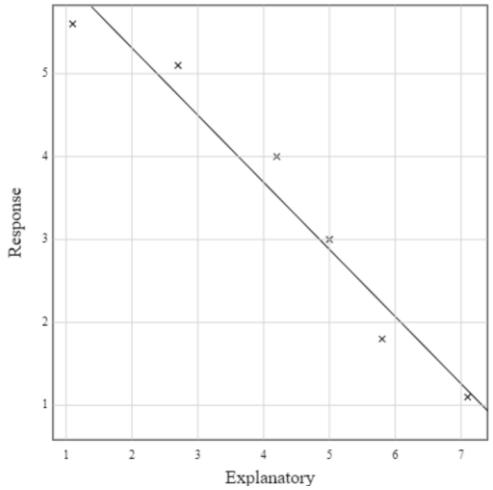
The important thing to remember is that, just because you have established a correlation between two variables, this does not mean that one variable's differences are causing the differences in the other variable.

Do not apply linear regression to non-linear data. We can calculate a regression line for any set of data consisting of two paired numerical variables. This only makes sense, however, if the data displays some linearity. The regression line is useless as a predictor when the correlation is low.

Be cautious if your data includes outliers. Correlation and linear regression are not resistant – the presence of outliers can alter the results substantially. Here is an extreme example. In the graph below, notice that there is a single outlier at approximately $x = 30$, $y = 50$. With this outlier, the regression line has *positive* slope. The correlation is *positive*, with value $r = 0.9577$. Notice that the other data points do exhibit some linearity, but with a decreasing direction.



Now here is the plot with that outlier removed. The regression line matches the data reasonably well, and it has negative slope. With the outlier removed the correlation is *negative*, with value $r = -0.9764$.



One thing to be aware of in this context is that an outlier might indicate an error in data entry; if feasible, that possibility should be examined. For example, in a recent survey of statistics students, the question was how many hours per week were spent listening to music. One student replied 180, which is of course impossible as there are only 168 hours in a week. Perhaps that student was thinking in term of minutes rather than hours? In any case, for any study involving that variable to be of any use that student's responses would have to be removed.

In fact, for the graphs given above, something similar happened. Students were measuring two quantities in centimeters, but one student mistakenly did the measurement in millimeters, creating the misleading outlier.

3.6 – Data File Analysis, Part 2

As we mentioned earlier, the author of the lessons has provided two calculators for your use. In this section we continue our description of the second calculator, which is designed to work with data files similar to those that might be created by spreadsheet software. The data file we use is the same used in Lesson 2. You should have it saved to your own computing device, but in any case here again is a link to that file:

First day survey

As a reminder, the file contained student responses to a first day survey containing these questions:

1. What is your gender? (M) Male (F) Female
2. What is your class year? (FR) Freshman (SO) Sophomore (JR) Junior (SR) Senior
3. How many states have you visited?
4. Do you currently smoke? (Y) Yes (N) No
5. How tall are you (in inches)?
6. How many days per week do you read a newspaper?

In the Data File Analysis, Part 1 (in Lesson 2), we used statistics and graphs to examine several of the variables individually. Now we will use the calculator to examine possible associations between the variables. To begin, open the data file calculator using the following link, then use the *Load vertical file* button to load the data file containing the student responses.

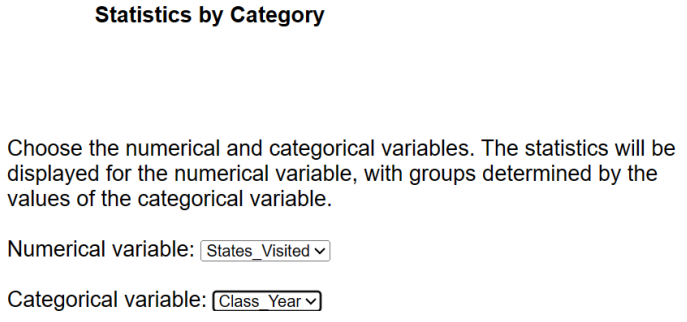
Data file calculator

Association between a categorical and a numerical variable

We will analyze the association between class year and states visited, both numerically and graphically. The class year will be the explanatory variable.

