Psychology 312: Lecture 9 Inferential Statistics & Threats To Internal Validity

Slide #1 **Inferential Statistics & Threats to Internal Validity** **Audio:** This lecture will address inferential statistics and threats to internal validity.

Slide #2
 	
Outline
• Statistical Hypothesis Testing.
    o Type I & Type II errors.
    o Jury example.
    o Alpha, p values, power.
• Threats to internal validity.
 	
Audio:
We will begin with a discussion on statistical hypothesis testing and a review of type I vs. type II errors. We will then address a jury example and compare it to the various outcomes that are possible in experimental design. Next we will address alpha, p values and the issue of power. We will conclude by addressing the common threats to internal validity that are encountered in experimental design.

Slide #3 **Statistical Hypothesis Testing** Image of a table of true state of affairs. **Audio:** This table represents the four possible outcomes in statistical hypothesis testing. The rows in the table represent the various decisions we could make representing regarding the null hypothesis. Are two options are to fail to reject the null or reject the null. The columns represent the true state of affairs for our data. That is chance alone responsible for the results we have observed or is chance not responsible and in fact we have a true treatment effect. Let's talk about each of these squares. Let's start in the top left square. In this situation we have a combination in which chance in fact by itself is responsible for the results of our experiment and we choose to fail to reject the null. This is a correct conclusion and in this instance we have avoided making the type I error. Now let's look to the bottom right corner. In this situation chance was not responsible for our results suggesting we have a treatment effect and we have rejected the null hypothesis. This would also be a correct conclusion. In this case we correctly reject the null and avoid making the type II error. The two remaining boxes represent instances in which we make incorrect conclusions. In the lower left square we have a situation in which chance was responsible for the results of our experiment. However we falsely we reject the null hypothesis and committed the type I error. We said there was an effect when in fact there was not one. In the top right corner we have a situation in which chance was not responsible for our results and we fail to reject the null hypothesis. In this instance we have committed the type II error by incorrectly retaining the

null hypothesis. That is we have failed to say that there is a treatment effect when in fact there actually was one.

Slide #4

Jury Example

Assume: innocent unless proven guilty.
H_0: innocent (no different from rest of population).
H_a: guilty (different from rest of population).

Audio:

We can compare the four possible outcomes in statistical hypotheses testing to a jury example. In this case we have a situation in which a jury must make a decision about the guilt or innocence of a particular defendant. In the table the rows represent the decisions about the defendant. The null hypothesis in this case would be to assume that the defendant is innocent. That is he or she is no different from the rest of the population. Whereas the alternative hypothesis would be to assume that the defendant is guilty. Thereby concluding that he or she is different from the rest of the population. So in the table the top row represents the acceptance of the null hypothesis that is to conclude that the defendant is innocent. Where the bottom row represents the rejection of the null hypothesis that is to conclude that he or she is guilty. The columns in the table represent the true state of affairs regarding the guilt or the innocence of the defendant. In the first column we have a situation in which the defendant is truly innocent and in the second we have a situation in which he or she is truly guilty. This means that each square in the table represents a particular outcome regarding the jury's decision of guilt or innocence of the dependent. In the first column we have a situation in which the defendant is truly innocent and in the second we have an instance in which he or she is truly guilty. This means that each square in the table represents a particular outcome regarding the jury's decision of guilt or innocence of our defendant. Let's talk about them each in turn. In the top left corner we have a situation in which the defendant is truly innocent and the jury in fact concludes that innocent. Thus making a correct conclusion. In this case correctly retaining the null hypothesis. In the bottom right corner we have a situation in which the defendant is truly guilty and the jury correctly concludes that he or she is guilty. Thus rejecting the null hypothesis. The other two boxes represent errors identical to those made in statistical hypothesis testing. In the bottom left we have a situation in which the defendant is truly innocent, but the jury incorrectly concludes that he or she is guilty and in the top right we have a situation in which the jury fails to convict the defendant even when the defendant is in fact guilty. These two boxes represent examples of type I and type II error respectively.

Slide #5

Which Error is Worse? Jury Example

- Which type of error is worse?
    - Jury: innocent but found guilty.
    - Science: type I error: rejecting the null then it is correct.
- So, in science, we set specific criterion for rejecting the H_0.

Audio:

I like this comparison for a couple of reasons. First I find that conceptually type I and type II errors often challenging for students. So anchoring them to something concrete can be helpful.

