




Research Methods in Human Relations

William Ray, Fall 2017



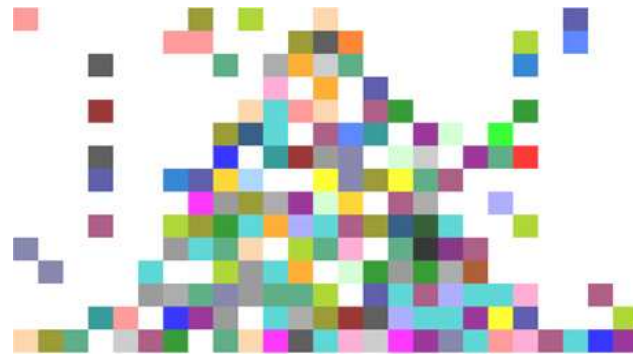
Copyright ©2010, 2011, 2012, 2013, 2017 by William O. Ray

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the author.

Table of Contents

Introduction	1	Confidence Intervals for Means	225
Graphs	14	Confidence Intervals for Proportions	234
Means	36	Hypothesis Tests for Means	245
Variance	46	Hypothesis Tests for Proportions	270
Proportions	60	Hypothesis Tests for One Sample	281
Normal Tables	65	Analysis of Variance	292
Outside-In Calculations	73	Correlation	317
Inside-Out Calculations	84	Linear Regression	334
Distributions	99	Multiple Regression	350
Foundations of Research	107	Inter-rater Reliability	362
Research Design	125	Two-way ANOVA	371
Experiments	163	Reliability and Validity of Scales	388
Sampling	179	One Way Tables	406
Survey Construction and Formatting	198	Two Way Contingency Tables	415

Research Methods in Human Relations



William Ray, Fall 2013

1. Introduction

1.1. Course Schedule

- *First Friday evening.* Introduction, graphs, group exercises.
- *First Saturday morning.* Descriptive Statistics. Spreadsheets, Normal Calculations.
- *First Saturday afternoon.* Foundations of Research, Research Design, teams.
- *First Sunday afternoon.* Survey Construction, Confidence Intervals for Means, teams.

- *Second Friday evening.* Distributions, Confidence Intervals for proportions, teams.
 - *Second Saturday morning.* Hypothesis tests for means and proportions.
 - *Second Saturday afternoon.* ANOVA, teams. **Prospectus due.**
 - *Second Sunday afternoon.* Linear Regression. Multiple Regression, teams, time permitting
-
- *Third Friday evening.* Sample Examination, Library faculty presentation, teams.
 - *Third Saturday morning.* Inter-rater Reliability, Multiple Regression, Chi-squared tests.
 - *Third Saturday afternoon.* Two-way ANOVA, Chi-Squared tests, Teams.
 - *Third Sunday afternoon.* Team presentations.


1.2. Grading

Graded Assignments

Gradebook Declarations	5%
Team Prospectus	15%
Team Research Presentation	25%
Team Contribution	5%
Final Examination	30%
Final Written Assignment	20%

Grading Scale

Last A	90%
Last B	80%
Last C	70%
Last D	55%



It is essential that you familiarize yourself with the course content on **CANVAS**.

Be sure to read the syllabus and related materials on the website **CANVAS** for details on

- the assignments;
- the gradebook declarations and the postings to the discussion boards;
- on academic honesty.

There is also a set of videos [Against All Odds](#) available in the OU-Tulsa library that provide further learning opportunities for many course topics.

1.3. Team Research Project

A major component of this course will be the team research project. Each team will be responsible for

- identifying a research problem;
- defining research objectives;
- defining the population to be studied;
- defining the variables;
- developing a model (or set of hypotheses);
- designing a research instrument to measure the variables;
- gathering data;
- analyzing the data;
- presenting the data to the class.

More details are available in the syllabus and online in the content area **Basic Research Concepts**.

1.4. Examples of Research Projects

- Can the wording of a survey influence results?

What kind of data do you think you would have to gather to answer this question? Does the subject matter of the survey matter?

- Do an opinion survey, with an analysis of the population characteristics that might influence the outcome.

For example, gender, age, income and education often influence political preferences. Would you include hair color? Political Affiliation?

- Who is more likely to ask for "paper" bags at the grocery store, men or women?

Are there any other factors that might influence this request? How would you gather data on this question?

- Who is more likely to come to a complete stop at a stop sign, men or women?


Are there other factors, such as location, time of day or weather, that could influence this behavior? What would be the advantages and drawbacks of doing a survey on this topic?

- What are the optimal hours for the library to be open?

What kind of data would you gather in order to answer this question? Do you think a student's major, age, work or family situation might influence their response?

- Does knowledge about a disease influence how people react when meeting someone with that condition?

Many studies show a fear of contagion even when the person knows that the disease cannot be transmitted by casual contact, or is not communicable at all.



Your topic does not have to be complex or deep. It should be one that is fun and interesting – or at least that you can convince your group is interesting! It should be **sharply focused** so that you can actually complete the entire project during the time allotted.

In all cases your project and presentation must

- avoid stereotypes;
- use inclusive language;
- show a sensitivity to diversity.

You are, after all, going to be professionals in human relations and your research presentation should reflect these basic principles.

This is a **team project**. That means that all team members share equally in the credit for the project. Team members should all put forward equal effort on the project, although this does not mean every person does exactly the same thing. For example, some of the team members may be able to observe more subjects than others, although each team member should observe a minimum of ten subjects. Some team members may be proficient at computers or video and make special contributions in those areas, while others are especially adept at public speaking or role-playing.

Working on a team also means compromise, finding common ground, accepting errors (especially first draft errors!), listening as well as speaking, shared responsibility and shared rewards.


After the final weekend team members will be asked to evaluate the contribution of other members of their team. These peer evaluations will be used in part by the instructor to assign a **team contribution** grade to each individual.

1.5. Peer Evaluation

Each student will be asked to rate the **other** members of their team using, for example, the following four criteria.

- **Preparation.** Were they prepared when they came to class?
- **Contribution.** Did they contribute productively to group discussion and work?
- **Respect for others' ideas.** Did they encourage others to contribute their ideas?
- **Flexibility.** Were they flexible when disagreements occurred?

The overall ratings you assign to your team members must average ten points – with some team members getting higher ratings and some lower. You will be expected to explain your ratings.



Teams may develop their own criteria for peer evaluation, but teams should agree among themselves **this weekend** if they plan to use different criteria than the ones listed above. Changing the rules mid-course would be seen by most people as unfair.

Each person will be assigned an average peer evaluation score, omitting the **lowest** and **highest** peer score. The instructor may further adjust the peer evaluation averages based on his observations and based on the reasons provided on the peer evaluation. Altogether the peer evaluations and the instructor's adjustments will comprise the **Team Contribution** portion of the grade, which counts 5% of the total grade.

It is important that you do your peer evaluations honestly so that the contributions of your fellow team members can be recognized.

There is a form on **LEARN.OU.EDU** to complete to assign the peer evaluation score to the other members of your team.

1.6. Absences

It is important that you come to class. Whether you are in class or not, you will be responsible for all material covered in class.

Throughout the scheduled class sessions time is set aside for the teams to work together on the research project. The instructor will rotate among the teams during these times to ask and answer questions. It is important that team members take advantage of this scheduled time.

In the real world, team members sometimes must be absent. The team is still expected to perform and everyone, including the absent member, benefits. If the team feels the absence is justified and if the absent member tries to make amends, most teams will gladly extend the benefit. However, if the team members feel that the absence is not justified or that the absent person is freeloading, there are likely to be consequences on the peer reviews. If you must be absent, consult with your team members early and work to make up your contribution.

1.7. Gradebook Declarations

As you have already seen, several assignments requiring posting something to a message board. Other assignments will involve doing one of the active learning modules, or reacting to another student's postings. All of these online assignments require that you take a "quiz" in which you affirm that you have done the assignment. The instructor will check at least 50% of all gradebook declarations for accuracy.

If you misrepresent what you have done in a gradebook declaration, you can be charged with academic misconduct. As a minimum sanction, you will receive a grade of "zero" for the **entire** 15% of your grade related to the declarations.

Honesty is both a fundamental scholarly value and a fundamental value of human relations professionals. Scholars rely on honest reports from other scholars to advance knowledge. Human relations professionals value fair play; personal advantage gained through dishonesty disrespects your fellow students.

2. Graphs



One of the most powerful ways to summarize data is a **graph**. A graph is a visual tool that lets the reader quickly discern trends and make comparisons.

The graph above is from [MS MONEY](#) and summarizes a one-year history of the Dow Jones Industrial Average. What do you think of this graph?

The [GRAPHS.XLSX](#) spreadsheet accompanies this section.

2.1. Histograms

Suppose that you administer an IQ test to 128 high school students and obtain the scores at right.

Since the scores are sorted from lowest to highest, we can see that they range from 57 to 141. What other patterns can you see from this list of scores?

There are so many scores there is not very much **information**.

57	84	90	96	100	104	106	114
65	84	90	96	100	104	107	117
67	84	90	97	101	104	107	117
71	84	92	97	101	104	107	118
76	84	92	97	101	104	109	119
76	85	93	98	101	104	109	120
76	85	93	98	102	105	110	121
77	85	93	98	102	105	111	124
78	86	93	99	102	105	111	125
78	86	94	99	102	106	111	127
79	86	94	99	102	106	112	127
80	87	95	99	103	106	112	129
82	88	95	99	103	106	112	131
83	88	95	99	103	106	112	133
83	88	95	100	103	106	113	133
83	89	95	100	103	106	113	141

Our goal is to **summarize** the data in order to deduce patterns. Right now we can't see the forest for the leaves.

We will construct ranges or **cells** and count how many actual observations fall in each cell.

The purpose of this is to help us visualize what information is in the data.

The range of our observations is

$$141 - 57 = 84$$

If we wanted, say, **exactly** ten cells, then we would need cell to be 8.4 units long. But our goal is not to be exact but instead to visualize the data. So, instead of using 8.4 for each cell, round 8.4 to a more convenient number, say 10. Will we get exactly ten cells using cell lengths of ten? Does it matter?

Next decide the bottom – lower limit – for the first cell. For example, we could make the lower limit for the first cell 40.

A better choice might be 56 (so that the upper limit of the cell is 65, ending in 5).

This results in cells that look like

lower	upper	count
56	65	
66	75	
76	85	
86	95	
96	105	
106	115	
116	125	
126	135	
136	145	

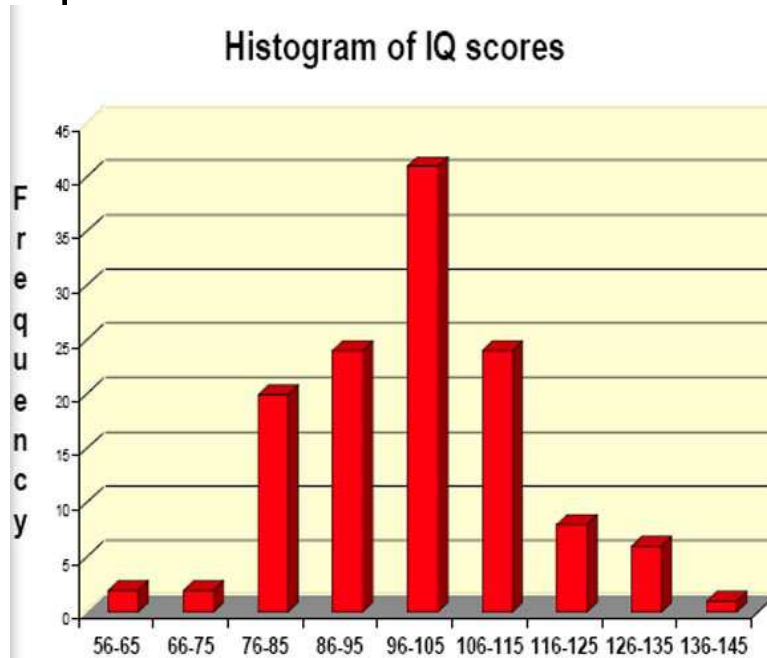
There are of course other possible choices—this is just one of many possibilities.

Filling out the table then gives:

lower	upper	count
56	65	2
66	75	2
76	85	20
86	95	24
96	105	41
106	115	24
116	125	8
126	135	6
136	145	1

57	84	90	96	100	104	106	114
65	84	90	96	100	104	107	117
67	84	90	97	101	104	107	117
71	84	92	97	101	104	107	118
76	84	92	97	101	104	109	119
76	85	93	98	101	104	109	120
76	85	93	98	102	105	110	121
77	85	93	98	102	105	111	124
78	86	93	99	102	105	111	125
78	86	94	99	102	106	111	127
79	86	94	99	102	106	112	127
80	87	95	99	103	106	112	129
82	88	95	99	103	106	112	131
83	88	95	99	103	106	112	133
83	88	95	100	103	106	113	133
83	89	95	100	103	106	113	141

The final step is to construct a [bar chart](#) or [histogram](#) that graphically represents the data



The height of each bar corresponds to the count in each category.

Some other observations that are now apparent:

- The observations appear to have a “bell-shaped” distribution.
- Most of the observations are in the middle range. In fact, 89.5% of the observations fall between 86 and 115:

$$89.5\% = \frac{24 + 41 + 24}{128} \times 100\%$$

- The observations appear to be symmetrically distributed, centered roughly at 100.

Fortunately, Excel makes it easy to do histograms, provided that the **Data Analysis Tool** tool has been **installed**. While this is a standard component of Excel, it does not install by default. Once installed, it's easy to use.

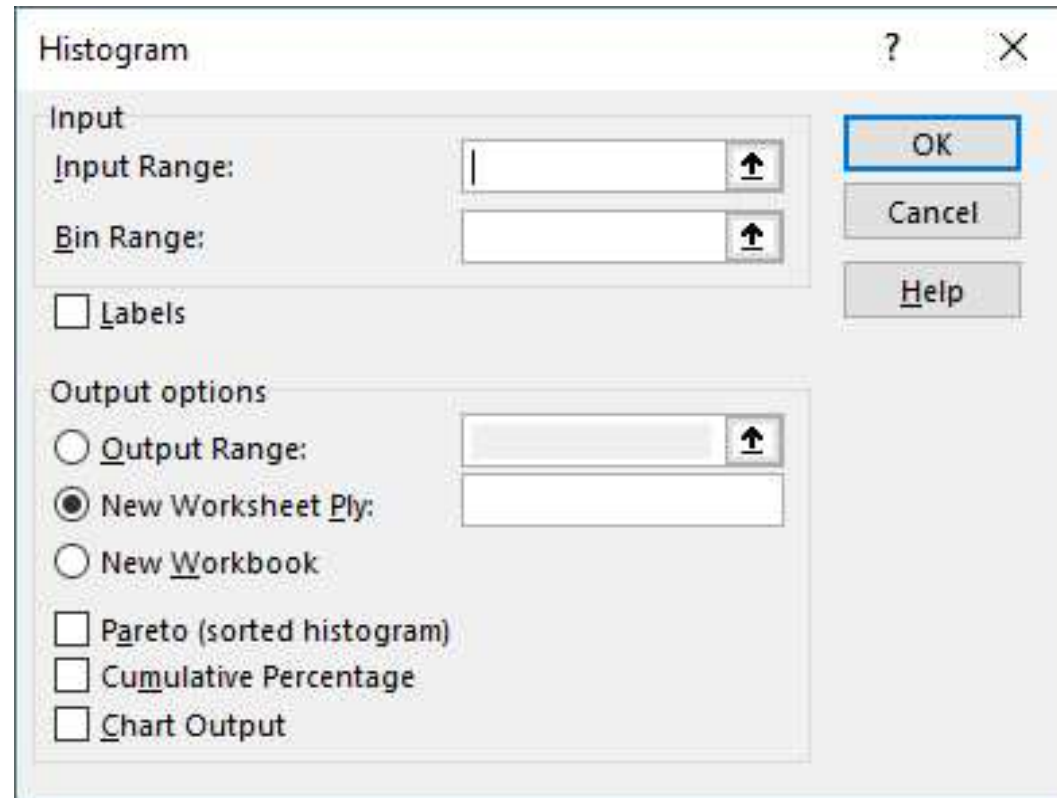
Solution Template

Step 1. Enter your data in a spreadsheet. The data do not have to be ordered.

Step 2. Figure out your **bins**—these are the intervals into which you will sort your data.

Step 3. Each bin has a lower and upper limit. Enter the **upper limits** into the spreadsheet.

Step 4. Use Data → Data Analysis → Histogram to create your histogram.



In the tool, fill in the cell ranges for your data and your bins. You can optionally choose a location for the output. Check Chart Output to have it produce a chart.



Step 5. Construct the histogram.

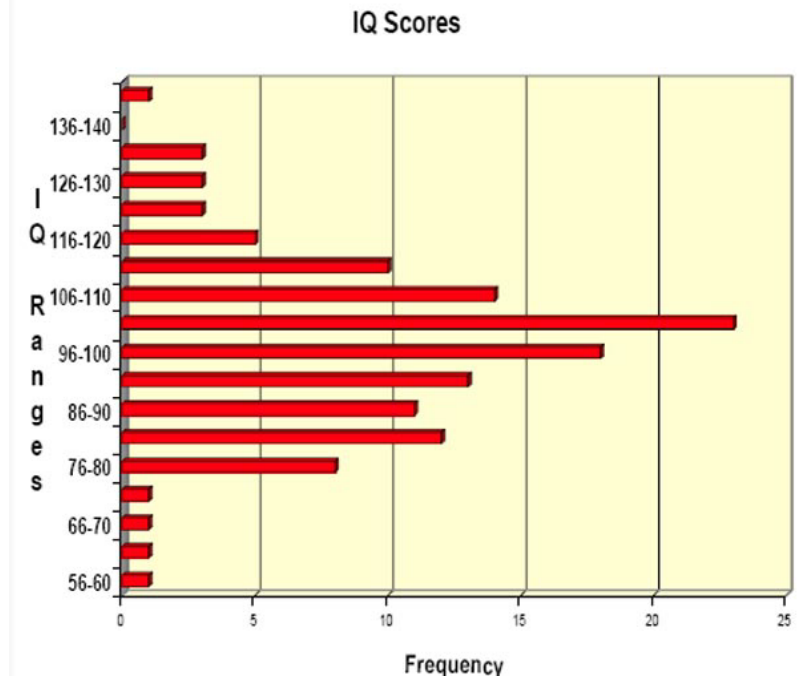
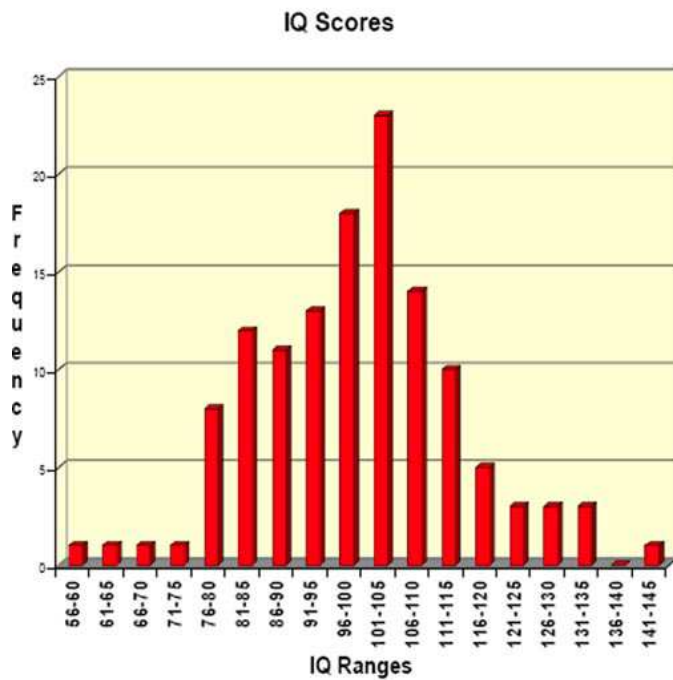
_____ **End of Solution Template** _____



You do it.

Starting with the same data as in the previous example, construct a histogram with approximately fifteen cells.

57	84	90	96	100	104	106	114
65	84	90	96	100	104	107	117
67	84	90	97	101	104	107	117
71	84	92	97	101	104	107	118
76	84	92	97	101	104	109	119
76	85	93	98	101	104	109	120
76	85	93	98	102	105	110	121
77	85	93	98	102	105	111	124
78	86	93	99	102	105	111	125
78	86	94	99	102	106	111	127
79	86	94	99	102	106	112	127
80	87	95	99	103	106	112	129
82	88	95	99	103	106	112	131
83	88	95	99	103	106	112	133
83	88	95	100	103	106	113	133
83	89	95	100	103	106	113	141



Which histogram is easier to read?

2.2. Nominative Scales

There are four kinds of scales, or ways of measuring subjects, used in behavioral sciences:

- Nominative scale
- Ordinal scale
- Interval scale
- Ratio scale

Nominative or **naming** scales assign labels to subjects, usually based on **attributes**. These can divide subjects into **categories**, such as gender, place of birth, or some other non-quantitative characteristic. Sometimes nominative scales involve numbers, but they are used as labels. For example, offensive and defensive linemen on a football team wear uniforms numbered in the 60s or 70s, while quarterbacks wear uniforms numbered between 1 and 19. The relative value of that number has no meaning other than as a label.

An **ordinal scale** is similar to a nominative scale, in that it divides the

subjects into categories. However, ordinal scales include the notion of **ordering**. Examples might be grades, socioeconomic status, or rank.

An **interval** scale is similar to nominative and ordinal scales, but the magnitudes between adjacent intervals are the same. Temperature measured in degrees Fahrenheit or Celsius is an example. Clearly ninety degrees is hotter than fifty degrees (ordering), and the difference between forty degrees and fifty degrees is the same as the difference between twenty degrees and thirty degrees. Interval scores don't have a true zero, although there may be an artificial zero such as the freezing point of water. **Time** and **date** are interval scales. Interval scales may or may not be infinitely divisible, so GRE scores are interval scores.

Finally, **ratio** scales add a true zero to the mix. Physical measurements like height, weight, blood pressure, and distance are ratio scales. **Elapsed time** which measures the time from a particular event has a true zero. The zero in a ratio scale represents the absence of what is being measured. Thus, temperature in degrees Kelvin which uses absolute zero—the total absence of heat—is a ratio scale. GRE scores are



not ratio scales since all GRE scores range between 200 and 800—there is no zero.

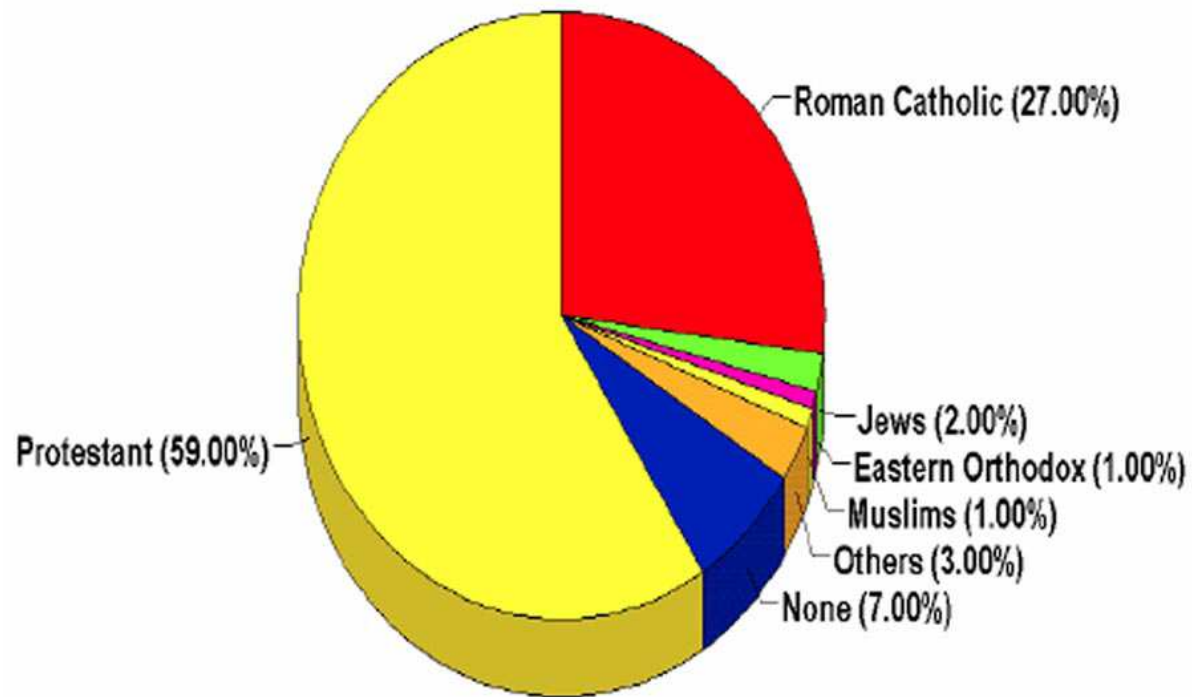
Because interval and ratio scales involve **magnitudes**, they are **quantitative** scales, while nominative and ordinal scales are **qualitative** or **attribute** scales. Quantitative and attribute scales are summarized in different ways.

Pie charts are the appropriate visual presentation for nominative scales and for some ordinal scales.

For example, a recent poll asked the respondents to self-identify their religious affiliation. The results were as follows:

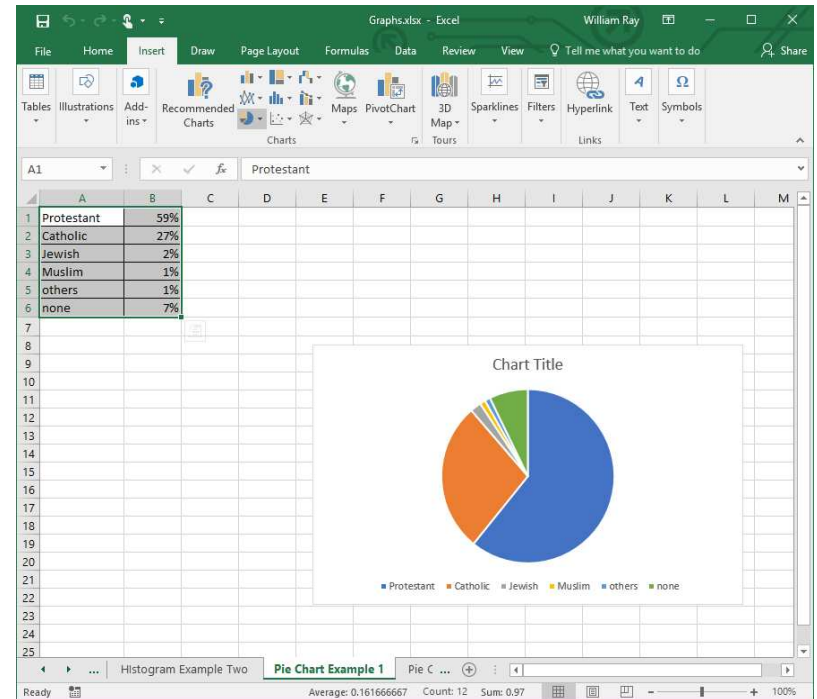
Protestant	59%
Catholic	27%
Jewish	2%
Muslim	1%
others	1%
none	7%

The resulting pie chart is:



Who invented pie charts?

Excel makes it easy to produce pie charts, too. First, create a table that includes your categories. Then highlight the table and INSERT → CHARTS and chose a pie chart.



For another example, research suggests that there are three factors that contribute to long-term well-being or happiness:

Environmental factors	10%
Genetic factors	50%
Intentional activities	40%

We can easily use Excel to make a pie chart of the above table.

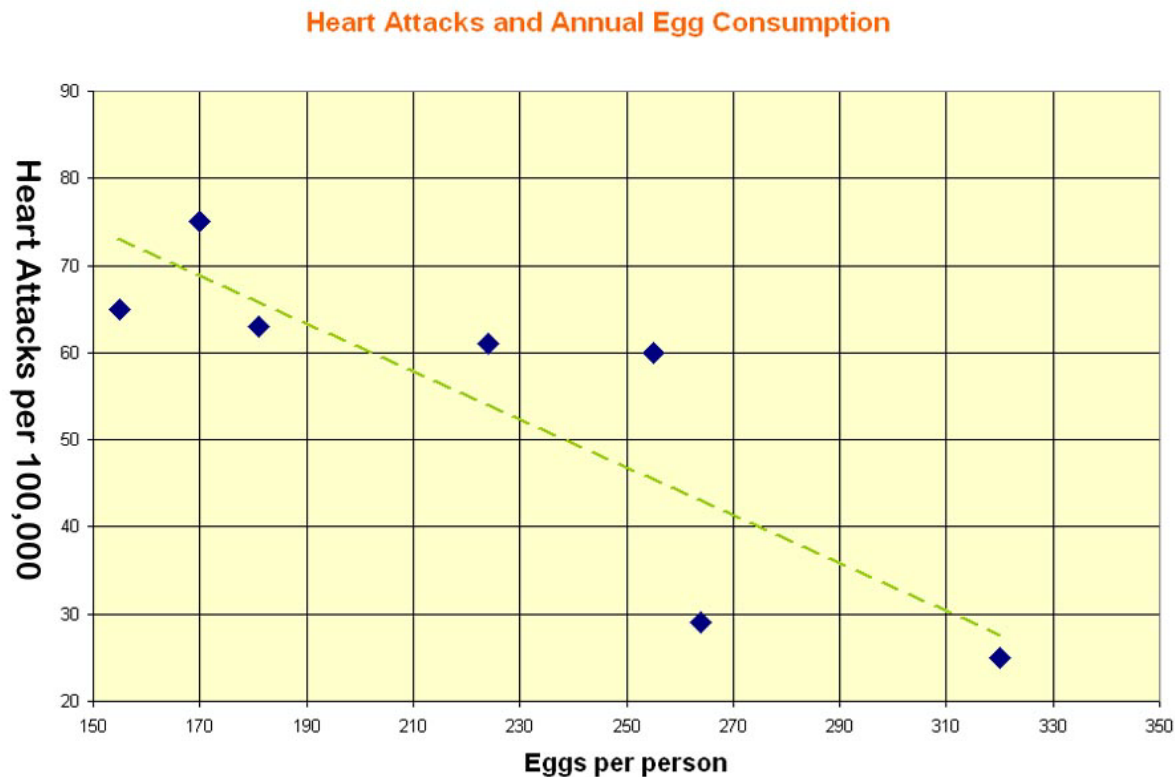
2.3. Scatter Plots

Sometimes you will gather two or more quantitative measures on each subject. In these cases you are often interested in determining if there is a relationship between the two variables (height and weight for example). The visual presentation that can help you understand this is the scatter plot.

Suppose for example you gather data on annual egg consumption and annual heart attack rates for several countries:

Country	Annual Eggs	Mortality
Australia	155	65
UK	170	75
Canada	181	63
Germany	224	61
US	255	60
France	264	29
Japan	320	25

A scatter plot for this data is



Note that we added a **trend line** to this scatter plot. What conclusions might you draw from this plot?

Excel makes it easy to do scatter plots, too. For another example, consider the following data on deaths in the workplace per 100,000 workers.

Year	Rate per 100,000
1955	8.6
1960	7.7
1965	7.3
1970	6.8
1975	6
1980	5.8
1985	4.8
1990	4
1995	1.9
2000	1.8

3. Means

3.1. Example.

Suppose that we gather the following data on the savings (in \$1,000's of dollars) for the nine OU employees who retired last July.

2	2	2
10	17	24
26	34	45

Question. Is this census data or sample data?

Question. What are the "average" savings?

Note that there are several possibilities for the average:

- The most frequent observations is two (the *mode*).
- The 50th percentile is seventeen (the *median*).
- The *midrange* is 23.5 (average of highest and lowest).
- The **arithmetic average** or **mean** is eighteen.

$$\text{mean} = \frac{2 + 2 + 2 + 10 + 17 + 24 + 26 + 34 + 45}{9}$$

The **mean** is the kind of average that you will most frequently encounter; in this course we will use “mean” and “average” interchangeably.

If we have **census data** then the symbol for the mean is μ and we say we have found the **population mean**. If we have **sample data** then the symbol for the mean is \bar{x} and we say we have found the **sample mean**. Whether we have census or sample data the mean is computed the same way:

$$\text{mean} = \frac{\text{sum of observations}}{\text{number of observations}}$$

- Given a choice, which you rather know: the **sample mean** or the **population mean**?
- Are you studying the **population** or the **sample**?

You are studying the sample in order to draw **conclusions** about the **population**.

Whenever you gather data you will be dealing with a population.

- For *census data* you have complete information on the entire population;
- For *sample data* you have information on only part of the population.

Thus

- The *population mean* μ involves **no uncertainty**.
- The *sample mean* \bar{x} involves **uncertainty**.

Other things being equal census data is preferable to sample data. In the real world, however, often only sample data is available. Since sample data must always involve error – being incomplete – it is important to develop strategies to minimize error. Statistical tools permit the researcher to determine how large the error might be.

3.2. Example.

Suppose that nine OU employees retire in October with the following accumulated savings (again listed in \$1,000's of dollars):

11	15	16
17	18	20
20	20	25

Find the mean.

Solution. To find the mean, first find the sum of all the numbers. In this case, that sum is 162. Now divide by the number of observations (in this case 9):

$$\text{mean} = \frac{162}{9} = 18.$$


This is easy to do with a hand calculator or with spreadsheets.

MEANS.XLSX pre-loads the Excel formulae for calculating means and related statistical measures.

We have computed the means of two data sets so far:

2	11
2	15
2	16
10	17
17	18
24	20
26	20
34	20
45	25

Each has nine observations and each has a mean of 18. Yet one data set seems to be somewhat more *variable* than the other. Measuring



how variable a set of observations are from their mean is our next topic. Before doing that we will briefly introduce some mathematical notation.

If you take observations on n subjects ($n = 9$ in our previous two examples) then you have a list of n numbers:

$$x_1, x_2, \dots, x_n$$

In the example we just completed, the x_i 's have the particular values

x_1	2
x_2	2
x_3	2
x_4	10
x_5	17
x_6	24
x_7	26
x_8	34
x_9	45

Using this notation, then we can write the mean as

$$\text{mean} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

A shorthand notation for

$$x_1 + x_2 + \cdots + x_n$$

is

$$x_1 + x_2 + \cdots + x_n = \sum_{i=1}^n x_i$$

and so we can also write

$$\text{mean} = \frac{1}{n} \sum_{i=1}^n x_i$$

This shorthand notation frequently appears in statistics books. Sometimes calculators will use the capital sigma Σ to indicate a function summing a series of numbers. There is Σ button in Excel for summing a list of numbers.

To summarize, if you have sample data

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

and if you have census data

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i$$

4. Variance

In addition to measuring the mean of a collection of numbers, it is usually also necessary to measure how much *variability* there is in the numbers. Since the mean is the “average” of the numbers, we might first calculate how much each observation differs from the mean:

x_i	$x_i - \mu$	
2	2-18	-16
2	2-18	-16
2	2-18	-16
10	10-18	-8
17	17-18	-1
24	21-18	6
26	26-18	8
34	34-18	16
45	45-18	27

If you add up how much each observation varies from the mean, *you will get zero*. In fact, you will get zero *every time no matter what the original numbers*. Since

$$\text{mean} = \frac{\text{sum of observations}}{n}$$

it follows that

$$n \times \text{mean} = \text{sum of observations.}$$

Adding the **positive** terms and the **negative** terms together in the table then exactly cancel out!

x_i	$x_i - \mu$	
2	2-18	-16
2	2-18	-16
2	2-18	-16
10	10-18	-8
17	17-18	-1
24	21-18	6
26	26-18	8
34	34-18	16
45	45-18	27

Because this naive approach fails, it is standard to instead *square* the differences between the observations and the mean.

x_i	$x_i - \mu$		$(x_i - \mu)^2$
2	2-18	-16	256
2	2-18	-16	256
2	2-18	-16	256
10	10-18	-8	64
17	17-18	-1	1
24	21-18	6	36
26	26-18	8	64
34	34-18	16	256
45	45-18	27	729

This approach gives the following average:

$$\text{sum of the differences squared} = 1918$$

$$\text{average of the differences squared} = 213.1$$

This average is called the *mean square error* or the *variance*.

Remember that these data are on savings of retirees. Thus we have computed a “variability” of

$$213.1(\text{thousands of dollars})^2$$

When we squared the differences to do our computations, we also squared the labels (dollars). Since “dollars²” has no intuitive meaning, we should probably now take the square root (so that the units return to dollars, the same as those used for the mean). The result is a number called the *standard deviation*:

$$\begin{aligned}\text{standard deviation} &= \sqrt{\text{variance}} \\ &= \sqrt{213.1} \\ &= 14.6\end{aligned}$$

The *standard deviation* for the July retirement data is thus \$14,600.

The steps we just went through found first the *average of the squared differences* or the **variance**. Then we took the square root of the variance to find the **standard deviation**.

For the mean, the calculations for the *sample mean* and the *population mean* are identical. *This is not true for the variance and the standard deviation: the calculation is different depending on whether we started with census data or sample data.*

The calculations we just did were for *census data*. To summarize these:

$$\text{population variance} = \sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

and

$$\text{population standard deviation} = \sigma = \sqrt{\sigma^2}$$

For **sample populations**, the calculation is almost the same. First, for sample data the symbol for the variance is

$$\text{variance} = s^2$$

and the symbol for the standard deviation is

$$\text{standard deviation} = s.$$

The formulae are:

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

and

$$\text{sample standard deviation} = s = \sqrt{s^2}$$

In practice, calculators and spreadsheets have the formulae for the mean, the population standard deviation and the sample standard deviation built in. Thus it is usually not necessary to use these formulae directly.

4.1. Example.

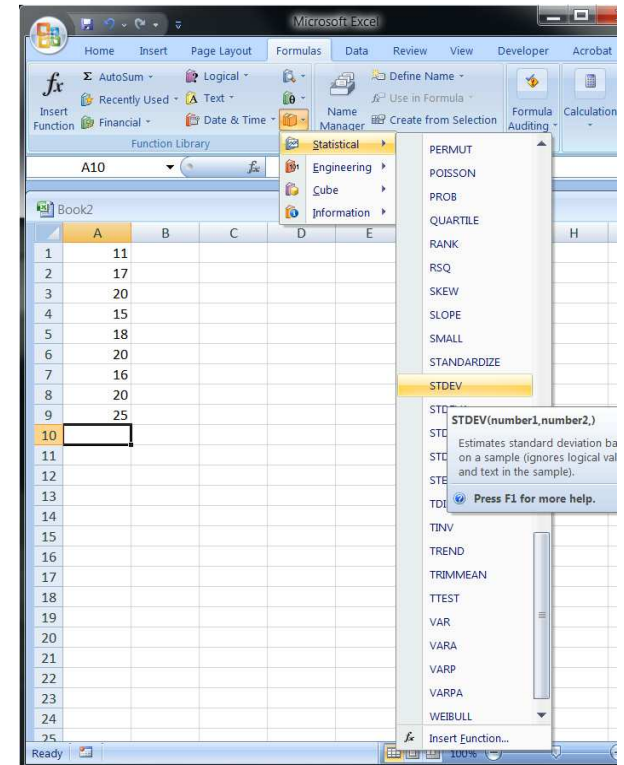
Using the built-in functions of Excel, find the mean, standard deviation and variance of the following sample.

11	15	16
17	18	20
20	20	25

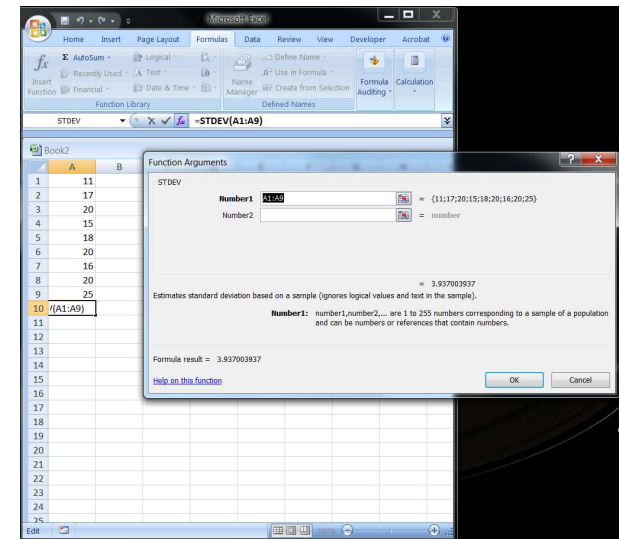
Solution. **Step 1.** Enter the nine numbers into nine different cells on an Excel spreadsheet.