Example. Generate a table that shows the number of students, the five-number summary, the mean, and the standard deviation for states visited, separately for each class year.

Solution. Use the menu option *Descriptive statistics*, submenu option *Statistics by category*. This option will display various statistics for a numerical variable, with the responses grouped by the values of a categorical variable. In this case our numerical variable is States_Visited, with categorical variable Class_Year. Choose those variables as shown here:



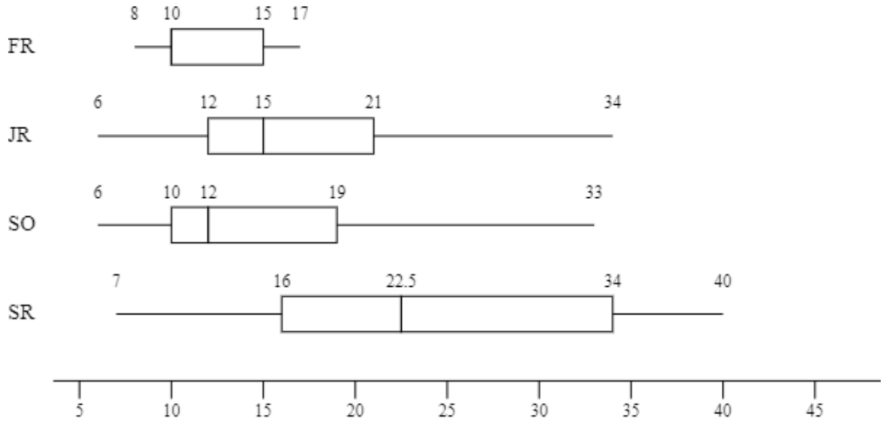
When you choose the *Computations* button, the statistics screen will be displayed. Use the checkboxes to choose the requested values: the count, mean, standard deviation, and five-number summary. These are the results:

Statistics for the *States_Visited* variable, with groups determined by the values of the *Class_Year* variable

	FR	JR	SO	SR
Size	7	23	27	6
Mean	12	16.7826	14.3333	23.6667
Std. dev.	3.3166	7.0386	6.3063	12.4365
Min	8	6	6	7
Q1	10	12	10	16
Median	10	15	12	22.5
Q3	15	21	19	34
Max	17	34	33	40

Example. Create a side by side boxplot for the *States_Visited* variable, with separate plots for each value of the *Class_Year* variable.

Solution. Return to the menu. To obtain the side by side boxplots, use *Graphs*, submenu *Side by side boxplots*. Just as in the first example, choose the desired numerical variable and categorical variable, then click *Computations*. Here is the graph – notice that we have checked the box to include labels on the plots.



States_Visited, for each value of the *Class_Year* variable

Conclusion. The graphical analysis (the side by side box plots) makes it especially clear that for these students there does seem to be some association between class year and states visited – in particular, seniors are more likely to have visited a larger number of states, and freshmen a smaller number. A quick look at the numerical analysis supports this conclusion – notice in particular that the mean for the seniors is almost twice as large as that for the freshmen, and the median is more than twice as large.

Association between Two Categorical Variables

We will analyze the association between a student's class year and whether or not they smoke, both numerically and graphically. Once again the class year will be the explanatory variable.

Example. Create a contingency table for the *Class_Year* and *Smoke* variables, with *Class_Year* as the explanatory variable.

Solution. Use menu option *Descriptive statistics*, submenu *Contingency table*. Choose Class_Year as the explanatory variable, Smoke as the Response variable, and press *Computations*. By default the display includes row percents, which is what we have used in our analysis of association, but you may change that to column percents or overall percents as needed.