Secondly the comparison allows us to highlight a very important aspect of decision making both in the jury situation and in the science situation. It is generally believed that it is more problematic for a jury to convict an innocent individual then it is to allow a guilty individual to go free. We have a parallel situation in science. In science we typically think that the type I error is the more significant error relative to the type II error. That is it is potentially more problematic to see that you have a treatment effect when you do not have one then it is to fail to detect a treatment effect that actually exists. This means that in science we tend to structure things in such a way that we will reduce the likelihood of making a type I error and we do this specifically by setting some criteria for rejecting the null hypothesis.

Slide #6

Statistical Hypothesis Testing

- Alpha level
  - Probability of committing a Type I Error
  - Set at .05 ($p \leq .05$)
    - 5 in 100 chance of committing a type 1 error.
    - 95% of time we will not make type I error.

Audio:
To understand the criteria for avoiding a type I error we have to discuss the issue of alpha level. Alpha level represents the probability of committing a type I error. It is customary in science to set alpha at .05. In inferential statistics we see alpha represented or written as p less than or equal to .05. What this statement means is that there is a five and one hundred chance of committing an error based on the results of this particular experiment. Said another way 95% of the time we will avoid making the type I error.

Slide #7

Statistical Hypothesis Testing

- Beta level
  - Probability of committing a type II error.
  - It represents the inverse of power (probability of avoiding a type II error).
    - Power = 1 – beta
    - Want power to be high

Audio:
In contrast to alpha level beta level represents the probability of committing the type II error. Formally it is the inverse of power. Power being the probability of avoiding the type II error. Mathematically therefore power equals one minus beta and ideally we want to design an experiment in which power is high.

Slide #8

Statistical Hypothesis Testing

- What do we do to avoid type I errors?
  - Eliminate confounds!
- What do we do to decrease type II errors?
  - Increase sample size (which increases power).

- Design study to minimize factors that may reduce the size of our F ratio.
  - Things that either…
    - Increase within-grp variance.

      AND/OR
    - Decrease between-grp variance.

Audio:
If both type I and type II errors are problematic how do we go about avoiding them. You avoid the type I error by avoiding potential confounds in your experiment. Recall that a confound is something other than your independent variable that could produce systematic changes in your dependent variable across conditions. Therefore making it appear as if you have a treatment effect when you actually do not. So we want to get rid of these. We avoid type II errors either by increasing sample size, which will increase the power in our experiment and help us avoid making a type II error or by using particular design arrangements that minimize factors that may reduce the size of our F ratio. Specifically these are things that would either within group variance and or decrease between-group variance. Remember that we want the opposite situation. We want to design an experiment in which we have low within-group variance and high between-group variance.

Slide #9

Threats to Internal Validity

- Major confounding variables.
  - Primarily affect between-group VAR
  - Note: can also affect within-grp VAR depending on how they interact with other variables.
- Let's look at some examples….

Audio:
Because type I errors are considered more serious then type II errors in experimental design we are going to spend the rest of this lecture focused on threats to internal validity. In doing so we will highlight the major confounding variables encountered in experimental design. Recall that a confounding variable is a variable other than the independent variable that will contribute to between-group variance. While we often time discuss confounding variables in terms of between groups variance please recognize that these variables can also affect within-group variance depending upon how they interact with others in the experiment. Now we are going to look at some specific examples. These examples are intended to highlight different examples of potential confounding variables.

Slide #10

Hypothetical Study #1

- Examine whether children of different ages perceive different colors similarly or differently than adults.
  - 3 yr old versus 10 yr old children.
  - Show them 15 chips that represent various shades of a given color.

Audio:
In our first hypothetical study we are going to examine whether children of different ages.

Perceive different colors similarly or differently than adults. In this study we will recruit three year old children and compare those children to 10 year old children. We intend to show each child fifteen chips that represent various shades of a given color.

Slide #11

Hypothetical Study #1
Image of different color circles.

Audio:
For example something like this. So we have a color wheel and we select a particular color on that color wheel and then we have a serious of chips that represent different cues along each of those colors.

Slide #12

Hypothetical Study #1

- Participants are asked to arrange the chips in the appropriate order.
- Each participant is tested several different times with different individual colors.
- Record number of errors made in the order for each color.

Audio:
In this experiment each child will be asked to arrange the chips in the appropriate order. Each participant is tested several different times with different individual colors. A dependent variable in this experiment will be the number of errors that each child makes while establishing the order for each color.

