Inferential Statistics

I. Logic

How likely are the results if only random fluctuations (chance effects) are occurring? The *Null Hypothesis* is the hypothesis that differences in the data reflect only random variation.

If the likelihood of the results being due to chance is sufficiently low, then we reject the *Null Hypothesis* (that the results are due to chance) and accept that systematic variation (the *Research Hypothesis*) produced the observed results.

We *can not test* the Research Hypothesis directly because we do not know the true state of the world. We can infer that the Research Hypothesis is likely to be correct if we can show that the Null Hypothesis is *very unlikely to be correct*.

When we say that a result is statistically significant, what we mean is that it is **VERY UNLIKELY** that the results are due to chance.

II. A Simple Example

In a simple test of reading, participants read a passage and then answer questions about what they had read. One problem in developing such a test is that it may be possible to correctly answer the questions without reading the passage. It is also possible to guess the correct answer.

To assess this, we would give people the questions and ask them to choose the correct answers without letting them read the passages.

If the questions are 4 alternative, multiple choice then chance (guessing) on each question is 0.25 (one in four). So, in the long run (averaged over many participants), we expect people to get 25% (25 out of every 100 or 5 out of every 20) questions correct by chance.

When we run a few participants with the questions, we are only getting a limited set of data. We want to know the odds that the set of data that we obtain reflects chance performance.

A) Simple Case

One person answers 10 questions. If they got three of 10 correct, we probably wouldn't be very worried. After all, this is very close to the long term expected 2.5 of 10 (1 out of 4) by chance alone.

If this one person gets 8 of 10 correct, we worry. This is because the odds of getting 8 or more (that is, 8, 9, or 10) correct are less than 0.001. That is, if this person did the 10 questions 100,000 times, they would only score 8 or better on 100 trials by chance. It is VERY UNLIKELY that someone would get 8 of 10 correct by chance.

We get this probability from the binomial distribution. Basically, we calculate all the different ways that the person could answer the 10 questions. The proportion of these that yield 2 correct answers represents the probability of getting two correct. The odds on getting 8 or more correct are very small.

B) A slightly more complex case

Instead of 1 person, we have 20 people each answer the 10 questions. They average 4 correct. While the odds of one person getting 4 or more correct are greater than 1 in 10, what about with 20 people averaging 4 out of 10?

The odds of 20 people averaging 4 correct are less than 5 in 100. If we use more samples (people), we get a more accurate picture of what is really going on. This, in turn, means that the size of the difference between our result (4 of 10 in this case) and chance (2.5 of 10) can be smaller yet still be significant.

In order to evaluate whether our results are due to chance, we need to include the number of observations (in this case, participants) in our evaluation.

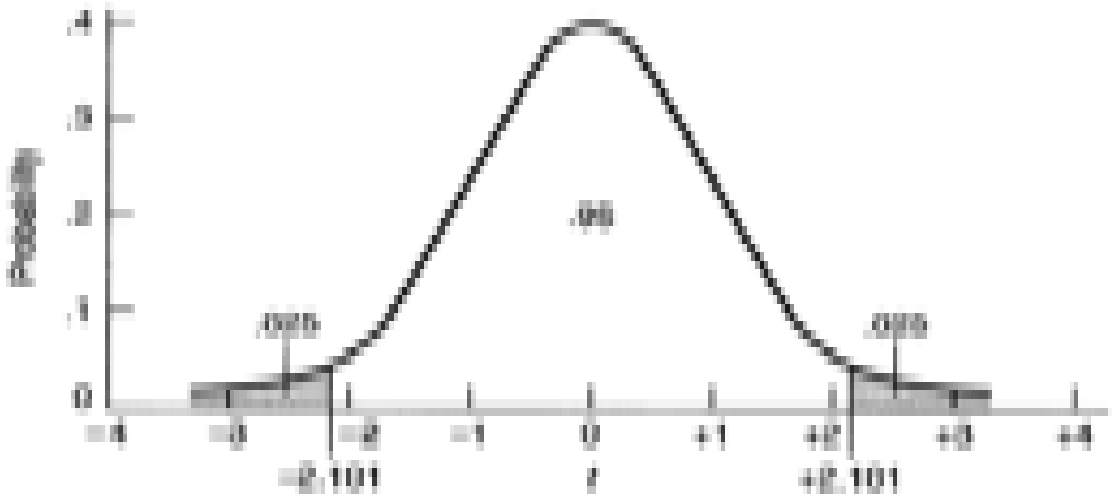
C) A more realistic case

Our last example uses the t-test to assess significance. The t-test is a test of means. Basically, it is used to assess whether two means are the same (Null Hypothesis of no difference) or not (Research Hypothesis of a difference).

