IAT 432
STATISTICS OVERVIEW
Reading: Chapter 12 (Analyzing Quantitative data) & Appendix A (Number Scales) from Martin, *Doing Psychology Experiments;*
Software: You can find software to perform a t-test or ANOVA in Excel (under data analysis if that module is loaded), using MatLab (in the student labs) or using SPSS (also in some student labs). Excel is easiest to use.

REVIEW: Assignment 3 Controlled Experiment
Independent variable: mouseboard layout Phone Pad or Alphabetic.
Dependent variables: typing speed (for three tasks/sentences) and error.
Basic question we want to answer is: What effect did the independent variable have on each of the dependent variables?
Null Hypothesis 1 -- There is no difference between people's mouse-typing speed when using an alphabetic or a phone pad layout on a simulated mouseboard.
Null Hypothesis 2 -- There is no difference between people's number of errors when using an either an alphabetic or a phone pad layout on a simulated mouseboard.

This is one of the simplest forms of a controlled experiment. There is one independent variable having two levels and two dependent variables.

Some notes about the Assignment
Note that typing each sentence is a trial. You'll do 3 trials for each of the two layout styles.
We're going to compare average time for trials 2&3 depending on mouseboard.
We're also going to compare average # of errors for trials 2&3 depending on mouseboard.

We don't use the data from trial 1 so that we get better quality data that takes into account the effects of learning. That is, the first trial is the "learning" trial which might have slower times and more error as subjects get used to the layout. We have three trials with different sentences, each with the same number of letters to add validity. Any effects which might be due to the exact sentence are cancelled out by using different sentences.

Remember to convert typing time to typing speed: example
46 (char in sentence) x 60 (seconds/minute)/ 40 typing time (seconds) = 69 characters/minute

Average time for trials 2&3 in chars/minute for the <u>first</u> mouseboard used	Average time for trials 2&3 in chars/minute for the <u>second</u> mouseboard used	Average #errors for trials 2&3 for the <u>first</u> mouseboard used	Average #errors for trials 2&3 for the <u>second</u> mouseboard used

After all teams have submitted their data by email (by Friday Oct 24), the TA will email you (in an excel sheet) the data for your section (by Monday Oct 27). Each studio section should have about 6x3-4 people per teams so you'll have data for about 20-24 subjects.

There are several ways to analyze data that can help us answer the question: What effect did the independent variable have on each of the dependent variables?

DATA TYPES

Numbers -- Quantitative data involves numbers – not all numbers are the same – numbers can be used in different ways and the kinds of operations you can do on them, depends on the kinds of numbers they are. For example – if I say its 20 km to school, then it also makes sense to say, half way to school is 10 km. But if I say the first baseman is number 28 and the second baseman is 14, is does not make sense to say the second baseman is only half the first baseman.

You have to know what kinds of numbers you're dealing with before you can decide how to graph and do statistical calculations on them.

Nominal Scale

Numbers used simply to name or identify something are said to be on the nominal scale
Example: # on a jersey for sports
Nominal numbers have no real quantitative properties … can't add or subtract them
Only stat you can do is count instances of them e.g., there were two shirts with #28 on them.

Ordinal Scale

Numbers that can be ordered or ranked are said to be on an ordinal scale
Example: First place, second place, third place
We know that runner who came first performed better than runner that came second etc but we don't know by how much time she was faster (e.g., .5 second or 10 seconds) … we only know order.
We can calculate statistics that deal with the rank but not the difference between numbers.

Interval Scale

Numbers where the difference between numbers is meaningful are said to be on an interval scale.
Example: temperature measured in Celsius (or F)

> 20 ° is 10 degrees hotter than 10 °
> 10 ° is 10 degrees hotter than 0°
> BUT We CANNOT say that 20 ° is twice as hot as 10°.

Number of errors is interval … 5 errors is 4 more than 1 and 9 is 4 more than 5 but 10 errors isn't necessarily twice as bad as 5 errors.

Ratio Scale

Numbers where the ratio of two numbers is meaningful are said to be on the ratio scale
Example: 20 km is twice as far as 10 km
A ratio scale has an absolute zero point (e.g., weights, distances, times)
An interval scale does not – 0 ° F has no real meaning other than it is 32 degrees below the freezing point of water.
Task time is ratio … 5 seconds is twice as fast as 10 seconds etc.