Slide #13

Hypothetical Study #1
Image of a diagram of a study done on 10 yr olds and 3 yr olds.

Audio:
The overall design of the experiment will look something like this. We have are two age groups ten year olds and three year olds. Each child in each of those age groups will be asked to order a set of chips for three colors yellow, then red, then blue and in each instance we will count the number of errors the child makes when attempting to complete that ordering.

Slide #14

Hypothetical Study #1

- Children showed the greatest error with the color blue and the two age grps differed from one another.
  - 11% of 10 yrs olds made errors.
  - 50% of 3 yr olds made errors.

Audio:
Let’s imagine that the results of that experiment show that children showed the greatest error with the color blue and the two age groups differed from one another. That is 11% of the ten year olds errors in comparison to 50% of the three year olds.

Slide #15
Hypothetical Study #1

- Conclude that the color blue is perceived differently by 3 yr olds versus 10 year olds (which both differ from adults).
- Accurate? Or did something go wrong?

Audio:
We conclude based on these results that the color blue is preserved differently by three year olds vs. ten year olds with in turn differ from adults. So the question is that an accurate conclusion or did something go wrong in this experiment?

Slide #16

Threats to Internal Validity

- Maturation Confound.
    - Problems of interpretation resulting from participant's maturation, either between tests or over time.
        - Growing older, stronger, healthier, more tiered, more bored, etc.
        - Fatigue effects.
            - Current exp: 3 y olds get tired faster.
        - Solution?
            - Randomize the order of the conditions.

Audio:
The results of the current experiment may have been the results of maturation confound. Maturation confounds refer to problems of interpretation resulting from participant's maturation, either between various tests in the experiment or over time. My maturation we mean things like growing older, stronger, and healthier or perhaps more tired or more bored over time. In the current example it would appear that fatigue effects were particularly possible. That is to say our three year old may have gotten tired faster than their ten year old counterpart. Because the color blue was the final condition their performance may have suffered in that condition not due to perceptual changes, but due to fatigue. One potential solution for this potential problem would be to randomize the order of the color conditions across our different participants.

Slide #17

Hypothetical Study #2

- Interested in examining how emotion is expressed on either side of the face.

Audio:
Now let's turn to another example. In hypothetical study number two lets imagine that we are interested in examining how emotion is expressed on either side of the face.

Slide #18

Hypothetical Study #2

- Previous research has shown that people can readily identify facial expression for 6 basic emotions.

Audio:

We are interested in this question because previous research has shown that people can readily identify facial expression for 6 basic emotions. You may have learned something about this in your other psychology courses.

Slide #19

Hypothetical Study #2

- But is that emotion better expressed by one side of the face than the other?
- Examine using a composite photo of either side of the face.

Audio:
However it is unclear whether emotion is better expressed on one side of the face than the other. We think that this is a reasonable question and we would like to try to examine it using a particular manipulation. In this case we plan to use what is called a composite photo of either side of the face.

Slide #20

Hypothetical Study #2
Image of pictures of different facial expressions of people.

Audio:
Composite photos are created by taking a picture of an individual's face cutting it down the middle and then using either side to replicate the other side of the face. Here are a couple of examples, because most of us have asymmetrical faces are composite photos do not look like our original photo and you can see in this instance that for both president Bush and for Paris Hilton the composite photos of the left and right sides of the face look dramatically different.

Slide #21

Hypothetical Study #2
Image of pictures of different facial expressions of people.

Audio:
Here are two more examples using Marilyn Monroe and Keith Richards.

Slide #22

Hypothetical Study #2

- Show participants three different types of photos (IV) of the same person expressing different emotions.
    - Original photo.
    - Left composite.
    - Right composite.
- Asked participant to rate the emotional intensity (DV) of each photo.

Audio:
Let's imagine that we plan to show participants three different types of photos of the same person expressing different emotions. So our three different photos which will serve as the independent variable in this experiment are the original photo, left composite photo and a right

composite photo. We plan to ask participants to rate the emotional intensity of each of these photos and that rating will serve as our dependent variable.

Slide #23

Hypothetical Study #2
Image of diagram scores of emotion.

Audio:
The overall design of our experiment would look something like this. From a population we intend to select a sample those participants will then be shown a set of left composite photos and asked to score them on emotionality. Then a set of original photos and score those on emotionality. Finally a set of right composite photos and score those on emotionality.