Step 2. Now position the cursor in a cell that does not include data—for example, the cell just below where you’ve entered your numbers.

Step 3. Select the *Formulas* tab at the top of the window. Then, click on the lower right icon in the function library. Select *Statistical Functions*, and a drop-down menu will appear. Select *StDev*.



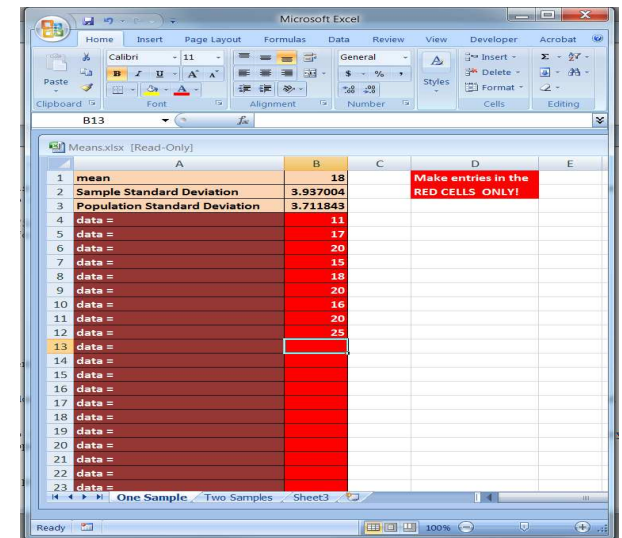
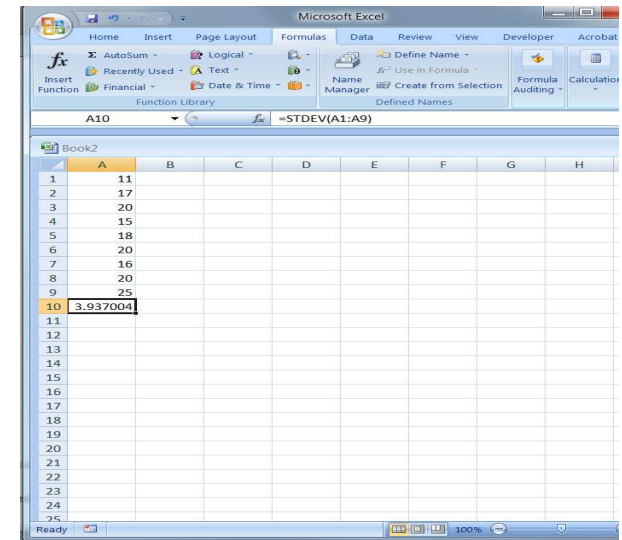
Step 4. When you select *StdDev*, a pop-up window appears asking you for the location of the cells that contain your data. Excel will pre-fill this with the non-zero cells adjacent to the location of the cursor. You can this by over-typing, or by selecting the appropriate cells with your mouse within the spreadsheet.



Step 5. Once you have the proper cells selected in the pop-up window, press enter or click on OK. The spreadsheet will then display the sample standard deviation. **Note:** *If you wanted the population standard deviation, you would select the function StDevP.*

Excel reports that the sample standard deviation is 3.937.

Alternatively, you could use the spreadsheet MEANS.XLSX found in the online course resources. This spreadsheet is pre-configured to calculate the mean, sample standard deviation and population standard deviation of up to 60 data points.



4.2. Example.

Use Excel to find the mean and standard deviation of the following census data.

8	6	10	12
11	12	15	3

Interpretation of Standard Deviation.

The standard deviation is just a measure of how much the data deviate from the mean. In general the standard deviation has no intrinsic meaning beyond the concept of “mean square error.” However, under “normal” conditions – we will define normal conditions shortly – there are some numerical inferences possible from the standard deviation.

For example

Approximately 68% of all observations will “normally” fall between

mean - one standard deviation

and

mean + one standard deviation

Similarly,

About 95% of all observations will “normally” fall between

mean - 2 × standard deviation

and

mean + 2 × standard deviation

4.3. Example.

GRE scores are “normally distributed” with a mean of 500 and a standard deviation of 100. Thus approximately 68% of all GRE scores fall between

$$500 - 100 \quad \text{and} \quad 500 + 100$$

or approximately 68% of all GRE scores fall between

$$400 \quad \text{and} \quad 600$$

Similarly, approximately 95% of all GRE scores fall between

$$500 - 200 \quad \text{and} \quad 500 + 200$$

or approximately 95% of all GRE scores fall between

$$300 \quad \text{and} \quad 700$$

4.4. Example.

The scores from the Stanford-Benet IQ test are “normally distributed” with a mean of 100 and a standard deviation of 15. Approximately 68% of all IQ scores fall between

$$100 - 15 \quad \text{and} \quad 100 + 15$$

or approximately 68% of all IQ scores fall between

$$85 \quad \text{and} \quad 115$$

Similarly, approximately 95% of all IQ scores fall between

$$100 - 30 \quad \text{and} \quad 100 + 30$$

or approximately 95% of all IQ scores fall between

$$70 \quad \text{and} \quad 130$$



5. Proportions

When your observations consist of *numeric* data the *numerical* summaries used are the mean and the standard deviation. When your observations consist of *attribute* data, the numerical summary is the *proportion*.

Sometimes it will help to think of membership in a category as a “yes” or “no” answer to the question: “Is this observation in this category?” The most obvious example of this is opinion polling: you ask each member of the sample, for instance, whether or not they are a registered Republican. Instead of a number for each observation you have a “yes” or a “no” response.

What you do is then place your data in categories – some of the subjects respond with a “yes” and some respond with a “no.” You can then compute a proportion. For census data this looks like

$$\begin{aligned}\text{population proportion} &= p \\ &= \frac{\text{number answering yes}}{\text{population size}}\end{aligned}$$

while, for sample data, it is

$$\begin{aligned}\text{sample proportion} &= \hat{p} \\ &= \frac{\text{number answering yes}}{\text{sample size}}\end{aligned}$$

The symbol p is used for the population proportion and the symbol \hat{p} (p -hat) is used for the sample proportion.

Note that the proportion (sample or census) must always be between 0 and 1; the corresponding percentages are

$$p \times 100\% = \text{population percent}$$

and for sample data

$$\hat{p} \times 100\% = \text{sample percent}$$

We will generally use the proportions rather than percentages because the formulae used later in the course rely on proportions, not percentages.

Usually the “yes” response is thought of as “success” and the “no” response is thought of as “failure.” Usually you set up your problem so that whatever you are studying is the “yes” or “success” response. When this terminology is used, “Success” and “Failure” have no intrinsic meaning beyond being a “yes” or “no” response.

5.1. Example.

Mechanics R Us hires two classes of employees: mechanics and managers. Last year, the franchise in Oklahoma hired a total of 425 male mechanics out of a total of 465 male applicants for mechanic positions. What proportion of male applicants for mechanic positions was hired?

Solution. The total number of observations is

$$n = 465.$$

We are studying “being hired” so the question is

- “Was the subject hired as a mechanic?”

The “yes” respondents to this question are the “successes.” There are

$$k = 425$$

successes.

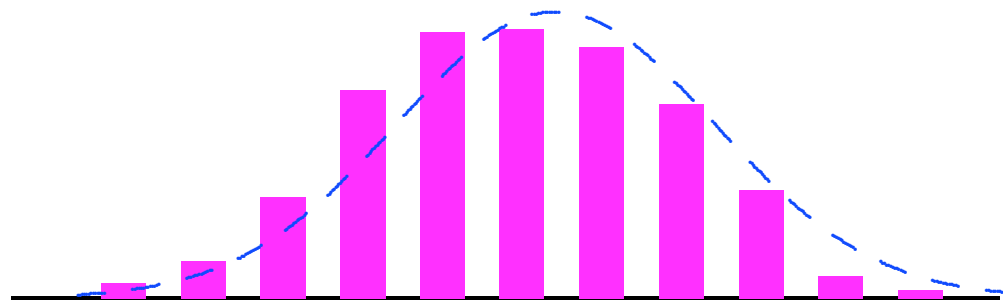
Thus the hiring rate for male mechanics is

$$p = \frac{425}{465} = 0.9139$$

The proportion is 0.9139; the percentage is 91.39%.

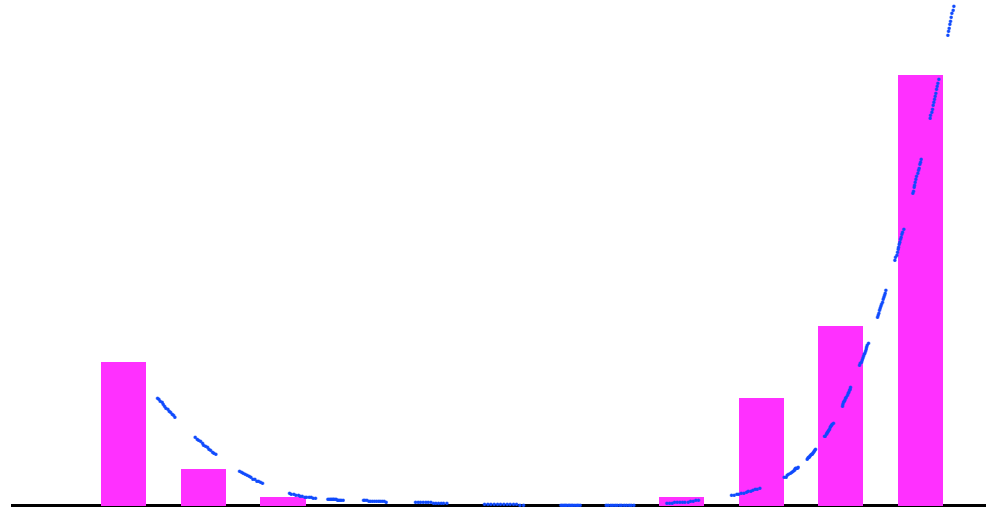
6. Normal Tables

Sometimes the graph of data will appear to fall into a regular pattern. There are certain kinds of patterns that occur over and over again in data. One of the most important of these recurring patterns is called the "normal distribution" or "bell-shaped curve."



Examples of data that might follow a normal distribution are height, weight, income, and IQ scores. Many standardized measurements turn out to be normally distributed.

However, just because this distribution *often* occurs does not mean it *always* occurs. For example, if there were a "handedness" scale it might look like



reflecting the fact that almost no one is truly ambidextrous: about 7% of the population is left-handed and about 93% is right-handed. Interestingly, "pawed-ness" in mice is normally distributed. (How do you think you might test for this in mice?)

Properties of normal curves:

- Normal curves are symmetric about the mean μ .
- About 95% of the observations fall between
$$\mu - 2\sigma \quad \text{and} \quad \mu + 2\sigma$$
- About 68% of the observations fall between
$$\mu - \sigma \quad \text{and} \quad \mu + \sigma$$
- The curves are completely characterized by the mean μ and the standard deviation σ .

The statement “the curves are completely characterized by the mean and standard deviation” means that the *formula* for the normal curve involves only these two parameters:

$$\text{normal curve} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{x-\mu}{\sigma}\right)^2}.$$

Normal curves were first understood as an "error function" – as the natural variation that occurs in all physical measurements. The German mathematician Karl Gauss was the first to see the normal curve as something that could be graphed. Because the German word for "numbers" is "zahlen" the values associated with the normal curve are sometimes call "*z-values*."

Prior to the introduction of the euro, the most common German banknote was the 10 deutsch mark note which included both Gauss' face and the above formula.



The *area under the curve* corresponds to population percentiles. For example, 95% of the *area* under the curve is between

$$\mu - 2\sigma \quad \text{and} \quad \mu + 2\sigma$$

and hence 95% of the observations in a normal population also fall in this range. There are numerical techniques for finding the area under curves – this is a major topic in calculus. Using these techniques it is possible to compute the percentile corresponding to an observation from a normally distributed population. Instead of doing this computation for each possible observation, statisticians instead build a table of possible observations and corresponding percentiles.

There are two categories of problems you will learn how to solve using normal tables: “outside-in” and “inside-out” problems. The reason for the words “outside-in” and “inside-out” has to do with how one uses the tables and will be clearer after we do some problems.

Finding percentiles (outside-in problems). In these you will be given:

- the mean
- the standard deviation
- a “raw” score (a measurement on a member of the population)

And you will be looking for

- the percentile which corresponds to the given measurement.

Finding measurements (inside-out problems.)

In these problems you will be given:

- the mean
- the standard deviation
- a percentile

And you will be looking for

- the “raw score” or measurement which corresponds to the given percentile.

These are the two classes of problems which you will learn how to solve in this segment.

As we have previously seen, percentiles by themselves are not particularly useful objects. However, the *techniques* will turn out to be fundamental in working other kinds of problems.

7. Outside-In Calculations

7.1. Example.

Suppose that a population is normally distributed with mean 100 and standard deviation 15 (IQ scores are so distributed). Find the percentile which corresponds to an IQ score of 114.

Solution.

Since the normal curves are completely characterized by their mean and standard deviation, all we have to do is look in a table which gives the percentiles for scores taken from a population that is normally distributed having mean $\mu = 100$ and standard deviation $\sigma = 15$.

However, no such table exists. There is only *one* normal table: the one

for populations that have mean $\mu = 0$ and standard deviation $\sigma = 1$ – the “standard normal” or z table.

To solve this problem, we need to convert the “raw” score of 28 to a “standard normal” or “ z ” score; then we can use the normal table which appears in your study guide.

Step 1. First *make a dictionary*:

mean	μ	100
StDev	σ	15
Score	x	114

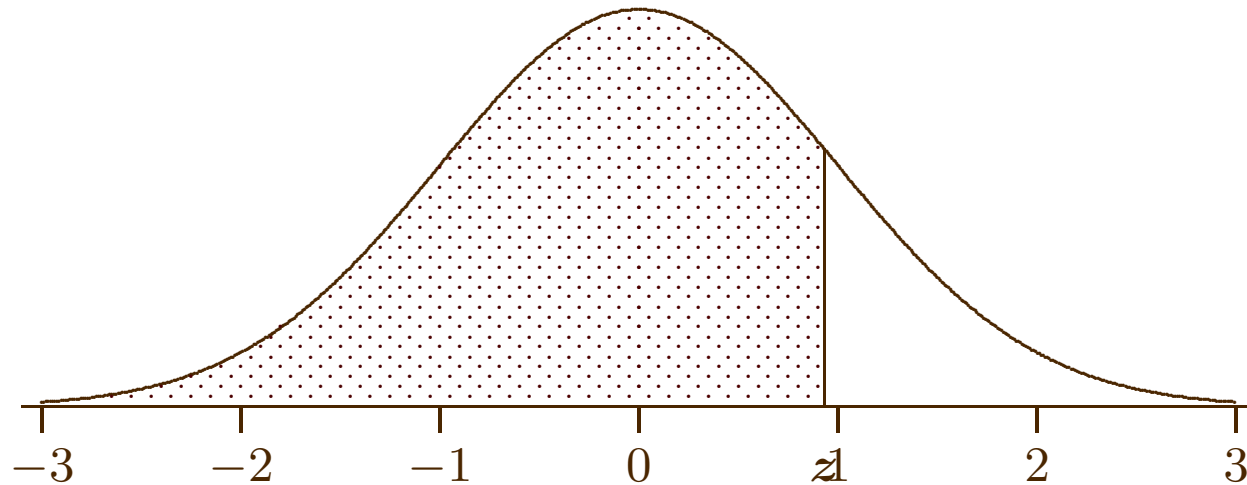
Step 2. Next convert the given score (or observation) to a *standard normal* or z score using the formula:

$$z = \frac{x - \mu}{\sigma}$$


In our example,

$$z = \frac{114 - 100}{15} = \frac{14}{15} = 0.93$$

Step 3. Now look at the normal table in your study guide. There are actually two normal tables, one for negative values of z and one for positive values of z . Since our $z = 0.93$, we will look in the part of the table corresponding to positive values of z . The table gives the proportion of observations which fall to the left of z :



In this case, we need to look up the proportion corresponding to an observation of $z = 0.93$. To do this, look in the left column of the table until you find the first two digits; you will actually see $0.9z$. Now move to the column corresponding to the last digit; the number which you find in



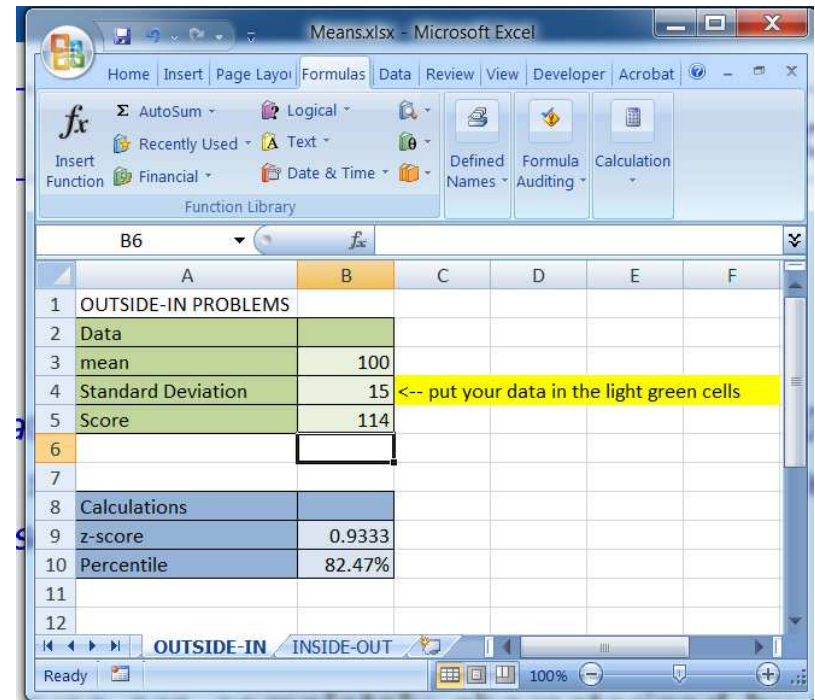
this column is the *proportion* of observations which are less than z .
The proportion which you should find in the table is 0.8238.

Step 4. (Optional) You can convert the proportion to a percentile by multiplying by 100%. The corresponding percentile is 82.38%, i.e., 82.38% of all IQ scores are 114 or lower.

There is a spreadsheet solution to this as well.

To use the spreadsheet, open MEANS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled OUTSIDE-IN. Enter the data.

From this, the corresponding percentile is **82.47%**.



Solution Template

Step 1. Make a list of what you are given. In some problems you will be given *census* data (as in the ACT problem above); in other problems you will only be given *sample* data, in which case you will use the

sample data to *estimate* the population parameters μ and σ .

mean	μ or \bar{x}
StDev	σ or s
Observation or score	x

Step 2. Use the formula

$$z = \frac{x - \mu}{\sigma}$$

to convert the observation x to a z score. If you are only given sample data, you will need to approximate the formula with:

$$z = \frac{x - \mu}{\sigma}$$
$$\approx \frac{x - \bar{x}}{s}$$

Step 3. Find the proportion corresponding the z score in step 2 using the normal table. You do this by locating the z score starting in the left hand column (“outside-in”).

Step 4. (Optional) Convert the proportion you find in step 3 to a percentile by multiplying by 100%.

————— **End of Solution Template** —————

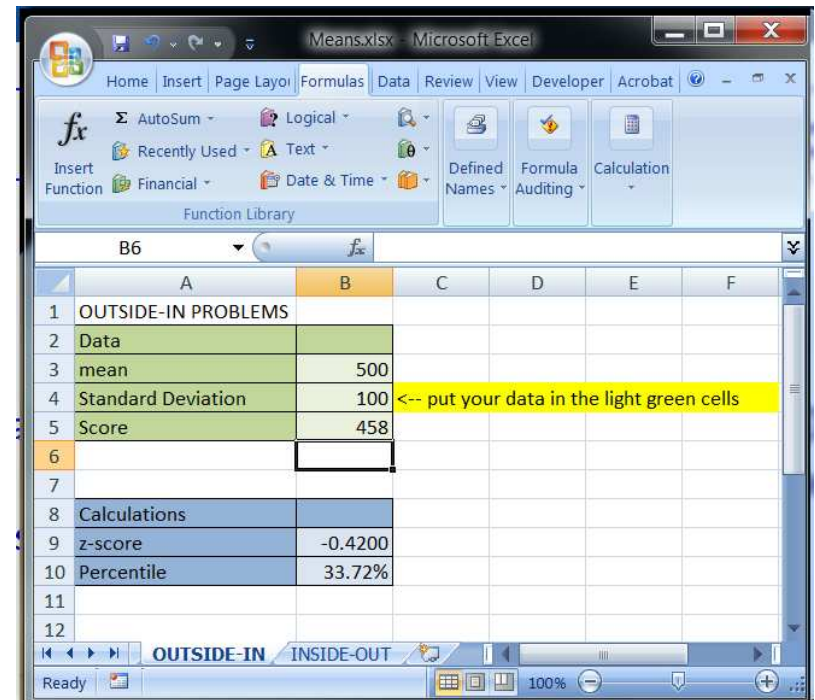
7.2. Example.

GRE scores are normally distributed with a mean of 500 and a standard deviation of 100. A student’s GRE score is 458; what is the corresponding percentile?

Solution. Of course, the easy way to do this is with the spreadsheet.

To use the spreadsheet, open MEANS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled OUTSIDE-IN. Enter the data.

From this, the corresponding percentile is **33.72%**.



For completeness, we'll include the methodology with the tables.

Step 1. In this problem

mean	$\mu = 500$
StDev	$\sigma = 100$
Observation	$x = 458$

Step 2. Find the z score;

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ &= \frac{458 - 500}{100} \\ &= -\frac{42}{100} \\ &= -0.42 \end{aligned}$$

Step 3. Looking “outside-in” (but this time in the *negative* part of the z table) you can find that the corresponding proportion is 0.3372.

Step 4. In other words, 33.72% of all scores will be less than the observed score of 458. ■

Question. How many GRE scores would you expect to be *larger than* 458?

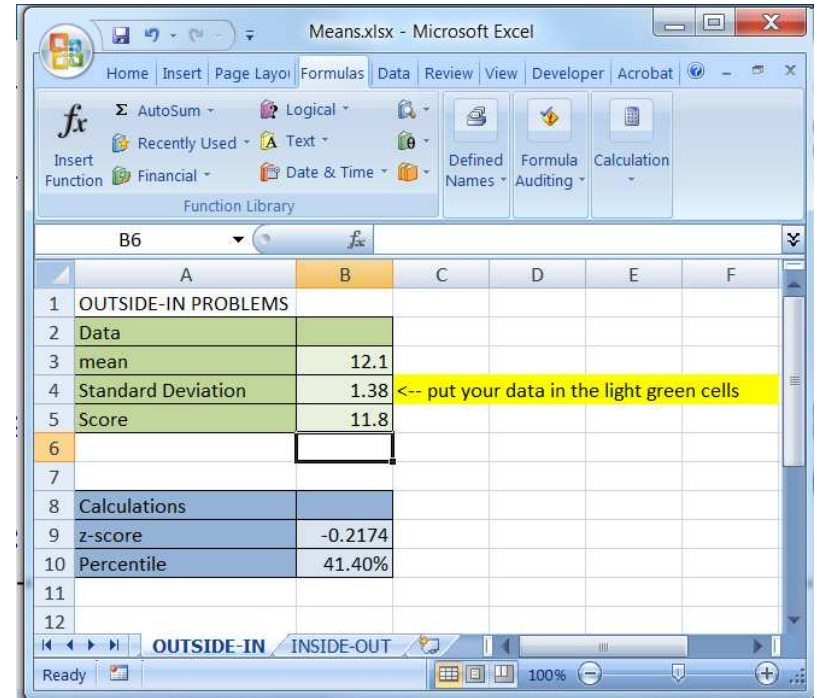
7.3. Example.

The Norman Speedskating Team has 18 members who skate at practices. For these 18 members, the average lap time is 12.1 seconds with a standard deviation of 1.38 seconds. Assuming that these data are from a normally distributed population, what percentile corresponds to a lap time of 11.8 seconds?

Solution. We'll do this one just with the spreadsheet.

To use the spreadsheet, open MEANS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled OUTSIDE-IN. Enter the data.

From this, the corresponding percentile is **41.40%**.



8. Inside-Out Calculations

So far all we have discussed is the first of the two types of normal computations: you have been *given* a raw score (such as a lap time) and have been *seeking* the percentile which corresponds to that score. You have been working “outside-in” with the normal tables since you have been finding the z score on the outside left edge of the table and then have been finding the corresponding proportion on the inside of the table.

Now we will consider the class of problem: you will be *given* a percentile and will be *seeking* the raw score which corresponds to that percentile. This class of problem will be “inside-out” since all of the steps are reversed.

8.1. Example.

Recall that GRE scores are normally distributed with a mean of 500 and a standard deviation of 100. What GRE score corresponds to the 80th percentile?

Solution. There is a spreadsheet solution to this as well, but we'll first do the longer way.

Step 1. Again, the first step is to make a list of what you know:

mean	$\mu = 500$
StDev	$\sigma = 100$
Proportion	0.80

Note that since the table deals with *proportions* we have converted the percentile to a proportion by dividing by 100%.

Step 2. Now we need to find the z score which corresponds to the given proportion of 0.80. To do this, we must first locate the proportion *inside* the body of the table.

Inside the table, you can't find 0.8000 exactly; you *can* find

$$\begin{array}{ccc} 0.7996 & \text{and} & 0.8023 \\ \updownarrow & & \updownarrow \\ z = 0.84 & & z = 0.85 \end{array}$$

The z -score corresponding to 0.8000 is somewhere between $z = 0.84$ and $z = 0.85$. Since the table is only accurate to two decimal places, we can't do much better than this. We can see, though, that $z = 0.84$ corresponds more closely to 80% than does $z = 0.85$ (0.7996 vs. 0.8023). Hence we will use $z = 0.84$.

Step 3. The next step is to convert this z score to a corresponding GRE score ("raw" score). To do this, you will need to use the formula

$$x = \mu + z \times \sigma$$

which is just the formula

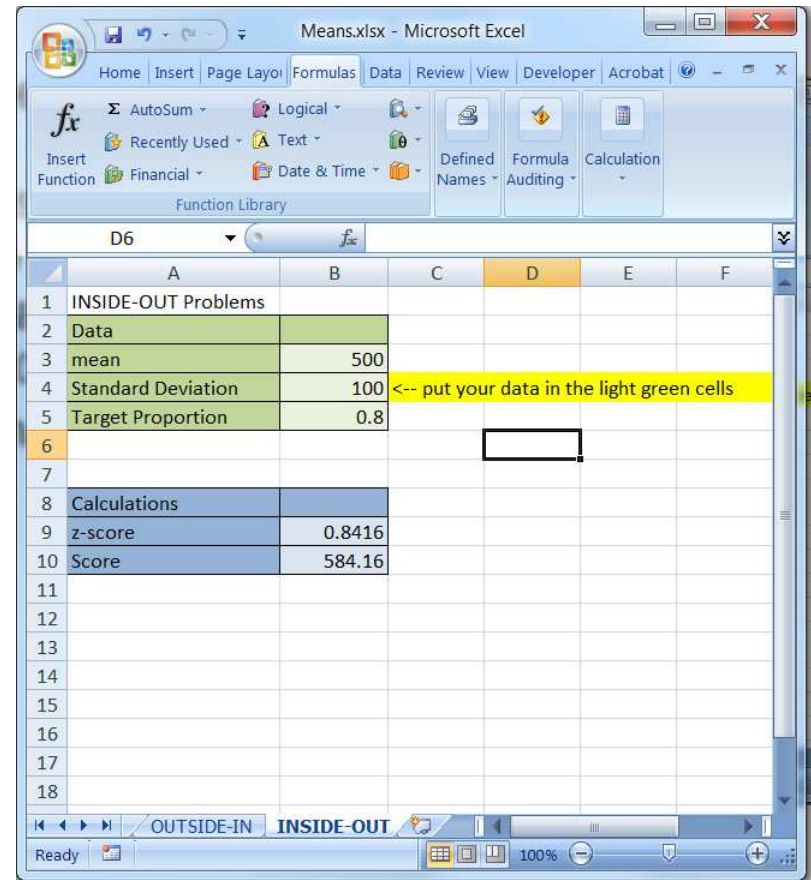
$$z = \frac{x - \mu}{\sigma}$$

solved algebraically for x . In our problem, this becomes

$$\begin{aligned}x &= \mu + z \times \sigma \\ &= 500 + 0.84 \times 100 \\ &= 500 + 84 \\ &= 584\end{aligned}$$

and so a GRE score of 584 corresponds to a percentile of 80%.

To use the spreadsheet, open MEANS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled INSIDE-OUT. Fill in the data dictionary and read the result.



Solution Template

Step 1. Make a list of what is known. Once again, sometimes you will be given census data and sometimes you will be given sample data. In the latter case, you will use the sample estimates \bar{x} and s to estimate μ and σ .

Mean	μ or \bar{x}
StDev	σ or s
Proportion	p

Don't forget to convert the given *percentile* to a *proportion* by dividing by 100%.

Step 2. Locate – as closely as possible – the given proportion inside the normal tables. If the proportion is greater than 0.5, look in the “positive” part of the table. If the proportion is less than 0.5, look in the “negative” part of the table. Reading “inside-out” find the corresponding

z score.

Step 3. Compute the corresponding “raw” score by using the formula

$$x = \mu + z \times \sigma.$$

If μ and σ are not known, you must approximate μ and σ with the sample mean and standard deviation:

$$\begin{aligned} x &= \mu + z \times \sigma \\ &\approx \bar{x} + z \times s \end{aligned}$$

to obtain an approximate conversion to a raw score.

————— **End of Solution Template** —————

8.2. Example.

Find the ACT score which corresponds to the 25th percentile. (Recall that ACT scores have a mean of 20 and a standard deviation of 5.)

Solution. **Step 1.** In this problem

Mean	$\mu = 20$
StDev	$\sigma = 5$
Proportion	$p = 0.25$

Step 2. We must look *inside* the table for 0.2500. The “positive” part of the table gives z scores greater than zero and hence proportions greater than one half; the “negative” part of the table gives z scores less than zero and hence proportions less than one half. Thus, we can only find 0.2500 inside the negative part of the normal table.

We can't find 0.2500 exactly but we can find

$$\begin{array}{ccc} 0.2514 & \text{and} & 0.2482 \\ \downarrow & & \downarrow \\ z = -0.67 & & z = -0.68 \end{array}$$

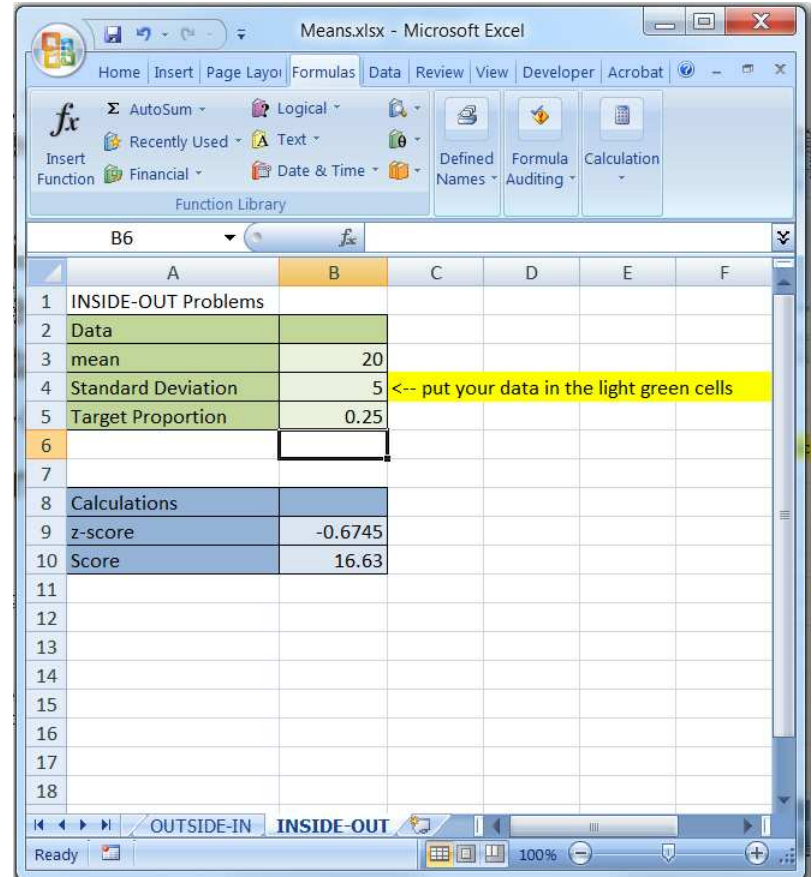
Since the proportion corresponding to $z = -0.67$ is slightly closer, we will use $z = -0.67$.

Step 3. Now we can find the corresponding ACT score:

$$\begin{aligned} x &= \mu + z \times \sigma \\ &= 20 + (-0.67) \times 5 \\ &= 20 - 3.35 \\ &= 16.65 \end{aligned}$$

and so a score of 16.65 corresponds to the 25th percentile.

Alternatively, using the spreadsheet:
To use the spreadsheet, open MEANS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled INSIDE-OUT. Fill in the data dictionary and read the result.



8.3. Example.

There are eight members of the Norman Speedskating team who have amateur cards and compete in regional meets; their lap times are

<i>9.8</i>	<i>10.2</i>	<i>10.8</i>	<i>11.8</i>
<i>12.3</i>	<i>12.5</i>	<i>13.1</i>	<i>13.8</i>


(a) Find the mean and standard deviation of this sample.

Solution.

To use the spreadsheet, open MEANS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled INSIDE-OUT. Enter the data.

From this, the mean is $\bar{x} = 11.79$ and the standard deviation is $s = 1.41$. (What did you do wrong if you got 1.322 for the standard deviation?)

	A	B	C	D
1	mean	11.7875		
2	Sample Standard Deviation	1.413645		
3	Population Standard Deviation	1.322344		
4	data =	9.8		
5	data =	10.2		
6	data =	10.8		
7	data =	11.8		
8	data =	12.3		
9	data =	12.5		
10	data =	13.1		
11	data =	13.8		
12	data =			
13	data =			
14	data =			
15	data =			
16	data =			
17	data =			
18	data =			



(b) Assuming that the data are from a normally distributed population, approximate the percentile which corresponds to a lap time of 11.3 seconds.

Solution. This is an outside-in problem.

Solution.

To use the spreadsheet, open MEANS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled INSIDE-OUT. Enter the data.

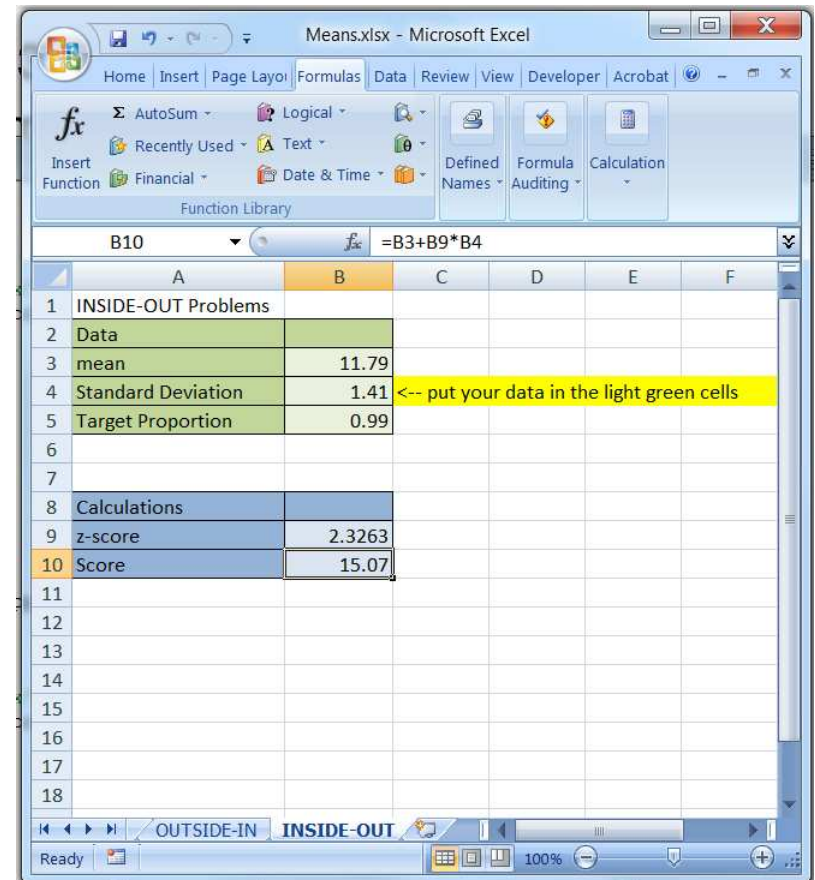
From this, the corresponding percentile is **36.41%**.

	A	B	C	D	E	F
1	OUTSIDE-IN PROBLEMS					
2	Data					
3	mean	11.79				
4	Standard Deviation	1.41				
5	Score	11.3				
6						
7						
8	Calculations					
9	z-score	-0.3475				
10	Percentile	36.41%				
11						
12						
13						
14						
15						
16						
17						
18						

(c) Assuming that the data are from a normally distributed population, approximate the time which corresponds to the 99th percentile. This is an inside-out problem.

To use the spreadsheet, open MEANS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled INSIDE-OUT. Enter the data.

From this, the corresponding lap time is **15.07 seconds**.



9. Distributions

	<i>Group I</i>	<i>Group II</i>	<i>Group III</i>	<i>Group IV</i>
<i># of Packets</i>				
<i>Mean # of M&Ms</i>				
<i>St. Dev of # of M&Ms</i>				
<i>Overall # of M&Ms</i>				
<i>Overall # Non-Red M&Ms</i>				
<i>Proportion Non-Red</i>				

- What is the population from which these samples are drawn?
- Why is there a different answer for each group for the proportion of non-red M&M's?
 - What do you suppose the results would look like if we did this experiment with 20,000 groups?

Experiment. Suppose that the true population proportion is $p = 0.75$. Take a sample of size 100 from this population and compute \hat{p} . Record the result.

Now take another sample of size 100 from the population; again compute \hat{p} and write down the result. You now have computed two (probably different) values for \hat{p} .

Repeat the process a third time, getting a third computed value for \hat{p} . Continue until you have 20,000 computed values for \hat{p} , each based on a randomly selected sample of size 100.

Make a frequency table of the results.

What do you suppose that a graph of the results would look like?