Rows: *Class_Year*
Columns: *Smoke*

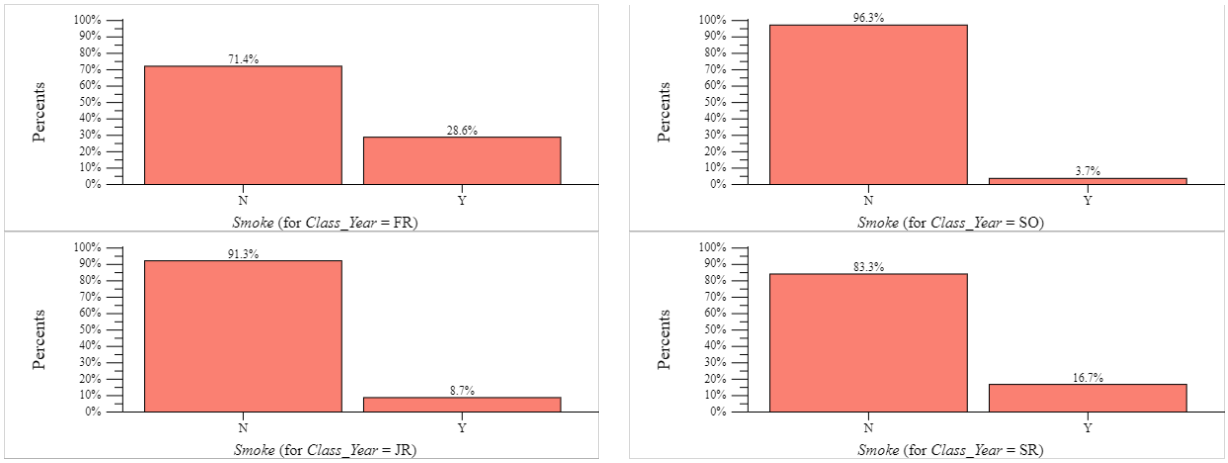
	N	Y	Totals
FR	5 71.43%	2 28.57%	7 100.00%
JR	21 91.30%	2 8.70%	23 100.00%
SO	26 96.30%	1 3.70%	27 100.00%
SR	5 83.33%	1 16.67%	6 100.00%
Totals	57 90.48%	6 9.52%	63 100.00%

In each cell, the count is given, followed by the "row percent." In particular this table tells us the following, and similarly for each of the other cells:

- In the cell FR, N, there are 5 freshman nonsmokers; 71.43% (that is, 5/7) of the freshmen do not smoke.

Example. Create side by side bar charts for the same variables, again with *Class_Year* as the explanatory variable.

Solution. Return to the menu. Use option *Graphs*, submenu option *Side by side bar charts* and choose the same explanatory and response variables to obtain these four plots – to make it easier to fit in this space we have used crop, copy, and paste to move the bottom two graphs beside the top two here.



Conclusion. Both the contingency table and the graph make it clear that the percentage of smokers for freshmen is more than that for the other classes.

Association between Two Numerical Variables

Is there any association between how often a student reads the newspaper and how many states they have visited? We will analyze this numerically and graphically, again with states visited as the explanatory variable.

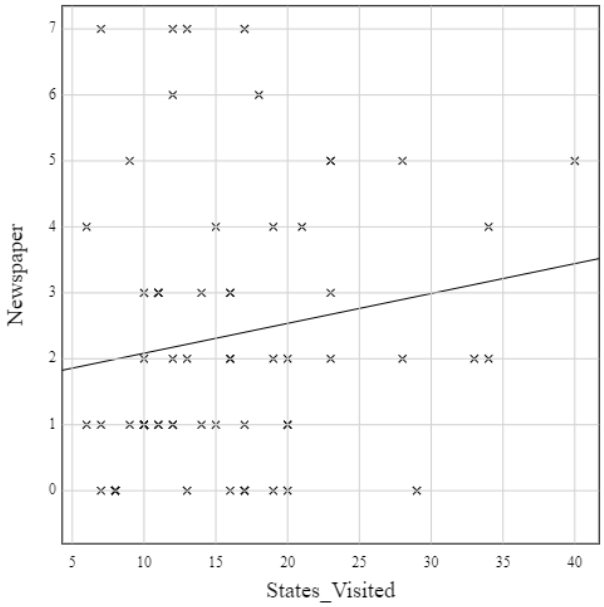
Example. Find the correlation between *States_Visited* and *Newspaper*. Include a regression line, with *States_Visited* as the explanatory variable.