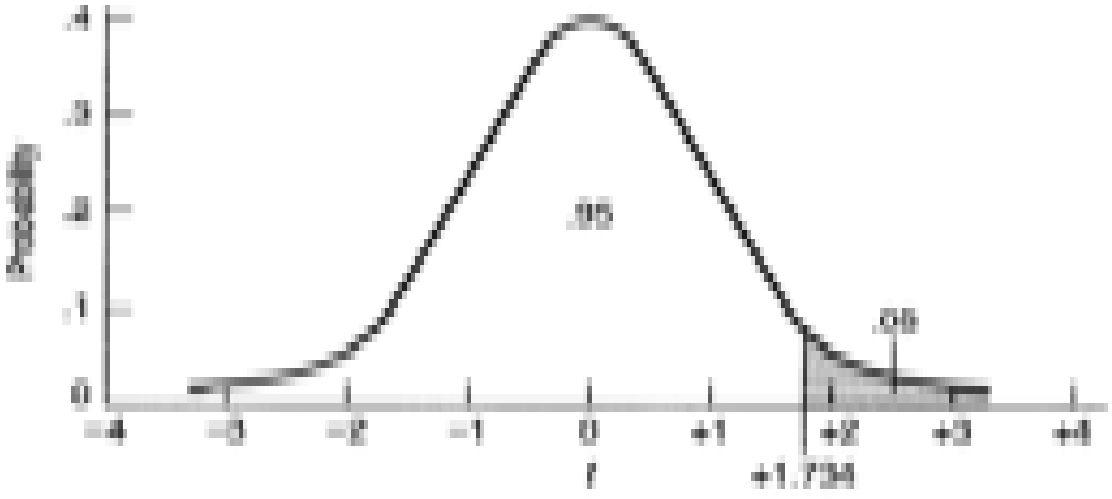
In our example, we are comparing the performance of our 20 participants against "chance" and asking whether the participants are consistently scoring above chance. We would subtract the expected value due to chance from each participants score and compute the average difference (relative to chance) for our participants. This is divided by a measure of the variability in the data.

This computation yields a "statistic" that has a know distribution if the effects are due to chance. The statistic that we have computed indicates where our data fall in this distribution.

This is the distribution of the t statistic.



Critical value for two-tailed test with .05 significance level



Critical value for one-tailed test with .05 significance level

Using the distribution shown above, 2.5% (.025) of the differences between two means should have a *t*-value greater than 2.101 by chance.

Similarly, 2.5% of the differences should be less than –2.101

That is, if our results of computing *t* produce a value greater than 2.101 or less than –2.101, then our results are extreme (relative to chance). Consequently, the probability that our results are due to chance is low (less than .05 or 5% in this case). We would reject the Null Hypothesis.

This example uses what is called a two-tailed test because we will accept an extreme difference at either end of the distribution. (The distribution at the top shows the extremes for both directions.)

Degrees of Freedom. All statistical tests use degrees of freedom as part of the determination of the probability that the data reflect chance (random variation).

Degrees of freedom is determined from the number of independent observations. Basically, it reflects the number of participants in the study (more participants means more degrees of freedom).

III. Type I and Type II Errors

A) The Basics

Abbreviate Null Hypothesis as H_0 and our Research Hypothesis that the manipulation has an effect or that there is a relationship among the variables as H_1.

Then, depending upon the decision we make about accepting or rejecting H_0, there are two ways we can be right, and two ways we could be wrong.

	Reality	
Decision	H_0 true	H_1 true
Accept H_0	*correct*	Type II error
Accept H_1	Type I error	*correct*

B) Type I Errors

The probability of a Type I error is known. This comes from the knowledge of the distribution of effects due to chance.

The likelihood that the results are due to chance is called the alpha level. If alpha is less that .05, (one in twenty), we reject the Null Hypothesis. If alpha is greater than .05, we do not reject the Null Hypothesis. We also do not accept the Null Hypothesis, because we do not know the probability that we could be wrong (the probability of a Type II error).

B) Type II Errors

The probability of a Type II error is unknown.

In order to know it, we would have to know how the variables (in our data) are related in the population. However, if we already knew this, we wouldn't be doing the research!!