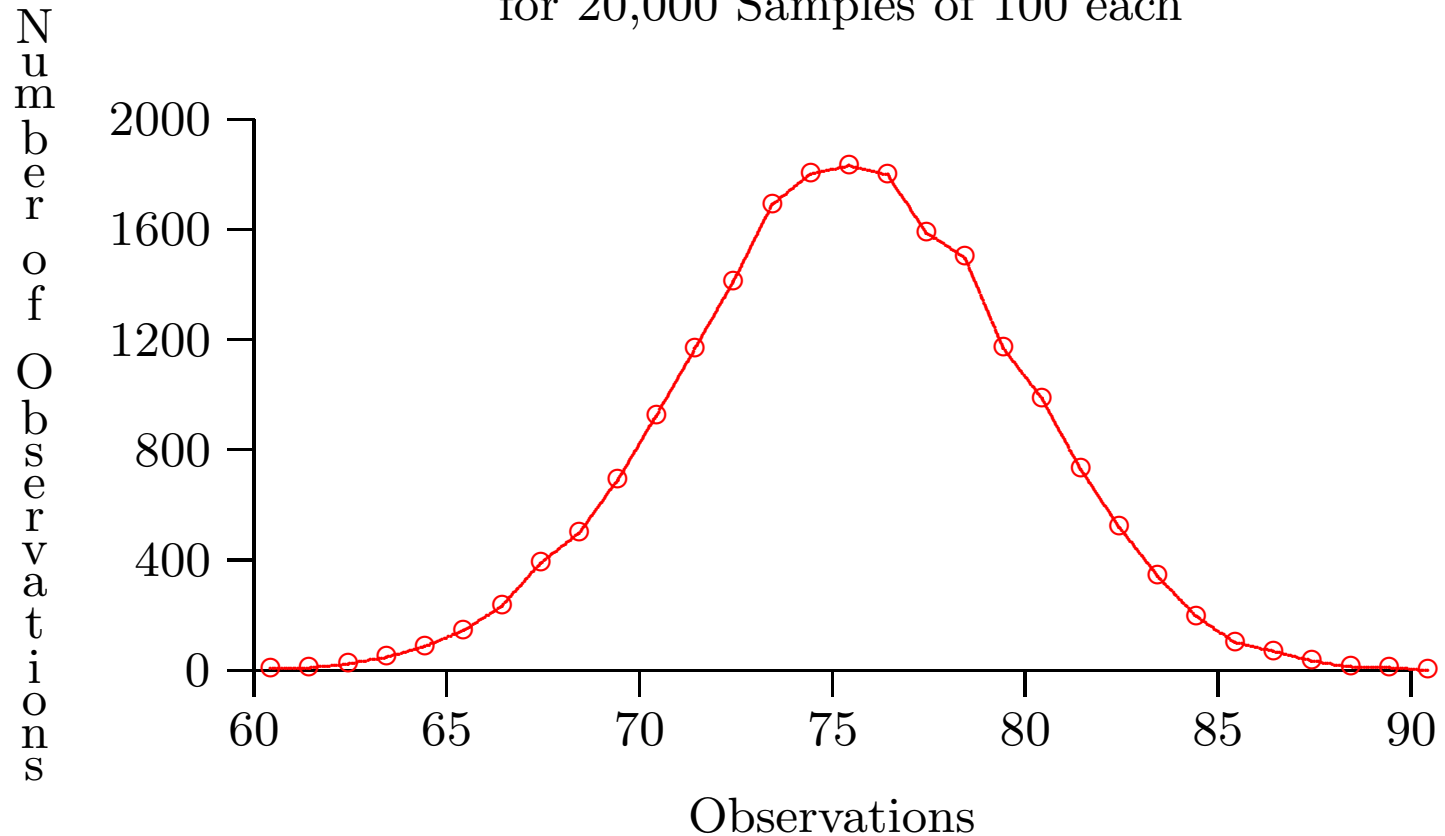
I performed this experiment (using a random number generator on a computer to simulate taking the samples and computing \hat{p}). I got the following frequency table:

observed %	frequency
60	6
61	9
62	23
63	47
64	86
65	144
66	233
67	389
68	498
69	690
70	924
71	1166

observed %	frequency
72	1410
73	1689
74	1801
75	1831
76	1798
77	1588
78	1498
79	1168
80	986
81	730
82	520
83	341
84	195
85	101
86	69
87	34
88	12
89	10

A plot of this data reveals

Computation of sample proportion \hat{p}
for 20,000 Samples of 100 each

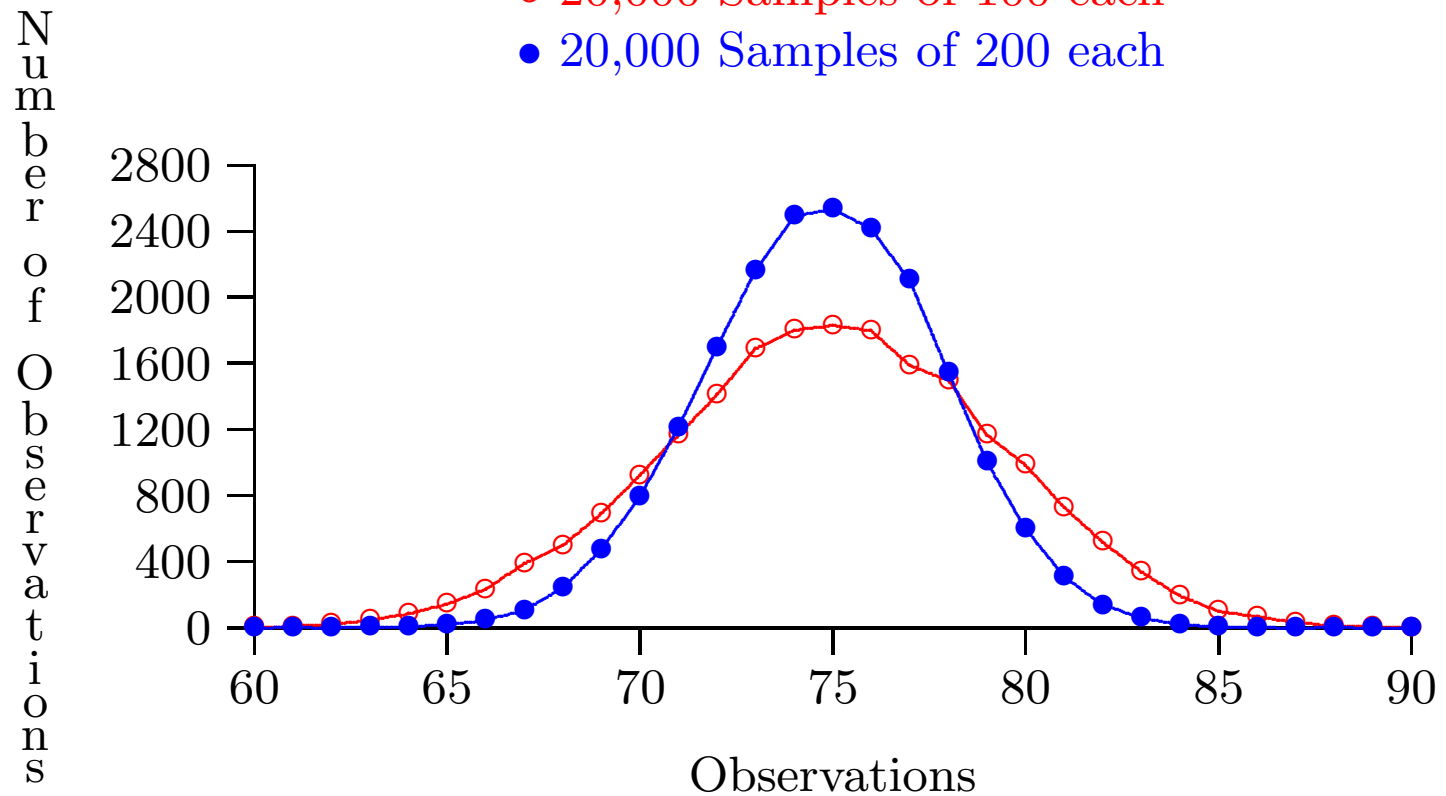


Suppose that you double sample size from 100 to 200, repeating again with 20,000 sample?

Computation of sample proportion \hat{p}

○ 20,000 Samples of 100 each

● 20,000 Samples of 200 each



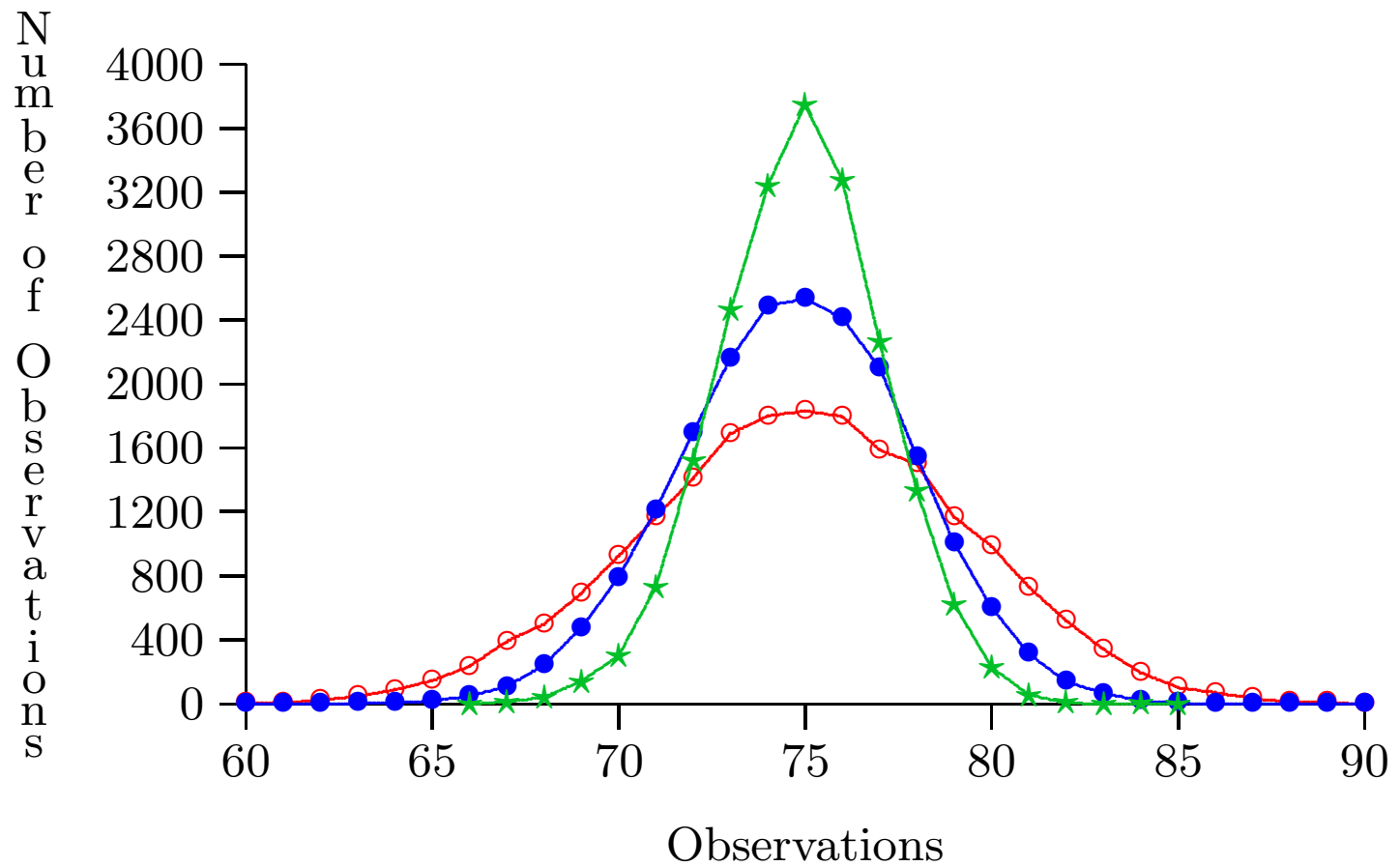
Note that the larger samples are **correct** more often (taller in the middle) and **incorrect** less often (lower in the tails). However doubling the sample size does not increase the peak by two.

Computation of sample proportion \hat{p}

○ 20,000 Samples of 100 each

● 20,000 Samples of 200 each

★ 20,000 Samples of 400 each



Observations:

- The distribution of repeated samples is approximately normal.
- Larger sample sizes are “more accurate” than smaller sample sizes.
- Larger samples have less variability than smaller samples.
- In order to **double** the accuracy you must **quadruple** the sample size.

Note that we have found a **sample** of 20,000 **sample proportions**. In principle it is possible to keep doing this forever. We have discovered that the **population** of all possible sample proportions is approximately normally distributed. Since it is normally distributed it must have an associated **mean** and **standard deviation**. We’ve also discovered that there is less variability in larger samples and quadrupling the sample size doubles the accuracy.

The **Central Limit Theorem** summarizes these results.

9.1. Central Limit Theorem

For large sample sizes, the distribution of the sample mean and the sample proportion are approximately normal. In particular, for large sample sizes the sample mean \bar{x} is approximately normal with mean μ and standard deviation

$$\text{sampling standard deviation of } \bar{x} = \frac{\sigma}{\sqrt{n}}$$

while \hat{p} is approximately normal with mean p and

$$\text{sampling standard deviation of } \hat{p} = \sqrt{\frac{p(1-p)}{n}}$$

10. Foundations of Research

Ethics in Research: a lesson from history.

In 1932 the US Public Health Service began a study of syphilis, using an initial group of 600 impoverished, African-American sharecroppers from Macon County, Alabama. At first, the men received free medical care, meals, and burial insurance. Of the original 600, 399 had previously contracted syphilis.

The participants were not told of their diagnosis. They were also not told when funding for medical treatment ran out, and were instead told they were being treated for “bad blood.” Even after penicillin became

available as a successful therapy in the late 1940s, both treatment and information about the diagnosis were withheld from the subjects.

In 1974, a whistle blower revealed the severe ethical lapses of this study. Known today as the **Tuskegee Syphilis Study**, it is the arguably the most egregious violation of ethics in the history of US biomedical research.

Congress responded to the revelation of this study by passing the National Research Act. Among other things, this law established the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The law mandated that the Commission determine the basic ethical principles that should underlie research in the behavioral and medical sciences. The Commission was further charged with establishing guidelines to assure that research is conducted in compliance with those principles.

The Commission first met in 1976 in the Belmont Conference Center of the Smithsonian Institution. Intensive meetings, discussions, and public hearings continued for three years, and in April of 1979 the Commission published what is now known as the Belmont Report. This report



provides the foundation for the oversight and regulation of research involving human subjects. The report is available [online](#).

Today, the Office of Human Research Protection (OHRP) in the Department of Health and Human Services promulgates regulations regarding federally funded research. Every US institution that receives federal funds and undertakes human subjects research must establish an **Institutional Review Board (IRB)**. The IRB reviews all research involving human subjects for compliance with , and no research may be undertaken without Board approval.

There are a few exceptions, such as the kinds of research projects you will do in this class. They are exempt since they are not invasive, risks are minimal, and don't involve protected classes such as children or prisoners.

Belmont Report: Basic Ethical Principles

We quote directly the three basic ethical principles identified in the Belmont Report.

1. **Respect for Persons.** Respect for persons incorporates at least two ethical convictions: first, that individuals should be treated as autonomous agents, and second, that persons with diminished autonomy are entitled to protection.
2. **Beneficence.** Persons are treated in an ethical manner not only by respecting their decisions and protecting them from harm, but also by making efforts to secure their well-being.
3. **Justice.** Who ought to receive the benefits of research and bear its burdens? This is a question of justice, in the sense of "fairness in distribution" or "what is deserved." An injustice occurs when some benefit to which a person is entitled is denied without good reason or when some burden is imposed unduly. Another way of conceiving the principle of justice is that equals ought to be treated equally.

Belmont Report: Ethical Principles applied to Research

1. **Informed Consent.** To quote the Belmont report, “Respect for persons requires that subjects, to the degree that they are capable, be given the opportunity to choose what shall or shall not happen to them. This opportunity is provided when adequate standards for informed consent are satisfied.”

This includes **information** on such things as the research procedure, the purposes of the research, anticipated risks and benefits, alternative procedures (where therapy is involved), and a statement offering the subject the opportunity to ask questions and to withdraw at any time from the research.

The researcher needs to be sure the subject **comprehends** the information given. It shouldn't be rushed, and it should be given in language

the subject can understand.

The subject must be a volunteer, free from undue pressure or constraints. The researcher needs to pay special attention to persons with diminished mental capacity, minors, or those in prison. Anything that reduces the voluntary character of participation violates informed consent. **Coercion** happens when there is a direct threat of harm to the subject. **Undue influence** occurs if the inducements to participate are excessive or improper.

2. **Assessment of Risks and Benefits.** The onus here is first on the researcher to balance the benefits of the research against the risks to subjects. If risk of harm to subjects is excessive relative to the benefits, the research should not be undertaken. It's also the responsibility of the researcher to give the subjects realistic information about the likelihood and character of harm. Harm could include physical discomfort or pain, but could also include psychological, social, legal, or other harm. Similarly, benefits generally relate to public health or welfare, but are uncertain while the research is on-

going.

3. **Selection of Subjects.** There should be fair procedures and outcomes with respect to the selection of subjects. For example, using only prisoners, or low-income persons in risky drug trials imposes higher risks on those populations and would be unfair and violate the principle of justice.


Research Questions

The launching pad for research is to start with a general area of interest and start asking questions.

For example, you might have noticed you have one friend who is terrified of spiders, while another thinks spiders are cool. So, **fear of spiders** might be a general area of interest.

Thinking further, you might wonder why people have different reactions to spiders. This leads some possible questions.

1. Are there some **characteristics** that people who are fearful of spi-

- 
- ders share with each other but not with people who are not fearful of spiders:
- (a) maybe one gender is more likely to be fearful of spiders than the other;
 - (b) maybe people who fear spiders are more prone to anxiety in general;
 - (c) maybe there is a difference in **risk-taking** between people who are afraid of spiders and those who are not.
2. Are people who are afraid of spiders more likely to be afraid of other things:
- (a) maybe people who are afraid of spiders are more likely to be also be afraid of insects than people who are not afraid of spiders;
 - (b) maybe they are more likely to be afraid of other things, like high places, or loud noises, than people who do not fear spiders;
3. Is there a common kind of traumatic personal experience, such as a spider bite, that people who fear spiders are more likely to share

than those who do not fear spiders?

Each of these questions are more specific than the original general topic. The sub-questions are more specific yet. We are **refining** our research question.

In developing a research question regarding a treatment or therapy, PICOT provides a set of questions to help guide and structure the process.

P	Population	What population are you studying or what Problem are you trying to solve
I	Intervention	What intervention do you plan
C	Control	What is your control or what will you use to compare your intervention against
O	Outcome	How will you measure the outcome of your intervention
T	Time	How much time will this take?

To settle on a **research question**, we need to consider some specific features:

- our question should be **realistic**;
- our question should be **quantifiable**;
- our question should be **falsifiable**;

- it should be **feasible** to answer our question.

Realistic Questions.

A **realistic** question builds on prior observations or theories. Realistic questions pose reasonable connections. It's not reasonable, for example, to suppose that shoe size would be connected with fear of spiders, although it's possible that age or education might.

Quantifiable Questions

To be **quantifiable** means you should be able to gather quantitative evidence to answer the question. Each of the above three questions satisfy this requirement. Notice question two is a more specific form of question one in that it asks if people who are afraid of spiders have other specific fears.

Notice, too, that all of the questions involve **comparing** those who fear spiders with those who do not. Comparing two groups is a fundamental feature of most deductive research and one hallmark of quantifiable research questions. This leads to the concept of independent and dependent variables.

Independent and Dependent Variables

We can imagine that we would gather several pieces of information on research subjects for the spider project:

- Does the subject fear spiders?
- What is the gender of the subject?
- Does the subject have other anxieties?
- How open is the subject to risk-taking?

Each of these pieces of information represent **variables** that we will need to measure. Typical, this measurement would involve designing a questionnaire to ask each subject. In some instances, such gender, we can just ask the subject directly. In other cases, such as risk-taking or anxiety, we might use a **scale** or set of questions that measure the trait indirectly. We'll talk more about scales in a later lecture.

The point here, though, is that the context of our project imposes a fundamental difference on the above variables. Our basic question is "How are people who fear spiders different from those who do not?" Broadly speaking, we suspect that "fear of spiders" **depends** on these on these other factors. So, in this example,

“Fear of spiders” is a **dependent variable** and the other variables like gender, anxiety, and risk-taking are **independent variables**.

The distinction between dependent and independent variables is fundamental to empirical deductive research. We’ll return to these concepts repeatedly in the rest of the course.

In the context of research objectives, we are formulating questions that describe relationships between **independent variables** and **dependent variables**.

Falsifiable Research Questions

On its simplest level, a **falsifiable** question is one that can be **disproved**. For example, the following would not be a falsifiable research question.

Do people who are fearful of spiders share common characteristics?

If this were false, then people who are fearful of spiders would share **no** common characteristics, a patently absurd assertion. They would all have tongues, for example, or ears. For the same reason, this is not a **reasonable** question.

For research question one, we are instead asking if people who are fearful of spiders are more likely to share a set of traits—fearful of ants, for example—than people who are **not** afraid of spiders. In other words, we would be **comparing** those who are fearful of spiders with those who are not—see the PICOT structure above.

Our question implicitly asks if we can **make predictions** about who is likely to be fearful and who is not based on a set of traits. **Falsifiable** questions **imply conjectures or predictions** that we can test by gathering evidence. The sub-questions for question one are more specific and are both quantifiable and falsifiable.

Feasible Questions

Finally, it must be possible to actually gather evidence to answer the question. Question three, about similar traumatic life experiences, is probably too broad and general to meet this criterion.

The [Raven Paradox](#) is an example of a research question that meets the above criteria but also exposes some paradoxes.

Research Hypotheses

Formulating research hypotheses is the final foundational step in research. A well-posed research question is most often phrased as a **question**. For example, we might ask

Is there a difference in education between people who are fearful of spiders and those who are not?

Or possibly

Is there a difference in age between people who are fearful of spiders and those who are not?

The transition from a research question to a research hypothesis is then easy.

A research hypothesis is a statement that speculates about the relationship between dependent and independent variables.

Thus, the hypotheses that correspond to the above question involve replaying “is there” with “We hypothesize there is...”

We hypothesize there is a difference in education between people who are fearful of spiders and those who are not.

We hypothesize there is a difference in age between people who are fearful of spiders and those who are not.

We could even be more specific about the relationship between the **dependent variable** fear of spiders and the **independent** variables, education and age.

We hypothesize that those with less education are more likely to be fearful of spiders.

We hypothesize that a younger people are more likely to be fearful of spiders.


To summarize this section, **ethical research projects** must include

- Informed consent of the subjects;
- An assessment of potential risks and benefits; and
- Fair procedures and outcomes with respect to the selection of subjects.

Research objectives involve formulating a set of questions regarding possible **relationships** between **dependent** and **independent** variables.

Research objectives should be

- realistic;

- 
- quantifiable;
 - falsifiable; and
 - feasible.

Research hypotheses formalize research objectives as affirmative speculations.

11. Research Design

Over the course of history humans have evolved many different ways of understanding the world. **Epistemology** is the branch of philosophy that deals with the nature and foundation of knowledge. In Plato's dialogue *Theaetetus* knowledge is said to be **justified true belief**. Thus there are three attributes of knowledge: one believes a statement is true, it really is true, and the belief that it is true is justified.

The question of **justification** then becomes central to any theory of knowledge. Research is a particular theory of knowledge based on **empiricism**, i.e., that theories are tested against reality and are accepted or rejected depending on how well they correspond to evidence.

Empiricism is simply one of many different ways of justifying knowledge. Some others include:

- *An appeal to authority.*
- *Direct observation of the senses (naive empiricism).*
- *Mysticism.*
- *Logic or rationality.*

Scientists will sometimes use non-empirical methods. Scientists rely on the prior work of other scientists, which is one kind of appeal to authority. The very belief that there is an external world in which cause-and-effect are real phenomena might be thought of as a mystical belief. In fact, in some settings, such as the sub-atomic world described by quantum mechanics, this phenomenological approach fails to pass the test of correspondence to the facts.

However, the basic premise of science, that theories are accepted or rejected depending on how well they correspond to the facts, remains at the heart of scientific inquiry.

Most research projects pass through four basic phases:

- A. Organization and Planning Phase
- B. Data Collection Phase
- C. Data Analysis Phase
- D. Reporting Phase

The balance of these notes describes these phases in greater detail.

©Cartoonbank.com



"That's the gist of what I want to say. Now get me some statistics to base it on."

11.1. Steps in Phase A, Organization and Planning.

A.1. Background Research

A.2. Define the terms

A.3. Define the variables

A.4. Develop a model

A.5. Determine Objectives

A.6. Design measures

A.7. Select the strategy

A.8. Design the instruments

A.9. Design the Sample

A.10. Develop the budget

The steps are not necessarily done in the above order and often decisions in one step force the researcher to revisit decisions made in earlier steps. Further there are many other ways of dividing up the preliminary design steps: no one methodology can apply to all situations. The above taxonomy is provided as one of many possible templates.

A.1. Background Research.

Background research consists of learning what is already known about your research topic. Most generally this will consist of building an **annotated bibliography** of research articles already published on your topic. There are many tools available today to assist in bibliographic searches. A reference librarian can assist you in learning these tools (but you should not ask a librarian to do your search for you). Once you have constructed a list of articles, the next step is to read as many of them as you can, taking notes on the content. Those that are relevant to your research topic, along with your summaries, become your annotated bibliography. The bibliographic entries should conform to an accepted standard, such as the APA standard. In this way others can also find your bibliography useful, plus using the standard will assure that you have enough information to relocate the document.

A.2. Definitions.

The basic **terms** needed to carry out the research must be clearly defined. These definitions must be

- appropriate to what is being studied; and
- applicable in field research.

In particular, the person collecting data must be able to use the definition to decide whether or not it applies to whatever is being observed.

Another critical definition is the **population** that is being studied. This must also be clear, appropriate and applicable in field research.



In defining the population there are often two conflicting goals.

- Make the study as broadly applicable as possible.
- Make the study narrowly focused on a particular group or question.

The former goal leads to a different, and more expensive, kind of population than does the latter goal. Sometimes, even with sufficient resources, it is more effective to conduct several narrowly focused studies rather than one study on a large and complex population.

We will discuss this in greater detail in the section on sample design.

A.3. Variables.

A **variable** is any attribute that varies in quantity or quality from subject to subject.

Examples:

- Age
- Gender
- Sexual Orientation
- Weight
- Shoe Size
- IQ
- Hair Color
- Income

Selection of variables will depend on their relevance to the topic

Which variables you choose to measure is often driven by a theory about which variables matter. For example, in a study of social behaviors one would not ordinarily include eye color or hair color as variables. However, if the study dealt with the role of serotonin in social behaviors, one might include both eye color and hair color since natural serotonin levels are different in certain blue-eyed, blonde-haired populations. Variables can also be classified as **dependent** (sometimes called response variables) and **independent** (sometimes called control variables).

Dependent variables are generally the primary effect or outcome that you are studying.

Independent variables generally influence the primary effect or outcome.

Thus if you are studying how well people cope with stressful social situations, you might use a standardized instrument to measure tolerance to social stress. This instrument would measure the **dependent variable**. If you think serotonin levels influence the ability to tolerate social stress, a **independent variable** might be serum levels of serotonin. Alternatively, since certain blonde-haired and blue-eyed populations tend to have depressed serotonin levels, you might use eye color and hair color as dependent variables.

In this example, we have theorized that the independent variable (serotonin level) influences or controls the dependent variable (tolerance to social stress). Hence the terms *control* and *response* variables are sometimes more intuitive than *independent* and *dependent* variables. Implicit in this selection of variables is the *theory* that serotonin levels influence one's tolerance to social stress. In most situations the selection of variables is intimately tied to an underlying theory. This theory is the foundation for the research questions.

A different way in which variables can be classified is by what they measure:

- **Attribute variables** refer to a characteristic of the subject.
- **Quantitative variables** refer to those attributes for which there is a natural numerical measurement.

Examples of attribute variables might be hair color, religious affiliation, political affiliation or gender.

Examples of quantitative variables might be class rank, GRE score or annual income. Quantitative variables fall into three categories:

- **Ordinal** where measures are larger or smaller but magnitude is not specified (class rank is an example).
- **Interval** where a magnitude is specified but there is no natural zero (most standardized psychometric instruments are interval measures).
- **Ratio** where there are both magnitudes and a true zero (a measurement such as annual income is an example of a ratio variable).

A.4. The Model.

At the most fundamental level, the **model** describes the manner in which the variables interact with each other and the context in which that interaction occurs.

Thus when you select your variables, some of which will be dependent and some of which will be independent, you have already described at least in part how the variables interact. The context for the interaction includes the population you are studying and a reasoned explanation of why the variables might interact in the theorized way. This simple model is often sufficient for social science research projects.

Models can be made more complex by adding more detail or structure to the theorized interaction. For example, in a project that studied **income** we might include **annual income** as the dependent variable and **gender** as an independent variable. This selection of variables theorizes that there is relationship between gender and income without specifying the particulars. A more detailed model might theorize that income is lower for females than for males. This provides more detail to the relationship between the two variables.

Sometimes your model can be even more specific. Some of our statistical methods (regression) will let us actually decide whether or not the independent variable can reliably predict the value of the dependent variable. In fact, statistical regression would actually deduce a mathematical formula relating income and gender (where gender were coded as "0" for males and "1" for females, for example). The formula would include an error estimate as well, to give the researcher an idea of the strength of the relationship.

Of course, the **context** of the research will determine when a variable is independent and when it is dependent.

For example, if one were studying voting patterns, then the dependent variable might be **party affiliation** while dependent variables might include both **income** and **gender**. In this example, income is no longer the dependent variable but instead, in this new context, an independent variable.

When defining variables the researcher should keep in mind in mind that each individual has a unique perspective. Diverse research teams can assist bringing varied perspectives to the process, reducing the likelihood that questions will be mis-interpreted.



"Pi what squared? Long John, you should be able to get this."


A.5. Research Objectives.

Your research objective describes what you propose to learn by doing your project. Often research objectives are expressed as hypotheses.

A **hypothesis** is a provisional idea whose merit is to be evaluated.

A well-formed hypothesis should be **falsifiable**. In order for a hypothesis to be falsifiable

- there must be certain explicit and observable predictions that can be deduced from the hypothesis; and
- it must be possible to gather data, usually through observations, to see whether or not the predictions are correct.



Thus a goal of a research project is often to gather evidence about the predictions of a model. In our prior example on serotonin levels and tolerance for social stress, we hypothesized that changes in serotonin levels would change tolerance for social stress. This is falsifiable since this conjecture meets the two conditions above.

A more detailed conjecture, such as depressed serotonin levels depress tolerance to social stress, is likewise falsifiable.

We are then in a position to gather data and see if the basic predictions of the hypothesis are borne out. If they are, then the hypothesis still is not necessarily proved. It is still provisional since we have failed to disprove it, which is different from proving it! Most hypotheses will go through many years of research and testing before being generally accepted by the scientific community as proved.

Even widely accepted hypotheses can be disproved by the appearance of new data. A striking example was the sudden appearance in 1938 of the coelacanth, a species of fish thought to have been extinct for over 65 million years. The hypothesis that the coelacanth was extinct was falsifiable and, since there were no recent fossils and no known living specimens, the consensus was that the extinction hypothesis was true. The appearance of the fish in the ocean off of South Africa, however, provided evidence proving the hypothesis false.

The lesson of the coelacanth is that scientific knowledge is implicitly provisional. There is always the potential that new data can either refine or overturn the existing scientific consensus.



A.6. Define Measures for the Variables.

At first glance it may appear that the definition of your variables and how you will measure them are the same step. Sometimes this is true and sometimes not.

For example, your variable might be **annual income**. At first glance this would appear to be a quantitative variable of the ratio type. This even has a well-defined measurement provided by everyone's income tax filings.

However, subjects will likely be quite reluctant to reveal to even the most trusted researcher the exact dollar amount of their annual income. This reluctance would greatly complicate data collection and could introduce considerable bias in the study.

A better approach might be to choose to measure this as an ordinal rather than ratio variable by asking subjects "which statement best describes your annual income" and giving them a set of responses:

- $\leq \$0.00$
- $\$0.00 < \$20,000$ annually
- $\$20,000.01 < \$40,000$ annually
- $\$40,000.01 < \$60,000$ annually
- $\$60,000.01 < \$80,000$ annually
- $\$80,000.01$ or more annually

Subjects are likely to be more comfortable with this approach. Thus the researcher will most likely get more willing and truthful respondents than with the more intrusive question "what did you report to the IRS last year for your annual income."

A.7. Research Strategy.

The research strategy plots out the specific ways in which data is gathered. The strategy includes for example

- *how the researcher will interact with subjects, both before and after gathering the data;*
- *protecting the rights of the subjects (ethical conduct);*
- *the method by which the researcher will gather data;*
- *whether data will be gathered in an experiment, through observation or by survey.*

One fundamental question to resolve is whether or not the researcher will **observe behavior**, **interview subjects**, or do both. Many research projects will both conduct interviews and observe behaviors. In this case special care must be taken to avoid bias!

Sometimes the act of observing can influence behavior. For example, in conducting a research project to see if people obey the stop signs on campus, would it be a good idea to stand next to the stop sign with a clip board to gather your data?

Some things to consider for observational studies:


- *Will the researcher be a participant in what is being observed?*
- *Will the subjects know that they are being observed?*
- *Will the observations be in a naturally occurring setting or in a contrived setting?*
- *Will a checklist or other structured tool be used to record the observations?*
- *Will observations be direct or indirect (for example, using video or audio recordings)?*

Clearly there are many ethical issues to consider in observational studies since it is possible that the subjects will not have had an opportunity for informed consent.

In your research projects, you may not use minor children in any way without first having the approval of the Institutional Review Board. IRB approval generally takes a minimum of 6-8 weeks to obtain.

Any research involving groups with limited capacity for informed consent (examples would be prisoners or the mentally impaired) must have prior IRB approval before any contact with subjects is undertaken. In all cases care should be taken to protect the confidentiality of respondents





If the researcher chooses to interact with subjects there are many different ways to accomplish this. Qualitative researchers will generally gather data from such sources as focus groups, case studies, in-depth interviews or expert panels. Quantitative researchers will rely on formal interview techniques, either in a personal interview, in a telephone interview or with a written survey.

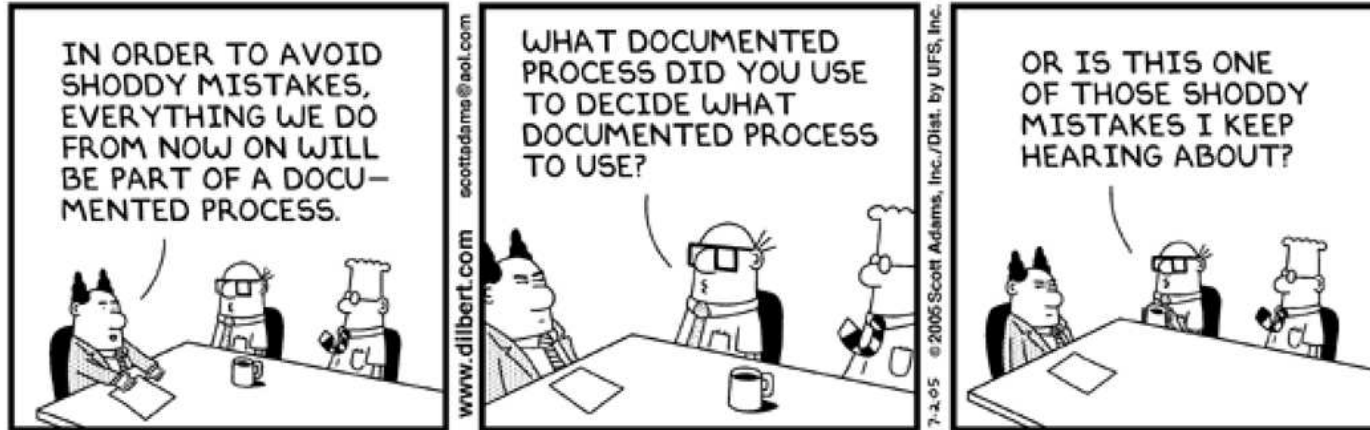
If the researcher interacts with the subjects, then special care should be taken to avoid **bias**. Many researchers will carefully script all interactions with subjects, including those interactions prior to administering the survey instrument, to assure that unplanned interactions do not influence subject responses.

Finally, some studies are **cross-sectional** and some are **longitudinal**. A cross-sectional studies gathers point-in-time data while a longitudinal study follows the same group of studies for an extended period of time. The Framingham Heart Study is a famous example of a long-running longitudinal study.

A.8. Research Instruments.

Your research instrument is the tool you use to record your observations. In an observational study this might be a checklist or other structured way of standardizing the observations. In a survey, instrument is the list of questions that you ask.

The researcher should be sure that the instrument measures every variable required for the study. The instrument should not measure variables not needed for the study.



© Scott Adams, Inc./Dist. by UFS, Inc.

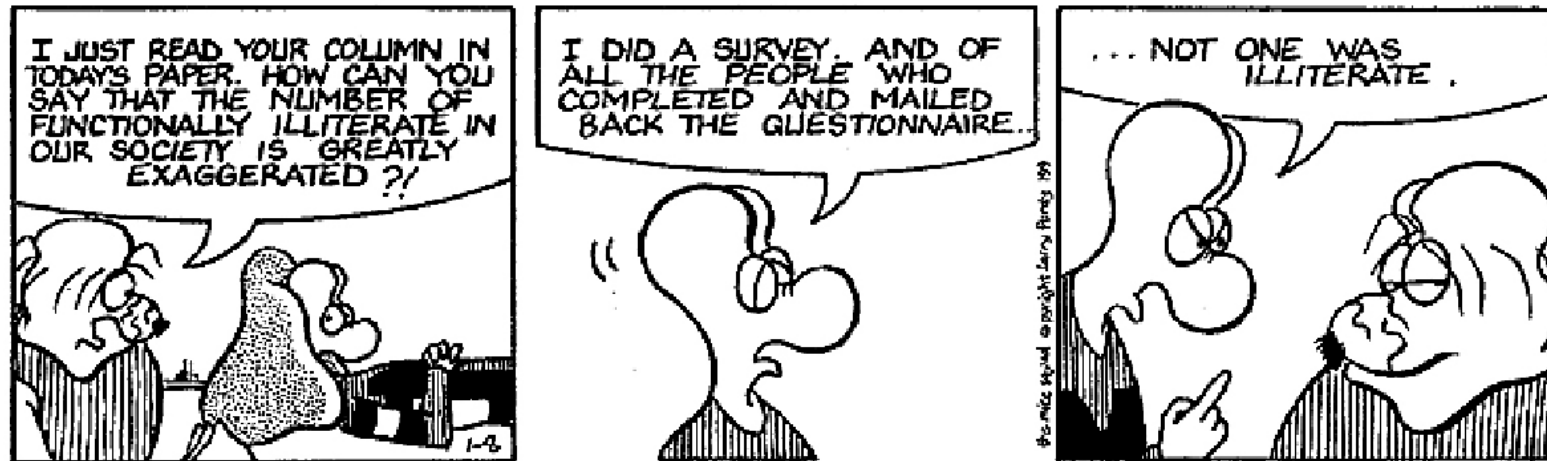
When a researcher does a survey there is necessarily interaction with the subjects. Care must be taken to avoid the instruction of bias in this interaction. When writing survey questions the researcher should

- *Be sure that the survey measures every variable required by your design.*
- *Keep each question focused on one variable (avoid double-barreled questions).*
- *Keep your questions simple. Remember, if a question can be misinterpreted, it will be!*
- *Avoid vague questions.*
- *Avoid leading questions. Sometimes even question order can influence responses!*
- *Make sure that the respondent has enough information to answer.*
- *Be sure that choices are mutually exclusive and collectively exhaustive.*
- *Minimize the number of open-ended questions.*

A.9. Sample Design.

Read the section in the **Study Guide** on this material.