Ways to Analyze Quantitative Data

To analyze your quantitative data you can use graphing as well as two basic kinds of statistics: descriptive and inferential.

Visual – Graphs
Descriptive statistics (e.g., mean, variance, std deviation)
Interferential statistics (i.e., what can we infer from data about hypothesis? Is there a significant difference of means between groups?)

GRAPHS: PLOTTING FREQUENCY DISTRIBUTIONS

Graphing frequency distributions allows you to visually explore your data and look for meaningful patterns.

Frequency is a count of number of instances of each “score” or number value.

A graph has two axes. The vertical axis is called the y-axis or ordinate. The horizontal axis is called the x axis or the abscissa.

You can use a bar or line graph for a frequency distribution. A bar graph is common.

Plotting Frequency Distributions

When you first look at your raw data, you will have it in tables. Maybe something like this:

TYPING SPEED

Phone Pad Layout		**Alphanumeric Layout**	
Participant Number	Average Typing Speed (char/min)	Participant Number	Average Typing Speed (char/min)
1	60.1	1	60.4
2	67.0	2	80.9
3	76.1	3	79.3
4	50.1	4	57.9
Etc		Etc	

NUMBER OF ERRORS

Phone Pad Layout		**Alphanumeric Layout**	
Participant Number	Number of errors	Participant Number	Number of errors
1	3	1	3
2	4	2	3
3	2	3	2
4	5	4	3
Etc		Etc	

We want to know: What effect did the independent variable have on each of the dependent variables? It’s hard to tell if there’s any difference between the two groups from looking at the tables.

We need to rearrange the data so that we can interpret the raw data more easily. One way to first look at your data is to plot frequency distributions, which is a plot of how frequently each value appears in the data for each level of the independent variable (e.g., mouseboard layout) and for each of the dependent variables (e.g., typing speed and error).

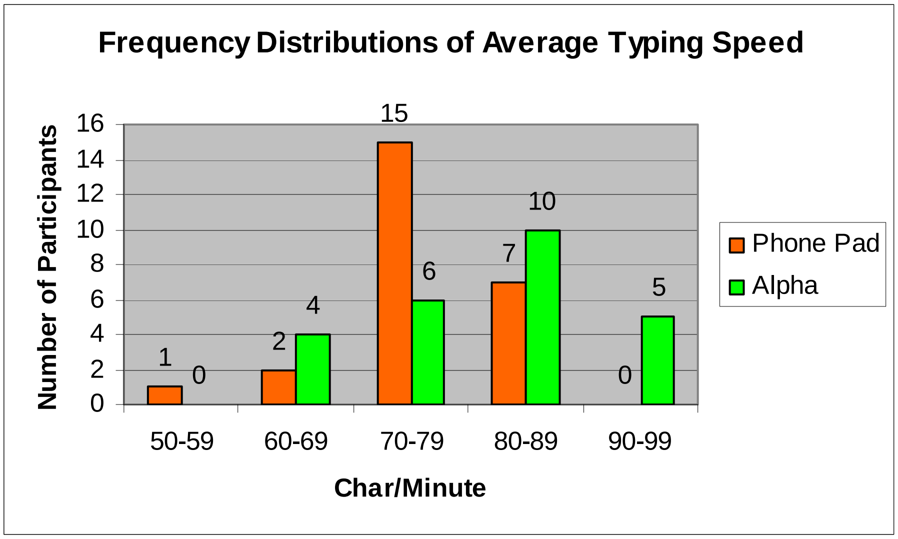
That is, you plot how frequently each value appears. You may have data where no value occurs more than once. So you need to put the data into categories. For example, if typing speeds range from 50-100 char/minute, you could use categories: 50 – 60 char/minute; >60 to 70 char/minute; >70 to 80 seconds char/minute; >80 to 90 char/minute; >90 to 100 char/minute. In some cases there are meaningful categories, in other cases, just splitting the data range into equal categories is sufficient. We call the categories “bins.”

These bin values go along the x or horizontal axis. For error, you should use the actual number of errors (averaged over the two trials). For example, if error count is between 1-10 you can have a category for each [1,2,3,4,5,6,7,8,9,10]. Need to make sure you get the categories right. E.g., if you have 11 errors for one subject then this 1-10 categories won’t work. You’d need 11 too. The dependent variable goes along the x axis.