Slide #24

Hypothetical Study #2

- Let's say results show that participants are most accurate reading emotionality in the "right" composite photos relative to the "originals" or the "left" composites.
- We conclude that emotionality is expressed differently on different sides of the face.
- Accurate conclusion? Or did something go wrong?

Audio:
Let's imagine that we conduct this experiment. In fact the results show that participants are most accurate reading emotionality in the "right" composite photos relative to the "originals" or the "left" composites. We conclude therefore that emotionality is expressed differently on different sides of the face. Is that a correct conclusion or did something potentially go wrong?

Slide #25

Threats to Internal Validity

- Testing/practice confound.
    - Problems that result from repeated measurement of same individual.

Audio:
It is possible that results of the present experiment are a result of a testing or practice confound. As its name suggests testing and practice confounds occur in situation where participants are being asked to complete the same task again and again and as a result their performance on the dependent variable improves over time.

Slide #26

Threats to Internal Validity.

- Solution?
    - Randomize order of conditions
    - Use between-subject design rather than within-subject.
        - Within: same participants in all conditions.
        - Between: different participants in different conditions (left, right, original).

Audio:
Design and practice effects can be controlled by using a number of different methodological strategies. One of those would be to randomize the order of conditions across participants. In doing so participants would perceive the different types of photos in different orders thus helping to control for practice effect over time. A second strategy would be to use a between subject design arrangement rather than a within-subject design arrangement. The first two experiments ive used here to discuss confounding variables. Both are examples of within-subject design arraignments. In this type of design arrangement the same participants serve in all conditions. Whether that was the different types of colors in the first experiment or the different types of photos in the second experiment. In the present experiment involving photos we could redesign the experiment so it was run as a between-subject design arrangement. In this instance different participants would view different types of photos. So one group would see only left composite photos, another only right composite photos and finally a third only original composite photos. We would then compare the performance across those three conditions.

Slide #27

Threats to Internal Validity

- History Confound.
    - Events that take place between measurements in pretest-postest design (not related to IV).
        - EX: record mood at pretest, administer drug, then test again 6 months later.
            - Pretest during winter months.
            - Posttest during summer months.

Audio:
Another common confound in experimental design is the history confound. History confounds occur when an event takes place between measurements and experimental design. Particularly when that design is a pretest/posttest design. That event is unrelated to the independent variable, but none the less has the potential to affect the dependent variable. Let's say for instance we are interested in testing a new anti depressant drug. To do this we plan to recruit a set of participant's record their mood at a pretest, administer our drug and then test them again six months later. A history confound could exist for this experiment if we chose to conduct our pretest during the winter months and our posttest during the summer months. Our participants mood might improve over time, but not necessarily because of our drug, but because the weather is also improving.

Slide #28

Hypothetical Study #3

- Examine the effects of authoritarianism on task mastery.
- Hypothesize that people who rate high in authoritarianism will perform more poorly on complex tasks, due to their likelihood to employ simple solutions.

Audio:
Now let's consider a third hypothetical study. In this study we are interested in examining the effects of authoritarianism on task mastery. Let's imagine that previous research has encouraged us to hypothesize that people who rate high in authoritarianism will perform more poorly on

complex tasks, due to their likelihood to employ a simple problem solving strategy.

Slide #29

Hypothetical Study #3
Image of a diagram.

Audio:
Let's further imagine that we plan to use the following design to test this hypothesis. We will recruit a set of participants and pre-screen those individuals in terms of authoritarianism. We will need some participants that rank quite high in this characteristic and others that rank quite low. We will then ask all participants to complete two tasks an easy task and a complex task. We can counterbalance the ordering of these two tests across participants to control for any order of that. We are predicting that individuals who rate high in authoritarianism will perform quite well on the easy tasks, but not as well on the complex test. We will predict the opposite pattern of results for those who rate low authoritarianism. Those individuals should perform quite well on the complex task, but not as well on the easy task. On the surface this may appear like a relatively reasonably design arrangement. Unfortunately what we will discover is that authoritarianism is highly correlated with another participant characteristic. Specifically it is highly correlated with intelligence. Now we have a situation in which intelligence scores are contaminated our potential results and serving as a confound. In this case if we observe some differences in performance at the end of our experiment across the various tasks it may be that difference is due to intelligence and not to differences in authoritarianism.

Slide #30

Threats to Internal Validity

- Selection Confound.
    - Participants vary on one or more variable (other that the IV) across the conditions.
    - Particularly likely when using an "experimenter-selected" IV.