THE MICE SQUAD



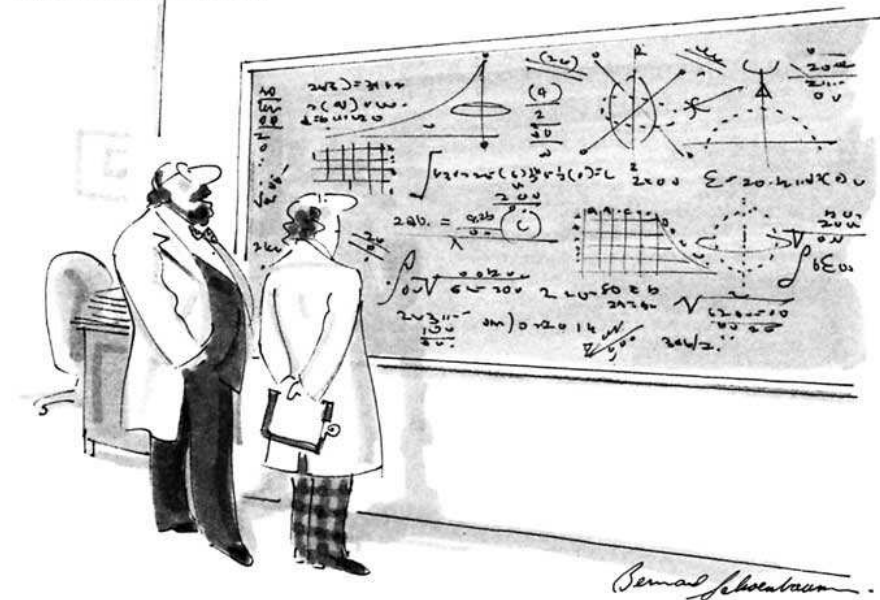
A.10. Budget.

At a minimum your budget should include consideration of

- *Finances*
- *Personnel*
- *Access to subjects*
- *Time constraints*

If your resources are insufficient in any of these critical areas you will need to redesign your project.

©Cartoonbank.com



"Oh, if only it were so simple."

11.2. Steps in Phase B, Gathering and Summarizing the Data.

- B.1. Pilot Study
- B.2 Gathering the Data
- B.3 Summarizing the Data

Generally the researcher will first do a pilot study to test research instruments, basic research design and other assumptions. After any necessary adjustments, then the data collection begins, using all of the elements of the design. As the data is gathered it is often summarized. The summary can include graphical demonstrations, means, standard deviations, proportions or more advanced measurements such as correlations. Generally the researcher will explore many different summaries before settling on those that most clearly summarize the data.

11.3. Steps in Phase C, Analysis and Conclusions.

- C.1. Statistical Analysis
- C.2 Conclusions
- C.3 Limitations
- C.4 Future Research

The **statistical analysis** is designed to help the researcher form **conclusions**, along with an estimate of how likely the conclusions are to be true. The analysis of data will be covered in later lectures.

In this phase the researcher selects the most effective ways of summarizing the data. Note, however, that the original research plan will almost certainly have in mind applying particular statistical techniques to analyze the data.

One of the basic steps in a research project is to focus your research on a well-defined question, with a limited and well-defined population. While these steps often constitute good methodology, they also limit the scope of your project. In the course of gathering and analyzing the data the researcher may discover other, unplanned limitations of the study that become part of the research record. Limitations, of course, whether planned or unplanned are opportunities for future research!



**"It's my fervent hope, Fernbaugh,
that these are meaningless statistics."**

11.4. Steps in Phase D, Reporting the Results.

The **research report** is an organized recapitulation, in narrative form, of the research project. Many professional organizations have particularized formats for the research report. The elements discussed in this section are generic for the purposes of this course and are intended as an exemplar rather than as definitive.

- D.1. The Title
- D.2. The Abstract
- D.3. Introduction
- D.4. Research Objectives
- D.5. Methods
- D.6. Descriptive Statistics
- D.7. Analysis
- D.8. Conclusions
- D.9. Discussion
- D.10. Appendices
- D.11. References



D.1. Title.

Your research project should have a title that is short, descriptive and captures the reader's attention.

D.2. Abstract.

Your abstract should generally be limited to 200 words (approximately one double-spaced typed page). It should outline your research objectives, methods and conclusions. Bear in mind that many readers will only look at your abstract.

D.3. Introduction.

This section gives the background for your research and should include

- *a summary of past results;*
- *provide a context for the research (why should anyone care about your topic?);*
- *define the fundamental concepts you will use;*
- *define the variables;*
- *define the population;*
- *define the sample.*

D.4. Research Objectives.

This section lists the research objectives. These are listed as conjectures or hypotheses with the intent of developing evidence to determine whether or not the conjecture is supported by objective data.

The purpose of a research project is never to "prove" or "disprove" a hypothesis. Objectives phrased in that way show that the researcher is biased toward one outcome or another.

D.5. Methods.

This section describes

- *strategy for sample selection;*
- *research protocols (how the researcher interacted with the subject);*
- *how the variables were measured;*

D.6. Descriptive Statistics.

This section includes summaries of the data, including for example

- *histograms*
- *pie charts*
- *scatter plots*
- *means*
- *proportions*
- *standard deviations*
- *correlation coefficients*

or whatever else makes the data easier to visualize and understand.

D.7. Analysis.

The statistical analysis is discussed in this section. The specifics of statistical testing and analysis will be covered in subsequent lectures.

D.8. Conclusions.

The main conclusions are often determined by the results of the statistical tests and so the "analysis" and "conclusions" sections are sometimes conflated.



D.9. Discussion.

This section should recapitulate the context, assumptions, hypotheses, methods and conclusions of the study. This section should also discuss limitations to the research and possibilities for future research.

D.10. Appendices.

It is unusual for a published research paper to include the complete research instruments or all of the data that was studied. Indeed, subject confidentiality may require destroying or protecting some of the data. However representative examples from the instruments and data are sometimes included in an appendix.



D.11. References.

The bibliography should include only those items actually cited in the study. The researcher should, of course, cite any reference that is used directly or indirectly in implementing the project or reaching conclusions. References should follow the style appropriate to the discipline, such as the APA style. Most journals and professional societies have specific style manuals authors are expected to use.

When in doubt about references, a librarian is always reliable source of information.

12. Experiments

Experiments are the single most important tool available to the researcher. While not every research project can be fully framed as an experiment, the basic principles of experimental design can often still be applied to increase reliability.

The essence of an experiment involves exposing subjects to a treatment and observing the results.

The **treatment** is often fundamentally connected to the researcher's **hypothesis**. The hypothesis is falsifiable, so there are explicit predictions that can be deduced from the hypothesis. The treatment is then selected to see if those predictions are correct.

There are three fundamental principles of experimental design.

- Control
- Randomization
- Replication

Every well-designed experiment will attend to each of these principles. Each principle contributes in a powerful way to the reliability of the conclusions. In order to provide a context for the application of the principles, we will consider the following hypothesis.

12.1. Example.

Hypothesis: Low carbohydrate diets result in greater weight loss for males over age 30 than do low fat diets.

12.2. Control.

Notice that our hypothesis actually involves comparing two different diets: low carbohydrate diets and low fat diets. This is not an accident. Almost every well-formed hypothesis will involve some kind of comparative conclusion, whether between treatments (as in this case) or between treatment groups. For example, another possible hypothesis might be

Hypothesis B: Low carbohydrate diets are more effective for males than for females.

In this latter case our hypothesis there is only one treatment, but the conjecture is that it has different effects on different groups.

Even a seemingly direct hypothesis such as

Hypothesis C: Low carbohydrate diets result in weight loss.

is really a comparative statement. If we simply expose subjects to a low carbohydrate diet and observe weight loss, we don't know that the weight loss would not have occurred anyway, for example it might be an artifact of the subjects being observed (the **Hawthorne effect**). In order to properly test this hypothesis, we should have a second group exposed to no particular diet, and compare the results of both groups. The only difference between the first group (the **the experimental group**) and the second group (the **control group**) is the special diet to which the first group is exposed. If we see a difference between these two groups, which are the same in every way except the diet, then and only then do we have reliable evidence that supports Hypothesis C.

Thus using **experimental** and **control** groups is an important feature of experimental design and one of the ways in which the researcher exercises **control**.

Sometimes the experimental and control groups use the same subjects. In our original hypothesis

Hypothesis: Low carbohydrate diets result in greater weight loss for males over age 30

we might divide the subject pool into two groups, A and B. We might then expose subjects in Group A to a low carbohydrate diet and subjects in Group B to a low fat diet for three months. At the end of the first three months, we could then reverse the diets, with Group A exposed to a low fat diet and Group B exposed to a low carbohydrate diet. In this way we guarantee that any difference between the outcomes in "low fat" and "low carb" diets is due to the diet and not due to differences in the subjects.

Other ways in which the researcher exerts **control** over the research process involve the **research protocol**. The protocol is method by which subjects are exposed to the treatments, how the researcher interacts with the subjects, and how the measurements are taken.


Part of any protocol with human subjects involves assurances of ethical treatment. The most basic elements of ethical treatment include

- **Informed Consent.**
- **Information about potential risks and benefits.**
- **Ability to withdraw from the study at any time without penalty.**
- **Basic information about the purpose of the experiment and what the subjects will experience.**

Of course the researcher should take steps to avoid exposing subjects to undue risk, and should consider whether the potential benefits of the proposed research justify any potential risks to the subjects. Universities are required by law to have independent review boards that approve all research involving human or animal subjects.

In our diet example, all of the publicity about the low-carb Atkins diet might influence the outcomes. The subjects' expectations about weight loss might influence their compliance with the diet or might influence weight loss all by itself (the so-called **placebo effect**). Thus in our example we might not inform the subjects about the sequencing of the diets.

Presumably in our example the dependent variable would be **weight loss**. This means that a member of the research team will weigh the subjects at least at the beginning and end of each phase of the study. It is possible that the expectations of the researcher could also influence the measurements. This is particularly true in the case of drug trials where the control group is often receiving a placebo and the experimental group is receiving the experimental drug. Thus the member of the research team taking the measurements also is usually not informed as to which group is being observed.



This kind of design is said to **double-blind** since neither the subjects nor the observers know which treatment is being measured. This is another fundamental way in which the researcher exerts control over the experiment.

In order to assure that the experiment measures the difference between the treatment groups, the researcher will often take steps to standardize all interactions with subjects. Intake interviews, exit interviews, processing questions and all other interactions with the subjects are often carefully scripted and members of the research team are not permitted to deviate from the script. Even the physical setting – subject sitting or standing – can influence results and is therefore standardized. This is another aspect of **control**.

12.3. Randomization.

Experiments will almost always involve samples rather than census data. When dealing with samples, error is unavoidable since the researcher necessarily has incomplete information. Good experimental design avoids **bias** or systematic error. Systematic error favors one outcome over another in the experiment and thus can lead to false conclusions.


Random error however does not favor one outcome over another, but is neutral with respect outcomes. Thus in our diet example we would randomly select the test subjects.

In a random sample, every member of the population has an equally likely chance of being selected for the sample.

There are many challenges with constructing a truly random sample. Properly speaking the researcher should have a complete list of all members of the population and then randomly select the sample from that list: similar to a giant lottery.

There are many ways in which sampling bias can occur. For example, running an ad in a newspaper might result in persons more motivated to lose weight or to persons who are otherwise not representative of the population. Similarly, randomly selecting potential subjects from a phone directory limits the subject pool to those who have listed telephone numbers, missing those who only own cell phones, who do not have a phone or whose numbers are unlisted. These sampling methods do not involve **randomization** and are hence subject to bias.

Other sampling techniques involve **stratified random samples**, **cluster samples**, and **multi-phase sampling**. All of these are designed to increase the likelihood that the sample is similar to the population being studied and discussed more fully on the course website.



In our diet example, one proposed strategy was to divide the subjects into two groups, alternating the diet plans between the two groups. The division into the groups could be done randomly – for example by flipping a coin. This element of randomization is much easier to manage since we are now dealing with a smaller group, the sample. Note that this approach has the effect of randomizing the sequence in which any individual subject is exposed to the two diet plans.

By randomizing the sequence in which the subjects are exposed we are also exerting control over the sequence. It is possible that one diet is more effective if followed by the other, so having half our subjects randomly selected to be exposed to the diets in inverted order controls for this.

Blocking is a concept similar to stratified random samples. In stratified random samples, the population is divided into strata and then the sample is constructed by sampling from each strata. The strata are defined in ways that are relevant to the variables: for example, subject weight might be useful strata in our example.

In blocking, the sample is already constructed but there might still be differences in the subjects that could influence the results. In our example, early weight loss tends to be higher for persons with higher weight. Thus if one group started with more persons of higher weight, this could bias the outcomes. Thus the researcher might **block** the sample by initial weight, then randomly sequence the diets in each block. Ultimately the researcher still has two groups which are exposed to the diets in inverse order, but the groups are constructed in a way that makes them more similar.

Once again, the goal is to assure that our measurements are sensitive to differences in the treatments rather than unplanned differences in the subjects.

12.4. Replication.

This is the simplest principle: make the sample as large as possible.

This will make your measurements more sensitive to differences in the treatments and less sensitive to differences in the subjects.

As we have already seen, larger samples have smaller sampling variance. This is another way of stating the above observation. While larger samples are certainly more reliable, we shall see that relatively small samples can provide highly accurate and reliable conclusions. Properly constructed samples and surveys have repeatedly proven to provide reliable and accurate predictions regarding many phenomena, including elections.

12.5. Clinical Trials.

Clinical trials are a particular kind of experiment. These trials, which are required for new pharmaceuticals, are designed to occur in three phases, each testing a different hypothesis:


- **Phase One Trials** only look for harmful side-effects.
- **Phase Two Trials** test for efficacy.
- **Phase Three Trials** are longer term and test for both harmful side-effects and efficacy.



Phase One trials tend to be *cross-sectional studies*, relatively short-term and involve relatively small samples. Phase one trials look only for harmful side-effects. Aspirin would most likely not be approved for sale if it were introduced today due to harmful side-effects, namely aspirin allergy in a significant part of the population.

Phase Two trials tend to be also be *cross-sectional studies*, somewhat longer-term and can have samples that are quite large (at least 10,000). Phase Three trials are *longitudinal studies* and often have extremely large sample sizes.

Typically Phase Two and Phase Three trials are expected to identify harmful side-effects that affect as few as 0.5% of the population with at least 99% reliability. This level of accuracy and reliability requires samples of at least 10,000.



In 1950 Jonas Salk spent 18 months in human trials before the Salk Polio vaccination was approved for use in the general population. Today it typically takes as much as 18 **years** for all three phases of a clinical trial to complete and a new drug to be approved for sale. Fewer than one drug in one thousand that enters Phase One trials successfully completes Phase Three trials.

Clinical Trials were introduced in the 1960's. What happened between the Salk vaccine trials in the 1950's and the introduction of clinical trials 1960's that led to the introduction of more stringent protocols? Why is this being re-thought today?

13. Sampling

One of the first steps in research design is to define the population you wish to study. There are then two possible strategies for learning about this population:

- **Census** – gather complete information on everyone; or
- **Sample** – gather information on only part of the population.

Most of the time you will necessarily gather a sample. Because a sample is *incomplete information*, your data and your conclusions will necessarily be subject to error.

This unit is about sampling. The incomplete information contained in a sample means that there will be *error*. Thus reducing *systematic error* or *bias* becomes critical.

Sampling has two competing goals:

- Reduce costs; and
- Minimize error.

Error is minimized by:

- making the sample as representative of the population as possible and hence reducing *bias* and
- increasing sample size (the *replication* principal of experimental design).



There are a number of different kinds of samples. Some of the various types are:

- Convenience Samples;
- Self-selected Samples;
- Simple random samples;
- Systematic Samples;
- Stratified random samples;
- cluster samples;
- multistage samples.

The first two are inexpensive and easy to produce. They are also very prone to error. Self-selected samples are especially prone to bias, since you tend to sample the respondents who have strong motivations to participate.

Systematic samples often sound appealing, but also can lead in unexpected ways to error. For example, suppose you were to conduct an “occupancy survey” of commercial properties in a particular neighborhood based on sampling, say, every tenth lot. It is possible that every tenth lot could turn out to be a corner lot, hence more desirable and less likely to be vacant. This would result in *systematic error* or *bias* in your sample.

Simple Random Samples.

Because samples necessarily contain error, the goal of the research is minimize that error. Random samples are a technique designed to reduce bias in the sample.

In a random sample, every member of the population has an equally likely chance of being a selected for the sample.

The random character of the error introduced means that it is not *systematic* and hence the resulting error (which must necessarily be present since a sample has incomplete information) will not be *systematic error*, i.e., the sample will not be *biased*.

Simple Random Samples. In a simple random sample, every member has an equally likely chance of being selected for the sample – much like a huge lottery. In order to construct a simple random sample, you must:

- Construct a list of all members of the population;
- Randomly select members from the list.

To do the random selection, think of rolling dice or flipping a coin for each member on the list to decide if they are in the sample or not. That way, whether or not any particular member of the population has an equally likely chance of “winning” the roll and being in the sample. Tools like Excel include random number generators that let you assign random

numbers to each member of the population to facilitate sampling.

Selecting a sample based on who happens to be available at the time of collection is not a random sample. It's a convenience sample and is susceptible to bias.


Notice that your sample is only as good as your listing of the population. If your list is biased, then so is your sample. Generating the list can be expensive, difficult, and sometimes even impossible.

Non-respondents. Some people who are selected to participate may choose not to do so. You need to have at least a 50% response rate in order to use your sample. You should strive for an 80% or higher response rate. The Nielson organization (the company that does television ratings) routinely gets response rates of over 95%

Stratified Random Samples.

Because a simple random sample is often difficult to construct, another strategy is called the *stratified random sample*. This technique tries to make the sample representative of the Population by identifying traits important to the study and assuring that the sample and the population have a similar distribution of those traits.

The technique is similar to blocking, except that blocking is systematically applied after the sample is selected; stratification is applied as part of the sample design prior to selecting the sample.



When to Stratify. Sometimes you can identify attributes that are important to the response you are studying.

If you are studying voter preferences, then, for example, political affiliation is an important attribute of a subject. You would want your sample to have about the same proportion of Democrats, Republicans and Independents as the population you were studying.

Random sampling might result, by chance, in more of one party or another in the sample. This would result in bias. Stratified sampling is a strategy to avoid this.

Steps in Stratified Random Sampling.

1. Divide the population into Strata which are homogeneous with respect to attributes important to your study. (Don't use shoe size as a stratum in a presidential preference poll!)
2. Do a random sample from each stratum.
3. Pool the strata together to obtain the overall sample.

13.1. Example.

Suppose you are doing an opinion poll in Oklahoma and that you want to use gender and ethnicity as strata. Since there are two genders and six ethnicities (White, Black, Asian, Hispanic, Indian and Other) this results in $2 \times 6 = 12$ strata. Suppose in addition that you want a sample of size 700. The problem is to find the "right" number for each strata to make the

sample representative of the population:

<i>Cell Distributions</i>		
	<i>Male</i>	<i>Female</i>
<i>White</i>		
<i>African-American</i>		
<i>Asian-American</i>		
<i>Hispanic</i>		
<i>Native American</i>		
<i>Other</i>		
<i>Total</i>	<i>350</i>	<i>350</i>

Here is the strata distribution in the Oklahoma Population:

<i>Cell Distributions</i>		
	<i>Male</i>	<i>Female</i>
<i>White</i>	<i>29%</i>	<i>29%</i>
<i>African-American</i>	<i>7%</i>	<i>7%</i>
<i>Asian-American</i>	<i>2%</i>	<i>2%</i>
<i>Hispanic</i>	<i>4%</i>	<i>4%</i>
<i>Native American</i>	<i>7%</i>	<i>7%</i>
<i>Other</i>	<i>1%</i>	<i>1%</i>

Solution.

Cell Computations Compute the cell sizes according to the population percentages (recall the sample was to be 700).



For example,

$$\begin{aligned}\# \text{ of white males} &= 29\% \times 700 \\ &= 203\end{aligned}$$

or

$$\begin{aligned}\# \text{ of Hispanic females} &= 4\% \times 700 \\ &= 28\end{aligned}$$

Repeating these computations for each cell gives sample sizes for each

stratum:

Cell Distributions		
	Male	Female
White	203	203
African-American	49	49
Asian-American	14	14
Hispanic	28	28
Native American	49	49
Other	7	7
<i>Total</i>	350	350

Sampling Fractions. This example assumes *proportionate sampling Fractions*. Unequal sampling fractions can also be done, although then the computations the sample statistics are slightly more complex. For example, means are then computed for each strata and a weighted av-

erage of the strata means – using the population proportions – is computed.

Unequal sampling fractions are sometimes more appropriate. For example, there are only 4 female admirals in the US Navy, so a stratified random sample would be unreasonable.

Other Considerations Suppose the Oklahoma researchers added three income levels to their study:

- low (say under \$25K/year);
- middle (say \$25001-\$75000 / year); and
- high (say higher than \$75K/year).

This results in $2 \times 6 \times 3 = 36$ strata, clearly a more complex sampling problem. To preserve confidentiality, each cell in the sample should have at least three subjects. All of this could result in large samples – often beyond a reasonable budget.

Summary.

In stratified random sampling the heterogeneity of the sample is obtained by combining internally homogeneous strata.

This approach is time consuming and expensive, leading a search for other alternatives

Cluster Sampling.

13.2. Example.

Suppose you are doing a dietary survey of fourth graders in Oklahoma. Some control variables might be

- *gender of the child;*
- *family income;*
- *residency (urban, suburban or rural);*

- *ethnicity.*

A stratified random sample using these criteria is probably not possible – even getting a list of all fourth graders in Oklahoma might not be possible.

If you think of the child as the "sampling unit," then this problem is difficult. There is another possible "sampling unit:"

- The **school** might be a sampling unit!

It would certainly be possible to obtain a list of all schools in Oklahoma. One could then randomly select from this list of "clusters," then do a census in each school.

With this technique, *the heterogeneity of the sample is obtained by combining internally heterogeneous clusters.*

Multistage Sampling.

This more modern approach combines stratified random sampling and cluster sampling techniques. For example, in our dietary survey we might:

Stratify the schools (urban, suburban, rural) before sampling students within in each school. When sampling the students within the schools, we could further stratify according to income, ethnicity and gender.

At the school level this latter stratification is more manageable. This is how modern polling organizations operate.

More Examples.

13.3. Example.

In 1936 the American Mercury Magazine polled over 10,000 household on the presidential election that year. They used a carefully selected random sample based on phone book listings. The poll predicted a landslide victory for Alf Landon.

Of course, Franklin Roosevelt won by the largest margin in US history up to that time.

What did they do wrong?

13.4. Example.

***The Hite Report.** Shere Hite mailed out over 100,000 surveys to a very carefully designed stratified random sample of US women. When she analyzed her nearly 9,000 replies she found stunning results. For example,*

- *70% reported extra-marital affairs;*
- *80% reported their marriages were a mistake.*

Other researchers were unable to repeat these results. Why?

Conclusions

- Sampling inevitably results in error.
- Sampling techniques are designed to minimize bias by making the sample as representative of the population as possible.
 - Sampling has two competing goals – minimizing costs and maximizing accuracy.
 - Poorly designed samples are the cause of many flawed research conclusions.

14. Survey Construction and Formatting

The step of gathering your data involves interacting with subjects. This section discusses **surveys** as a data-gathering technique, but many of the same observations apply other techniques. For example, if you **directly observe** the behavior of subjects instead of questioning them, you would create an **observational checklist** using similar principles.

The first step is to write down your basic research question. For example, your basic question might be

What influences viewer satisfaction with downloaded movies?

In this case, **viewer satisfaction** is the **dependent variable** and the **influences** are the **independent variables**. At a minimum, you need a question on your survey that **measures** satisfaction. One possibility might

be:

Which of the following best describes your most recent experience watching a downloaded movie? Check one box only.	
<input type="checkbox"/>	Extremely Satisfied
<input type="checkbox"/>	Satisfied
<input type="checkbox"/>	Neither satisfied nor dissatisfied
<input type="checkbox"/>	Dissatisfied
<input type="checkbox"/>	Extremely dissatisfied

There are some things to observe about this question. First, because the choices are listed vertically it's **clear which box to check**. Listing them in a row can be confusing.

Second, this list of answers **permits a neutral response**. If we used an **even number of responses** we would **force respondents to choose** between being satisfied and dissatisfied.

Since the basic question deals with possible influences on viewer satisfaction, we'll need to construct a list of what these might be. Examples include things like:

- respondent age;

- respondent educational level;
- expense;
- convenience;
- video/audio quality.

Each **influence** corresponds to a **independent variable**. For each independent variable, we need a question on the survey what will **measure it**. For example, for age we might ask

Which of the following best describes your age?	
<input type="checkbox"/>	At least 18 but younger than 28
<input type="checkbox"/>	At least 28 but younger thn 38
<input type="checkbox"/>	At least 38 but younger than 48
<input type="checkbox"/>	At least 49 but younger than 58
<input type="checkbox"/>	58 or older

Continuing in this way constructs your survey.

Generally speaking, every question on the survey corresponds to a variable, and every variable corresponds to a question. The questions **measure** the variable.

One exception to the above is a **filter question**. In order to be satisfied or dissatisfied with downloaded movies, the subjects must have **downloaded a movie**, i.e., our population consists of

The population in this example consists of people who have downloaded and watched a movie, for example in the last three months.

Finally, when we ask our questions, we won't ask the respondents to **summarize their behavior**. Instead, we will ask them about **the last time they downloaded a movie**.

Part of good survey design includes how researchers will interact with their subjects before and after administering the survey. Typically, the researcher will want to

- identify themselves to the subject;
- inform the subject why the researcher is doing the survey;
- inform the subject what the survey is about;
- inform the subject of how long the survey will take;
- inform the subject of safeguards regarding their confidentiality;
- ask them to participate.

After administering the survey, a thank you is always appropriate. Sometimes researchers will also include a token of appreciation. For example, families participating in Nielson surveys receive a token payment of \$15 per month.

The way that you interact with your subjects is your **protocol**.

14.1. Example.


A recent (2015) *poll* showed that 45% of Americans believe that aliens/extraterrestrials have visited the Earth. Suppose you are interested in studying what influences people to have this belief. For example, the poll showed that women are more likely than men to disbelieve in alien visitations. To study these influences, you would need to

- Define your *population*;
- define *dependent* a variable regarding belief in alien visits; and
- create a list of factors or *independent* variables that influence this belief.


Group Assignment. Identify four possible independent variables besides gender that might influence belief in alien visitations. Design a questionnaire that measures the dependent and independent variables. Include your *protocol*. Also be sure to identify the *population* you intend to study and a filter question if required.

Sometimes your research objectives will involve questions that are **too complex** for a self-administered survey. Sometimes the objectives will require **interactive questions** which necessitate the intervention of a trained interviewer. Sometimes the questions may be **personal in character**. Some personal questions are better administered in a personal interview in which the respondent and the researcher have an opportunity to develop a trusting relationship; in other cases the anonymity of a self-administered survey may elicit more reliable responses.

Your **budget** includes not only **financial** resources but also the **availability of trained interviewers** and the **available time** frame for responses. **Self administered questionnaires** require mailing lists, printing and postage costs. In addition you must wait for your respondents to return the forms and may incur the expense of follow-up mailings. **Telephone surveys** can have a quicker turn around but add the expense of trained interviewers and (except for local surveys) long distance tolls. **Personal interviews** are the most expensive in terms of all resources – financial, personnel and time.



Another issue is **sample design**. Self administered surveys generally require mailing lists. Telephone surveys require phone directories. Personal interviews will require either addresses or phone numbers or both to establish the interview appointment. In each case the sample must be selected from the available list and can only be as reliable as the original list.



Each of the survey methods provide potential sources of bias. A self administered questionnaire may be misinterpreted or may elicit biased responses due to question wording or order. Since your only contact with your respondents is the survey instrument, you may never discover these problems. Similar problems can arise with telephone and personal interviews, but because the interviewer can interact with the respondent the researcher has an opportunity to discover errors or sources of bias in the survey instrument. Of course, because the interviewer is interacting with the respondent there is potential that this interaction will bias the results; special care must be taken that the interviewer maintain a completely neutral demeanor and tone. Another problem with self administered and telephone surveys is that the respondent may not be the person you selected for your sample.

Finally the researcher needs to choose a method which has a reasonable response rate. In general you should strive for an 80% response rate to your survey. *A survey with a response rate of less than 50% is not valid for statistical inference* unless there is strong evidence that the non-respondents are similar to the respondents in all essential features. Since the self administered questionnaire can be filled out according to the respondent's schedule this will enhance the chance of reply (and of a thoughtful and accurate reply); of course, it also enhances the chance that the survey will be discarded or lost. Telephone surveys rely on being able to contact the respondent; this is exacerbated by the increasing use of answering machines and other call-screening devices. Personal interviews tend to have the highest response rate.

Response rates can be enhanced by contacting the respondent with a post card or letter to advise respondents that they have been selected for a survey, inform them of the purpose of the survey and let them know when the survey instrument (survey, phone call, interviewer) will arrive. Each question on your survey should be **purposeful** – do not ask question just because you think it might be interesting. In a survey on a presidential election, don't ask what kind of car the respondent drives. Unnecessary questions waste your time and are inconsiderate of the respondent – and probably will depress the response rate. A question does not have to address a research objective to be purposeful: you might ask a question whose only purpose is to motivate response. (“Do you think that the national news represents your interests?”)

Often surveys will gather demographic information (gender, ethnicity, age, marital status, occupation) on the first page of the survey. For this reason these are sometimes called *face sheet* questions. Such demographic information is often gathered just to validate the sample; often the researchers will also check the response variables against these

controls to ascertain if there is a difference in common demographic groups.

Your questions can assume several formats: short answer, open-ended, check lists, yes/no, Likert scales or some other form. Short answer and open-ended questions have the advantages of not suggesting responses (and hence less bias) however they also tend to be much more difficult to tabulate.

Be careful with your checklists: what is wrong with the following questions (taken from student surveys)?

What is your age?

0-25 25-35 35-45 45-55 55+

Of course, the choices are not mutually exclusive, so, for example, a respondent who is 35 years old could give either of two answers.

What is your age?

0-24 25-34 35-44 45-54 55+

This one has been fixed so that the responses are mutually exclusive, but it is difficult to read. It would be better if a list were used:

What is your age?

- 0-24
- 25-34
- 35-44
- 45-54
- 55+

Beware of unwarranted assumptions! What's wrong with the following?

Which of the following best describes your experience with the Armed Forces?

- On Active Duty
- Retired
- Husband on Active Duty
- None

Checklists need to be mutually exclusive and collectively exhaustive. Be sure to include an “other” response with room for the respondent to fill in their own answer on any checklist. When you construct your checklist, you are making a list of what you *think* are all possible answers. You can be undone if what you think you know is wrong.


Likert scales can either provide for a continuous response

Strongly Agree

Strongly Disagree

or for graduated response:

- (a) Strongly Agree
- (b) Agree
- (c) Neutral
- (d) Disagree
- (e) Strongly Disagree



The above graduated scale permits a neutral response; an even number of choices forces a choice:

- (a) Strongly Agree
- (b) Agree
- (c) Disagree
- (e) Strongly Disagree

Notice that the graduated Likert responses will be easier to tabulate than continuous Likert scales.

You should endeavor to save your respondents unnecessary reading. One way to do this is with *filter questions*. If your survey deals with CD purchases (for example) your first question might be:

Do you own a CD player? Yes No.
If your answer is “yes” please continue. If your answer is “no” please be sure to check the “no” circle and return the questionnaire. Your answer is very important to the accuracy of our research. Thank you very much for your assistance.

Internal Consistency. Surveys will often contain the same question worded in several different ways and placed in different positions in the questionnaire. This is done to check for consistency in responses (as well as to check for bias in the questions themselves). If you do this, there are statistical tests (such as Cronbach's alpha) to help you decide if the responses to a set of questions are consistent. Because this adds to the expense and length of the survey, many surveys omit such consistency checks in the final version of the survey instrument.

The formatting of the questionnaire – especially for self administered surveys – can be very important. A survey needs to both *appear* to be simple and short to fill out as well as actually *be* simple and short. Long and complex surveys will have poorer response rates; in addition, the accuracy of the response will degrade toward the end of the survey due to fatigue on the part of the respondents.

Your cover letter, survey title and instructions not only convey information about the survey but can serve to motivate the respondent to answer your questions. Part of your goal is to persuade your respondent to assume some ownership in your research project, investing time and consideration in the responses. While such motivational efforts are important, care should be taken to avoid biasing the responses: a cover letter from the National Committee of a political party would almost certainly bias the responses regardless of political persuasion.


In addition to question formatting, the paper size can influence the response rate, with Monarch size (7×10 inch) obtaining more responses than letterhead size ($8 \frac{1}{2} \times 11$ inches), which in turn does better than legal size ($8 \frac{1}{2} \times 14$ inches).

As a general rule, the fewer pages the higher your return will be. Of course, there will be a tradeoff between paper size and number of pages, but often a Monarch sized four page survey will draw as well as a two page survey printed on letterhead stock.

If you print your survey on colored stock it may be confused with an advertisement, depressing the response rate. If you need to use colored stock (for example, to help code different sample groups), use light colors. The paper weight should be sufficiently heavy to accept printing on both sides without showing through; avoid too heavy a stock since this will increase mailing costs.

You should select a consistent and legible typeface. *You should avoid fancy or peculiar typefaces since they are difficult to read. Even italics should be avoided – people are generally able to read italic type at only about 60% of the rate for non-italic type.*


The most legible typefaces tend to be sans serif (like your lecture notes) as opposed to fonts like the traditional Times Roman used in newspapers. Typefaces should be at least ten point size (twelve pitch on a typewriter) but twelve point (ten pitch on a typewriter) is generally



more legible. A line length of five inches is generally best to avoid eye strain and fatigue.


Your survey should not look cramped or cluttered. Liberal use of white space and vertical – as opposed to horizontal – lists can make your survey more attractive as well as clearer. The survey should look professional (but not too slick: you don't want to be confused with an advertisement, nor do you want to alienate the respondents by appearing too affluent). It should always be clear where to respond (and whether the response is a check box or written). In the case of written responses, the amount of room you leave will suggest the length of response which you desire.

The questions on the survey should be neutrally worded. Care should be taken not to suggest answers, either in the format, wording or positioning of the question. Surveys can be biased in subtle ways.



For example, if you are asking a question about education and expect that all of your respondents will have at least a high school education, include a choice “grammar school only” so that respondents will not feel they are falling in the lowest category.


Some questions may embarrass the respondent. Questions about sexual activity or preference are notorious for unreliable responses. Even if the respondent is certain that the answers will be confidential, untrue or missing responses are inevitable for sensitive questions unless the respondents are able to both establish a trusting relationship with the researcher (for example through personal interviews) and are convinced as to the importance of the project. The Kinsey studies, for example, were preceded by involving community leaders (religious leaders, newspapers, etc.) in giving public expressions of support for the project. Then the subjects were interviewed several times to assist in developing a trusting relationship with the researcher.



The order of questions can influence results by placing the respondent in particular frame of mind. The question “Do you support the use of public funds to pay for abortions” is, itself, phrased in a neutral fashion. However, if you precede this question with a sequence of questions about the respondent’s own children or about child care you will get one set of answers. Preceding the question with a sequence of questions about women’s rights and freedom of choice will produce a different set of responses.

The phrasing and structure of the questions, and the layout and character of the responses should be designed with problems of tabulating in mind.

As you construct your questions, keep in mind that you will eventually need to tabulate and summarize the results of your survey. You will also probably want to be able to statistically analyze the results in some manner. The final steps in your survey, the analysis and interpretation of results, can be a nightmare unless you have planned ahead.



Before you administer your survey, it's a good idea to make some trial graphs, tables and charts to describe how you will summarize your data. Also think about the numerical summaries and cross-tabulations which you will want to make (how did women answer questions 12-15 differently from men). Finally, think about the statistical tests which you will need to perform in order to realize your research objectives and answer your hypotheses. Make sure that you can readily capture the required data from your questionnaire.

Sometimes you will want to consolidate several questions into a single group in order to get an overall response. For example, if you are interested in attitudes about women in combat, you may have a set of questions dealing with women serving in various specific roles (fighter pilot, air evac pilot, tank commander, CIC officer, etc.) which can be consolidated in a single scaled response. If you do this, make sure that the scales from question to question are consistent!

Finally, you will need to pay attention to how the responses are *coded*. Using a computer will facilitate doing summaries, but computers only understand numbers, not “attribute” responses. Thus when measuring “gender” for example, you might want to code one gender as “0” and one gender as “1”:

- “males” = 0; and
- “females” = 1.

On the other hand, if your response includes more than two categories, you should regard *each category* as a separate “yes/no” responses. For example, if your survey has two questions:

1. What is your gender? ○ male ○ female
2. Which best describes your highest educational level?
 - (a) do not have a high school diploma
 - (b) high school diploma or GED
 - (c) two years of college or less
 - (d) four years of college or less
 - (e) more than four years of college

In this case you might have the following five respondents:

#	Gender	Education
1	Female	(b)
1	Male	(a)
1	Female	(e)
1	Female	(c)
1	Male	(d)

In this case you could code “gender” with a single column of “0’s” and “1’s.” But “education” becomes *five* columns, with each of the choices coded “0” for “no” and “1” for “yes.” Any particular respondent will have exactly one “1” response for the education choices.

respondent #	Gender	5Education				
		(a)	(b)	(c)	(d)	(e)
1	1	0	1	0	0	0
2	0	1	0	0	0	0
3	1	0	0	0	0	1
4	1	0	0	1	0	0
5	0	0	0	0	1	0

Questions. Why is only one column, not two, needed for “gender?” Do you really need to use five columns for Education?

Later lectures will deal with the **reliability** and **validity** of surveys, including **inter-rater reliability** and the **internal consistency** of survey items.

15. Confidence Intervals for Means

Idea: The sample mean \bar{x} estimates the true population mean μ . A confidence interval assigns an “error bound” so that the true mean is probably between

$$(\bar{x} - \text{ERROR}) \text{ and } (\bar{x} + \text{ERROR})$$

Confidence intervals give a way of selecting the “error” so that a fixed percentage (say 95%) of all possible intervals

$$(\bar{x} - \text{ERROR}) \text{ to } (\bar{x} + \text{ERROR})$$

contain the true population mean μ .

The interval

$$\bar{x} \pm \text{ERROR}$$

is called a *confidence interval*. Since this interval is based on sample data, we can't be *certain* that the population mean is inside this interval – we can only be “confident.”

If you know the sample size, the sample mean, and the sample standard deviation, the Excel spreadsheet `Formulas.XLSX` in the resources section of `LEARN.OU.EDU` for this course will calculate the value of the error term for you according to the formula:

$$\text{ERROR} = \pm \frac{zs}{\sqrt{n}}$$

where z is a cut-off found from the normal tables.

© Cartoonbank.com



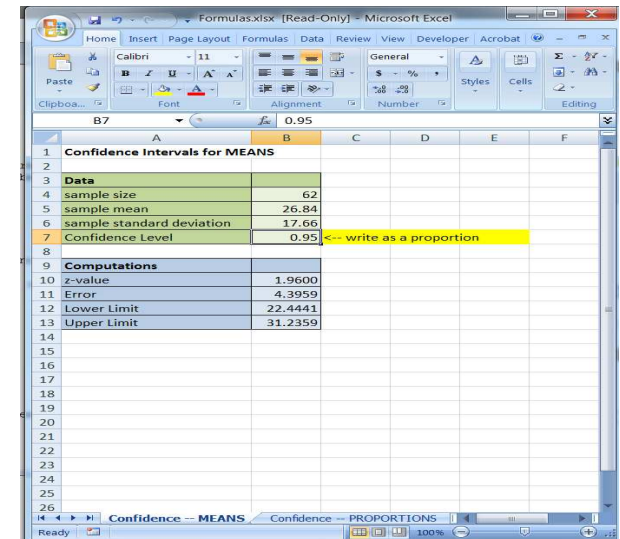
15.1. Example.

Suppose that 62 night shift workers are randomly selected and surveyed to find the weekly hours spent on child care. In this sample, it is found that the mean is 26.84 hours and the standard deviation is 17.66 hours. Construct a 95% confidence interval for the average weekly time spent in child care by night shift workers.

Solution. **Step 1.** First make a list of all variables in the problem:

\bar{x}	26.84
s	17.66
n	62

Step 2. Now enter the list into the spreadsheet FORMULAS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled Confidence-MEANS. Note also that you should convert the 95% to a proportion, 0.95.