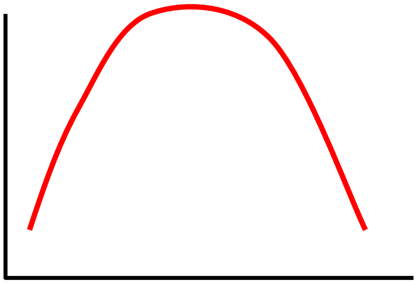
Then you plot the number of values (participants) you have in each category (i.e., frequency which means count) on the y or vertical axis. This won’t be meaningful if you only had a few participants. But if you have around 20 or more, it can be useful. If the independent variable has only a few levels you can plot the data for each level on the same graph. You should use a different colour for each level of the independent variable to that each plot is visually distinct. For Assignment 3, you can do plots for each of the two levels of the independent variable (e.g., layout is phonepad or alphanumeric) and then you can compare the shapes of the distributions by looking at them overlaid. Distribution means how many scores fall in each bin and the shape of the curve that is made if you connect the points (or tops of bars in a bar graph).

For assignment 3 you’ll have two bar graphs, one for each dependent variable (typing speed & number of errors) and for each graph you’ll have two sets of data plotted over top of each other (i.e. one plot for phone pad and one for alphabetic).

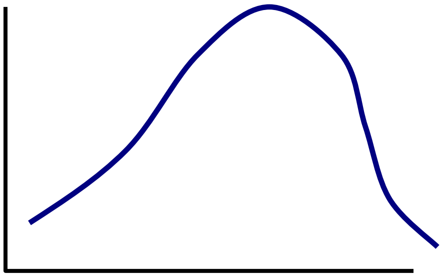
In this figure below, there is a graph of frequency distribution where both levels of the independent t variable (i.e., both mouseboard layouts) are plotted together (i.e., two frequency distributions). The independent variable is mouseboard and has levels of phonepad or alphanumeric. The dependent variable is “typing speed in char/minute” and has been categorized into 50-59, etc. The frequency (y axis) is the number of participants in each category of “typing speed” By looking at this you can see that each layout has a different frequency distribution (or shape).



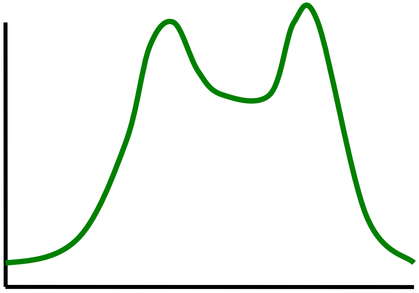
In statistics, there are different names for different shapes of distributions.



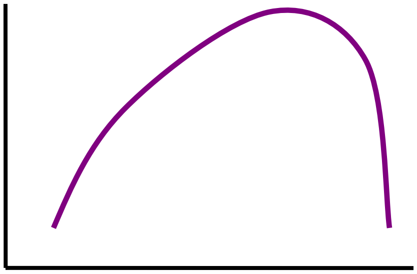
Normal



Skewed



Bi-Modal



Truncated

Normal – bell shape; data must take this shape in order to be able to use many statistical tests of inference. If we take a sample of a population for some test, the test results are assumed to be normally distributed around a central mean. Many people get about the same score, with a few getting really good scores, and a few getting poor scores. You can apply this idea to test times, and error counts as well.
Skewed – asymmetrical with more scores on one side (the sides that trail off are called “tails”) (e.g., IQs scores of university graduates will have more scores higher than average)
Bi-Modal – has two most frequent categories rather than one (e.g., heights of men and women; spatial test scores for men and women)
Truncated – one tail looks completed chopped off (e.g., reaction times -- there is a limit to how fast a human can react. Another example is the results of ratings from a Likert scale

survey because there is a lower and higher limit to scores which makes each "tail" drop off or be truncated).

Now, we have looked at the individual points in each data set. From this we can begin to interpret our data sets for phone pad and alphanumeric. In this case, we see that the data for each group has slightly different shape. The phone pad data is normally distributed. However, the alphanumeric data is a normal distribution, skewed slightly to the right.

While a frequency plot is a good start to summarizing our data, and making it easier to interpret, it's still rather rough. The next step is to summarize the data so that we can describe it more concisely.

DESCRIPTIVE STATISTICS

Descriptive statistics allow you to summarize data in various ways in order to describe characteristics of the data rather than having to report every data point.