We do know that as we make the probability of a Type I error smaller by using a lower alpha, the probability of a Type II error increases. Consequently, if we set our significance level too strictly, we risk missing real effects.

The significance level of .05 is set as a reasonable compromise between making Type I errors and making Type II errors.

We also know that as we increase the number of participants in our study, the probability of a Type II error decreases. This is because a larger sample is more likely to be similar to the population.

C) Real World Example 1 - Medicine

You see the doctor about a pain in your lower abdomen. The doctor is faced with a decision about which of two states of the world is true *given* your symptoms (the data):

1. You have appendicitis
2. You do not have appendicitis

Of course, the doctor needs to make a decision and recommend whether to operate to remove your appendix. There are four possible outcomes:

	Reality	
	H_0 true	H_1 true
Decision	(no appendicitis)	(appendicitis)
Accept H_0 (no operation)	correct	Type II error (patient dies)
Accept H_1 (operate)	Type I error (risk dying)	correct

Clearly, in this example, one error has a larger negative consequence than the other. *If you do have appendicitis and you do not have treatment (the operation to remove your appendix), you will die*. Doctors generally err on the side of caution. That is, they risk a Type I error to avoid Type II errors.

D) Real World Example 2 – The Legal System

A jury hearing a case is charged to decide whether a defendant is guilty (beyond a reasonable doubt) or not guilty. Again, there are two states of the world and two decisions leading to one of four possible outcomes.

	Reality	
Decision	H_0 true (innocent)	H_1 true (guilty)
Accept H_0 (not guilty)	correct	Type II error (guilty goes free)
Accept H_1 (guilty)	Type I error (innocent convicted)	correct

The instructions to juries are designed to reduce or minimize Type I errors. This approach is similar to science where we also set a low level for Type I errors.

IV. Related Issues

A) How Many Participants are Enough?

This can be determined two ways:

1. What have previous studies used? Was it sufficient to find significant differences? Is this study using a more or less sensitive dependent variable, a more or less sensitive design? (With greater sensitivity, fewer participants are needed.) Is this study doing a stronger or weaker manipulation of the IV? Are the effects that we are looking for small (more participants) or large (fewer participants).

2. Power analysis. Based on the likelihood of a Type II error that you choose and the likely size of the effects that you are investigating, a formula can be used to determine the number of participants.

B) Why do we use a **.05** level for determining that results are significant?

This level is a compromise between the likelihood of a Type I error and a Type II error. We do not know the probability of a Type II error. However, we do not want to reject the Null Hypothesis when it is reasonably possible that it is true. We use the relatively conservative significance level of .05 so that we do not reject the Null Hypothesis prematurely.

If we decide not to reject the Null Hypothesis, we can always run the study again to see if we get the same results (*Experimental Reliability*).

Note, in contrast, that the appendicitis and jury cases above do not have the option of replication.

C) Interpreting Non-significant and Significant Results

1. Non-significant Results.

In general, when results are not significant, we do not conclude that the Null Hypothesis is true. This is because *we do not know* the probability that we could be making a Type II Error.

Studies can fail to find significant effects because of confusing instructions, unmotivated participants, weak IV manipulations, an insensitive DV, etc.

However, when a lack of significant effects is confirmed by replication so that we see a pattern of non-significant effects, then we are fairly confident that there is no effect.

Finally, what are the costs of not making a decision versus accepting the Null Hypothesis? (See the examples of Type I and Type II errors above.)

2. Significant Results

Just because the effect of a variable is significant does NOT mean that the relation is strong or important.

A low but significant correlation means that the relation is weak and other factors are probably more important. A new medical treatment can improve health by a small amount but be very costly. Is the treatment worth the cost?

Conversely, small effects can be very important. Suppose that improvements that cost little in a benefits package at a large company decrease employee turnover by 1%. This is a small effect. However, if the company has 10,000 employees who leave every year and it costs $10,000 to train a new employee, then reducing the turnover by 1% saves $1,000,000 per year (100 employees times $10,000 per employee). If the benefits changes cost less than $1,000,000 per year then the company saves money.

V. Statistical Tests

A) The sign test.

This is the test used in our first example. It assesses whether a set of observations deviate from the rate expected by chance.

For our example, we'll take a set of data that are based on nominal scale measurements. A Psychology Department has 27 faculty, 16 males and 11 females.

If the larger population of Ph.D.s in psychology is 50% female and 50% male, then we might expect that the faculty at the university would be 50-50 also. Is the actual proportion reliably different from 50-50?