Once you have entered the dictionary, the spreadsheet calculates the confidence interval for you:

22.41 to 31.24.

We are “95%” confident that the true average hours spent on child care by night shift workers is somewhere between 22.41 hours and 31.24 hours.

Questions. Suppose that you know that day shift workers spend, on average, 29.71 hours in child care. (Since we “know” this, 29.71 must be the true mean μ for day shift workers.) Is the above experiment evidence that night shift workers spend less time on child care than do day shift workers? Suppose that day shift workers spent 35 hours per week? What are some uncontrolled variables in this experiment?

Solution Template

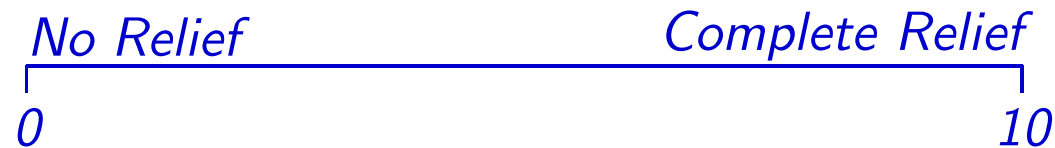
Step 1. Make a dictionary of the data given in the problem:

sample mean	\bar{x}
sample standard deviation	s
sample size	n
confidence level	C

Step 2. Use the Excel spreadsheet to calculate the confidence interval.

15.2. Example.

In a study on using hypnosis to relieve pain, 58 subjects were asked to fill out the following *Likert Scale*:



(The total length of the scale was 10cm.) For this sample, the mean was 6.2 and the standard deviation was 4.18. Find a 98% confidence interval for the perceived pain relief.

Solution. **Step 1.** First make a list of all variables in the problem:

\bar{x}	6.2
s	4.18
n	58


Step 2. Now enter the list into the spreadsheet FORMULAS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled Confidence-MEANS. Note also that you should convert the 98% to a proportion, 0.98.

The resulting 98% confidence interval is

4.92 to 7.48

	A	B
1	Confidence Intervals for MEANS	
2		
3	Data	
4	sample size	58
5	sample mean	6.2
6	sample standard deviation	4.18
7	Confidence Level	0.98 <small>write as a proportion</small>
8		
9	Computations	
10	z-value	2.3263
11	Error	1.2768
12	Lower Limit	4.9232
13	Upper Limit	7.4768
14		
15		
16		
17		
18		
19		
20		
21		
22		
23		
24		
25		
26		

Question. suppose we know that people with similar pain who take aspirin have mean pain relief of 7.42, as measured by this scale. Do we have convincing evidence that aspirin does better than hypnosis?



Suppose ibuprofen results in a score of 7.81; would we be able to conclude ibuprofen is better than hypnosis? Do we have convincing evidence that ibuprofen is better than aspirin? What kind of data would we have to gather to compare ibuprofen and aspirin?

16. Confidence Intervals for Proportions

Confidence intervals for proportions have the same basic underlying concepts as confidence intervals for means. The only difference is in the calculation that Excel uses for the error term:

$$\text{ERROR} = z \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

where n is the sample size and \hat{p} is the sample proportion:

$$\hat{p} = \frac{\text{number of successes}}{\text{sample size}}$$

Remember: “success” is the outcome which is the focus of the research and not necessarily any conventional notion of success or failure.

16.1. Example.

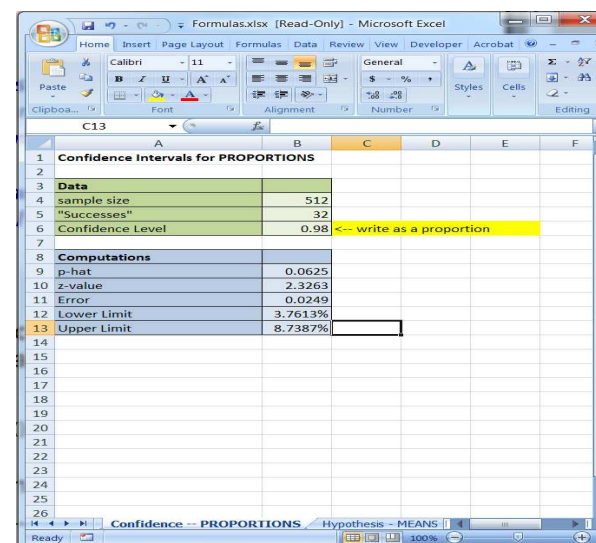
Currently all lab samples from a physician's office are sent to *Tests R Us*, a commercial lab specializing in analyzing and producing pathology reports. The physician suspects that *Tests R Us* may be cutting corners, and decides to double check their results against the state laboratory which has essentially a 100% accuracy rate. Of 512 samples, a *Tests R Us* incorrectly identifies 32. Find a 98% confidence interval for the proportion of incorrectly identified samples.

Solution.

Step 1. First find the sample proportion; since problem is focusing on incorrect identifications, “success” is an *incorrectly* identified tissue sample.

sample size n	512
"successes" k	32
confidence level	98%

Step 2. Now enter the list into the spreadsheet FORMULAS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled Confidence-PROPORTIONS. Note also that you should convert the 98% to a proportion, 0.98.



In particular, the confidence interval for the proportion is 3.75% to 8.47%.

Question. The physician could send all samples to different commercial

lab, *Test Depot*. A recent analysis of their results indicated that *Test Depot* will incorrectly identify samples 4% of the time. However, *Test Depot* will charge nearly double per test compared with *Tests R Us*. Do we have sufficient evidence to justify spending more on *Test Depot*?

Remark. There are two kinds of errors which the lab could make:

- They could report no disease when the patient is really ill.
- They could report disease when the patient is healthy.

Question. Which kind of error do you think is more serious? Is it important to know which kind of error the lab is making?

→ In our next topic – hypothesis testing – there will always be two kinds of error which are possible. In general a researcher will be able to control the chances of only one of the two types of error. Part of the research design is to decide which kind of error is more important to control.

Solution Template

Step 1. Make a dictionary of the data given in the problem:

“successes”	k
sample size	n
confidence level	C

The formulae used by the spreadsheet require that both $n\hat{p}$ and $n(1-\hat{p})$ be at least five. In general, you should check this before proceeding. *However, these requirements will be satisfied in all problems encountered in this class.*

Step 2. Enter the data in Formulas.XLSX and read the results.


End of Solution Template

Given a desired confidence level, it is possible to control the error term in confidence intervals for proportions (but not in confidence intervals for

means) by controlling the sample size. In general, increasing the sample size will decrease the magnitude of the error term (thereby decreasing the length of the confidence interval). The following table shows the sample sizes required for various confidence levels and error terms.

	94%	95%	96%	99%
$\pm 4\%$	552	600	657	1032
$\pm 3\%$	982	1067	1167	1835
$\pm 2.5\%$	1415	1537	2086	2653
$\pm 2\%$	2209	2401	2627	4128
$\pm 1\%$	8836	9604	10506	16512

The first row shows the confidence level. The first column shows the magnitude of the error term. The interior of the table shows the required sample size. Note that halving the error term requires quadrupling the sample size. Small error terms are often too expensive to justify the cost of the requisite extremely large samples. Opinion polls often always use



confidence limits of $\pm 2.5\%$ with 95% confidence, giving a sample size of 1537. The *Newsweek* poll on the health care proposals use a 98.5% confidence interval, with error of $\pm 4\%$.

The researcher is free to decide on a confidence level. Most frequently these levels will be 90% or higher. If the researcher chooses, for example, an 80% confidence level, then the chances of an error – that the confidence interval does not include the true value – are 20%. For most applications, this is an unacceptably high chance of error. If the researcher chooses a confidence level less than 90%, the researcher needs to justify why.

16.2. Example.

An attorney suspects that the prescription drug Seldane may have harmful side effects. In order to accumulate evidence on this suspicion in support of a possible class action lawsuit, the attorney surveys 75 Seldane users. The survey instrument informs the respondents of the purpose of the survey and then asks if they have experienced harmful side effects. Of the 75 individuals surveyed, 58 respond affirmatively. Find an 80% confidence interval for the proportion saying they suffer harmful side effects. If you were on a jury would you find this evidence convincing? Why or why not?

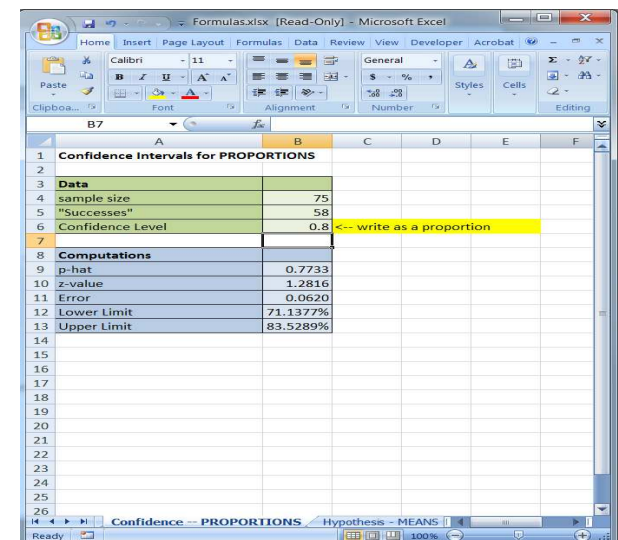
Solution.

Step 1. First find the sample proportion; since problem is focusing on harmful side effects, “success” is answering “yes” to the question about

side effects. Thus

sample size n	75
"successes" k	58
confidence level	80%

Step 2. Now enter the list into the spreadsheet FORMULAS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled Confidence-PROPORTIONS. Note also that you should convert the 98% to a proportion, 0.80.

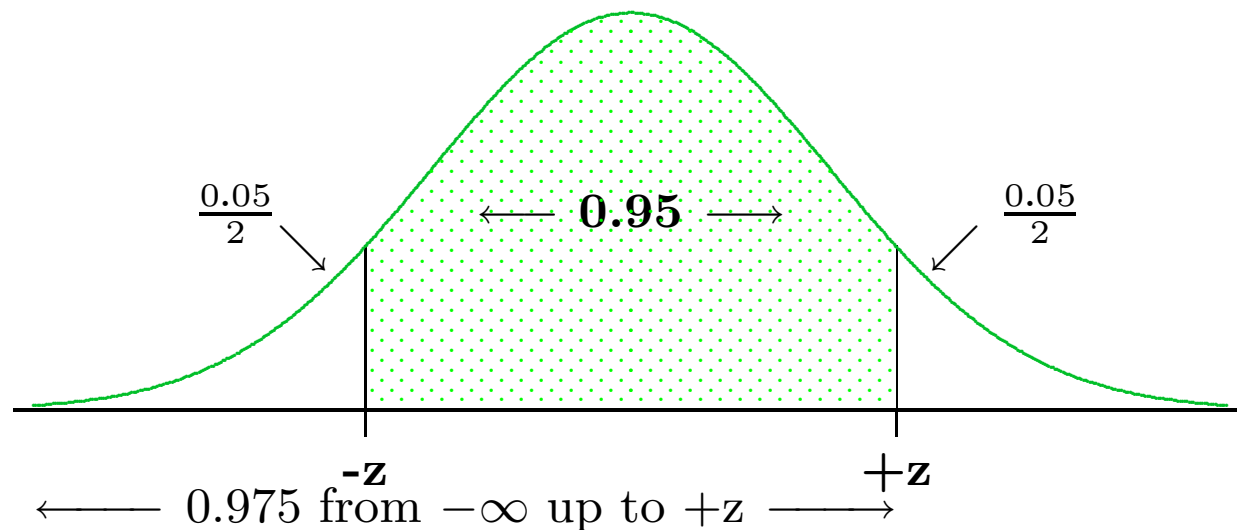


Thus, the confidence interval is:


71.3% to 83.53%

What if you don't have an excel spreadsheet? You can still work the problem, but you will need to use the "inside-out" techniques you learned for normal tables, and the formulae mentioned earlier for the error terms.

Suppose, for example, that you did not have access to Appendix B and needed to find the cut-offs for a 95% confidence interval. You would need to first find z so that 95% of the area under a normal curve falls between $-z$ and $+z$:



If there is 0.95 of the area between $-z$ and $+z$, this leaves 0.05 for the



two tails collectively or 0.025 for each tail individually. Then up to $+z$ there is 0.975 ($=0.95+0.025$) of the area. We need to compute this last proportion since the normal table is designed to give us areas *up to* a specific z -value.

From the normal table, the z -value which corresponds to 0.975 is $z = 1.96$.

You will always be able to use the Excel spreadsheet on your examination problems.

17. Hypothesis Tests for Means

17.1. Example.

A health care worker conjectures that vitamin supplements given to expectant mothers will increase the birth weights of the resulting newborns. To test this, she randomly selects 100 expectant mothers and then randomly assigns them to one of two, equally sized groups. Expectant mothers in Group A receive supplements starting in the first trimester of their pregnancy, while expectant mothers in Group B receive a sugar pill. The researcher records the birth weights of the children for all 100 mothers.

Results. *In Group A, there were 51 children for whom the average birth weight was 3.55kg with a standard deviation of 0.62kg.*

In Group B, there were 50 children with an average birth weight of 3.39kg and a standard deviation of 0.22.

Research Question. *Is this significant evidence of an increase in birth weights?*

Note that there are two possible conclusions:

The supplements had the desired outcome of increasing birth weights,

OR

Small sample sizes and random differences in subjects created the appearance of a difference in outcomes where none really exists.

In hypothesis testing, the word “significant” is a technical term with a particular meaning that relates to these two contrary conclusions. We will return to this problem shortly and choose one of the above conclusions and reject the other.

17.2. General situation for hypothesis testing.

You will usually be testing to see if an experiment has generated sufficient evidence to support a conjecture. Your conjecture is called the

Alternative Hypothesis

Typically we will be testing to see if the mean in the experimental group differs from the mean in a control group. Sometimes the research will have access to retrospective historical data and then general population might serve as the control group.

- Control Population
- Mean μ_C not known, but is estimated by the sample mean of the control group \bar{x}_C .
- Not subject to the experimental treatment.
- “Null” since nothing is done to the group.

- Experimental Population
- Mean μ_E *not* known but is estimated by the sample mean \bar{x}_E of the experimental group.
- Subject to the experimental treatment.

The untreated control population is generally the status quo. In principle, we could do a census and find the true mean μ_C for the control population. The treated experimental population is only an imagined future population since the experimental treatment has not yet actually been applied to the entire population.

If we actually knew μ_E (the true mean of the experimental population) there would be no need for a statistical test. Instead, all we know is an estimate of the true mean which is based on our sample. Since there are inherent uncertainties in sampling, we will need a way to decide if our results are due to the random error implicit in sampling or actually represent a consequence of the treatment.

Based on the sample mean, we want to decide whether

$$\mu_E = \mu_C \quad \text{or} \quad \begin{cases} \mu_E \neq \mu_C \\ \mu_E < \mu_C \\ \mu_E > \mu_C \end{cases}$$

If we decide $\mu_E = \mu_C$ then the experimental treatment made no difference. This is called the *Null Hypothesis*.

$$H_0 : \mu_E = \mu_C$$

Recall that μ_C is the mean of the control (untreated) population and μ_E is the mean of the experimental (treated) population. Instead of knowing the actual, census value for these two means, you will know estimates of \bar{x}_C and \bar{x}_E .

Depending on the character of the experimenter's conjecture, the *Alternative Hypothesis* could be any one of the following:

$$H_A : \mu_E > \mu_C$$

$$H_A : \mu_E < \mu_C$$

$$H_A : \mu_E \neq \mu_C$$

The experimenter's conjecture will dictate which one of the above will be the alternative hypothesis. Only one of the above can be the alternative hypothesis.

Hypothesis tests are designed to give a systematic way to choose between the null and the alternative hypotheses. This is accomplished by controlling the likelihood of making an error. There are two kinds of error which are possible.

	H_0 true	H_0 false
Accept H_0	OK	Type II Error
Reject H_0	Type I Error	OK

Hypothesis tests are very conservative. You will change from the status quo (reject H_0) only if the evidence is overwhelming. In other words

You want to make sure you do *not* reject H_0 when H_0 is really true.

or, in terms of Type I and II error:

You want the chances of Type I Error to be as small as possible.

The *significance level* is the maximum chance of Type I error that the researcher is willing to accept. The statistical tests will produce an “observed” significance level, also called a *p*-value. If the *p*-value is less than the pre-set significance level, then the researcher rejects the null hypothesis. Otherwise, the researcher accepts the null hypothesis. The researcher is technically free to decide what constitutes an unacceptably high chance of Type I Error, i.e., an unacceptably high *p*-value. However, there are some accepted standards. The researcher who deviates from those standards needs to be prepared to justify why the deviation is appropriate. The usual nomenclature is:

$p - \text{value} \leq 1\%$	Highly Significant results
$1\% < p - \text{value} \leq 5\%$	Significant results
$5\% < p - \text{value}$	Results are not significant

With this background, we can now return to the original example which started this section.

17.3. Example.

A health care worker conjectures that vitamin supplements given to expectant mothers will increase the birth weights of the resulting newborns. To test this, she randomly selects 100 expectant mothers and then randomly assigns them to one of two, equally sized groups. Expectant mothers in Group A receive supplements starting in the first trimester of their pregnancy, while expectant mothers in Group B receive a sugar pill. The researcher records the birth weights of the children for all 100 mothers.

Results. *In Group A, there were 51 children for whom the average birth weight was 3.55kg with a standard deviation of 0.62kg.*

In Group B, there were 50 children with an average birth weight of 3.39kg and a standard deviation of 0.22.

Research Question. *Does this provide evidence of a difference in birth weights between the treated and untreated groups, using a significance level of 4%?*

Solution.

Step 1. First make a list of all the relevant variables.

	Experimental	Control
sample mean	3.55	3.39
standard deviation	0.62	0.22
sample size n	51	50

Remember that μ_C is the mean of the control population and μ_E is the mean of the experimental population.

Step 2. Next write down the null and alternative hypotheses:

$$H_0 : \mu_E = \mu_C$$

$$H_A : \mu_E > \mu_C$$

We use “ $\mu_E > \mu_C$ ” since we conjecture that the treatment increases birth weights, and hence improves the health of the newborns.

Step 3. Now enter the list into the spreadsheet FORMULAS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled hypothesis-MEANS x 2. You should use Sample 1 to record the experimental data and Sample 2 to record the control data.

	Sample 1	Sample 2
Mean	3.55	3.39
Standard Deviation	0.62	0.32
Sample size	51	50
test statistic	1.634248704	
p-value LEFT TAILED	94.8897%	
p-value RIGHT TAILED	5.1103%	

Step 4. The spreadsheet calculates two p -values: a “left-tailed” and a “right-tailed” value. If you glance back at the alternative hypothesis

$$H_A : \mu_E > \mu_C$$

you can see that the inequality is like an arrowhead that points to

the right. This indicates that you should use the right-tailed p -value—assuming that Sample 1 is the experimental data and Sample 2 is the control data.

Step 4. Thus, the p -value for this data is 5.11%. Since this exceeds the pre-set significance level of 4%, we are unable to reject the null hypothesis and must, therefore, accept the null hypothesis. Thus, at a significance level of 4%, we believe the null hypothesis. ■

Question. Suppose that the significance level had been 5%? In this case, we would have **rejected** the null hypothesis, since the p -value is *less than* 5%. In this case, we would believe the *alternative hypothesis*, that the supplements do increase the birth weights. The p -value, 4.14%, is the chance that belief is wrong.

Solution Template

Step 1. Make a dictionary assigning values to each of the variables:

	Experimental	Control
sample mean	\bar{x}_E	\bar{x}_C
standard deviation	s_E	s_C
sample size	n_E	n_C
significance level	α	

Step 2. Write down the null and alternative hypotheses. The null hypothesis will always be:

$$H_0 : \mu_E = \mu_C$$

while the alternative hypothesis will be one of the following:

$$H_A : \mu_E < \mu_C \quad (\text{a left tailed test})$$

$$H_A : \mu_E > \mu_C \quad (\text{a right tailed test})$$

$$H_A : \mu_E \neq \mu_C \quad (\text{a two tailed test})$$

Step 3. Now enter the list into the spreadsheet FORMULAS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled hypothesis-MEANS x 2. You should use Sample 1 to record the experimental data and Sample 2 to record the control data.

The screenshot shows a Microsoft Excel spreadsheet titled 'Formulas.xlsx - Microsoft Excel'. The active sheet is 'hypothesis - MEANS x 2'. The spreadsheet contains the following data:

Hypothesis tests - MEANS, TWO Samples		
Data		
	Sample 1	Sample 2
Mean	3.55	3.39
Standard Deviation	0.62	0.32
Sample size	51	50
Calculations		
test statistic	1.634248704	
p-value LEFT TAILED	94.8897%	
p-value RIGHT TAILED	5.1103%	

Below the calculations, a yellow highlighted cell contains the text: "You must decide whether to use the left-tailed p-value or the right-tailed p-value".

Step 4. Use the form of the alternative hypothesis to select the appropriate p -value from the spreadsheet. If the pre-set significance level is *larger* than the observed p -value, then you can *reject* the null hypothesis, and the p -value represents the probability of a Type I Error. Otherwise, you *accept* the null hypothesis.

End of Solution Template

The spreadsheet calculates the value of a *test statistic*


$$\frac{\bar{x}_E - \bar{x}_C}{\sqrt{\frac{s_E^2}{n_E} + \frac{s_C^2}{n_C}}} \quad (1)$$

which is, in turn, used to calculate the *p*-value.

If it is reasonable to assume that the experimental and control populations have the same standard deviation, then the appropriate test statistic would use the pooled sample standard deviation s_{pooled} :

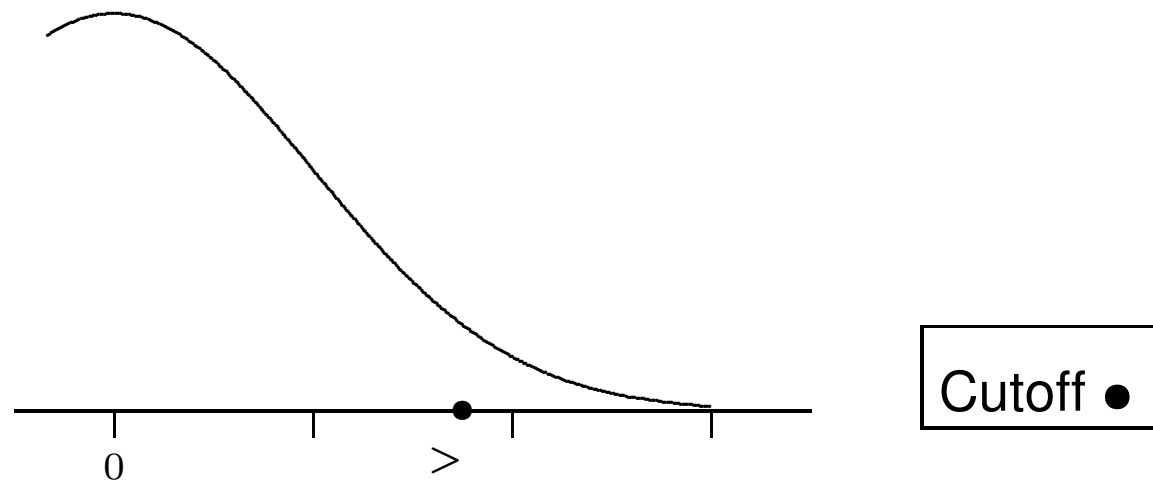
$$\frac{\bar{x}_E - \bar{x}_C}{s_{pooled} \sqrt{\frac{1}{n_E} + \frac{1}{n_C}}}$$

The spreadsheets for this course make the more general assumption that the experimental and control populations have different standard deviations and so formula (1) above.



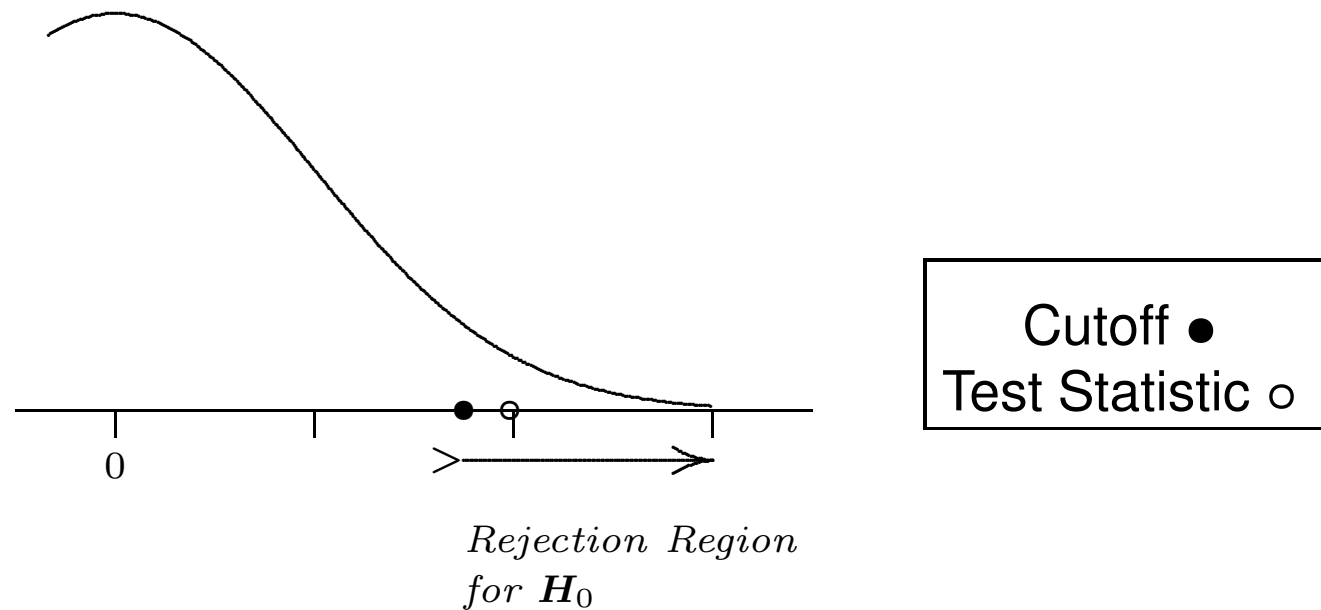
This test statistic is generally assumed to be normally distributed if the sample size is larger than 30; otherwise it has something called a Student's t -distribution (not covered in this course). The value of the test statistic determines the decision that you make. The rationale for the terminology “left-tailed” and “right-tailed” tests derives from the associated graphs of the distribution of the test statistic.

Suppose, for example, the pre-set significance level is 5%. One could do an inside-out problem to find the z-value associated with 5% of the area in the right-hand tail. If you then plotted that value on a number line, you'd get something like the following:



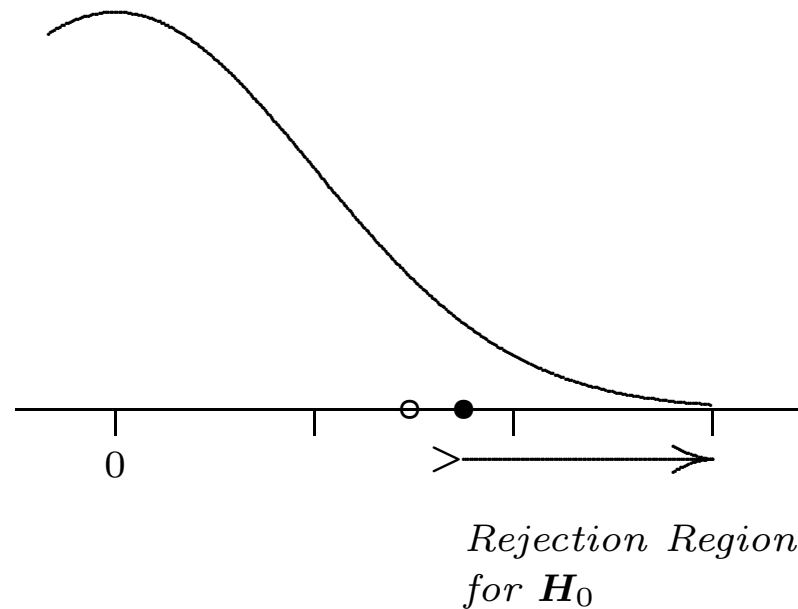
Notice that the area to the right of the solid dot ● is then 5%, or whatever the pre-set significance level happened to be.

Next, plot the calculated test statistic as a hollow dot (\circ). You might get the following



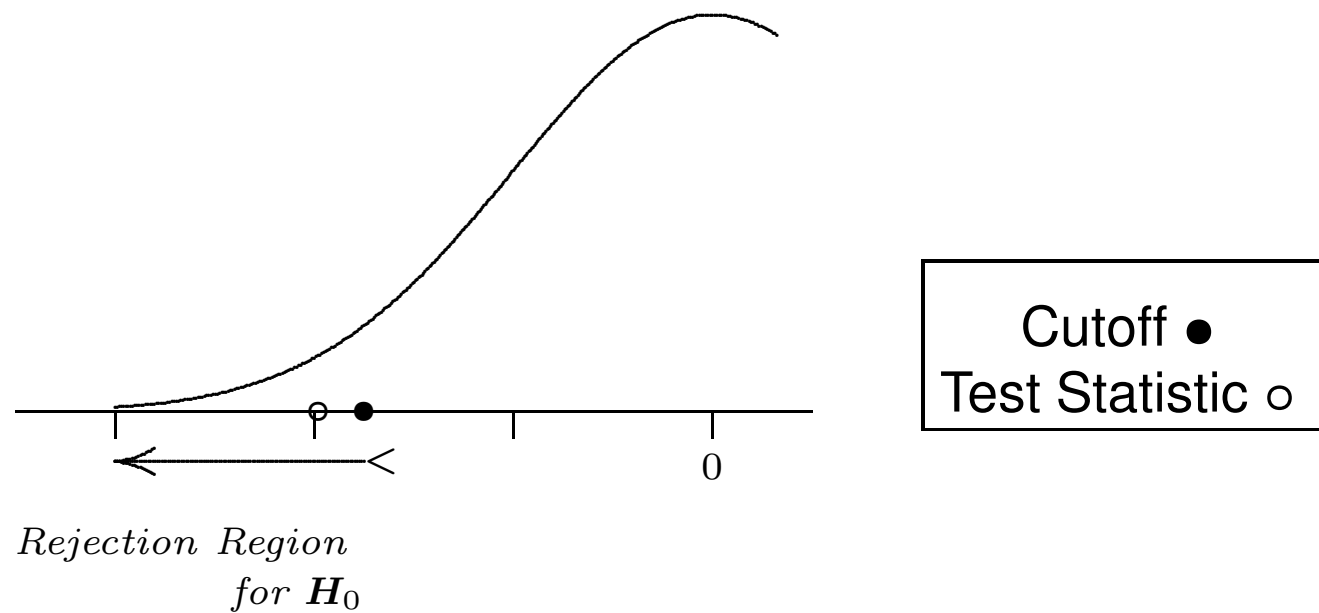
In this case, the test statistic is to the right of the calculated z-value for the pre-set significance level. The area to the right of the test statistic is *less* than the area to the right of the solid dot. Hence, we'd *accept* the null hypothesis.

If, in contrast, you got the following graph

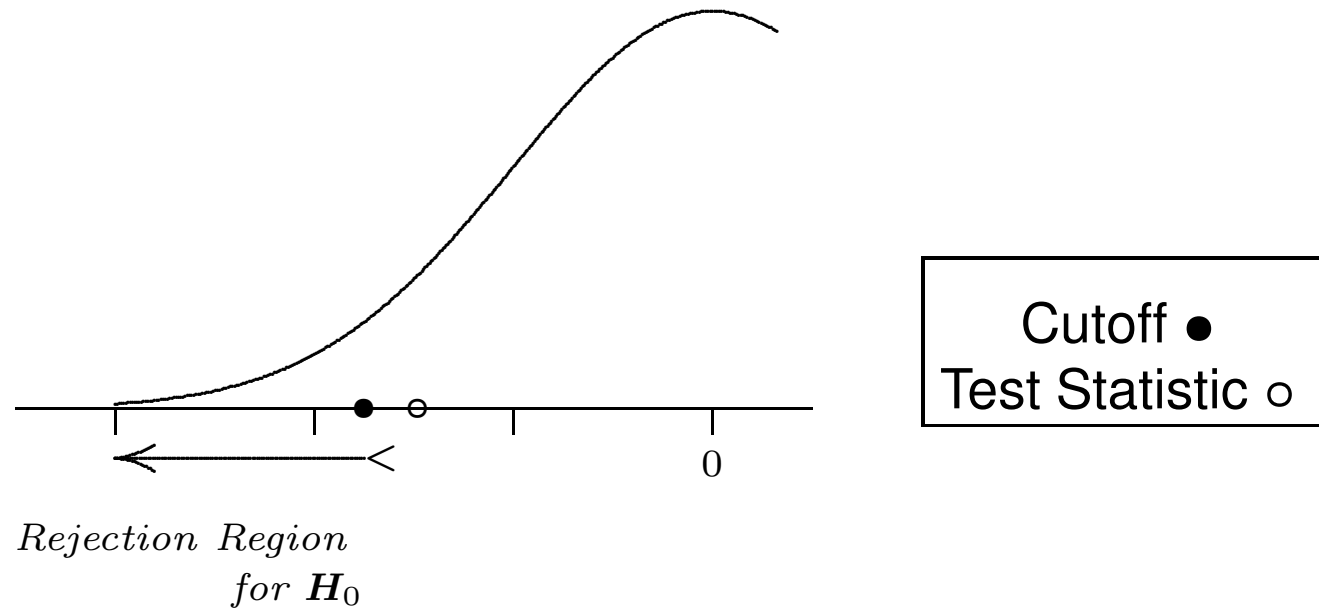


then you'd accept the null hypothesis.

A similar picture for left-tailed tests might be



in which case we'd reject the null hypothesis, or



in which case we'd accept the null hypothesis.

The point of the graphs is to illustrate how the spreadsheet calculations interact with the normal tables.

17.4. Example.

Surgery, planned or unplanned, subjects patients to great psychological stress. A counselor in a surgery ward suspects that patients who receive pre-operative briefings explaining to them exactly what to expect following surgery will experience less stress. To test this hypothesis, the researcher selects a random sample of 80 patients and randomly divides them into two treatment groups, each of size 40. In Group A, the patients receive a pre-operative briefing, while Group B patients receive no briefing.

Following surgery, the researcher administered the “State Anxiety Inventory” (SAI) was given to each patient, although five patients in Group A and two in Group B dropped out of the study. SAI scores range from 20 to 80 with lower scores indicating less anxiety. In Group A, the average anxiety score was 60.5 with a standard deviation of 5.6. In Group B, , the average anxiety score was 63.5, with a standard deviation of 8.3. Is this significant evidence (at the $\alpha = 4\%$ level) that pre-operative briefings reduce patient anxiety following surgery?

Solution.

Step 1. First make a list of all the relevant variables.

	Experimental	Control
sample mean	60.5	63.5
standard deviation	5.6	8.3
sample size	35	38
significance level	4%	

Step 2. Next write down the null and alternative hypotheses:

$$H_0 : \mu_E = \mu_C$$

$$H_A : \mu_E < \mu_C$$

We use “<” since we conjecture that anxiety levels will be *less* in the experimental, treated group.

Step 3. Now enter the list into the spreadsheet FORMULAS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled hypothesis-MEANS x 2. You should use Sample 1 to record the experimental data and Sample 2 to record the control data.

	Sample 1	Sample 2
Mean	3.55	3.39
Standard Deviation	0.62	0.32
Sample size	51	50
test statistic	1.634248704	
p-value LEFT TAILED	94.8897%	
p-value RIGHT TAILED	5.1103%	

You must decide whether to use the left-tailed p-value or the right-tailed p-value

Step 4. Since this is a left-tailed test (we have a less-than sign in the alternative hypothesis), we can use the p -value from the spreadsheet is 3.42. Since this is *less than* the pre-set significance level of 4%, we reject the null hypothesis and believe the alternative, i.e., we believe that the briefings reduce anxiety. The chance of Type I Error is 3.42%, the



p-value.



18. Hypothesis Tests for Proportions

Conceptually, these are exactly the same as hypothesis test for means. The only differences are that we find sample proportions \hat{p}_E , \hat{p}_C , and \hat{p}_T , the total proportion of successes in the pooled sample. The spreadsheet then calculates the test statistic:

$$\text{test statistic} = \frac{\hat{p}_E - \hat{p}_C}{\sqrt{\hat{p}_T(1 - \hat{p}_T) \left(\frac{1}{n_C} + \frac{1}{n_E} \right)}}$$

where

$$\hat{p}_T = \frac{\text{total successes in pooled sample}}{\text{total number in pooled sample}}$$

and we test

$$H_0 : p_E = p_C \quad \text{against} \quad H_A : \begin{cases} p_E > p_C & \text{or} \\ p_E < p_C & \text{or} \\ p_E \neq p_C \end{cases}$$

18.1. Example.

A researcher surveyed 1,600 randomly chosen females, aged 40-60. In order to participate, the subjects must either be currently married or divorced. The researcher gathered data about whether or not the subjects had cohabited prior to their first marriage and whether or not that marriage ended in divorce. There were 732 who cohabited prior to marriage, and 345 of this group were divorced. There were 868 who did not cohabit prior to marriage, and 348 of this group were divorced.

Using a significance level of 5%, does this provide significant evidence that cohabitation prior to marriage is associated with a higher divorce rate?

Solution.

Step 1. First make a dictionary of the information given in the problem; this problem focuses on divorce, so “divorce” constitutes “success.”

	Cohabited	Did not Cohabit
n	732	868
Divorced	345	348

Step 2. The treatment in this case is cohabitation, and the experimental outcome is divorce. Since the problem asks whether cohabitation increases the chances of divorce, our hypotheses are:

$$H_0 : p_E = p_C$$

$$H_A : p_E > p_C$$

Step 3. Now enter the list into the spreadsheet FORMULAS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled hypothesis-PROPORTIONS x 2. You should use Sample 1 to record the experimental data and Sample 2 to record the control data.

	sample 1	Sample 2
sample size	732	868
"successes"	345	11
Calculations	sample 1	sample 2
sample proportion	0.471311	0.012673
Pooled Proportion	0.2225	
test statistic	21.97409	
p-value LEFT TAILED	100.00%	
p-value RIGHT-TAILED	0.00%	
You must decide whether to use the left-tailed p-value or the right-tailed p-value		

Step 4. From the alternative hypothesis, this is a right-tailed test. Since the p -value is less than the target significance value, we *reject the null hypothesis* and *accept the alternative*. This means that we believe that cohabitation is associated with higher divorce rates. The

probability that we have made a Type I Error is the p -value, 0.23%.

Solution Template

Step 1. Make a dictionary assigning values to each of the variables:

	Experimental	Control
sample size	n_E	n_C
successes	k_E	k_C
significance level	α	

In order to use the spreadsheet, we must have both

$$np_0 \geq 5$$

and

$$n(1 - p_0) \geq 5$$

This requirement will always be fulfilled in problems and examination questions in this class. There are other techniques one can use (χ -squared tests) when this requirement is violated.

Step 2. Write down the null and alternative hypotheses. The null hypothesis will always be:

$$H_0 : p_E = p_C$$

while the alternative hypothesis will be one of the following:

$$H_A : p_E < p_C \quad (\text{a left tailed test})$$

$$H_A : p_E > p_C \quad (\text{a right tailed test})$$

$$H_A : p_E \neq p_C \quad (\text{a two tailed test})$$

(The reason for the terms right, left and two tailed tests is the same as in hypothesis testing for means.)