Central Tendency (Average)

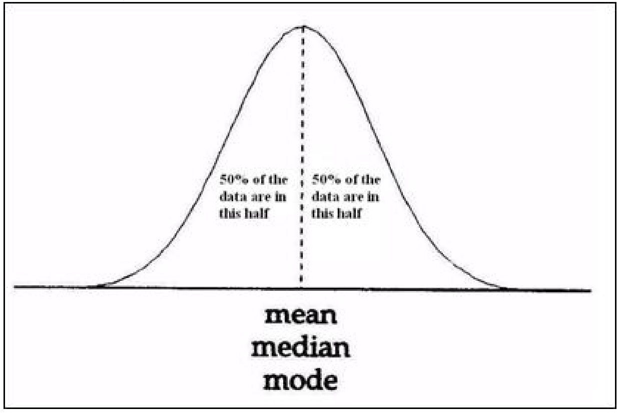
One important way to describe a data set is to describe the middle of the distribution. For assignment 3, the middle of the distribution will describe the way a typical participant behaved (how fast they types and how many errors in each mousepad layout group). There are three measures of central tendency: mode, median and mean. They all are ways to compare groups. What is the average of each group? Is it the same or different?

The *mode* is the value that occurs most frequently. The mode doesn't take into account any other the other data points than the ones that occur most frequently.

The *median* is the middle value; that is it has equal numbers of values lower and equal numbers of values lower than it.

The *mean* is the weighted average of the values; that is the sum of all the value divided by the total number of values you added together.

Which of mean, median or mode should you use? This depends on the distribution of data values. For example if you have a normally distributed group of values (this means that when you plot them and connect the dots, you get a bell shaped curve as in the figure below), then the mode, median and mean are the same.


50% of the data are in this half
50% of the data are in this half
mean
median
mode

When your distribution of values is not normally shaped then the mode, median and mean may be different. In this case, it may be important to report using all three.

For assignment 3, you first calculate the average value or mean typing speed and number of errors for each participant for trails 2 & 3. For example …

Average for EACH participants
Phone pad --
Ave typing speed participant 1 using phone pad = (56 char/minute + 67 char/minute) / 2 = 61.5 char/minute
Ave typing speed participant 2 using phone pad = (67 char/minute + 73 char/minute) / 2 = 70.0 char/minute
Etc …
Ave number of errors participant 1 using phone pad = (3 errors + 2 errors) / 2 = 2.5 errors
Same for Alphanumeric

You also calculate the average typing speed and number of errors for all the participants in each of the two mouseboard layouts. For example …

Average for ALL participants
Phone pad --
Ave speed participant 1 + ave speed participant 2 + …. + ave speed participant 24 = (1680)/24 = 70 char/minute
Ave error participant 1 + ave error participant 2 + …. + ave error participant 24 = 60/24 = 2.5 errors
Same for alphanumeric.

Dispersion (Variation)

An average describes one aspect of the data set. A second statistic that helps describe a data set is a measure of dispersion which means how spread out the values are (or how much variability there is in the data set). Measures of dispersion are standard deviation, range and variance.

Example: average typing speed in char/minute

	Phone	**Alpha**
Average/Mean	75.5	83.5
Mode	80.0	89.0
Median	75.0	84.0

There’s some difference we don’t know much. Looking at dispersion gives us more information.

You can think of *standard deviation* as the extent of the error you are making by using the mean to represent the whole data set. A high standard deviation means that there is a lot of variation in the data set – which the mean doesn’t represent the data set very well. A low standard deviation means that the values are all clustered around the mean and that the mean is an adequate way to describe the data set. For a normal distribution about two thirds of the values will fall within one standard deviation on either side of the mean.

In this example (below), the means are the same but we see more variation in the alphabetic data set. What do you think this means? One thing we can say is that participants performed with the same average typing speed on both mouseboard layouts BUT that there was much more variation for participants using the alphanumeric layout – some participants typed quite slowly and others quite fast – the results are less consistent.