Solution. Use menu option *Descriptive statistics*, submenu choice *Correlation and linear regression*. Choose *States_Visited* as the explanatory variable, *Newspaper* as response, then click *Computations*. Here are the results.

r :	0.167
r^2 :	0.0279
Explanatory (x) variable :	*States_Visited*
Response (y) variable :	*Newspaper*
Regression line :	$\hat{y} = 1.6314 + 0.0453x$

Example. Create a scatterplot for the same situation.

Solution. All we need to do is use the provided radio button to display the scatterplot with the regression line, shown here:



Conclusion. We see only a small positive correlation; although the regression line has been calculated, it is nearly useless as a predictor – it only "explains" less than 3% of the variability in the *Newspaper* responses ($r^2 = 0.0279$). These observations are reinforced visually by the scatterplot – there seems to be virtually no linear pattern in the points plotted.

Exercise 10: Do the following, using the data file referred to in the previous examples:

a. Create a table showing the five-number summary for the *Newspaper* variable, separated by the gender of the students.
b. Create side by side boxplots of heights for the males and females in the class. Comment on the connection to Exercise 15 in Lesson 2.
c. Create a contingency table with *Gender* as explanatory variable and *Smoke* as response. Does there seem to be a connection between the variables for these students?
d. Find the correlation and regression line with *Newspaper* as explanatory variable and *States_Visited* as response – this is the opposite of what we did in the example above. Is the correlation coefficient the same?
e. Refer to part (d). When we had *States_Visited* as explanatory and *Newspaper* as response, the regression line was
$$y = 1.6314 + 0.0453x$$
Can we simply solve for *x* to obtain the regression line with the explanatory and response variables switched?

Exercise 11: In Exercise 16 of Lesson 2 you created a data file containing the data presented originally in Lesson 1, and again at the beginning of this lesson. Use that data file for the following:

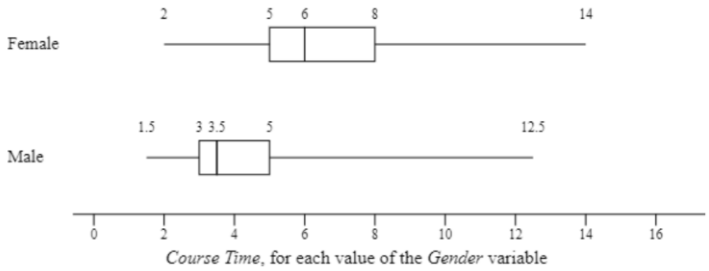
a. Analyze the association between the number of states visited and gender, by creating a table showing mean and median for each gender, and by creating side by side box plots and histograms.
b. Analyze the association between class year and having a job, by creating a contingency table with class year as explanatory variable.
c. Analyze the association between the *Politics* variable and the *Political party* variable by creating side by side bar charts with *Political party* as explanatory variable.
d. Analyze the association between time spent on the course and time spent watching TV, by calculating correlation and regression line, with a scatterplot.

Solutions to Exercises

1: For each question, identify by name the two variables the question relates to. Also, identify the two variables as N/N (both numerical), or C/C (both categorical), or C/N (one categorical and the other numerical).

a. Have the women in that class visited more states than the men? Gender, States – C/N
b. Are the men in that class taller than the women? Gender, Height – C/N
c. Are smokers in the class more likely to have a job than non-smokers? Smoke, Job – C/C
d. Are students in the class who have visited a lot of states more likely to be conservative? States, Politics – C/N (although it could be analyzed as N/N since Politics has values on a scale)
e. Do students in the class who spend more time watching TV spend less time studying? TV, Course Time – N/N
f. Are the juniors and seniors in the class more likely to be Democrats than the freshmen and sophomores? Class, Political Party – C/C

2: Based on this graph for the students in that same class, does there appear to be an association between *Gender* and *Course Time* (the amount of time the students expect to spend per week on the statistics class)?



Yes. The amount of time *depends on* the gender. (It is higher for females than for males)

3: A recent study reported that employed women averaged 16 hours of housework per week, and employed men averaged 11 hour per week. For the people in the study, compare the mean amount of time spent on housework for men and women. Use both subtraction and division. Give a sentence reporting each result "in the context of the problem," similar to how it might be reported in the media. Also, report your conclusions using words such as *association*, *depends on*, *independent*.