Our H_1 is that the proportion is different from 50-50 and our H_0 is that it is not different from 50-50.

What we want are the odds that a split of 16 and 11 is different from 50-50? If the probability for any one faculty member being male or female is 50-50, this is like flipping a coin. We can re-phrase our statistical test as "What are the odds that a fair coin, flipped 27 times, would come up heads on 11 (or fewer) flips?"

This probability can be calculated. It is based on permutations and combinations and comes from the binomial distribution.

In this case, the probability that 16-11 is different from 50-50 is greater than .10 so that the evidence says that we can not conclude that the department is significantly different from 50-50.

B) Parametric versus Nonparametric Tests

There are two classes of tests: parametric and nonparametric.

The parametric statistics (*t*-test, Analysis of Variance, chi square) make assumptions about the shape of the distribution of results based on chance. Most assume interval or ratio scale data. The chi square can be used with any scale.

Nonparametric tests (e.g. Mann-Whitney U, sign test) do not make the same distribution assumptions and some can be used with nominal and ordinal data.

C) Two parametric tests are the *t*-test and Analysis of Variance:

1. The *t*-test is designed to indicate whether two means are the same. It does this by comparing the size of the difference between the means to the variability within each condition in the data. If the mean difference is large, relative to the variability, then the test will indicate that the means are significantly different.

There are two versions of this test. When the data for the two means come from different participants, this is the independent samples t-test. The other is for when each participant provides data for both means. In this case, the difference between the scores is determined for each individual and we test whether this difference is reliably different from zero. This is the dependent (repeated measures) t-test.

The two tests compute the variability in the data differently.

2. Analysis of Variance (ANOVA) involves comparing variability between different conditions (*systematic* variance) to the variability within conditions (*error* variance). When the ratio (also called F) is high (systematic variance is larger than the error variance) we have significant effects.

The advantage of an ANOVA is that it can compare more than two conditions at one time. In essence, the test is examining all of the conditions on one variable at one time. If there is a reliable difference across them, then there will be a significant effect for that variable.

The ANOVA can also be used to assess whether interactions between variables are reliable (factorial designs).

D) Other tests

1. The Chi-Square can be used with any data scale. It assess whether the frequency (or proportion) of responses in each category matches a particular profile.

We could use a Chi-Square to assess whether the 16-11 split of male and female faculty in the psychology department deviates significantly from the hypothetical 50/50 ratio.

2. For a factorial design, the Chi-Square can be used with nominal scale data. There is no test for ordinal data. For an interval or ratio scale, analysis of variance is used.

VI. Reliability, Replication, and Converging Operations

The alternative to using inferential statistics to establish reliability is to repeat the study.

A) Reliability and Replication

If someone conducted an opinion survey on preferences for president, opinion on tax cuts, etc., what would give you confidence in the results?

1. A larger sample. More likely to generalize to the target population.

2. Similar results from a second sample (repeatability).

Repeating an experiment, survey, etc. and getting the same or similar results increases our confidence that the results are reliable.

Since we can not repeat a study exactly in most cases (we usually can not run the same subjects again), repeating a study shows if the results are consistent across sets of participants. If we get the same or similar results, we have demonstrated *experimental reliability*.

Experimental reliability is preferred over statistical reliability. Statistical reliability is based on the odds that our results are due to chance. Put another way, even significant effects can still occur by chance.

If we replicate a study, it is unlikely that all of the other, extraneous (irrelevant) factors will be exactly the same. The results the second time are likely to be somewhat different. However, if both results are very similar, then they probably are an accurate reflection of reality.

Thus, *experimental reliability*, in which the pattern of results from one study recurs in additional studies, is preferred to statistical reliability.

Of course, there is a cost to experimental reliability. We have to run the study a second (or more) time(s).

There are three types of replication of a study:

1. Direct (repeat the same study)
2. Systematic (vary other factors)
3. Conceptual (use new operational definitions)

We will defer conceptual replication to the next topic (chapter).

B) Direct replication. Repeat a previous study with new subjects. Keep the changes between studies to a minimum.

Example: Luchins (1942) studies on "set effects" (Einstellung) in problem solving. In these studies, Luchins studied whether people would use already known solutions to deal with new problems, even when a better or simpler solution was possible. That is, would subjects develop an approach or "mind set" about the problems which would prevent them from finding other solutions?