Example: Mean and Standard Deviations

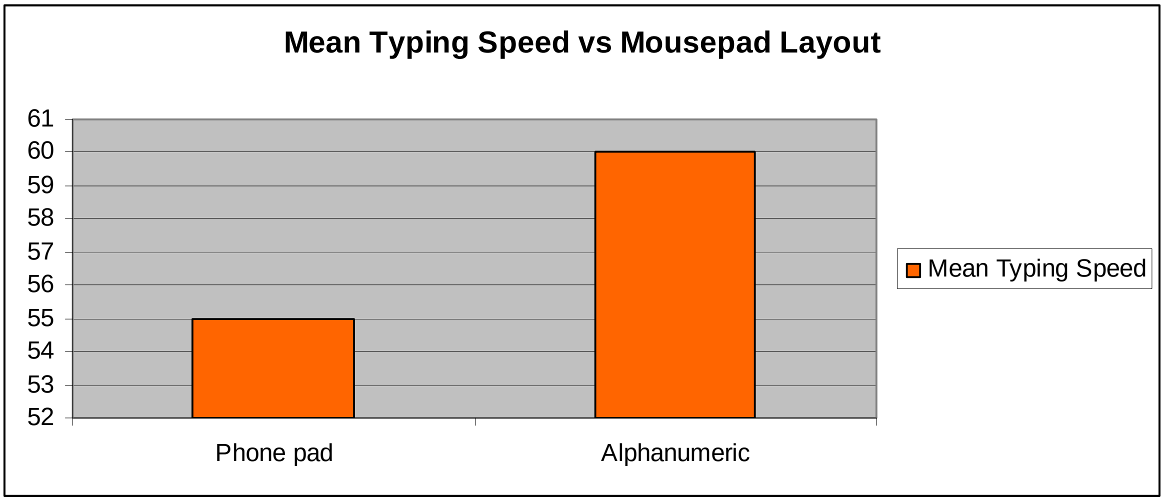
Typing Speed	**Phone** (Char/min)	**Alpha** (Char/min)
Mean	60	60
Std Deviation	5.3	12.9

“The tyranny of mean” is an expression that refers to the tendency people have to only use the mean to represent a data set and instead of describing the variation in the data set and interpreting the results in the context of their original questions.

Remember that your descriptive statistics are only meaningful if you have a controlled experiment and a reasonable number of participants in your study. In studies with three or four subjects, you might do better to just look at the individual data points and see what they tells you about your designs. You need to keep in mind that this information is very limited and may not hold if you test more people.

Graphing mean scores

We can plot the independent variable on the x axis and the mean value for the dependent variable on the y axis. For assignment 3, the independent variable is nominal. Either phone pad or alphanumeric. For an independent variable that is nominal, we can use a bar chart.



When independent variable is different data type, use different types of graphs.
Ordinal or Interval → bar graph
(Ratio → Line graph)

So far, we have been able to summarize some aspects of our data which helps us interpret the data. But we have not yet addressed our original hypotheses. To do this, we need to turn to inferential statistics.

INFERENTIAL STATISTICS

Two find out if the data collected for dependent variables for the two or more levels of the independent variable differ we can use frequency distribution and mean plus standard deviation. By comparing the frequency distribution or means we can see if there is a difference. But what do we mean by difference? How different do the two groups have to be to be “different”? Let’s back track for a minute. The goal of our experiment was investigate a hypothesis. The null hypothesis is the hypothesis of no difference. If a hypothesis was “Users will type faster with alphanumeric” then the null hypothesis is “Users will type the same speed regardless of mouseboard layout.” So what we are after in this case was to examine if there was a difference between how long it took the users to type the sentences with the phone pad and the alphabetic layout.

But it’s not just the sample of users we tested we’re interested in. We want to know something greater than this. We want to know if the results would hold if all users of this design participated in our experiment. The potential group of all users is called a population of users. That is, we want to know if it is likely that there is a difference in task performance between a population of all users using one mouseboard design and a population of all users using another design. Since we can’t test everyone, we use samples. Inferential statistics help us determine how likely it is that the mean we found using a sample would also be found if we used a different sample or even (potentially) tested all users. It also helps us compare the means of two samples to determine if they are statistically different.

The “infer” part means that these tests help you determine or infer if there is a difference between two populations based on the means for the two samples – i.e., the users using one design and the users using another design. We do this by testing the null hypothesis. That is, we test how likely it is that we would get the data we got for the two groups if the two groups were the same. In essence inference tests ask: How likely is it that the populations which these two sample groups represent are the same? If it is unlikely that they are the same, then you can say there is a difference.