Subtraction: 16 – 11 = 5. On average, the women in the study spent 5 hours more per week on housework than the men in the study.
Division: 16/11 = 1.45. On average, the women spent almost one and a half times as many hours per week doing housework as the men (or 45% more hours per week).
Or: 11/16 = 0.69. The men, on average, spent less than three fourths as much time doing housework as the women (or only 69% as much time).
Based on these results, we would say that for the people in the study there is an association between gender and amount of time doing housework. The amount of time doing housework depends on the gender. Gender and amount of time doing housework are *not* independent.

4: Create a contingency table for the Gender and Political Party variables in the data, with Gender as the explanatory variable. (Just ignore the person who didn't answer the Political Party question.)

	Party			
Gender	D	R	I	Totals
Male	8	5	2	15
Female	4	9	1	14
Totals	12	14	3	29

5: Use the contingency table from Exercise 4, reproduced here, to calculate these proportions. (Again we totally ignore the one person who didn't answer both questions on the survey.)

	Party			
Gender	D	R	I	Totals
Male	8	5	2	15
Female	4	9	1	14
Totals	12	14	3	29

a. The proportion of males who are Republican. 5/15 = 33.33%
b. The proportion who are male Democrats. 8/29 = 27.59%
c. The proportion of the females who are Democrats. 4/14 = 28.57%
d. The proportion of the Democrats who are female. 4/12 = 33.33%

6: Use the contingency table from Exercise 4, reproduced here, to calculate these probabilities. (Again we totally ignore the one person who didn't answer both questions on the survey.)

	Party			
Gender	D	R	I	Totals
Male	8	5	2	15
Female	4	9	1	14
Totals	12	14	3	29

a. The probability that a randomly selected student is a male Democrat. 8/29 = 0.2759
b. The probability that a randomly selected male is an Independent. 2/15 = 0.1333
c. The probability that a randomly selected Independent is male. 2/3 = 0.6667
d. The probability that a randomly selected individual is male, given that they are Republican. 5/14 = 0.3571
e. P(Republican | male) 5/15 = 0.3333

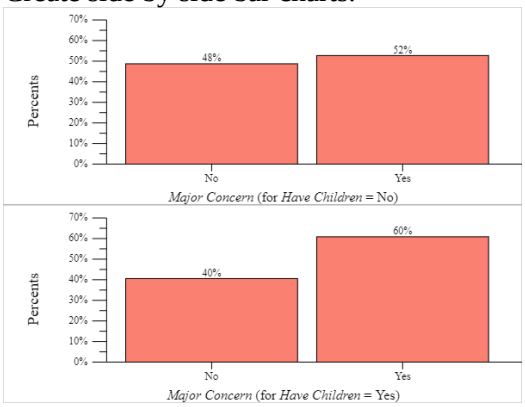
7: Here is a contingency table summarizing answers in a small town to a survey with two questions: "Do you have children in the family?" and "Is paying bills a major concern in the family?"

	Bills a major concern?		
Have children?	Yes	No	Totals
Yes	510 60%	340 40%	850
No	473 51.98%	437 48.02%	910
Totals	983 55.85%	777 44.15%	1760

Do the following using the data:

a. Using "do you have children" as the explanatory variable, add the conditional proportions for each value of the explanatory variable; the sum of the rows should be 100%. See the table above.

b. Create side by side bar charts.



c. Use subtraction and division to compare the proportions answering "yes" to the question about bills, for the two groups.

60% - 51.98% = 8.02% = 0.0802. The proportion that views bill-paying as a major concern is 0.0802 higher for households with children.

60%/51.98% = 1.15. Households with children are slightly more likely (15% more likely) to view bill-paying as a major concern.

d. In that small town, is there an association between having children and viewing bills as a major concern? There may be an association, but it is not as obvious as in some of our other examples.