Luchins used water-jar problems in his studies. In these problems, the person must obtain the correct amount of water by filling jars of a known capacity. For example, if given jars of capacity 2, 7, and 12 pints, and asked to come up with 10 pints, you could fill the 12 pint jar and then pour from it into the 2 pint jar and fill it (the 2 pint jar). The water remaining in the 12 pint jar is now 10 pints.

In Luchins' experiments, subjects watched the researcher solve a problem then they did a series of problems.

	Empty Jars			
Problem	A	B	C	Goal
1	29	3		20
2	21	127	3	100
3	14	163	25	99
4	18	43	10	5
5	9	42	6	21
6	20	59	4	31
7	23	49	3	20
8	15	39	3	18
9	28	76	3	25
10	18	48	4	22
11	14	36	8	6

In the demo, A - 3B yields the goal.

The subject discovers that B - A - 2C solves problem 2, and also problems 3, 4, 5, and 6.

The subjects also use this for problems 7 and 8 even though A - C and A + C give more direct solutions.

Finally, even after solving problem 9 (which requires a different solution), subjects returned to the old solution for problems 10 and 11.

If the same problems are presented in a different order, then different solutions were found. For example, a control group run with problems 7 and 8 first found the more efficient solutions and none of the subjects used the less efficient solution of the experimental group.

A group of subjects run exactly the same as the experimental group except for writing the phrase "don't be blind" before problem 7 produced intermediate results (between the control and experimental groups).

Luchins original report included NO inferential statistics. It did include a number of direct replications with different participants, so the results are reliable.

C) Systematic replication.

How strong (or fragile) is the set effect (above)? Does it hold up for different populations of individuals and different specific problems?

Systematic replication is an attempt to establish reliability and generality by varying factors that are not thought to make a difference.

Luchins did find that the order of presentation of problems made a difference. Hence, the notion of a set effect in problem solving is linked to having a particular approach work repeatedly.

For Luchins' study, we might test different groups of individuals (ages, cultures) or we might vary the instructions. If the results are consistent across these variations, then we have a robust finding that does not depend upon particular, idiosyncratic details.

This combination of direct and systematic replication would give us much more confidence in the results than statistical tests. It also shows us some of the limits of the results. That is, systematic replication can help us to understand the generalizability of the results.

Summary

Inferential statistics are designed to assess the probability that the data reflect chance (random) variation (the *null hypothesis*).

If the probability that the data reflect chance is sufficiently low (less than .05 or 5%), then we reject the null hypothesis and accept the research hypothesis.

Given that either the *research hypothesis* (H_1) is true or that the null hypothesis is true (H_0), then if we decide to accept H_1 we can be correct or wrong (a *Type I error*). If we decide to accept H_0 then we can be correct or wrong (a *Type II error*). The probability of a Type I error is known. It is the significance level (probability) from our inferential statistical test. The probability of a Type II error is unknown.

Knowing the probability of a type II error would require that we know the "true" state of the world so that we could calculate the odds of obtaining our particular set of results given the true state of the world. Since we do not know the true state of the world, we can not compute the probability of a Type II error.

Adding more participants to the study means that the data come closer to reflecting the true state of the world. This will reduce the probability of a Type II error.

If a set of results is statistically significant, this means that the odds of the results being entirely due to chance are low. We refer to these results as *statistically reliable*.

As an alternative, we could run the study again with a new set of participants. If we got the same (similar) results, then we would refer to the data as having *experimental reliability*.

Experimental reliability is preferred to statistical reliability.

When a study is repeated, we refer to this as a replication. A *Direct Replication* repeats the original study as closely as possible with a new set of participants. Direct Replication establishes experimental reliability.

A *Systematic Replication* involves making changes in the original study that (in principle) should not alter the results. If similar results are found, this establishes experimental reliability and contributes to our understanding of the generality of the results.

Chapter 12 Sample Questions Answers

1) – d; 2) – b; 3) – b; 4) – d

Chapter 13 Sample Questions

1) Rejecting the null hypothesis when it is actually true is: a) an event that can occur with a specifiable probability b) one of two ways of being wrong using inferential statistics c) a Type I error d) all of the above

2) Statistical reliability determines whether results are: a) internally valid b) produced by subject bias c) likely to have occurred because of chance d) a & b above

3) In general, as the sample size increases: a) the power of a statistical test increases b) the influence of a few extreme scores in the data increases c) the probability of a type II error increases d) all of the above

Chapter 13 Sample Question Answers

1) – d; 2) – c; 3) – a