Levels of Significance

Now – how significant must a difference be to be important? Most scientists agree that for a difference to be significant, that the likelihood of getting the observed difference in samples due to chance should be less than 1 in 20 (or 5 in 100 which is .05). Some scientists are even pickier and say that the likelihood of getting the observed difference in samples due to chance should be less than 1 in 100 (.01). This is called a “p” value and it means level of significance. A p value of .01 is more significant than a p value of .05.

Choosing the Right Test

Next, you need to determine what statistical test to use.

You need to consider several things to decide which inferential test(s) to use. First, look at how you designed your study:

1. Number of independent variables and number of levels for each

   For Assignment 3, we have one independent variable which can be of two levels.

2. Number of dependent variables and data type of each.

   For Assignment 3, we have two dependent variables, each of which is interval data type.

3. The experimental design: between or within.

4. If your hypothesis predicts the direction of difference (e.g., phone pad will be faster typing) [one tailed] if you don’t know [two tailed]

Second, look at the data → Assumptions

1. How the data is distributed for each level of the independent variable and each dependent variable. If it's normally distributed can use a kind of test called parametric tests. Otherwise, you need to use non-parametric tests.
2. Do data sets have equal variance?

For assignment 3, you need to use a class of inferential statistical tests which will determine if there is a significant difference between the means for typing speed and number of errors for the two mouseboard layout groups: phone pad and alphabetic.

If you have one independent variable this is called a single factor design. The data collected for both of the dependent variables is interval data (e.g., ave typing speed, ave number of errors). The experimental design was within subjects, meaning that each participant typed using both layouts, so we'll used a paired test. We say from frequency plots that the data was roughly normally distributed, so we can use a paired t-test or a one-way ANOVA (ANOVA = analysis of variance). Depending on the number of levels there are for the independent variable, either a t-test or a ANOVA can be conducted. If there are only two levels of the independent variable then a t-test can be used, but if there are more than two levels a one-way ANOVA should be used. So we can do the t-test.

A t-test is used to determine the likelihood that an observed difference between two groups occurred by chance. If the test returns a very low value, we can say it is very unlikely that the differences seen in the data for the two groups occurred by chance. Thus we can say there is a significant difference between the groups and the null hypothesis is false.

If you have more than two levels of the independent variable, you can also use multiple t-tests between each set of two groups – however this gives you less information than an ANOVA. For example, since you recorded subject's task time for a phone pad and alphabetic mouseboard layouts, you'll use an unpaired t-test. If you recorded subject's task time for QWERTY, phone pad and alphabetic mouseboard layouts, you would to use a one-way ANOVA or three unpaired t-tests.

You can find software to perform these tests in Excel (under data analysis if that module is loaded), using MatLab (in the student labs) or using SPSS (also in some student labs). Excel is easiest to use.

EXAMPLE from excel sheet.
The t-test returns for two tailed, paired test the p value of 0.0032 which is less than .01. So this means that there is only 1 in 100 (actually about 3 in 1000) chance that you'd get this data if groups were the same.

Note – input format for the t-test is {first data set, second data set, tails = 2 type = 1} where tails = 2 means that you don't know if the mean from one data set will be higher or lower than the other (on the other hand if you expect an effect e.g., one mean should be higher than other, then you use tails = 1), and type = 1 for paired (i.e., paired means within design where each subject has a pair of data values, one from each layout; type = 2, 3 are for between designs. See excel help for more information).

INTERPRETATION

Lastly, three things about interpreting your results

First, when a statistical test fails to show a significant difference in the levels of an independent variable (e.g., between two different designs), it does not mean, necessarily that the two are the same. So if you do not find a significant difference, you can’t state for sure that they are the same. You just know that it is unlikely that they are different. This is why you can never prove the null hypothesis but you can disprove it by showing there is a significant difference. You can’t prove the original hypothesis either. You can just support it by disproving the null hypothesis.

Second, the p levels of .01 and .05 are not cast in stone. You need to think about what it means to for something to be significant at a p level. For a particular design situation, it might be fine that your results are significant at a .06 level.

Third, do not confuse statistical significance with practical significance. A difference is a difference only if it makes a difference. You should always refer back to the situation that prompted you to run an experiment and logically determine what your results mean in the context of this situation.

References

You can find more information on all this by reading the sections in either of these two books:

Doing Psychology Experiments, Martin → read Chapter 12 & parts of Appendix A [on reserve]

How to Think About Statistics, Phillips → read Chapter 9. [available in library]