8: Consider the survey of the previous exercise. In another town, the results were somewhat different, as shown here:

	Bills a major concern?		
Have children?	Yes	No	Totals
Yes	492 60%	328 40%	820
No	459 60%	306 40%	765
Totals	951 60%	634 40%	1585

Do the following using the data:

a. Using "do you have children" as the explanatory variable, add the conditional proportions; the sum of the rows should be 100%. See the table above.
b. In that small town, is there an association between having children and viewing bills as a major concern? For this set of data, there is definitely *no association* – both groups have exactly the same percentages for the answers for the response variable.

9: Suppose a survey question has three possible answers . Call them choice (a), choice (b), and choice (c). Of the 1200 males surveyed, 52% choose (a), 37% (b), and the rest (c). For the 900 females, the percentages are the same. Fill in the counts in this contingency table. Then use the results to calculate the percentages for the entire set of data.

	(a)	(b)	(c)	Totals
Males	624	444	132	1200
Females	468	333	99	900
Totals	1092 52%	777 37%	231 11%	2100

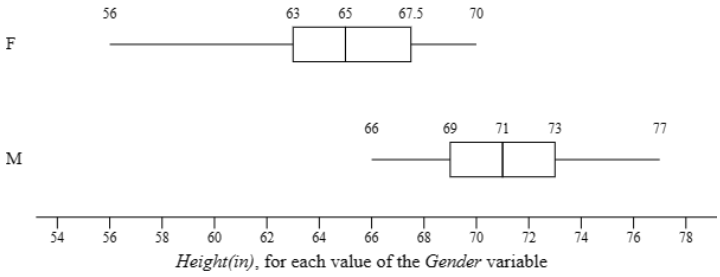
10: Do the following, using the data file referred to in the previous examples:

a. Create a table showing the five-number summary for the *Newspaper* variable, separated by the gender of the students.

Statistics for the *Newspaper* variable, with groups determined by the values of the *Gender* variable

	F	M
Min	0	0
Q1	1	1
Median	2	2
Q3	3	4
Max	6	7

b. Create side by side boxplots of heights for the males and females in the class. Comment on the connection to Exercise 15 in Lesson 2.
In the side by side box plots below, both graphs are on the same scale, making it clear that the men are taller – unlike the situation in Exercise 13 where the plots were done separately and were on different scales.



c. Create a contingency table with *Gender* as explanatory variable and *Smoke* as response. Does there seem to be a connection between the variables for these students?

Rows: *Gender*
Columns: *Smoke*

	N	Y	Totals
F	22 91.67%	2 8.33%	24 100.00%
M	35 89.74%	4 10.26%	39 100.00%
Totals	57 90.48%	6 9.52%	63 100.00%

The percentage of smokers for men is slightly higher, but not much. There isn't a strong connection between the variables.

d. Find the correlation and regression line with *Newspaper* as explanatory variable and *States_Visited* as response – this is the opposite of what we did in the example above. Is the correlation coefficient the same?
The correlation coefficient is the same.

r: 0.167
r^2: 0.0279
Explanatory (x) variable: *Newspaper*
Response (y) variable: *States_Visited*
Regression line: $y = 14.4099 + 0.6161x$

e. Refer to part (d). When we had *States_Visited* as explanatory and *Newspaper* as response, the regression line was

$$y = 1.6314 + 0.0453x$$

Can we simply solve for x to obtain the regression line with the explanatory and response variables switched?
If we solve for x, here are the steps:

$0.0453x = y - 1.6314$ subtract 1.6314 from both sides
$x = 22.0751y - 36.0132$ divide both sides by 0.0453

This is not the same as the regression line in part (d). In general, this is what will happen; the line in part (d) is the best line for predicting *States_Visited*, given *Newspaper*. Solving for x does give a line we can use to predict *Newspaper*, given *States_Visited*, but there is no reason to believe this line will necessarily be the ***best*** line for doing so.

11: In Exercise 16 of Lesson 2 you created a data file containing the data presented originally in Lesson 1, and again at the beginning of this lesson. Use that data file for the following:

a. Analyze the association between the number of states visited and gender, by creating a table showing mean and median for each gender, and by creating side by side box plots and histograms.

Statistics for the *States* variable, with groups determined by the values of the *Gender* variable

	Female	Male
Mean	17.8	15.6
Median	16	13

There is not much association, the figures are approximately the same for each group.