Step 3. Now enter the list into the spreadsheet FORMULAS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled hypothesis-PROPORTIONS x 2. You should use Sample 1 to record the experimental data and Sample 2 to record the control data.

	sample 1	Sample 2
sample size	732	868
"successes"	345	11
Calculations	sample 1	sample 2
sample proportion	0.471311	0.012673
Pooled Proportion	0.2225	
test statistic	21.97409	
p-value LEFT TAILED	100.00%	
p-value RIGHT-TAILED	0.00%	
You must decide whether to use the left-tailed p-value or the right-tailed p-value		

Step 4. Select the appropriate p -value from the spreadsheet using the direction of the inequality in the alternative hypothesis. If the p -value is less than the pre-set significance level, then you *reject the null hypothesis* and *accept the alternative*. Otherwise, you accept the alternative hypothesis.

End of Solution Template

18.2. Example.

A researcher gathers data on 1988 students in a large, urban high school. In this school, 123 students have a history of incarceration in a temporary detention center, while 1865 have no such history. Among those who have been incarcerated, the researcher determines that 22 have a diagnosis of a personality disorder, while 72 of the non-incarcerated group have a similar diagnosis.

Is this significant evidence that detained youth are at greater risk for a personality disorder than students who do not have a history of incarceration?

Solution.

Step 1. First make a list of all the relevant variables.

	Incarcerated	non-Incarcerated
sample size	123	1865
number of “successes”	10	72
α	5%	

Step 2. The treatment in this case is a history of incarceration, and the researcher conjectures that incarcerated youth have a higher incidence of disorder, so:

$$H_0 : p_E = p_C$$

$$H_A : p_E > p_C$$

Step 3. Now enter the list into the spreadsheet FORMULAS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled hypothesis-PROPORTIONS x 2. You should use Sample 1 to record the experimental data and Sample 2 to record the control data.

	sample 1	Sample 2
sample size	123	1865
"successes"	10	72
Calculations	sample 1	sample 2
sample proportion	0.081301	0.038606
Pooled Proportion	0.041247	
test statistic	2.306258	
p-value LEFT TAILED	98.95%	
p-value RIGHT-TAILED	1.05%	
You must decide whether to use the left-tailed p-value or the right-tailed p-value		

Step 4. This is a right-tailed test from the direction of the inequality in the alternative hypothesis. Since the p -value of 1.05% is less than the pre-set target of 5%, we reject the null hypothesis and accept the alternative, namely that incarcerated youth are at greater risk for personality



disorders. The chance of a Type I Error is 1.05%.



19. Hypothesis Tests for One Sample

Sometimes you will have access to retrospective census data and can use that information for your control group. Where this is possible, and where it doesn't degrade the experimental design, this is preferable since you can then reduce the uncertainty in your conclusions. Suppose, for example, we consider the following situation.

19.1. Example.

It is known that the presence of pets, such as cats and dogs, in nursing homes will reduce the loneliness of patients by 12.280 points on a standardized instrument that measures loneliness. The research that established this exposed the patients to the pets in group settings on the reasoning that the

presence of the pet would provide social lubrication for patient-to-patient interaction.

A researcher wonders if removing the socialization aspect would result in a difference in the outcomes of this treatment. To study this, the researcher randomly selects 40 nursing home patients and administers the standardized loneliness scale to this experimental group. This group then interacts daily for one hour with a dog in their room but away from other patients. At the end of a 21 day test period, the researcher then re-administers loneliness scale and finds an average reduction in loneliness of 10.9 standard deviation of 5.2. Can the researcher conclude, using a significance level of .05 that nursing home patients who interact with a dog away from other patients experience less reduction in loneliness?

Solution. **Step 1.** The dictionary in this case is slightly simpler, since the control population has a known mean. We also don't need to know the standard deviation of the control group.

	Control	Experimental
sample size	NA	40
Mean	$\mu=12.28$	$\bar{x}_E=10.9$
Standard Deviation	NA	$s_E=5.2$
α	5%	

The test statistic in this case is

$$T = \frac{\bar{x}_E - \mu}{\frac{s_E}{\sqrt{n}}}$$

Step 2. Now enter the list into the spreadsheet FORMULAS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled hypothesis-MEANS x 1. There is only one sample, and hence only one sample mean, sample standard deviation, and one sample size. There is only one population mean, the retrospective census data.

	A	B	C	D	E
1	Hypothesis Test - Means - One Sample				
2					
3	Data				
4	sample size	40			
5	Sample mean	10.9			
6	Sample Standard Deviation	5.2			
7	Population Mean	12.28			
8					
9	Computations				
10	Test Statistic	-1.6784			
11	p-value - LEFT TAILED	4.663065%			
12	p-value - RIGHT TAILED	95.336935%			
13					
14	You must decide whether to use the left-tailed p-value or the				
15	right-tailed p-value				
16					
17					
18					
19					
20					

Step 3. From the spreadsheet, we can see that the p -value is 4.55%, and hence we can (just barely) support the conjecture that pet interaction alone is less effective than pet interaction in the presence of other residents.

Questions.

- Did this project employ all three principles of experimental design? For each of the three principles, describe how this research did or did not use the principle. If it did not, describe how to refine the project to apply the principle.
- Could these results be confounded by some other effect?

There is also a similar, simplified version for one-sample hypothesis tests for proportions.

19.2. Example.

Researchers have conjectured that infectious diseases such as mononucleosis can lead to the onset of general depression. To test this a researcher studied a sample of 841 subjects with infectious mononucleosis and found that 56 met the criteria for general depression. Approximately 0.048 of the general population shows symptoms of general depression. Is this convincing evidence, using a significance level of .01 that infectious mononucleosis is

associated with an elevated incidence of general depression?

Solution. **Step 1.** Again, in building the dictionary there is just one sample.

	General Population	sample with Mono
sample size	NA	841
population proportion	.048	NA
number of “successes”	NA	56
α	1%	

Step 2. Now enter the list into the spreadsheet FORMULAS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled hypothesis-PROPORTIONS x 1. There is only one sample, and hence only one sample mean, sample standard deviation, and one sample size. There is only one population mean, the retrospective census data.

	A	B	C	D
1	Hypothesis Test - Proportions - One Sample			
2				
3	Data			
4	sample size	841		
5	"Successes"	56		
6	Population Proportion	0.048		
7				
8				
9	Computations			
10	p-hat	0.0666		
11	Test Statistic	2.5216		
12	p-value - LEFT TAILED	99.415900%		
13	p-value - RIGHT TAILED	0.584100%		
14				
15	You must decide whether to use the left-tailed p-value or the			
16	right-tailed p-value			
17				
18				
19				
20				

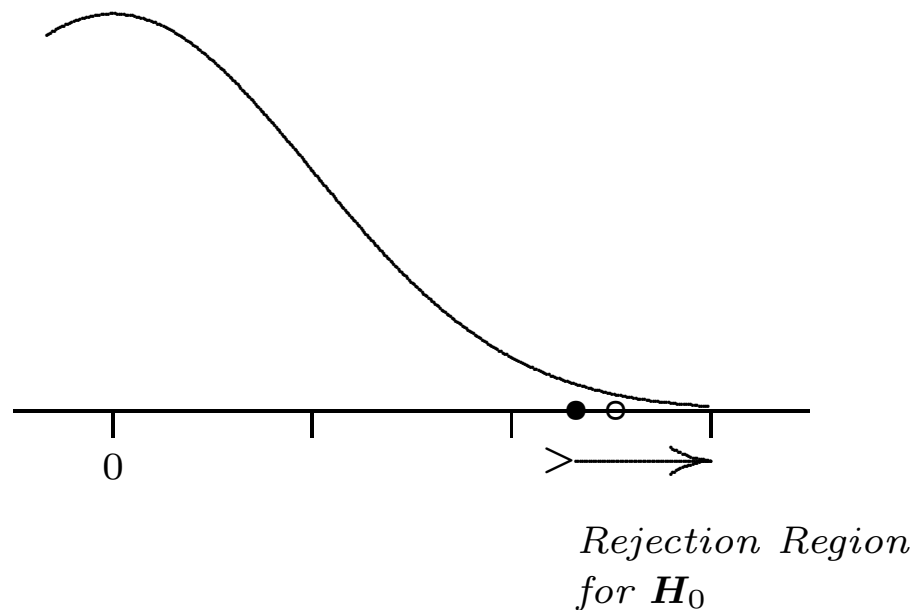
Step 3. Note that the observed value for \hat{p} in the sample is .066, which is larger than the corresponding value of .048 in the general population. Thus, the data seems to provide evidence in support of the conjecture. Since the p -value is 00.58%, we can in fact support the alternative hy-

pothesis with a probability of Type I Error of less than 1%.


While we have the p -value with the pre-set significance level, another approach is possible. Associated with each pre-set significance level there will be a cut-off in the normal tables. We could find this cut-off in exactly the way we worked inside-out problems. Some cut-offs for typical significance levels are tabulated in Appendix B in the study guide:

Significance Levels	> Right Tailed Tests	< Left Tailed Tests	≠ Two Tailed Tests	Confidence Levels
25%	0.674	-0.674	± 1.150	75%
20%	0.841	-0.841	± 1.281	80%
15%	1.036	-1.036	± 1.440	85%
10%	1.281	-1.281	± 1.644	90%
9%	1.340	-1.340	± 1.695	91%
8%	1.405	-1.405	± 1.750	92%
7%	1.476	-1.476	± 1.811	93%
6%	1.555	-1.555	± 1.881	94%
5%	1.644	-1.644	± 1.960	95%
4%	1.750	-1.750	± 2.053	96%
3%	1.881	-1.881	± 2.170	97%
2.5%	1.960	-1.960	± 2.241	97.5%
2%	2.053	-2.053	± 2.326	98%
1%	2.326	-2.326	± 2.575	99%
0.5%	2.575	-2.575	± 2.807	99.5%
			Confidence Limits	

In the above example, the cut-off associated with the significance level of $\alpha = .03$ is 2.326. The value of the test statistic, also reported in the spreadsheet, is 2.526. Since this is a right-tailed test, and since the test statistic is larger than the cut-off, we reject the null hypothesis and accept the alternative.



Cutoff ●
Test Statistic ○



The active learning modules on LEARN.OU.EDU all deal with **one sample** hypothesis tests, and use cut-offs from Appendix B in the above fashion.

20. Analysis of Variance

In our previous problems we had one sample and tested whether our sample differed from a known population mean. In the analysis of variance we will test multiple samples – and hence multiple treatment groups – against each other. The basic idea used by the *AN*alysis *O*f *VA*riance (*ANOVA*) is that the total variability within the pooled sample has three components:

- the variability within each treatment group;
- the variability between the groups; and
- the residual variability—everything else.

ANOVA tests to see if the "between groups" differences are large enough to conclude that the groups are really different.

For motivation, consider a couple of examples.

20.1. Example.

Suppose that there are three different diet plans each enrolling five clients. At the end of one month the following losses are recorded:

<i>Plan A</i>	<i>Plan B</i>	<i>Plan C</i>
5	15	10
5	15	10
5	15	10
5	15	10
5	15	10

The average loss in each plan is

$$\bar{x}_A = 5 \quad \bar{x}_B = 15 \quad \text{and} \quad \bar{x}_C = 10$$

Remember that the **population variance** involved calculating how much each observation differed from the mean, $x_i - \bar{x}$, squaring it $(x_i - \bar{x})^2$, and then averaging the squared differences

$$\frac{1}{n} \sum_i (x_i - \bar{x})^2$$

ANOVA involves **comparing sums of squared differences** like the above. The differences

$$(x_i - \bar{x})$$

represent the **error** between the observed value and the average value. With the diet plan data, the **sum of the squared differences** over the entire sample

$$\sum_{i=1}^{15} (x_i - \bar{x})^2 = 250$$

is the numerator in the variance

$$\sigma^2_{Total} = \left(\frac{\sum_{i=1}^{15} (x_i - \bar{x})^2}{15} \right) = 16.667$$

But there's another way to think about the variability. We could instead calculate the variability **between the treatments**, i.e., between the **plans**. Since the average loss in the three plans is

$$\frac{1}{3}(5 + 15 + 10) = 10,$$

the variability in the column averages is

$$\sigma^2_{plans} = \frac{(5 - 10)^2 + (15 - 10)^2 + (10 - 10)^2}{3} = \frac{50}{3} = 16.667.$$

In the above example, **the two numbers, σ^2_{Total} and σ^2_{plans} are the same.**

This isn't surprising since

In this example, *all* of the variability in the sample is due to the **differences in treatments** (column effects) and *none of the variability* is due to **individual differences in the clients**

In principle, there are three sources of variability in our diet plan example:

- the variability within each treatment group;
- the variability between the groups; and
- the residual variability—all other sources of variability having nothing to do with the groups.

In our idealized example, the only variability is between the groups.

In the real world you will never get such perfect data. There will always be some variability due to influences other than the treatments (column effects). In the diet plan, initial weight, gender, age, exercise and other uncontrolled variables (including measurement error) will result in other sources of variability. More realistic data might similar to that given in the next example.

20.2. Example.

Suppose that the data from the three diet plans had been:

<i>Plan A</i>	<i>Plan B</i>	<i>Plan C</i>
<i>5</i>	<i>10</i>	<i>5</i>
<i>5</i>	<i>15</i>	<i>5</i>
<i>5</i>	<i>15</i>	<i>10</i>
<i>10</i>	<i>15</i>	<i>10</i>
<i>15</i>	<i>15</i>	<i>10</i>

Using the two methods above, find the overall variance and the variance due

to the treatment (column) effects.

Solution. We still have the same fifteen numbers as in the first example, except now those numbers are distributed differently between the plans. However, this won't change the total variance, so

$$\sigma^2_{Total} = 16.667$$

as before.

The sample means are now

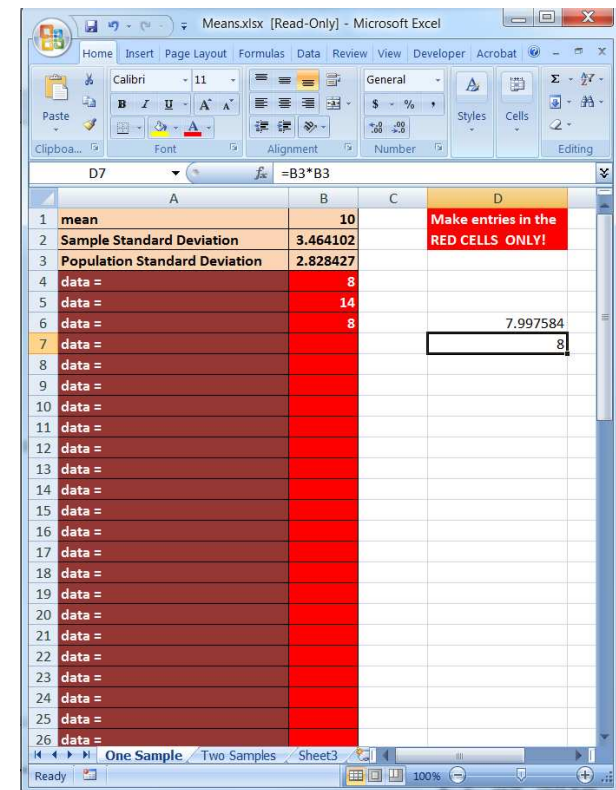
$$\bar{x}_A = 8 \quad \bar{x}_B = 14 \quad \text{and} \quad \bar{x}_C = 8$$

and so we can find the variance due to the column effects using the MEANS spreadsheet:

- enter the column means 8, 14 and 8 into the spreadsheet;
- find the population variance by squaring the population standard deviation. Since the spreadsheet calculates the value of the population standard deviation in cell B3, you can do this by entering

$$=b3*b3$$

into a cell on the spreadsheet and pressing enter. This gives you a value of $\sigma^2_{PLans} = 8$.



Thus about 50% ($\frac{8}{16.667}$) of the total variability in the population is due to the differences in treatments. The remaining variability is due to differences within the treatment groups and to residual effects not attributable to the diet plans.

ANOVA lets you decide whether or not the kinds of differences observed in the previous example are statistically significant. In particular, ANOVA will test

$H_0 : \mu_A = \mu_B = \mu_C$ against

H_A : the means are not all the same

Another way of saying the alternative hypothesis is:

$H_0 : \mu_A = \mu_B = \mu_C$ against

H_A : at least 2 differ one from the other

since if the means are not all the same, at least two must be different, one from the other.

The test will be accomplished by comparing the *variance* of the overall sample against the *variance* of the sample means. In practice, however, programs like the AnalyzeThis spreadsheet calculate sums of squared differences, as we did above. As we shall see, the report generated by

the ANOVA tab of AnalyzeThis includes sums of squared differences and averages—mean square differences—for each of the sources of variability mentioned above.

The basic assumptions of ANOVA are

- Each group is drawn from a normally distributed population;
- All populations have a common variance;
- all samples are independent;
- within each group sample, all observations are random and independent.

The test is fairly robust against violations of the normality assumption, provided that the sample sizes are equal, sufficiently large, and symmetrical. The test is less robust against violations the second assumption, that all the group populations have the same variance. It's also possible to test for violations of the above assumptions, although we won't cover that in this course.

Remark. Usually ANOVA is only done for three or more treatment groups. When you have just two groups, hypothesis tests for means

are theoretically equivalent to ANOVA and computationally simpler. If you are given the actual data, the [AnalyzeThis](#) spreadsheet will do all the necessary calculations and perform the ANOVA test for you. If you are just given summary information—for example on the final exam—you will use the [Formulas](#) spreadsheet and will be given the following information:

- Sample means for each treatment group.
- Overall sample standard deviation s_{Total} .
- Sample size n for each treatment group.
- Significance level.

20.3. Example.

An attorney is interested in whether knowledge of prior bad acts—criminal convictions—will influence potential juror's view of the guilt or innocence of a person accused of a crime. She randomly samples potential jurors and then divides her sample into three groups. Group I is told that the accused has a criminal record, Group II is told that the accused has no criminal record, and Group III is given no information on prior bad acts. Each subject then fills out a questionnaire measuring how likely it is, in the view of the subject, that the accused is guilty.

Group	I	II	III
mean	8	3.8	5
sample size	5	5	5

The pooled standard deviation was 2.8. At the 1% significance level do the data show that information on prior bad acts influences juror perceptions?

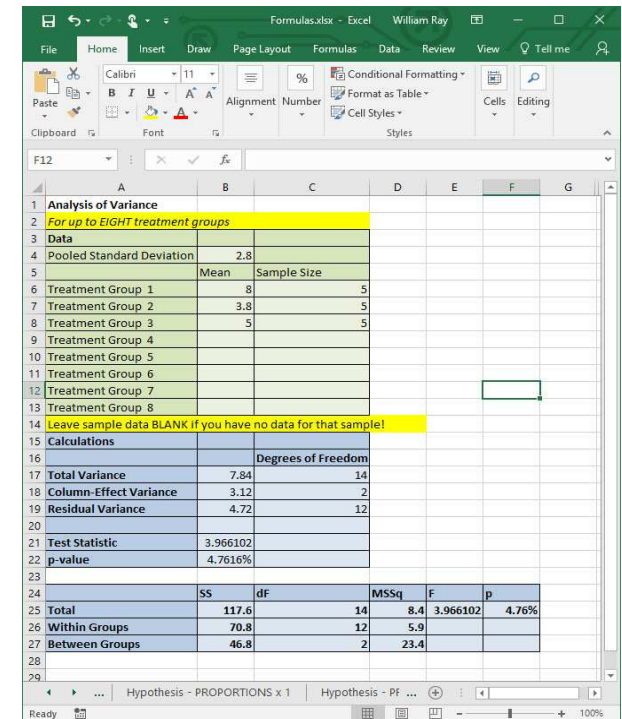
Solution.

Step 1. The above table, together with the pooled standard deviation of 2.8 provides the dictionary needed for the spreadsheet.

Step 2. Now enter the summary data and the pooled standard deviation into the spreadsheet FORMULAS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled ANOVA.

Step 3. Read the significance level from the spreadsheet. If it's less than the pre-set target, then we reject the null hypothesis, otherwise we accept the null hypothesis.

In this case, the pre-set significance level was 1% and the p -value is 4.76%, so we do NOT reject the null hypothesis. If we believe that the



The screenshot shows an Excel spreadsheet titled 'Formulas.xlsx' with the following data and calculations:

Analysis of Variance	
For up to EIGHT treatment groups	
Data	
Pooled Standard Deviation	2.8
Mean	Sample Size
Treatment Group 1	8 5
Treatment Group 2	3.8 5
Treatment Group 3	5 5
Treatment Group 4	
Treatment Group 5	
Treatment Group 6	
Treatment Group 7	
Treatment Group 8	
Calculations	
Degrees of Freedom	
Total Variance	7.84 14
Column-Effect Variance	3.12 2
Residual Variance	4.72 12
Test Statistic	3.966102
p-value	4.7616%
Summary	
Total	SS 117.6 df 14 MSSq 8.4 F 3.966102 p 4.76%
Within Groups	70.8 12 5.9
Between Groups	46.8 2 23.4

means are not all the same, then the chance of a Type I Error is 4.76%. Another way of stating this concluding is that the results are not highly significant but are significant since $1\% < p < 5\%$.

Remark. Note that an ANOVA is *always* a right-tailed test.

Solution Template

Step 1. Build your dictionary if it's not already in the problem:

pooled standard deviation	σ
Mean for treatment group 1	\bar{x}_1
Mean for treatment group 2	\bar{x}_2
Mean for treatment group 3	\bar{x}_3

The spreadsheet permits up to eight treatment groups, although ANOVA in general works for any finite number of treatment groups.

Step 2. Now enter the summary data and the pooled standard deviation into the spreadsheet FORMULAS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled ANOVA.

Step 3. Read the significance level from the spreadsheet. If it's less than the pre-set target, then we reject the null hypothesis, otherwise we accept the null hypothesis. ANOVA is always a right-tailed test

The screenshot shows an Excel spreadsheet with the following data:

	Mean	Sample Size
Treatment Group 1	8	5
Treatment Group 2	3.8	5
Treatment Group 3	5	5
Treatment Group 4		
Treatment Group 5		
Treatment Group 6		
Treatment Group 7		
Treatment Group 8		

	Degrees of Freedom
Total Variance	7.84
Column-Effect Variance	3.12
Residual Variance	4.72
Test Statistic	3.966102
p-value	4.7616%

	SS	df	MSSq	F	p
Total	117.6	14	8.4	3.966102	4.76%
Within Groups	70.8	12	5.9		
Between Groups	46.8	2	23.4		

End of Solution Template

The spreadsheet reports some other calculations in addition to the p -value.

First, it calculates the **total variance**, which is the square of the pooled standard deviation.

Second, it calculates the **column effect variance**, which is the amount of variability due to the treatments. The spreadsheet uses the following formula (in the case of four treatment groups):

$$V_C = \frac{n_A \bar{x}_A^2 + n_B \bar{x}_B^2 + n_C \bar{x}_C^2 + n_D \bar{x}_D^2}{n_A + n_B + n_C + n_D} - \bar{x}^2$$

where \bar{x} is the pooled average.

In the example with the lawyer and the potential jurors, all three groups had the same sample size—a **balanced sample**. This is not necessary for ANOVA—the groups can have different sizes.

Third, the spreadsheet calculates the **residual variance**, which is the variability due to all other sources than the treatments. This is the column variance subtracted from the pooled variance.

In addition to the above items, it calculates the **degrees of freedom** for each quantity. For the total variance, this is the size of the pooled

sample minus one. For the column effects, it is the number of treatment groups minus one. For the residual variance, it's the difference between these two numbers.

The above all needed to calculate the test statistic for ANOVA:

$$F = \frac{v_C / (\text{degrees of freedom for } v_C)}{v_R / (\text{degrees of freedom for } v_R)}$$

Similar to our initial motivating example, the F statistic compares two sources of variation in the data:

- Variation due to differences in the columns; and
- Variation due to residual effects.

Since the spreadsheet does all of these calculations for you, they are included here just for completeness.

Note that the spreadsheet also has a second table using "Sum of Squares" and "Mean Sum of Squares" statistic. The values for SS_{Total} and SS_{Between} are calculated as before, while

$$SS_{\text{Within}} = SS_{\text{Total}} - SS_{\text{Between}}.$$

The **degrees of freedom** are

SS_{Total}	total sample size - 1
$SS_{Between}$	number of groups - 1
SS_{Within}	difference of the above

while the **Mean Sum of Squares** column MSS_q is the **Sum of Squares** SS divided by the corresponding **degrees of Freedom** dF . The **F Statistics** F is

$$\frac{MSS_q \text{ Between Groups}}{MSS_q \text{ Within Groups}}$$

It turns out that the two approaches to calculating the F statistic are equivalent. The "sum of squares" approach takes somewhat better advantage of the precision of the computer and is slightly faster on large data sets, which is why it's the preferred method.

Thus, the F statistic, the p -value, and the conclusions are the same in this both tables.

The good news is that the spreadsheet does all of these calculations for you. You need to understand conceptually that the test is comparing sources of variability in the data and be able to interpret the p-value.

Finally, the AnalyzeThis spreadsheet produces the same results, but starts with the raw scores. The final numbers may be slightly different since AnalyzeThis uses the full precision of the computer while we rounded the test averages and pooled standard deviation in creating the Formulas version.

Variables are quantitative and generally continuous as opposed to discrete.

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8
Mean	8	3.8	5					
Sample Size	5	5	5					

Table I	Value	Degrees of Freedom
Total Variance	7.84	14
Column Effect Var	3.12	2
Residual Variance	4.72	12
Test Statistic F	3.966101695	
p-value	4.76%	

For completeness, this spreadsheet includes both the analysis in Table I and in Table II. Table II contains information on "sums of squares" and "mean square" statistics, but gives the same conclusions as Table I.

Table II	SS	dF	MSSq	F	p	Critical Value
Total	117.6	14	8.4	3.966102	4.76%	3.8853
Within Groups	70.8	12	5.9			
Between Groups	46.8	2	23.4			

Tests the null hypothesis that all the column means are the same against the alternative that they are not.

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8
Subject1	10	5	4					
Subject2	7	0	6					
Subject3	5	3	9					
Subject4	10	7	3					
Subject5	8	4	3					

20.4. Example.

A researcher is interested whether the presence of pets will influence the socialization of tenants in long-term care facilities. She selects four group homes and chooses a random sample of tenants from each home as follows:

<i>Group 1</i>	<i>10 tenants</i>	<i>group areas include dogs</i>
<i>Group 2</i>	<i>7 tenants</i>	<i>group areas in include cats</i>
<i>Group 3</i>	<i>15 tenants</i>	<i>both cats and dogs</i>
<i>Group 4</i>	<i>13 tenants</i>	<i>no pets</i>

After four weeks, the researcher scores each subject on a standardized socialization scale and obtains the following results.

Group 1	Group 2	Group 3	Group 4
67	69	80	64
78	65	78	81
82	83	62	57
79	66	67	65
84	64	74	84
100	69	71	80
83	71	87	78
75		78	66
73		70	71
75		86	74
		72	72
		76	69
		59	58
		81	
		71	

Can the researcher conclude that there is a difference in the socialization between the three groups?

Solution.

Notice that in this problem we have the raw data instead of summary data. Using the [AnalyzeThis](#) spreadsheet makes this easy—all we have to do is enter the data. See also the spreadsheet [ANOVAExample.XLSX](#), which has the data pre-loaded. Since the p -value is 4.39%, we have significant but not highly significant evidence that the treatment groups have different socialization levels.

The screenshot shows an Excel spreadsheet with the following data:

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7
Mean	79.6	69.57143	74.133333	70.69231			
Sample Size	10	7	15	13			

Table I	Value	Degrees of Freedom
Total Variance	73.64444444	44
Column Effect Var	13.05987247	3
Residual Variance	60.54704111	41
Test Statistic F	2.947871944	
p-value	4.39%	

Table II	SS	dF	MSSq	F	p	Critical Value
Total	3312.311111	44	75.279798	2.947872	4.39%	2.8327
Within Groups	2724.61685	41	66.45407			
Between Groups	587.6942613	3	195.89809			

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7
Subject1	67	69	80	64			
Subject2	78	65	78	81			
Subject3	82	83	62	57			
Subject4	79	66	67	65			
Subject5	84	64	74	84			
Subject6	100	69	71	80			
Subject7	83	71	87	78			
Subject8	75		78	66			
Subject9	73		70	71			
Subject10	75		86	74			
Subject11			72	72			

20.5. Example.

Researchers are interested in the relationship, if any, between alcohol consumption and blood pressure. A group of 32 non-smoking females, aged 20-65, were divided into four groups as follows

- Group A: no alcohol consumption for two weeks;
- Group B: 13 ml of red wine daily for two weeks;
- Group C: 13 ml of non-alcoholic red wine daily for two weeks;
- Group D: 1125 ml of beer daily for two weeks.

	Group A	Group B	Group C	Group D
change in blood pressure	-0.1	1.9 mm	1.9 mm	2.9 mm
# of subjects	8	8	9	9

For this sample, the pooled standard deviation was 1.21 mm. At the 5% significance level, do the above data show that there is a difference in the blood pressure response times between groups?

Solution.

Step 1. The above table provides the summary data for four treatment groups and for a pooled standard deviation of 1.21.

Step 2. Now enter the summary data and the pooled standard deviation into the spreadsheet FORMULAS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled ANOVA.

Step 3. Read the significance level from the spreadsheet. If it's less than the pre-set target, then we reject the null hypothesis, otherwise we accept the null hypothesis.

In this case, the pre-set significance level was 5% and the p -value is .0000%, so we reject the null hypothesis. The chance of a Type I Error

Analysis of Variance	
For up to EIGHT treatment groups	
Data	
Pooled Standard Deviation	1.21
Mean	Sample Size
Treatment Group 1	-0.1 8
Treatment Group 2	1.9 8
Treatment Group 3	1.9 9
Treatment Group 4	2.9 9
Treatment Group 5	
Treatment Group 6	
Treatment Group 7	
Treatment Group 8	
Leave sample data BLANK if you have no data for that sample!	
Calculations	
	Degrees of Freedom
Total Variance	1.4641 33
Column-Effect Variance	1.163495 3
Residual Variance	0.300605 30
Test Statistic	38.70508
p-value	0.0000%
	SS df MSq F p
Total	49.7794 33 1.508467 38.70508 0.00%
Within Groups	10.22058 30 0.340686
Between Groups	39.55882 3 13.18627



is less than 0.0000%.



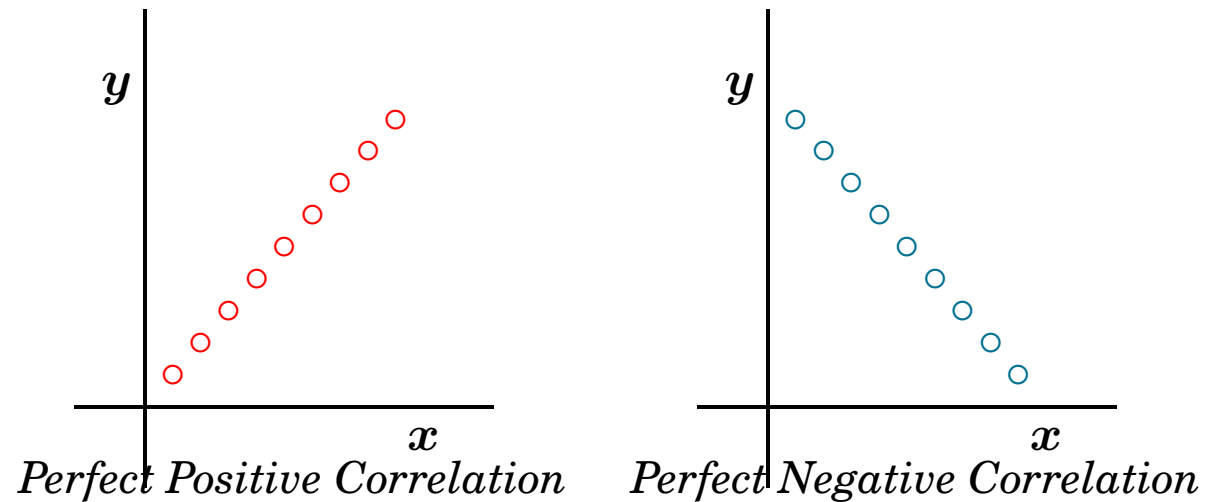
21. Correlation

Sometimes two measurements on a single individual will appear to be related:

Years of Education	income
height	weight
blood pressure	cholesterol

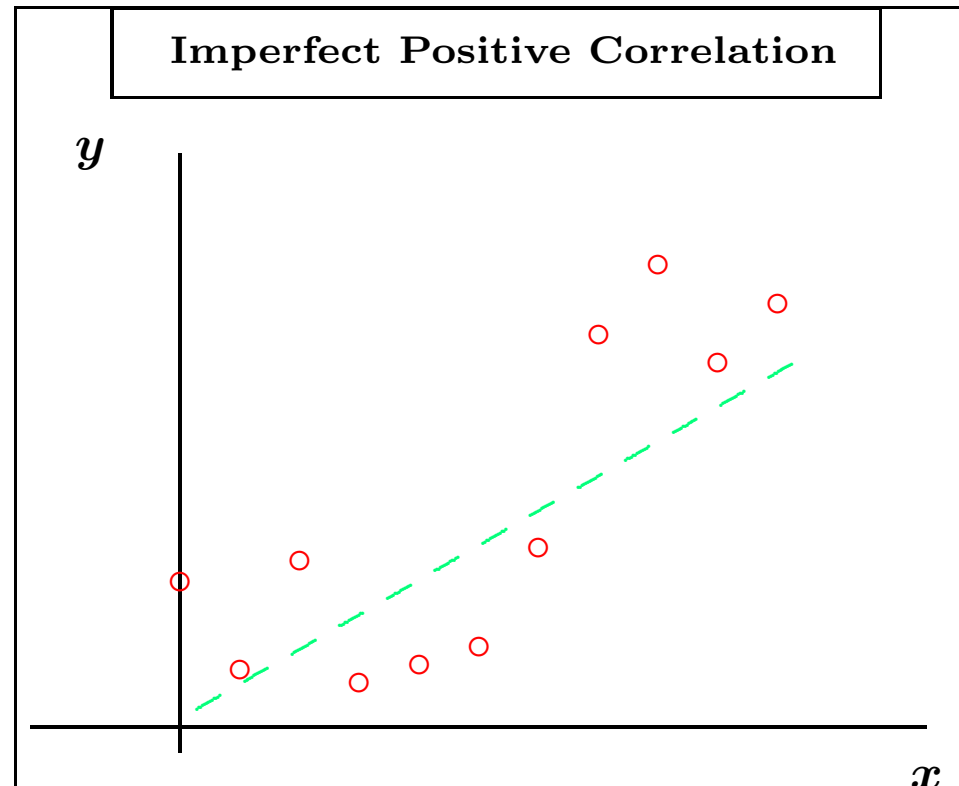
- In each case, we are taking two measurements on a single individual.
- Experience (or reasoning) suggests that the two measurements are not independent: whenever we observe a change in one we will also observe a change in the other.

- “Correlation” is a measure of straight-line relationships:



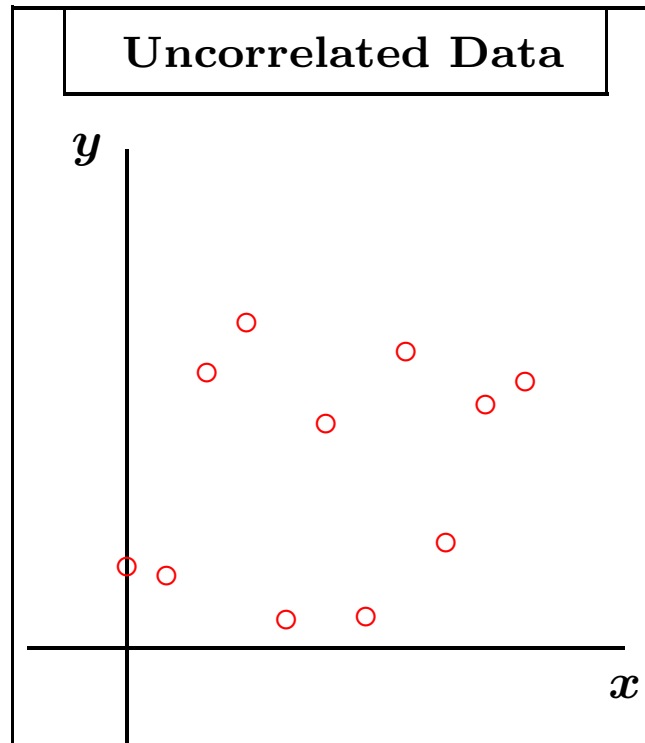
The circles (\circ) represent actual observations. In perfect positive correlation, as x increases, so does y . In perfect negative correlation, as x increases, y decreases.

In the real world, you never get perfect correlation of either type; at the least, there will be observational errors which cause some of the data points to slightly miss the straight line.

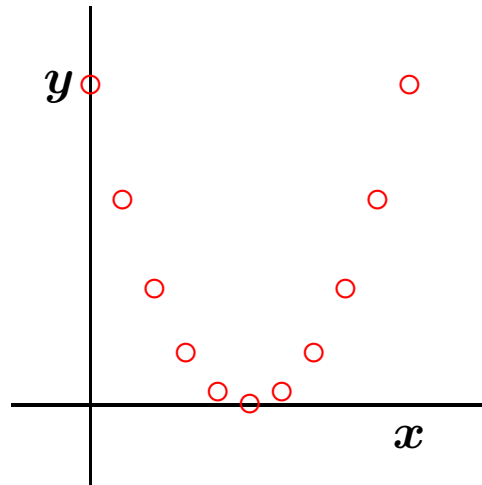


- Other times the observations may turn out to be completely unrelated or *uncorrelated* – we would expect that shoe size and income are

uncorrelated, for example.



- Correlations only measure straight line relationships:



The data at left are *uncorrelated* even though they are obviously related (parabolically).

Remark If we were to graph x against the logarithm of y ($\ln(y)$) we would get a straight-line relationship. This kind of transformation is often done to linearize nonlinear relationships.

Given a set of paired data points

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)$$

it is possible to compute a number, called the *correlation coefficient* ρ , which measures how closely the data fall on a straight line.

- The symbol ρ is used when **census data** are the basis for the calculation.
- The symbol r is used when **sample data** are the basis for the calculation.

The correlation coefficient ρ has the following properties:

- $-1 \leq \rho \leq +1$ and $-1 \leq r \leq +1$
- If $\rho = 0$ then the data are uncorrelated.
- If $\rho = +1$ then the data have a perfect positive correlation.
- If $\rho = -1$ then the data have a perfect negative correlation.

One of the things we can do is test the hypotheses:

$$H_0 : \rho = 0 \text{ against}$$

$$H_A : \rho > 0 \text{ or}$$

$$H_A : \rho < 0$$

to decide if the two variables are really correlated.

Suppose that the two variables really are correlated. Think of x as the input (or independent) variable and y as the output (or dependent) variable. Some of the variability in y is due the influence of x . Some of the variability in y is due to residual factors not measured by x (such as sampling error, measurement error, other things that might influence y). For example, we might gather data on each subject's income and educational achievement. We'd expect to see a positive correlation between the two: as education goes up, income goes up. We might even test the hypothesis that this is true, i.e., that

$$H_A : \rho > 0.$$

A second thing we might be interested in is whether or not **education predicts income**, i.e., whether there is a reliable formula that connects the two with reasonably small error. This is a slightly different hypothesis, since it is saying that education is a **primary** source of the variability in income, something that is probably untrue.

The value of ρ^2 measures the *proportion* of variability in y that is due to the influence of x .

$$\rho^2 \approx \begin{cases} \text{proportion of variability in } y \\ \text{due to the influence of } x. \end{cases}$$

While we will develop tests for how well the independent variable x predicts the dependent variable y , there is a rough standard that generally applies.

- *Accepted Practice:* you should not use x to predict y unless ρ^2 is at least 0.16 (or the absolute value of ρ ($|\rho|$) is at least 0.4).

There are several equivalent formulas for calculating the correlation coefficient. The simplest to write down is probably

$$r = \frac{\text{average of the products} - \mu_x \mu_y}{\sigma_x \sigma_y}$$

Conceptually, the products in the numerator measure the [interaction](#) between x and y . Dividing by the product of the standard deviations eliminates the "scale" or units from the number—in effect, standardizes it.