Audio:
This hypothetical example is intended to highlight a selection confound. Selection confounds occur when participants vary on one or more variable. Something other than the independent variable across conditions. In this case the presence of a selection confound produces systematic differences across our conditions at the outset of the experiment. Consequently those differences could contribute to between-group variance and make it appear as if we have a treatment effect when in fact we do not. Selection confounds are particularly likely when one is using a experimenter selected independent variable. You will recall that this type of independent variable is one that cannot be randomly assigned, but the researcher will use it as the basis for assignment across the different conditions in the experiment.

Slide #31

Another Example
Image of a diagram.

Audio:
One of the most common experimenter selected independent variables in social science research

is gender. This is a variable that is in fact quite complicated. Gender has both a biological comment in which we define biologically individuals as either male or female and a socialization history component in which individuals are expected or encouraged to conform to masculine or feminine behaviors. It is in fact even more complicated than that, because if we look closely at the biological component of gender we can break it into a number of sub-categories in which we have a genetic component, a hormonal components, differences in external and internal genitalia and finally secondary sex characteristics. All of which are collectively used to define an individual as biologically male or female. The point here is if we run an experiment in which gender is the independent variable and we are able to demonstrate a difference on some dependent measure on the end of the experiment it will be unclear to us which of these potential components is in fact driving that difference.

Slide #32

Threats to Internal Validity

- Regression toward the mean.
    - Extreme scores (DV) move toward the mean with repeated testing over time (unrelated to effect of IV).

Audio:
A fifth threat to internal validity is regression towards the mean. Regression towards the mean is observed when extremes scores that is scores on your dependent variable move towards the mean with repeated testing over time, but that change is unrelated to the independent variable. We often times see this in any situation in which individuals are being asked to repeatedly engage in an activity that will render some score on the dependent variable. We must recognize that overtime the variability of those scores may be reduced as scores move or gravitate towards the mean.

Slide #33

Threats to Internal Validity.

- Diffusion or imitation of treatments.
    - Individuals in an experiment may communicate with each other; this can reduce differences between conditions due to “diffusion.”
    - Solutions?
        - Make participants “blind” to the hypothesis.
        - Ask them not to talk about study.
        - Test over short period.

Audio:
Diffusion or imitation of treatment of confounds occur in situations in which individuals in the experiment have the opportunity to communicate with each other. In doing so they tend to reduce differences between the conditions due to the fusion of the treatment or the fusion of the manipulation of the independent variable. There are a number of strategies for attempting for attempting to reduce or control diffusion and imitation of treatment effects. They include things like: making participants blind to the hypothesis so that they do not know the true nature of the experiment, asking them deliberately not to talk to other individuals about the study or finally testing over a very short time period.

Slide #34

Threats to Internal Validity.

- Instrumentation.
  - Instrument (human or machine) used to measure DV is unreliable (gives different readings over time).
  - Solutions?
    - Training & practice.
    - Use multiple raters (check for inter-rater reliability).
    - Counterbalance order of conditions across raters.

Audio:
Instrumentation confounds refer to situation in which the instrument used to measure the dependent variable is unreliable. Meaning that it produces different readings over time, but those changes are not accurate reflections of changes in the dependent variable. These types of problems can exist when using both human and mechanical instruments. In a case of mechanical instrument we reduce these types of problems by carefully monitoring our equipment and checking it before using it for data collection. In the instance of the human instrument there are a number of best practices that are encouraged to reduce the likelihood of an instrumental problem. They include things like providing ample training and practice prior to true data collection. The use of multiple raters is often times also encouraged, because this will allow us to check for inter-rater reliability. Finally if we counterbalance conditions across raters we can help control for things like fatigue or boredom on their part.

Slide #35

Threats to Internal Validity.

- Mortality
  - Loss of participants due to drop out or refusal to participant.
  - Solution?
    - Removing similar participant(s) from other conditions.
    - Noting problem in results and discussion.

Audio:
The final threat to internal validity that we must discuss are mortality confounds. These types of confounds exist when the loss of participants due to drop out or refusal to participate in the experiment produce systematic differences across the various conditions in the experiment. The most common solution to the mortality problem is to remove similar conditions in an attempt to establish similarity across those conditions. If this cannot be done the researcher must be sure to note the problem in the result and discussion session of their formal research article.

Slide #36

Next Lecture

- That concludes this lecture.
- Next we will discuss “Control by Design.”

Audio:

That concludes this lecture. Next we will discuss control by design.