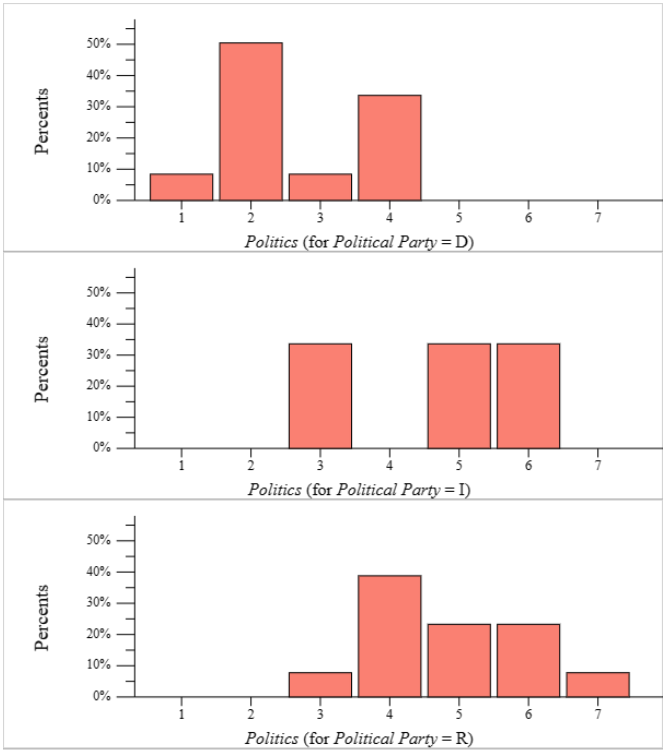
b. Analyze the association between class year and having a job, by creating a contingency table with class year as explanatory variable.

Rows: *Class Yr*
Columns: *Job*

	No	Yes	Totals
Freshman	4 44.44%	5 55.56%	9 100.00%
Junior	2 40.00%	3 60.00%	5 100.00%
Senior	0 0.00%	2 100.00%	2 100.00%
Sophomore	9 64.29%	5 35.71%	14 100.00%
Totals	15 50.00%	15 50.00%	30 100.00%

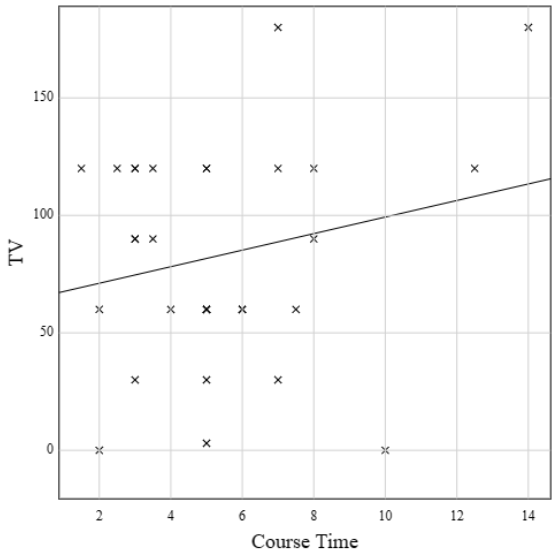
There is some association, seniors are more likely to have a job and sophomores less likely – for freshmen and juniors the figures are similar and somewhere in between.

c. Analyze the association between the *Politics* variable and the *Political party* variable by creating side by side bar charts with *Political party* as explanatory variable.
There is definitely an association. Keeping in mind that lower numbers indicate "liberal" and higher numbers "conservative," Democrats score in the liberal half of the scale and Republicans in the conservative half of the scale – not surprising, of course.



d. Analyze the association between time spent on the course and time spent watching TV, by calculating correlation and regression line, with a scatterplot.

r :	0.2209
r^2 :	0.0488
Explanatory (x) variable :	*Course Time*
Response (y) variable :	*TV*
Regression line :	$y = 64.0774 + 3.5227x$



There is not a strong correlation ($r = 0.2209$). This might be a bit surprising, one might expect that students who spend a lot of time on the course would have little time left for TV. Perhaps more surprising is that the correlation, weak as it is, is positive rather than negative.