A mathematically equivalent formula that is computationally less sensitive to rounding and hence more frequently used in textbooks and other settings is

$$r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{(n \sum x^2 - (\sum x)^2) ((n \sum y^2 - (\sum y)^2))}}$$

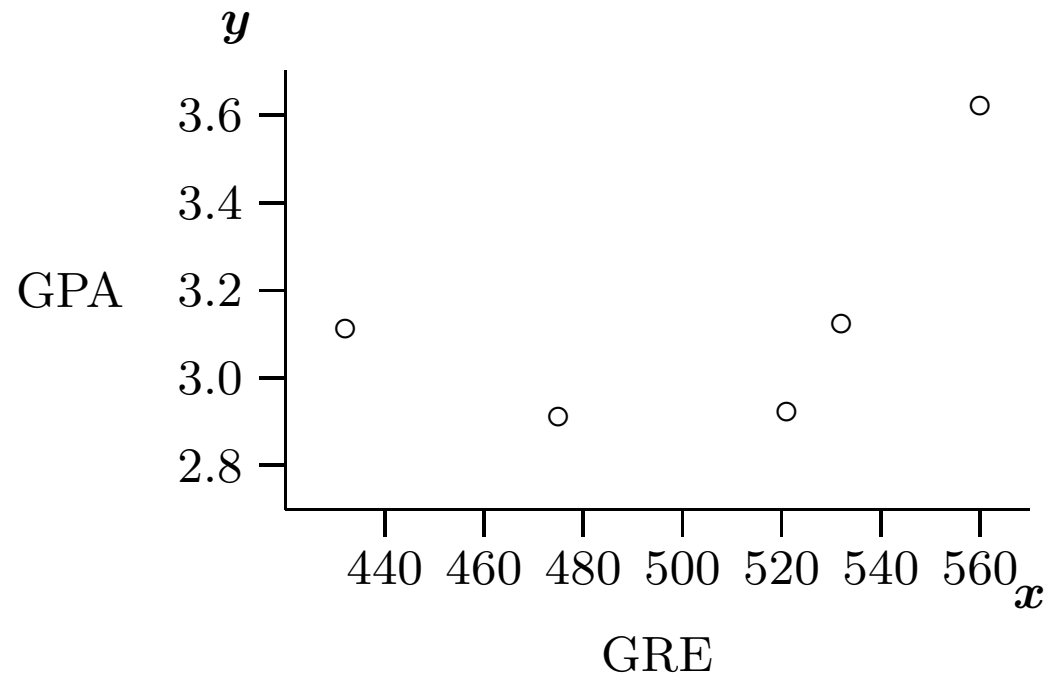
21.1. Example.

Consider the following data:

<i>GRE Scores</i>	<i>1st year GPA</i>
475	2.91
521	2.92
532	3.12
560	3.62
432	3.11

Find the correlation coefficient for the above data.

Note that the data have the following graph:



Solution. You can find the correlation coefficient r with the formula:

$$r = \frac{\text{average of the products} - \mu_x \mu_y}{\sigma_x \sigma_y}$$

where the “average of the products” is the average of the products of

the pair of observations (x, y) that you have for each subject. (When using this formula, it is essential that you use the key on your calculator which finds the “*population*” standard deviation (σ_n) .)

Step 1. To use the formula, make a “products” column with your table of data:

GRE Scores	1st year GPA	xy
475	2.91	1382.25
521	2.92	1521.32
532	3.12	1659.84
560	3.62	2027.20
432	3.11	1343.52

Step 2. Find the means for all three columns and the standard deviations for the first two:

	<i>GRE Scores</i>	<i>1st year GPA</i>	xy
<i>means</i>	$\mu_x = 504$	$\mu_y = 3.14$	$\mu_{xy} = 1585.83$
<i>st. dev</i>	$\sigma_x = 45.24$	$\sigma_y = 0.26$	–

Step 3. Apply the formula to find the correlation coefficient:

$$\begin{aligned} r &= \frac{\text{average of the products} - \mu_x \mu_y}{\sigma_x \sigma_y} \\ &= \frac{1585.83 - (504)(3.14)}{(45.24)(0.26)} \\ &= 0.538 \end{aligned}$$

(If you use the **rounded** values for the various calculated values reported in the table, you get $r = 0.34$. Again, these calculations are sensitive round-off error.)

In this problem, $r^2 = 0.289$, so about 28% of the variability in first year GPA can be predicted by the GRE. The remaining variability is due to residual factors (such as motivation, persistence, support, etc.)

Remark 1. The above formula is fairly simple, but is *extremely* sensitive to round-off error. For this reason, most books will use the more complex but equivalent formula given above.

Remark 2. As usual, there is a built-in Excel function to calculate the correlation coefficient. So, in this class, it's never necessary to do this calculation by hand. See the spreadsheet Means at right.

The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E
1	mean	510	3.136		Make entries
2	Sample Standard Deviation	40.72468539	268.5169391		RED CELLS O
3	Population Standard Deviation	36.42526596	0.258038757		
4	Correlation	0.651762427			
5	Slope (m)	92.00408507			
6	Intercept (b)	221.4751892			
7	input (x)	15			
8	predicted output (y)	1601.536465			
9	Standard Error	0.252650942			
10		X	Y		
11	data =	475	2.91		Note: you must have th the "X" and "Y" column
12	data =	521	2.92		
13	data =	532	3.12		
14	data =	560	3.62		
15	data =	462	3.11		
16	data =				
17	data =				
18	data =				
19	data =				
20	data =				
21	data =				

It is essential that you not confuse the notions of “correlation” and “cause and effect.”
High correlations do NOT necessarily imply a cause-and-effect relationship!!

21.2. Example.

Ice cream sales and deaths by drowning are correlated at $r = 0.83$.

- *Does this mean that ice cream sales cause deaths by drowning (for example, by causing stomach cramps)?*
- *Maybe it means that people get upset by seeing a drowning and so they assuage their grief by eating ice cream?*

21.3. Example.

Among elementary school children, math scores on a standardized test and weight are correlated at $r = 0.72$.

- *Should we encourage overweight children in order to improve math scores?*
- *Maybe kids who are good at math are unfit and fat?*

NOTE: High correlations DO NOT IMPLY that cause-effect relationships exist!!!!

22. Linear Regression

If observations are correlated (if the absolute value $|r|$ of the correlation coefficient is at least 0.4), we can use the input x to predict the output y .


22.1. Example.

Fat content of the body is of great medical and physiological importance, influencing for example death rates, the effectiveness of drugs and anesthetics. Fat content is generally calculated from body density, with higher fat values corresponding to lower body density. Body density is not easy to calculate; the most accurate method requires the subject to be submerged under water. Another method involves averaging certain skinfold measurements.

In a sample of 16 males aged 20-29 both skinfold and body density measurements were taken; the following data were gathered:

<i>Skinfold</i>	<i>Density</i>
<i>1.0</i>	<i>1.12</i>
<i>1.1</i>	<i>1.12</i>
<i>1.2</i>	<i>1.11</i>
<i>1.3</i>	<i>1.09</i>
<i>1.3</i>	<i>1.04</i>
<i>1.4</i>	<i>1.09</i>
<i>1.4</i>	<i>1.08</i>
<i>1.5</i>	<i>1.19</i>
<i>1.5</i>	<i>0.93</i>
<i>1.6</i>	<i>0.96</i>
<i>1.7</i>	<i>0.91</i>
<i>1.8</i>	<i>0.83</i>
<i>1.8</i>	<i>0.91</i>
<i>1.9</i>	<i>0.93</i>
<i>2.1</i>	<i>0.83</i>
<i>2.1</i>	<i>0.85</i>

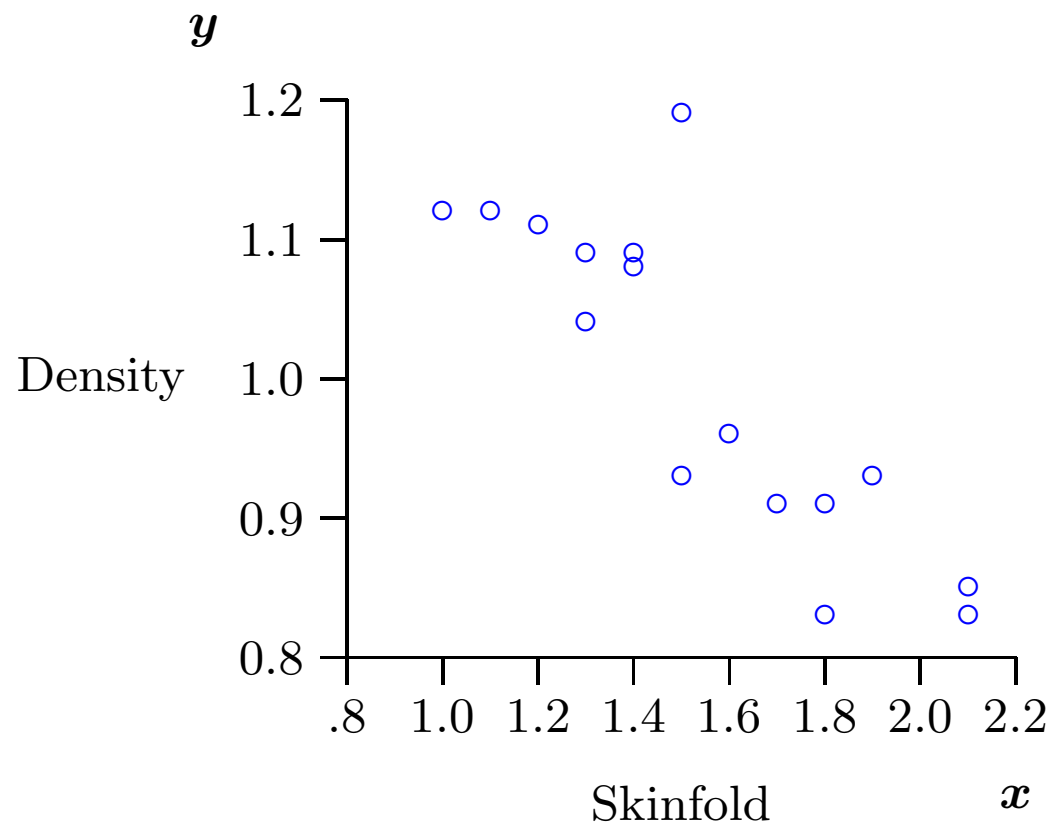
The average skinfold was found to be 1.55 with a standard deviation of 0.33.



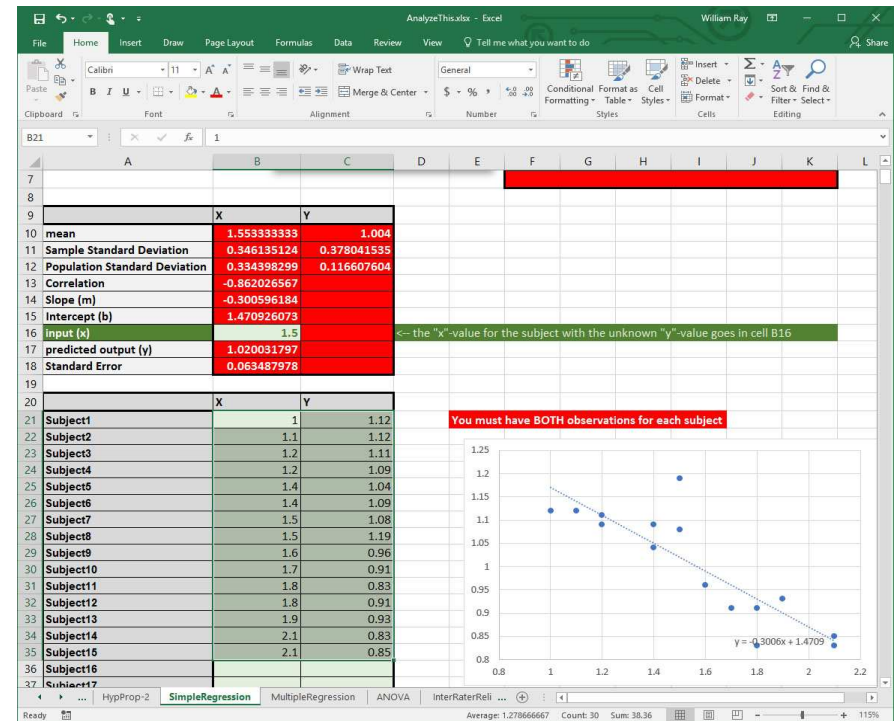
The average body density was found to be 1.00 with a standard deviation of 0.11. The variables "skinfold" and "body density" were shown to be correlated at $r = -0.86$.

You are confronted with a 23 year old male whose skinfold measurement is 1.5. Estimate his body density.

Note that the data have the following graph:

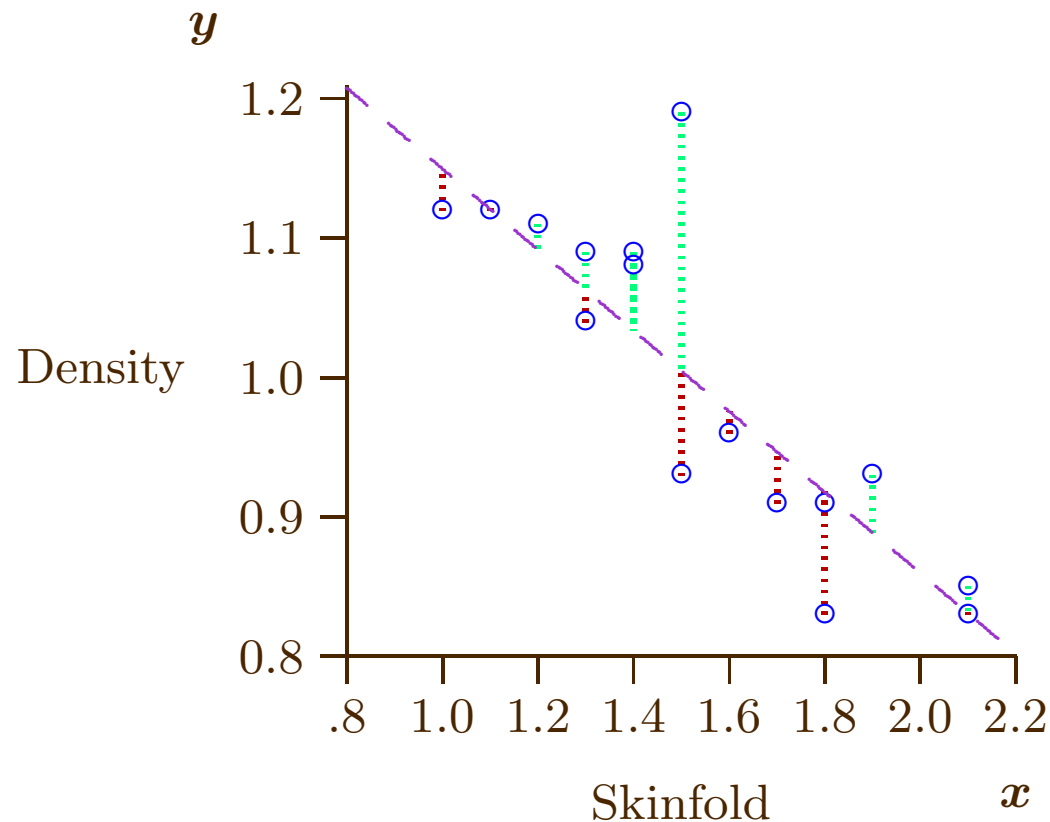


Remark 1. Since this problem gives you the actual data, you could use the Simple Regression tab of the AnalyzeThis spreadsheet to obtain the answer. Note that the spreadsheet even gives you the above scatterplot. However, since the problem also gives you the summary statistics, you can use the Formulas spreadsheet and avoid entering the data.



Solution. The idea is to find a formula for a line which best approximates the data. Since the correlation is not perfect, some of the data

points will miss the line which best fits the data:



The vertical (dotted) lines represent the “error” between the actual value of y (the data point \circ) and the predicted value of y which is on the

(dashed) line. The *regression line* is the straight line which results in the mean squared error (average of the squares of the error distances) being minimized. (For this reason the regression line is sometimes called the least squares line.)

The general formula for a straight line is

$$y = mx + b.$$

The method involves first finding the parameters m and b and then using the particular observation (the 1.5 skinfold for our 23 year old male) to predict body density.

Step 1. The first step is to make a list of the data. The hardest step is to decide which measurement will be x and which will be y .

The value you are going to *predict* will be y .

The value that *does the predicting* will be x .

In this problem we are going to try to predict body density; since this is what we are trying to predict, body density must be y . We will be

using skinfold measurements to do the prediction, so x must be skinfold measurements. This will be the summary data that we put in the spreadsheet, which will calculate values for m and b and then use those to do predictions based on skinfold measurements.

\bar{x}	1.55
s_x	0.33
\bar{y}	1.00
s_y	0.11
r	-0.86
x_0	1.5

The x_0 represents the particular observation given in the problem.

Step 2. Read the predicted results from the spreadsheet.

	A	B	C	D	E	F	G
1	Linear Regression						
2							
3	Data						
4	average for y	1					
5	standard deviation for y	0.11					
6	average for x	1.55	8				
7	standard deviation for x	0.33	8				
8	correlation	-0.86	9				
9	observation	1.5					
10							
11	Calculations						
12	value of "m"	-0.2867					
13	value of "b"	1.4443					
14	prediction	1.0143		135.0143			
15							
16	Note that the Active Learning Modules may round differently!						
17							
18							
19							

Question: why does this answer not agree with either of our observations for 1.5 skinfold? Why is it even different from the *average* of the two 1.5 observations?

Solution Template

Step 1. Make a dictionary which assigns values to the variables. Before making your table, check to see if you are trying to do a prediction. If so, the quantity you are trying to predict must be y and the quantity used to do the prediction must be x .

average for y	\bar{y}
st. dev. for y	s_y
average for x	\bar{x}
st. dev. for x	s_x
correlation coef.	r
indiv. x observation	x_0

The individual x_0 observation is the observation for a specific individual

which will be used to predict y .

The value you are going to *predict* will be y .

The value that *does the predicting* will be x .

Step 2. Now enter the list into the spreadsheet FORMULAS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled REGRESSION. You can then read the predicted value out of the spreadsheet, along with the calculated values for m and b .

————— **End of Solution Template** —————

22.2. Example.

A researcher administered varying doses of caffeine to twenty-four randomly selected male subjects aged 24-30, following which their blood pressure levels were measured. The average caffeine dosage was 320 mgs (about the equivalent of four cups of coffee) with a standard deviation of 240 mgs. The average increase in systolic blood pressure was 14 mm HG with a standard deviation of 4 mm. In this study caffeine consumption and increase in systolic blood pressure were found to be correlated with a correlation coefficient of $r = 0.74$. You are confronted with a subject whose systolic blood pressure is 134 mm HG. If this subject drinks 6 cups of coffee (which will contain 480 mgs of caffeine), predict the subject's blood pressure.

Solution.

Step 1. We will begin by predicting the subject's **increase** in blood pressure. Thus, in this problem, we are trying to predict increase in blood pressure; thus y = "blood pressure increase." We are using caf-

feine dosage as our predictor, so x =“caffeine.” Summarizing the data:

\bar{y}	14
s_y	4
\bar{x}	320
s_x	240
r	0.74
x_0	480

Step 2. Read the predicted results from the spreadsheet.

	A	B	C	D	E	F	G
1	Linear Regression						
2							
3	Data						
4	average for y	14					
5	standard deviation for y	4					
6	average for x	320	8				
7	standard deviation for x	240	8				
8	correlation	0.74	9				
9	observation	480	9				
10							
11	Calculations						
12	value of "m"	0.0123					
13	value of "b"	10.0533					
14	prediction	15.9733		149.9733			
15							
16	Note that the Active Learning Modules may round differently!						
17							
18							
19							

This gives the *increase* in blood pressure, but the problem asks for the blood pressure. Thus the answer

$$137 + 15.9733 = 149.9733.$$

This section dealt with **simple regression** because we used only one independent variable x . In the real world you will usually have several inputs contributing to a single output. For example, if the output is

$$y = \text{Income}$$

then some inputs might be

$$x_1 = \text{age}$$

$$x_2 = \text{education}$$

$$x_3 = \text{gender}$$

$$x_4 = \text{experience}$$

and the regression line would be

$$y = m_1x_1 + m_2x_2 + m_3x_3 + m_4x_4 + b.$$

This is called *multiple regression*; the concepts are the same but the computations are much more complex to do by hand. Spreadsheets and statistics programs make the calculations easy, of course. The spreadsheet [AnalyzeThis](#) includes tabs for both simple and multiple regression, with the latter permitting up to seven independent variables.

23. Multiple Regression

In many real-world situations researchers will have several **independent variable**. Spreadsheets for examples are here.

23.1. Example.

The human resources director for a chain of car dealers is interested in the attributes that influence sales. She randomly selects twenty sales people employed by the dealership and records their sales for the month of April, their scores on a standardized IQ test, and their scores on a standardized test for extroversion. She obtains the following results:

The researcher plans to use this information to rate applicants for sales jobs. If she has an applicant with an IQ of 110 and a score of 23 on the extrovert scale, what sales would she predict based on this data?

Sales	IQ	Extroversion Scale
\$2,625	89	21
\$2,700	93	24
\$3,100	91	21
\$3,150	122	23
\$3,175	115	27
\$3,100	100	18
\$2,700	98	19
\$2,475	105	16
\$3,625	112	23
\$3,525	109	28
\$3,225	130	20
\$3,450	104	25
\$2,425	104	20
\$3,025	111	26
\$3,625	97	28
\$2,750	115	29
\$3,150	113	25
\$2,600	88	23
\$2,525	108	19
\$2,650	101	16

Solution. The MultipleRegression tab of the AnalyzeThis spreadsheet answers these questions and more.

First we need to identify which variables are independent and which is dependent. In this example, the HR director is interested in what influences car sales, so this must be the **dependent** variable. The **independent variables** are then IQ and extroversion.

The **model** that the HR director proposes is that there is a linear relationship between sales, IQ, and extroversion:

$$y = c + m_1x_1 + m_2x_2$$

where

$y =$	car sales
$x_1 =$	IQ score
$x_2 =$	extroversion score

You can then enter the data into the cells B27:D46 of the spreadsheet. Note that y goes into the first column and the other two scores into the next two columns.

The analysis position of the spreadsheet gives you quite a bit of information.

Put your data in the green cells.
Do not edit any other cells.
Do not add or delete columns or rows.
There are hidden columns where most of the calculations are done.

Multiple regression line is of the form
 $y=c+m_1x_1+m_2x_2+m_3x_3+\dots+m_7x_7$

	SS	df	MS	F	p-value
ANOVA					
Total	2895750	19	152407.9	4.630316	0.024815
Regression	1021166.374	2	510583.2		
Residual	1874583.626	17	110269.6		

	c	m1	m2	m3	m4	m5	m6	m7
Coefficients	993.9245625	8.219912	49.70863					
Standard Error	788.0986054	7.01256	19.63374					
t	1.261167772	1.17217	2.531796					
p-value	22.43%	25.73%	2.15%					
Y		X1	X2	X3	X4	X5	X6	X7
Subject1		2625	89	21				
Subject2		2700	93	24				

The first section gives the regression statistics for testing

$$H_0 : r^2 = 0 \quad \text{against}$$

$$H_A : r^2 > 0$$

The key value is the p -value of 0.0248 or 2.48%, which means that we have significant but not highly significant evidence in support of H_A . From this we believe that the observed value of r^2 cannot be attributed to chance.

The second segment, labeled **ANOVA** we can skip for now.

The final section gives you the values of c , m_1 and m_2 in the above model, and so

$$y = 993.92 + 9.219x_1 + 49.70x_2.$$

From this, it's easy to substitute in $x_1 = 110$ and $x_2 = 23$ to predict sales of \$3,341 for the applicant with an IQ score of 110 and an extroversion score of 23.

But there is some additional information. For example, there are p -values for m_1 and m_2 . These relate to the hypotheses

$$H_1 : m_1 \neq 0 \quad \text{and} \quad H_2 : m_2 \neq 0.$$

Thus, if we believe m_2 is not zero, the chance we are wrong is 2.15%. On the other hand, if we believe m_1 is not zero, the chance we are wrong is 25%.

What does this say about using the model to predict sales?

In particular, this means **we should not use m_1** to predict sales, since we cannot assume its value is nonzero. Since we can't use m_1 , that means that the above prediction of \$3,341 is also not reliable, since it used m_1 . This suggests running another ANOVA using just extroversion and sales and omitting the variable IQ.

On the hand, the earlier p -value for r^2 lets us conclude that there is a connection between the variables. What the information on the coefficients means is that the connection, while real, is not strong enough to use for prediction.

The MultipleRegression tab of AnalyzeThis has many powerful features built into it, and tests more than one hypothesis. Multiple regression can even provide an alternative way of thinking about ANOVA.

23.2. Example.

There is a folk legend that if a mother drinks a beer prior to nursing her infant, the child will take in more breast milk. To test this, a nurse working in the maternity ward of a hospital randomly selected 40 nursing mothers and randomly divided them into four groups as follows:

- *Group I received instruction on breast feeding and ingested 10 oz of beer prior to nursing;*
- *Group II received the instruction and ingested 10 oz of a non-alcoholic beverage prior to nursing;*
- *Group III received instruction but was offered no beverage prior to nursing;*
- *group IV received neither instruction nor beverage prior to nursing.*

The researcher then weighed the infants before and after nursing and recorded the difference in weight, those differences being the amount ingested.

Solution. Using the ANOVA tab of AnalyzeThis...

We can **also** analyze the data using multiple regression by way of indicator variables:

$$m_1 = \begin{cases} 1 & \text{if the observation is in Group I} \\ 0 & \text{otherwise} \end{cases}$$

$$m_2 = \begin{cases} 1 & \text{if the observation is in Group II} \\ 0 & \text{otherwise} \end{cases}$$

$$m_3 = \begin{cases} 1 & \text{if the observation is in Group III} \\ 0 & \text{otherwise} \end{cases}$$

An observation is in Group IV exactly when

$$m_1 = m_2 = m_3 = 0$$

so we don't need an indicator variable for this group.
Using `MultipleRegression` gives identical results to ANOVA.

MultipleRegressionExamples.xlsx - Excel William Ray

File Home Insert Draw Page Layout Formulas Data Review View Tell me what you want to do Share

Clipboard Font Alignment Number Styles Cells Editing

H14

Multiple Regression

Put your data in the green cells.
Do not edit any other cells.
Do not add or delete columns or rows.
There are hidden columns where most of the calculations are done.

Multiple regression line is of the form
 $y=c+m_1x_1+m_2x_2+m_3x_3+\dots+m_nx_n$

Regression SS	4.10675								
Residual SS	25.563								
r^2	0.138415389								
SE s_y	0.842664425								
F Statistic, df	1.927825373	36							
p-value	0.142512788								

ANOVA	SS	dF	MS	F	p-value
Total	29.66975	39	0.760763	1.927825	0.142513
Regression	4.10675	3	1.368917		
Residual	25.563	36	0.710083		

	c	m1	m2	m3	m4	m5	m6	m7
Coefficients	3.89	0.53	-0.3	-0.2				
Standard Error	0.266473889	0.376851	0.376851	0.376851				
t	14.59805319	1.406391	-0.79607	-0.53071				
p-value	0.00%	16.82%	43.12%	59.89%				
	Y	X1	X2	X3	X4	X5	X6	X7
Subject1	4.9	1	0	0				
Subject2	4.1	1	0	0				

MultipleRegression-BreastFeedin Sheet1

Ready 100%

Notice that the **ANOVA** section in the spreadsheet replicates exactly the table from the ANOVA tab that we did earlier. In addition, the F -statistic is has the same value as the one that tests

$$H_0 : r^2 = 0; \quad \text{against}$$

$$H_A : r^2 > 0$$

Since the p -value is 0.1423 or 14.25%, we reject H_0 and believe that the value of r^2 cannot be attributed to chance. Similarly, the ANOVA statistic is telling us that the differences in the means cannot be attributed to chance. The approach using linear regression with indicator variables is thus seen to be statistically equivalent to the test comparing means.



24. Inter-rater Reliability

Researchers will often hire staff to gather their data. This can take many forms, including structured interviews, administering a **scale** or pre-determined set of questions, or observing and recording specific behaviors. Students doing a study of **who comes to a complete stop** needed to be sure that everyone in their group could consistently apply the definition of complete stop. These question arise in experimental design, and they all have to do with **inter-rater Reliability**. In this section, we'll discuss how to test for this when the research instrument is a **scale**.

24.1. Example.

A researcher is interested in how a scale that screens for substance abuse that is used in different environments: in an Emergency Room, a physician's office, and a DHS office. The researcher gathered data of ten subjects from each setting, with the data shown at right.

The researcher is interested in whether the scores depend on location and whether the scores differentiate between subjects.

	ER	Physician Clinic	DHS Clinic
	1	3	1
	2	9	2
	1	7	1
	8	2	2
	1	3	1
	2	2	1
	7	3	2
	5	1	2
	3	3	1
	3	2	1

Solution.

Two Way Analysis of Variance with no replacement

Put your data in the green cells.
Do not edit any other cells.
Do not add or delete columns or rows.
There are hidden columns where most of the calculations are done.

Total Ratings	30
Number of Raters	3
Number of Subjects	10

	SS	dF	MSq	F	p
SSt	143.8667	29	4.96092		
SSr	29.86667	9	3.318519	0.685539	71.29%
SSc	26.86667	2	13.43333	2.775057	8.90%
Sse	87.13333	18	4.840741		

<-- tests for whether columns are the same

<-- tests for whether the rows are the same

Reliability -0.45871 <-- Intra-class correlation

Averages	3.3	3.5	1.4					
Subjects	Groups							
	A	B	C	D	E	F	G	H
1	1	3	1					
2	2	9	2					
3	1	7	1					
4	8	2	2					
5	1	3	1					
6	2	2	1					
7	7	3	2					
8	5	1	2					
9	3	3	1					

The test for whether the scores are the same or differ by column (location) has a p -value of 71%, so we must conclude the tests are different. This isn't surprising since the average at DHS is so much lower than the other two locations.

The test for whether the row averages are the same has a p -value of 8.9%, so we can conclude that they are different. This means that the scale does differentiate between subjects.

The **intra-class correlation** describes how well values in the **same** group correlate with each other. A rough interpretation is that

less that 0.4	poor
between 0.4 and 0.50	fair
between 0.6 and 0.74	good
0.75 and higher	excellent

The spreadsheet calculates the ICC for the columns only. This means there is fair reliability within the locations.

The spreadsheet is essentially the same as the earlier ANOVA spreadsheet, except that it identifies an additional source of variability in the sample: the variability due to the individual rows. It still calculates the **sum of squares** for the columns in the same way, but now it does an ANOVA on the transpose of the rows:

Subjects	1	2	3	4	5	6	7	8	9	10
ER	1	2	1	8	1	2	7	5	3	3
Physician	3	9	7	2	3	2	3	1	3	2
DHS	1	2	1	2	1	1	2	2	1	1

This changes the calculation of the residuals accordingly. However, this is still fundamentally the one-way ANOVA spreadsheet from earlier, slightly tweaked.

The other added feature is the intra-class correlation, described in the above problem. It's calculated in a manner similar to the correlation coefficient.

24.2. Example.

The Human Resources Department at Mechanics R Us uses a standardized set of questions to assess job satisfaction. The VP for Human Resources wonders if the questions produce consistent results between interviewers and if the questions distinguish between different levels of satisfaction.

She randomly selects ten applicants and five interviewers. Each interviewer asks the ten applicants the questions and assigns a score to their responses. The results are at right.

Subjects	A	B	C	D	E
1	100	83	95	88	
2	90	86	94	80	
3	85	84	100	92	
4	85	90	91	95	
5	73	82	64	88	
6	72	74	79	77	
7	72	72	70	75	
8	84	85	90	77	
9	91	100	100	83	
10	81	83	85	91	

The VP's questions can be formulated as two hypotheses:

H_1 : The row averages are different

H_2 : The Column averages are the same

The hope is that the questionnaire *does* differentiate between the subjects, and thus that our evidence leads us to accept H_1 . At the same time, the hope is that the interviewers' use of the instrument is *reliable*, i.e., that we do **not** have evidence to reject H_2 . The second hypothesis is checking for **inter-rater reliability**.

Solution.

Two Way Analysis of Variance with no replacement

Put your data in the green cells.
 Do not edit any other cells.
 Do not add or delete columns or rows.
 There are hidden columns where most of the calculations are done.

Total Ratings	40
Number of Raters	4
Number of Subjects	10

	SS	dF	MSq	F	p
SSt	3193.1	39	81.87436		
SSr	1986.1	9	220.6778	5.24083	0.04%
SSc	70.1	3	23.36667	0.55493	64.93%
Sse	1136.9	27	42.10741		

<-- tests for whether columns are the same
 <-- tests for whether the rows are the same

Reliability **0.809191** <--Intraclass correlation

Averages	83.3	83.9	86.8	84.6				
Subjects	Groups							
	A	B	C	D	E	F	G	H
1	100	83	95	88				
2	90	86	94	80				
3	85	84	100	92				
4	85	90	91	95				
5	73	82	64	88				
6	72	74	79	77				
7	72	72	70	75				
8	84	85	90	77				
9	91	100	100	83				
10	81	83	85	91				
11								

In this case, we **accept** the hypothesis that the columns are the same, i.e., that there is no interviewer bias since the p -value is 0.04%.

We **reject** the hypothesis that the rows are the same, i.e., that the scale does differentiate between the subjects.

The ICC of 0.809 says that the reliability for the interviewers is excellent. ■

There are two **factors** or sources of variability that contribute to answering the VP's questions. The first factor is the variability due to the questions themselves. The second factor is the variability due to the subjects. The spreadsheet calculations are similar to those for one-factor ANOVA, except that all the calculations now take into account all the sources of interaction and are thus a bit more complex. In any case, for each factor, we will calculate the Mean-Square Error **between the components of the factor** and **within the components of the factor**.

25. Two-way ANOVA

The Analysis of Variance seeks to identify sources of variability in data with when the data is partitioned into differentiated groups. In the prior section, we considered two sources of variation that might contribute to the total variability in the sample:

- Variability between groups; and
- Variability within groups.

With **inter-rater reliability** we considered an additional source of variability by testing to see if the items differentiated between subjects.

In **two-way ANOVA**, we will have one **measurement variable** and a **nominal variable**, with the values of the nominal variable used to divide the sample into groups. The division might, for example, involve the experimental protocol as in the example on alcohol consumption and blood

pressure, or it might be the location of the subjects, as in the care facility example. In any case, it is some attribute that is important to the quantitative measurements. The nominal variable is sometimes called a **factor**.

Two-way ANOVA involves a **measurement variable** and two **factors** or **nominal variables**.

The test for inter-rater reliability we did in the prior section is an example of a special case of two-way ANOVA. In that test, the raters are an obvious category. However, we didn't have a second category to divide up the rows in the sample. This kind of two-way ANOVA is sometimes called **Two-factor ANOVA without replication** since we don't replicate our sample with another group. For example, if we divided our subjects (the rates) into male and female groups, we'd have **Two-factor ANOVA with replication**, which is the topic of this section.

As an example, let's return to the scale that screened for substance abuse.

Now the **location** where the screen is administered is one factor. The second factor we've added is the **gender** of the subject. If the researcher gathered data on five females and five males from each setting, the data might look something like the table at right.

	ER	Physician Clinic	DHS Clinic
Males	1	3	1
	2	9	2
	1	7	1
	8	2	2
	1	3	1
Females	2	2	1
	7	3	2
	5	1	2
	3	3	1
	3	2	1

There are several sources of variation in this example. For example, we could **disregard** location and just test to see if there is difference between genders, or we could **disregard** gender and just to see if there is difference between locations. Each of these would be a one-way ANOVA. Summary statistics for the columns and rows would look like

	Total	ER	Physician	DHS
Count	10	10	10	10
Sum	33	35	14	
Average	3.3	3.5	1.4	
Variance	6.455556	6.277778	0.266667	

	Males	Females
Count	15	15
Sum	44	38
Average	2.933333	2.533333
Variance	7.495238	2.695238

But the two **one-way** ANOVAS lose the ways that the factors of gender and location work together to influence the outcomes.

One-way ANOVA calculates the difference between the individual observations in, say, column one and the overall mean:

$$\sum (x_{1,j} - \bar{x})^2$$

One-way ANOVA assumes that the total variability comes from two sources: differences within the factors and the differences between the factors. It tests the null hypothesis that all the groups have the same mean against the alternative that they do not.

In two-way ANOVA, we can calculate the **row effects** of the vertical category by looking at the average data for males and females

males	females
2.933	2.533

Each of these averages are based on a sample of size fifteen and the overall average of all thirty subjects is 2.733. Just as with one-way ANOVA, then, the variability comes from

$$SS_{Rows} = 15 \times ((2.933 - 2.733)^2 + (2.533 - 2.733)^2) = 1.2$$

multiplying by 15 since there are fifteen subjects in each of the male and female samples.

In exactly the same way, we could find the variability due to the locations—the columns—using the overall average of 2.733 and the averages for each column

ER	Physician	DHS
3.3	3.5	1.4

and obtain

$$SS_{Cols} = 10 \times ((3.3 - 2.733)^2 + (3.5 - 2.733)^2 + (1.4 - 2.733)^2) \\ = 26.997$$

this time multiplying by ten, the number of subjects in each column.

The above accounts for the variability between locations (columns) and genders (rows), but does not account for the variability **within** our groups. The data is divided into **six** groups: three locations for each of two genders.

	ER	Physician Clinic	DHS Clinic
Males	1	3	1
	2	9	2
	1	7	1
	8	2	2
	1	3	1
Females	2	2	1
	7	3	2
	5	1	2
	3	3	1
	3	2	1

That means that we can find the average for each group, and then find the sum of squares within the groups. The calculations are in the table below.


	ER	Physician Clinic	DHS Clinic
Males	37.2	36.8	1.2
Females	16	2.8	1.2

so the total variability within the groups is the sum of the above entries, or 95.2.

We can account for the total variability in exactly the same way as with one-way ANOVA:

$$\sum (x_i - \bar{x})^2 = 143.8663.$$

Thus far we've accounted for



	SS
Rows	1.2
Cols	28.667
Within	95.2

The total we've accounted for so far is 123.2667, leaving

$$20.62 = 143.8663 - 123.2667$$

still not covered. This last variability comes from the [interaction](#) between the two factors.

While it's useful to see how the calculations work, Excel has built-in functions to do all of this. It's a little less straightforward to use than the built-in tools in the AnalyzeThis spreadsheet, but not terribly hard.

Before turning to Excel let's summarize our the general setup.

- We have a **numeric** data and **attribute** data on our subjects.
- We have two attribute variables which partition the data into groups of rows and columns.
- The groups are all the same size: the number of rows per column is the same for each column, and the number of columns is the same for each row.
- we test hypotheses on the row and column effects:

H_0 : means in column groups are all the same

H_0 : means in row groups are all the same

The alternative in each case is that the means are not the same.

- We also test for the interaction between the factors:

H_0 : there is no interaction

The alternative is that there is interaction.


The assumptions are:

- All samples are drawn from normally distributed populations;
- all populations have a common variance;
- all samples are drawn independently from each other;
- within each sample, the observations are sampled randomly and independently of each other.

25.1. Example.

The first example returns to the data involving the substance abuse scale given in different settings. We will perform a two a two-way ANOVA. This is with replication since we have categories for each of the factors.

	ER	Physician Clinic	DHS Clinic
Males	1	3	1
	2	9	2
	1	7	1
	8	2	2
	1	3	1
Females	2	2	1
	7	3	2
	5	1	2
	3	3	1
	3	2	1



Solution. Once the data analysis tool is installed, it's pretty easy to use. You can find it under the DATA tab in excel. Note that to use it, you must have the same number of subjects in each of the column groups, i.e., a balanced sample with respect to gender, divided equally between the locations.

We'll break the tool's analysis into two parts for convenience.

The first section reports summary statistics for factors. From this we can see that the overall average for the DHS clinic (1.4) is lower than the other two settings (3.3 and 3.5). We can also see differences in scores by gender at the DR and Physician offices, but not at the DHS clinic.

Anova: Two-Factor With Replication				
SUMMARY	ER	Physician	DHS Clinic	Total
<i>Males</i>				
Count	5	5	5	15
Sum	13	24	7	44
Average	2.6	4.8	1.4	2.933333
Variance	9.3	9.2	0.3	7.495238
<i>Females</i>				
Count	5	5	5	15
Sum	20	11	7	38
Average	4	2.2	1.4	2.533333
Variance	4	0.7	0.3	2.695238
<i>Total</i>				
Count	10	10	10	
Sum	33	35	14	
Average	3.3	3.5	1.4	
Variance	6.455556	6.277778	0.266667	

This gives the inferential statistics for the row factors (gender), column factors (location) and the interaction between the two factors.

Since all the p -values exceed 5%, we reject the relevant hypotheses in each case: there are no statistically significant differences by gender between the locations. Similarly, there are no statistically significant differences by location between the genders. Finally, there is no interaction between location and gender.

ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Sample	1.2	1	1.2	0.302521	0.587389	4.259677
Columns	26.86667	2	13.43333	3.386555	0.050638	3.402826
Interaction	20.6	2	10.3	2.596639	0.09531	3.402826
Within	95.2	24	3.966667			
Total	143.8667	29				

25.2. Example.

Suppose an HR director has four different training curricula covering the same subject. Each curriculum is designed so that it can be given either in a group classroom or as a self-paced course, with the self-paced course available in booklet form or online. Thus there are $4 \times 3 = 12$ different combinations of curricula and delivery. The Director wants to know if there are differences between the curricula, between the delivery methods, and if there is any interaction between the delivery method and the curriculum.

Setting	Training Curriculum			
	A	B	C	D
Self-paced, online	123	128	166	151
	156	150	178	125
	112	174	187	117
	100	116	153	155
	168	109	195	158
Self-paced, written	135	175	140	167
	130	132	145	183
	176	120	159	142
	120	187	131	167
	155	184	126	168
Group Class	156	186	185	175
	180	138	206	173
	147	178	188	154
	146	176	165	191
	193	190	188	169

Solution. Running the data analysis tool gives the summary statistics at right. Notice the averages for the curricula range from a low of 146 to a high of 167, and the averages for the delivery modes range from a low of 146 to a high of 174.

The ANOVA section of the tool's output lets us decide if these differences are statistically significant.

Anova: Two-Factor With Replication					
SUMMARY	A	B	C	D	Total
<i>Self-paced, online</i>					
Count	5	5	5	5	20
Sum	659	677	879	706	2921
Average	131.8	135.4	175.8	141.2	146.05
Variance	844.2	707.8	278.7	354.2	782.3658
<i>Self-paced, written</i>					
Count	5	5	5	5	20
Sum	716	798	701	827	3042
Average	143.2	159.6	140.2	165.4	152.1
Variance	498.7	978.3	165.7	217.3	511.0421
<i>Group Class</i>					
Count	5	5	5	5	20
Sum	822	868	932	862	3484
Average	164.4	173.6	186.4	172.4	174.2
Variance	443.3	428.8	212.3	175.8	330.6947
<i>Total</i>					
Count	15	15	15	15	
Sum	2197	2343	2512	2395	
Average	146.4667	156.2	167.4667	159.6667	
Variance	705.8381	871.0286	605.981	404.9524	


ANOVA						
<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Sample	8782.9	2	4391.45	9.933347	0.000245	3.190727
Columns	3411.65	3	1137.217	2.572355	0.064944	2.798061
Interaction	6225.9	6	1037.65	2.347138	0.045555	2.294601
Within	21220.4	48	442.0917			
Total	39640.85	59				

The analysis shows that the differences in the curricula and in the delivery are statistically significant, and that there is statistically significant interaction between the delivery and the curricula.

26. Reliability and Validity of Scales

When we looked at **inter-rater reliability** our focus was on differences in the raters and, to a lesser degree, whether the survey items differentiated between subjects. In this section we change focus to the research instrument itself. Our goal will be to see how well respondent answers “hang together.”

The goal of a survey is to measure a variable or variables that are relevant to your research objectives. Sometimes these variables are easy to **directly** measure—things like income, or blood pressure, or weight loss. Other times, though, variables are more subtle. This section is about using a **group** of survey questions, or **items**, to classify individuals according to psychological or social traits that can’t be directly measured.



When social scientists do this, they develop a set of questions that people in a particular group—say impulsive people or people sharing a particular socio-economic status—will answer in a similar way. Instead of measuring the variable **directly**, then, the measurement is **indirect**.

26.1. Example.


The Human Resources Department at Mechanics R Us develops a set of six questions related to job satisfaction.

	Agree Strongly	Agree	Disagree	Disagree Strongly
No one outside of work cares what I do here.				
I look forward to coming to work every day.				
Friday is the best day of the week.				
The most important thing about my job is that it pays				
At the end of the workday, I feel good about what I've				
People respect what I do at work.				

The Director is interested in whether employees consistently answer these questions.

Solution. In order to answer her questions, the HR Director randomly selects 20 employees and administers the questionnaire. Since some of the questions relate to positive views about work and some about negative views, she **scores** the answers so that **positive views** score higher.

	Agree Strongly	Agree	Disagree	Disagree Strongly
No one outside of work cares what I do here.	0	1	2	3
I look forward to coming to work every day.	3	2	1	0
Friday is the best day of the week.	0	1	2	3
The most important thing about my job is that it pays the bills.	0	1	2	3
At the end of the workday, I feel good about what I've accomplished.	3	2	1	0
People respect what I do at work.	3	2	1	0



The scores can then range from a low of zero to a high of eighteen.
With this scoring, she obtains the following results.

Subject	Q1	Q2	Q3	Q4	Q5	Q6
1	0	0	1	1	0	0
2	2	3	2	3	2	2
3	0	1	1	2	2	1
4	0	0	2	1	1	1
5	2	3	1	2	3	2
6	3	3	2	0	1	3
7	0	0	1	2	3	1
8	2	3	2	2	3	3
9	3	3	2	2	2	2
10	2	3	2	1	2	3
11	3	2	2	3	1	2
12	3	1	3	3	3	0
13	0	0	3	2	0	0
14	3	3	3	3	2	3
15	0	0	1	1	1	2
16	2	0	1	0	0	1
17	0	1	1	0	3	2
18	3	0	3	2	1	3
19	3	2	3	2	2	2
20	3	2	1	2	2	3

Cronbach's alpha is a measure of the internal consistency of this scale. It addresses the question of whether or not respondents are giving consistent answers.

Cronbach's Alpha=	0.705608099										Interpretation	
											Cronbach's Alpha	Internal Consistency
Number of Subjects	60										alpha >= 0.9	Excellent
Number of Questions	6										0.8 <= alpha < 0.9	Good
<p>Do not edit the RED cells. Put your data in the Green Cells. You may have up to 20 items and up to 1000 subjects. Leave BLANK columns for which you have no questions. EXCLUDE subjects who do not have complete responses (i.e., did not answer every question)</p>											0.7 <= alpha < 0.8	Acceptable
											0.6 <= alpha < 0.7	Questionable
											0.5 <= alpha < 0.6	Poor
											alpha < 0.5	Unacceptable
		Question 1	Question 2	Question 3	Question 4	Question 5	Question 6	Question 7	Question 8	Question 9	Question 10	
Item Mean	2.6	2.666666667	2.65	2.75	2.833333333	1.8						
Item StdDev	1.356465997	1.362187783	1.122868351	1.299038106	1.614173335	1.029563014						
Cronbach Alpha w/o this item	0.580655076	0.553811946	0.600484918	0.578961933	0.691063137	0.415061884						

In this case, we have an α of 0.705, which tells us that the internal consistency is acceptable.

Another question of interest to the HR director is whether or not the questions distinguish between the subjects. After all, if they all reported




the same job satisfaction, it wouldn't be a useful scale even it were internally consistent!


To check whether or not the scale differentiates between the employees, we could do an ANOVA, transposing the data so that the employee responses become the columns. If we do this, the resulting p -value is less than 0.01%, so we can be quite confident that the scale differentiates between employees. See the example spreadsheet for this section, which uses the data analysis tool to do the ANOVA.

When we use a scale, we say that the items on the instrument measure a **latent variable**.

As with any measurement, the broad goals in capturing latent variables are:

- 
- *Standardization*—all steps in data collection are standardized and consistent;
 - *Objectivity*—data gathering minimizes subjective biases from the observed and the observer;
 - *Test normalization*—test results from a large group provides the basis for comparison with individual results;
 - *Reliability*—multiple tests give the same conclusions;
 - *Validity*—the data actually measure the intended variable.

For example, suppose your variable is "introversion." You can't just ask someone if they are introverted, since that requires a subjective opinion on the part of the respondent. Instead, you might ask a *series of questions* designed to capture the concept of introversion. In this case, introversion is be a *latent variable* which is captured by a pattern of answers to the questions. Thus, in this case the individual questions, or items on the survey, don't represent variables at all. Instead, taken together, they collectively classify the extent to which individuals exhibit the latent variable. The set of questions used to capture the latent variable are sometimes called a *scale* since they are used to measure the degree to which the subject shares the trait with other people. Early scales for introversion asked questions like the following:

- 
- Do you suddenly feel shy when you want to talk to an attractive stranger?
 - Generally do you prefer reading to meeting people?
 - Do you prefer to have few but special friends?
 - Do you find it hard to really enjoy yourself at a lively party?
 - Do you like talking to people so much that you never miss a chance of talking to a stranger?
 - Can you easily get some life into a dull party?
 - Do you look forward to speaking in public?

You might expect an introvert, for example, to answer "yes" to the first four questions and "no" to the last three, so you would score the scale accordingly, giving a score of "1" to a "yes" on the first four questions and a score of "1" to a "no" on the last three. This set of questions thus appears to be **valid** since they all seem to deal with a particular social anxiety experienced by introverts. If you were testing for "extroversion,"

you'd just reverse the scoring, of course.

One of the ideas with writing scales for latent variables is to repeatedly ask similar questions phrased in different ways, some positive and some negative. Sometimes scales will include other questions as distractors, or even other questions looking for other variables.

Sometimes the latent variable turns out to consist of several less obvious sub-variables or *dimensions*. As an example, Sir Lawrence Olivier was well-known to be painfully shy, a characteristic shared by many introverts. Yet he clearly had no problem speaking in public—that was his career! This famous example illustrates that "communication anxiety" might be one dimension of introversion and "shyness" another. Indeed, it's not all uncommon for latent variables to have more than one dimension.

It's possible to analyze scales for dimensionality using something called factor analysis, but that's beyond the scope of this course.

It's also possible to test for validity. For example, simple inspection of our list of questions above verifies that it includes elements of intro-

verted/extroverted behavior—**content validity**.

Another kind of validity, **construct validity**, involves how well our scale actually measures the variable. A simple way to test for this might involve comparing our proposed scale with another, different scale for the *same trait*. A high correlation would increase our confidence in the validity of both scales.

Another way to test for construct validity might be to see how well responses to our scale correlate with responses to a scale designed to measure a *different* latent variable, say impulsiveness. Presumably, these are *different* variables, so we'd expect the correlation to be low: there's no relation between intro/extroversion and impulsiveness. A high correlation with a scale measuring an ostensibly different trait would thus call into question the validity of our proposed scale.

Finally, our scale should have **predictive** ability, or **criterion validity**. We should be able to use our scale to predict outcomes. A person scoring higher on an extroversion scale, for example, should seek out social situations that provide opportunities for lots of interaction and meeting

new people.

Cronbach's alpha is a particular test for the *reliability* of a scale. It is similar to the analysis of variance in that it compares the variability within the items to the overall variability of the entire scale. Generally speaking, the higher the value of alpha, the more reliable the scale. The generally accepted practice is

$\alpha \geq 0.9$	excellent
$0.9 > \alpha \geq 0.8$	Good
$0.8 > \alpha \geq 0.7$	Acceptable
$0.7 > \alpha \geq 0.6$	Questionable
$0.6 > \alpha \geq 0.5$	Poor
$0.5 > \alpha$	Unacceptable

For reference, the formula for Cronbach's alpha is:

$$\alpha = \left(\frac{k}{k-1} \right) \left(1 - \frac{\sum_{i=1}^k \sigma_{y_i}^2}{\sigma_x^2} \right)$$

where k is the number of items on the scale, $\sigma_{y_i}^2$ are the within-item variances, and σ_x^2 is the total variance.

Cronbach's alpha can be computationally complex, but spreadsheets or statistical programs make it relatively easy to calculate. The [AnalyzeThis](#) spreadsheet included in the course materials has a tab for using Cronbach's alpha to analyze data from a scale. The spreadsheet even includes a calculation of alpha where each item in turn is omitted from the scale. If the alpha is unchanged by omitting an item, it might be redundant and a candidate to remove from the scale.

26.2. Example.

A sample of 60 students take the following survey.

Please circle the answer that best describes your answer.


1. Statistics are dull.
 1. Strongly agree
 2. Agree
 3. Neutral
 4. Disagree
 5. Strongly disagree
2. Statistics are useful.
 1. Strongly agree
 2. Agree
 3. Neutral
 4. Disagree
 5. Strongly disagree
3. Statistics are important to understand.
 1. Strongly agree
 2. Agree
 3. Neutral
 4. Disagree
 5. Strongly disagree
4. People sometimes lie with statistics.
 1. Strongly agree
 2. Agree
 3. Neutral
 4. Disagree
 5. Strongly disagree
5. I love working with numbers.
 1. Strongly agree
 2. Agree
 3. Neutral
 4. Disagree
 5. Strongly disagree

*The results of this survey are stored in the spreadsheet **AnalyzeThis** on the tab **Cronbach**. Do the questions at left appear to all be dealing with the same latent variable? How reliable is this questionnaire? Suggest at least one way to improve the reliability.*

B1 $= (B4 / (B4 - 1)) * (1 - (B17 / B19))$

	A	B	C	D	E	F	G	H	I	J	K
1	Cronbach's Alpha=	0.689371957								Interpretation	
2										Cronbach's Alpha	Internal Consistency
3	Number of Subjects	60								alpha >= 0.9	Excellent
4	Number of Questions	5								0.8 <= alpha < 0.9	Good
5										0.7 <= alpha < 0.8	Acceptable
6										0.6 <= alpha < 0.7	Questionable
7										0.5 <= alpha < 0.6	Poor
8										alpha < 0.5	Unacceptable
6	<p><i>Do not edit the RED cells. Put your data in the Green Cells. You may have up to 20 items and up to 1000 subjects. Leave BLANK columns for which you have no questions. EXCLUDE subjects who do not have complete responses (i.e., did not answer every question)</i></p>										
12		Question 1	Question 2	Question 3	Question 4	Question 5	Question 6	Question 7	Question 8	Question 9	Question 10
13	Item Mean	3.133333333	3.266666667	3.016666667	3.283333333	3.4					
15	Item StdDev	1.071862346	0.997775303	1.072251007	1.126819516	1.704894914					
24	Cronbach Alpha w/o this item	0.618953448	0.58079096	0.577007769	0.561825518	0.839004724					
26		Question 1	Question 2	Question 3	Question 4	Question 5	Question 6	Question 7	Question 8	Question 9	Question 10
27	Subject1	4	3	4	4	1					
28	Subject2	4	3	3	3	5					
29	Subject3	5	5	5	4	5					
30	Subject4	3	4	4	5	5					
31	Subject5	3	4	3	3	1					
32	Subject6	3	3	2	3	5					
33	Subject7	2	3	2	2	5					
34	Subject8	4	5	4	5	1					
35	Subject9	4	5	3	5	5					
36	Subject10	5	3	3	5	5					
37	Subject11	2	2	1	1	1					

Ready Chi-square, 1-way Chi-square, 2-way **Cronbach** +



Cronbach's alpha is not a test for validity. It is also not a test for dimensionality. It only provides guidelines for the reliability of the scale.

27. One Way Tables

So far when we have considered proportions we have considered only two possible outcomes for our experiment: “success” and “failure.” Every member of our population must fall into one of these two categories. Often the population will be far more complex. For example consider the following table which shows the ethnic breakdown of the OU student population (in Fall 1992).

White	Black	Hispanic	Asian	Indian	Other
.758	.061	.023	.035	.047	.076

This is a *one way* table since we have only taken a cross-section of the student population in one way (ethnicity). We could have taken a cross-section in two ways (ethnicity and gender, for example) and produced a

two-way table:

	White	Black	Hisp.	Asian	Indian	Other
m	.402	.030	.014	.019	.022	.054
f	.356	.031	.009	.016	.025	.022
Tot	.758	.061	.023	.035	.047	.076

Initially we will deal only with one-way tables; the analysis of two way tables is similar. Consider the following example.

27.1. Example.

A random sample of 483 students is taken from the OU student body. It is found that this sample has the following ethnic breakdown:

	White	Black	Hispanic	Asian	Indian	Other
#	386	24	27	12	20	14
%	79.9	5.0	5.6	2.5	4.1	2.9

Does this sample differ significantly ($\alpha = 0.05$) from the overall student population with respect to ethnicity?

Notice that the sample has more whites and Hispanics than we might have expected and fewer of the other classifications. The question is whether this is due to some systematic error in the way the sample was taken or can be attributed to the natural random errors implicit in sampling. The Chi-squared test (χ^2 test) is a way of answering this question. In this case we are actually testing a hypothesis (note the word *significant*):

H_0 : sample is not biased with respect to ethnicity

against

H_A : sample is biased

We will step through the solution to this problem by way of introduction to the chi-squared test.

Solution. From our data we have an *outcomes* table:

	White	Black	Hispanic	Asian	Indian	Other
#	386	24	27	12	20	14

Based on what we know about the population, we can construct an *expectations* table:

	White	Black	Hispanic	Asian	Indian	Other
#						
	.758	.061	.023	.035	.047	.076

The first row in the expectations table is initially blank; the second row is the proportion of the of population which falls in each category. We know what proportion of the sample we would *expect* to fall in each category from the census data on the population. To fill in the first row (# row) in the expectations table, multiply the expected proportion times the total number in the sample. Thus the expected number of whites in the sample should be:

$$\text{expected whites} = 0.758 \times 483 = 366.11$$

and the expected number of Blacks in the sample should be:

$$\text{expected whites} = 0.061 \times 483 = 29.46$$

Continuing in this fashion, we can fill in the # row in the expectations table:

	White	Black	Hisp.	Asian	Indian	Other
#	366.11	29.46	11.11	16.91	22.70	36.71
	.758	.061	.023	.035	.047	.076

The chi-squared test now compares the actual outcomes with these expected outcomes by computing the following test statistic:

$$\chi^2 = \sum \frac{(\text{Observation} - \text{Expectation})^2}{\text{Expectation}}$$

Since we have six cells (ethnicities), the sum will have six terms. For

this problem the sum is:

$$\begin{aligned}\chi^2 &= \sum \frac{(\text{Observations} - \text{Expectations})^2}{\text{Expectations}} \\ &= \frac{(386 - 366.11)^2}{366.11} + \frac{(24 - 29.46)^2}{29.46} + \dots \\ &\dots + \frac{(27 - 11.11)^2}{11.11} + \frac{(12 - 16.91)^2}{16.19} + \dots \\ &\dots + \frac{(20 - 22.70)^2}{22.70} + \frac{(14 - 36.71)^2}{36.71} \\ &= 1.08 + 1.01 + 22.73 + 1.43 + 0.32 + 14.05 \\ &= 40.62\end{aligned}$$

As usual, though, we have a spreadsheet that does all of the above for us. In this case, the spreadsheet is FORMULAS.XLSX and the tab is Chi-square, 1-way. It's only necessary to enter the basic data and the

census data:

	White	Black	Hispanic	Asian	Indian	Other
sample	386	24	27	12	20	14
census	.758	.061	.023	.035	.047	.076

The first line represents the actual counts of each ethnicity in our sample, while the second line represents the proportion of each ethnicity in the actual population, based on a census. With this information, the spreadsheet computes everything for us, and provides a p -value to test the null hypothesis against the alternative hypothesis.


Enter the summary data into the spreadsheet FORMULAS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled Chi-square, 1-way.

The screenshot shows a Microsoft Excel spreadsheet titled 'Formulas.xlsx'. The active cell is B4, which contains the value 0.758. The spreadsheet displays the following data:


Observations	Class A	Class B	Class C	Class D	Class D	Class E	Totals
Sample	386	24	27	12	20	14	483
Census	0.758	0.061	0.023	0.035	0.047	0.076	1
Observed	0.799172	0.049689	0.055901	0.024845	0.041408	0.028986	1
Expected	366.114	29.463	11.109	16.905	22.701	36.708	483

Test Statistic	26.56911
Degrees of Freedom	5
p-value	0.0069%

The spreadsheet also shows a navigation bar at the bottom with tabs for 'Regression', 'Chi-square, 1-way', and 'Chi-square, 2-way'. The 'Chi-square, 1-way' tab is currently selected.



From the above, we see that the p -value is 0.0069%, so we have highly significant data to support the alternative hypothesis, that the sample is biased.



28. Two Way Contingency Tables

Sometimes the population can be partitioned on two directions. Consider the following table:

	Cancer	Heart Disease	Other
Smoker	135	310	205
Nonsmoker	55	155	140

The table lists the causes of death from 1000 randomly selected death certificates. Note that there are actually six categories instead of the two we have considered up to now. In addition, the categories themselves are in two groups: smokers versus nonsmokers and three different causes of death (so this is a 2×3 table). Each group (smok-

ers/nonsmokers and causes of death) are collectively exhaustive: each member of the population was either a smoker or not and (since we are sampling death certificates) each had a cause of death. We might be interested in testing the null hypothesis:

H_0 : Cause of death is unrelated to whether or not someone smokes.
against

H_A : Cause of death is related to smoking.

The Chi squared test (χ^2 test) is a vehicle for doing this. The basic setup is a contingency table like the one above with one or more rows and columns. The rows partition the population with respect to one variable (for example smoking) and the columns partition the population with respect to another variable (for example cause of death). The Chi squared statistic is a way of determining if the row effects and column effects are independent of one another. The more general form of the null and alternative hypotheses are:

H_0 : Row effects and column effects are independent.
against

H_A : Row and column effects are dependent on each other.
In order to understand the difference between “dependent” and “independent” events, consider some examples.

28.1. Example.

Suppose that a clinic treats a total of 150 patients this month. Some of the patients are adults (over 18) and some are children. Some of the conditions involve trauma (bruises, broken bones or other injuries) and some involve other conditions. A contingency table of outcomes might look like:

	<i>trauma</i>	<i>non-trauma</i>
<i>adult</i>	20	30
<i>non-adult</i>	40	60

Totaling the rows and columns gives a slightly better view of the data:

	trauma	non-trauma	total
adult	20	30	50
non-adult	40	60	100
	60	90	150

Forty percent of all patients ($\frac{60}{150} \times 100\%$) were seen for trauma; forty percent of all adults ($\frac{20}{50} \times 100\%$) were seen for trauma; forty percent of all children ($\frac{40}{100} \times 100\%$) were seen from trauma. If we randomly select a patient file and discover that the patient was seen for trauma, we can't deduce from this information that it was more or less likely that the patient was an adult. Similarly, if we randomly select an adult patient, we can't deduce whether or not the patient was seen for a trauma-related condition.

Now let's suppose that we select a different group of 150 patients with a slightly different distribution.

28.2. Example.

Suppose that the clinic treated a total of 150 patients two months ago. Some of the patients are adults (over 18) and some are children. Some of the conditions involve trauma (bruises, broken bones or other injuries) and some involve other conditions. Suppose that the contingency table of outcomes looks like:

	<i>trauma</i>	<i>non-trauma</i>	<i>total</i>
<i>adult</i>	10	40	50
<i>non-adult</i>	50	50	100
	60	90	150

Notice we have the same total number of patients, total number of adults, total number of non-adults, total number of trauma patients and total number of non-trauma patients. However, the distribution inside the cells of the contingency table is now different. Fifty percent of the

children are seen for trauma whereas only 20 percent of the adults are seen for trauma. In this second sample, *age* and *type of condition* are dependent.

Of course, data are rarely as decisive as in the above examples. The chi-squared test is a way to decide if an appearance of non-independence in the data is statistically significant. To see how the test works, let's reconsider the smoking data.

28.3. Example.

A researcher randomly selects 1000 death certificates and, after interviewing the attending physician, records the following information about the deceased:

	<i>Cancer</i>	<i>Heart Disease</i>	<i>Other</i>
<i>Smoker</i>	135	310	205
<i>Nonsmoker</i>	55	155	140

At a significance of level of 5%, do these data show that smoking and cause of death of dependent?

Note: the data can't show that smoking *causes* death since everyone in the sample is already dead. What the data can show is that dying of cancer or heart disease is related to whether or not the deceased smoked.

Following the same process that we used for the emergency room above, we could produce an *observations* table.

	<i>Cancer</i>	<i>Heart Disease</i>	<i>Other</i>	<i>totals</i>
<i>Smoker</i>	135	310	205	650
<i>Nonsmoker</i>	55	155	140	350
<i>totals</i>	190	465	345	1000

Next build an *expectations* table:

	<i>Cancer</i>	<i>Heart Disease</i>	<i>Other</i>	<i>totals</i>
<i>Smoker</i>				<i>650</i>
<i>Nonsmoker</i>				<i>350</i>
<i>totals</i>	<i>190</i>	<i>465</i>	<i>345</i>	<i>1000</i>
<i>proportions</i>	<i>.19</i>	<i>.465</i>	<i>.345</i>	

The *cells* of the expectations table are initially blank; you have to fill them in with computations. In addition, a new row (proportions) has been added. In the first column, this is the proportion of all deaths attributed to cancer ($\frac{190}{1000}$); in the second column, the proportion of all deaths attributed to heart disease ($\frac{465}{1000}$); in the third column, the proportion of all deaths attributed to other causes ($\frac{345}{1000}$). In each case, the proportion row is filled in with the formula

$$\frac{\text{column total}}{\text{overall total}}$$

You use the proportion row to fill in the cells. The number which goes in the cells is what *you would expect the result to be if the row and*

column effects were independent. Since 19% of all deaths were attributable to cancer, if “cancer” and “smoking” were unrelated, we would expect that 19% of all smokers’ deaths would be caused by cancer. Thus, the upper right cell in the expectations table is

$$19\% \text{ of } 650 = 123.5$$

Similarly, the upper middle cell in the expectations table is

$$46.5\% \text{ of } 650 = 302.25$$

and the upper left cell is

$$34.5\% \text{ of } 650 = 224.25$$

More generally, the cells in the expectations table are filled in as follows:

Expectations Table

	Cancer	♡ Disease	Other	totals
Smkr	$.19 \times 650$	$.465 \times 650$	$.345 \times 650$	650
NonSmkr	$.19 \times 350$	$.465 \times 350$	$.345 \times 350$	350
totals	190	465	345	1000
prop's	.19	.465	.345	

which results in an expectations table which looks like:

	<i>Cancer</i>	♡ <i>Disease</i>	<i>Other</i>	<i>totals</i>
<i>Smoker</i>	<i>123.5</i>	<i>302.25</i>	<i>224.25</i>	<i>650</i>
<i>Nonsmoker</i>	<i>66.5</i>	<i>162.75</i>	<i>120.75</i>	<i>350</i>
<i>totals</i>	<i>190</i>	<i>465</i>	<i>345</i>	<i>1000</i>
<i>proportions</i>	<i>.19</i>	<i>.465</i>	<i>.345</i>	

Notice that the rows and columns still add up to the marginal totals. This table gives what we would *expect* to observe if the row and column effects were independent. Notice that this differs from our actual observations:

	<i>Cancer</i>	<i>Heart Disease</i>	<i>Other</i>	<i>totals</i>
<i>Smoker</i>	135	310	205	650
<i>Nonsmoker</i>	55	155	140	350
<i>totals</i>	190	465	345	1000

Next we need a rule to decide if the differences between the observations and the expectations are statistically significant. The next step in the process is to compute the test statistic:

$$\chi^2 = \sum \frac{(\text{Observations} - \text{Expectations})^2}{\text{Expectations}}$$

the sum being taken over each data cell in the contingency tables. Fortunately, there is a spreadsheet to do all of this for us. For completeness, though, let's step through the computations that are hidden inside the spreadsheet. In our example there are six terms to sum:

$$\begin{aligned}
\chi^2 &= \sum \frac{(\text{Observations} - \text{Expectations})^2}{\text{Expectations}} \\
&= \frac{(135 - 123.5)^2}{123.5} + \frac{(310 - 302.25)^2}{302.25} + \dots \\
&\dots + \frac{(205 - 224.25)^2}{224.25} + \frac{(55 - 66.5)^2}{66.5} + \dots \\
&\dots + \frac{(155 - 162.75)^2}{162.75} + \frac{(140 - 120.75)^2}{120.75} \\
&= 1.07 + 0.199 + 1.652 + 1.989 + 0.36 + 3.069 \\
&= 8.349
\end{aligned}$$

As usual, we must now compare the value of the test statistic against a cutoff which we find in a table. The test statistic in this case is not normal however: it is a “chi-squared” statistic which is tabulated on page 666 (Table A-4) in your text. In order to use the table, you need to know the

degrees of freedom for the test statistic. This is computed by

$$\text{degrees of freedom} = (\# \text{ of rows} - 1) \times (\# \text{ of cols} - 1)$$

Thus in our problem the degrees of freedom are

$$(2 - 1) \times (3 - 1) = 2$$

The degrees of freedom tell you the *row* in the table in which you need to look. The entries across the top correspond (for this type of problem) to the significance level. Thus the cutoff for this problem is 5.991. This cut-off corresponds to the pre-set significance level of 5%, but our test statistic is larger, so the associated *p*-value would be less. As a consequence, we'd reject the null hypothesis that the variables are independent and conclude that they are dependent. This is, of course, much easier with the spreadsheet.

Solution.

Step 1. Enter the summary data into the spreadsheet FORMULAS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled Chi-square, 2-way.

Observations	Class A	Class B	Class C	Class D	Class D	Totals
Group 1	135	310	205			650
Group 2	55	155	140			350
Group 3						0
Group 4						0
Group 5						0
Group 6						0
Totals	190	465	345	0	0	1000
Proportion	0.19	0.465	0.345	0	0	1
0.10%						
Expected	Class A	Class B	Class C	Class D	Class D	Totals
Group 1	123.5	302.25	224.25	0	0	650
Group 2	66.5	162.75	120.75	0	0	350
Group 3	0	0	0	0	0	0
Group 4	0	0	0	0	0	0
Group 5	0	0	0	0	0	0
Group 6	0	0	0	0	0	0
Totals	190	465	345	0	0	1000
Test Statistic		8.348631				
Degrees of Freedom		2				
p-value		1.5386%				

Step 2. The p -value is 1.5386%, so we have *significant* (but not

highly significant) evidence that heart disease and smoking are *dependent* variables.

Notes:

1. The chi-squared statistic has other uses than the one described in this section. *Not every application of the chi-squared involves two-way contingency tables.*
2. In this unit our tests have not involved parameters (means, standard deviations) but instead have involved categories. The hypotheses related to issues of dependence or independence rather than magnitudes of parameters. For this reason, these kinds of tests are called *non-parametric*.

28.4. Example.

In a study of heart disease among males, the 356 subjects were classified according to socioeconomic status and smoking habits. The study recognized three levels of socioeconomic status (high, middle and low) and three smoking categories (current smoker, never smoked, former smoker). The data are summarized in the following contingency table:

	<i>high</i>	<i>middle</i>	<i>low</i>
<i>current</i>	51	22	43
<i>former</i>	92	21	28
<i>never</i>	68	9	22

At the 5% significance level do the data show that smoking habits and socioeconomic status are dependent or independent?

Solution.

Step 1. Enter the summary data into the spreadsheet FORMULAS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled Chi-square, 2-way.

Observations	Class A	Class B	Class C	Class D	Class D	Totals
Group 1	51	22	43			116
Group 2	92	21	28			141
Group 3	68	9	22			99
Group 4						0
Group 5						0
Group 6						0
Totals	211	52	93	0	0	356
Proportion	0.592697	0.146067	0.261236	0	0	1
Expected						
Group 1	68.75281	16.94382	30.30337	0	0	116
Group 2	83.57022	20.59551	36.83427	0	0	141
Group 3	58.67697	14.46067	25.86236	0	0	99
Group 4	0	0	0	0	0	0
Group 5	0	0	0	0	0	0
Group 6	0	0	0	0	0	0
Totals	211	52	93	0	0	356
Test Statistic		18.50974				
Degrees of Freedom		4				
p-value		0.0981%				

Step 2. The p -value is 0.0981%, so we have *highly significant ev-*

idence that smoking habits and socioeconomic status are *dependent* variables.

Two-way tables can also be used to do hypothesis tests for proportions:

$$H_0 : p_E = p_C \quad \text{against} \quad H_A : \begin{cases} p_E > p_C & \text{or} \\ p_E < p_C & \text{or} \\ p_E \neq p_C \end{cases}$$

In this case, we'd have two rows and two columns:

	<i>Experimental Group</i>	<i>Control Group</i>
<i>Number of Successes</i>		
<i>Number of Failures</i>		

and thus the test has one degree of freedom. This approach is slightly different from the one we used earlier, where we tested to see if two **parameters** were different, while the Chi-squared tests for **independence**.

28.5. Example.

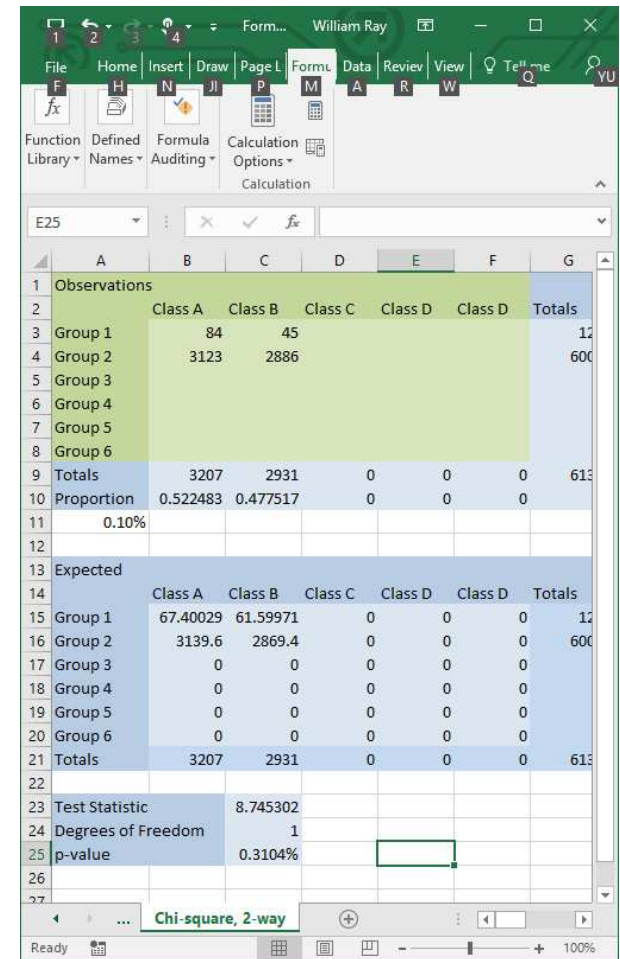
A large Midwestern hospital tracked the 12-month survival rates for persons who were treated for cardiac arrest in the hospital ER. The hospital gathered the following data.

	<i>Non-Smokers</i>	<i>Smokers</i>
<i>Survived at least 12 month</i>	84	45
<i>Deceased within 12 months</i>	3123	2886

Is there a statistically significant difference in the 12-month survival rates for smokers and non-smokers?

Solution.

Now enter the list into the spreadsheet FORMULAS.XLSX, using the tab at the bottom labeled Chi-square, 2-way. The reported p-value of 0.31% means we reject the null hypothesis that the proportions—i.e., survival rates—are the same for the two groups.



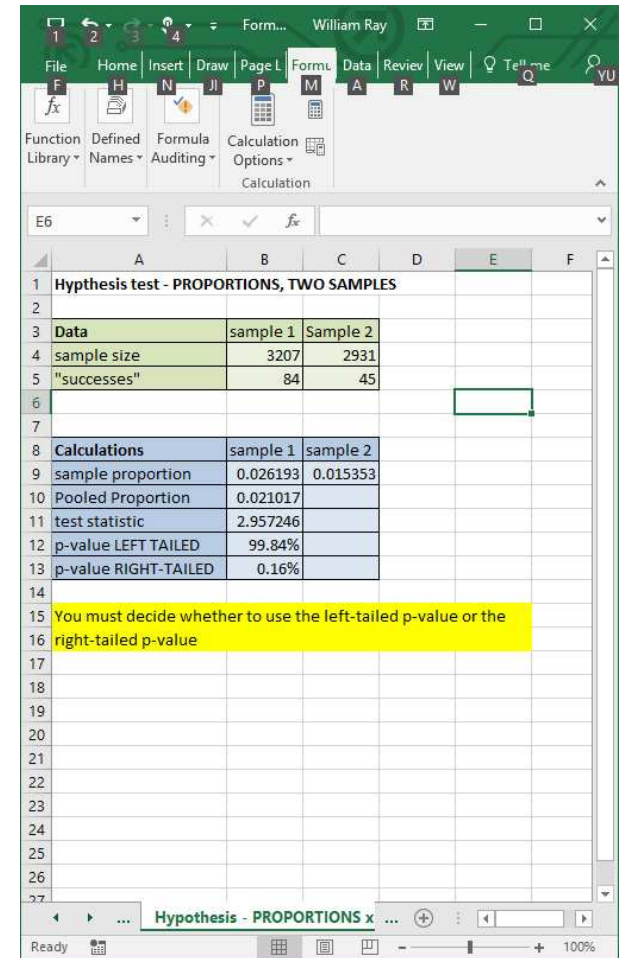
The screenshot shows an Excel spreadsheet with the following data:


Observations	Class A	Class B	Class C	Class D	Class D	Totals
Group 1	84	45				129
Group 2	3123	2886				6009
Group 3						
Group 4						
Group 5						
Group 6						
Totals	3207	2931	0	0	0	6138
Proportion	0.522483	0.477517	0	0	0	
	0.10%					
Expected	Class A	Class B	Class C	Class D	Class D	Totals
Group 1	67.40029	61.59971	0	0	0	129
Group 2	3139.6	2869.4	0	0	0	6009
Group 3	0	0	0	0	0	0
Group 4	0	0	0	0	0	0
Group 5	0	0	0	0	0	0
Group 6	0	0	0	0	0	0
Totals	3207	2931	0	0	0	6138
Test Statistic		8.745302				
Degrees of Freedom		1				
p-value		0.3104%				

You can do the same test using the tab labeled Hypothesis - PROPORTIONS x2 and obtain a similar but not quite identical solution. Remember, the tests are not quite the same, one being parametric and the other non-parametric. To do the earlier parametric test, you need to know the sample sizes rather than the number in success/fail category:

	<i>Non-smokers</i>	<i>Smokers</i>
<i>Sample size</i>	<i>3207</i>	<i>2931</i>
<i>Survival Rate</i>	<i>84</i>	<i>45</i>

From the spreadsheet, this gives a p-value of 0.16%.





Both the Chi-squared and the normal test *approximate* nominal (attribute) data with a continuous (numerical) distribution (the normal distribution). The relative accuracy of this approximation depends on several factors, including the expected and observed cell frequencies and how "far" the true value of the population proportion (assuming the null hypothesis) is from 0.5. There is a more exact test due to Fisher that is not covered in this class, but for most applications either the normal or Chi-squared approach provides satisfactory results. par