Hypothesis Testing

Chapter 8

Class Outline – 7-17-08

- Review of sampling distribution of the mean
- Hypothesis Testing – Chap. 8
- Questions about problem set?

Review (from Chap. 1)

- There are two major methods that researchers use to look for relationships between variables:
  - The correlational method – two variables are simply observed to determine if there is a relationship (example)
  - The experimental method – cause and effect relationship between two variables (changes in DV are caused by changes in the IV)
    - Must manipulate one or more variables
    - Must hold all other variables constant

Review (from Chap. 1)

- **Independent variable** – the variable that is manipulated by the researcher
- **Dependent variable** – the variable that we are measuring or
  - the variable that is observed for changes in order to assess the effect of the treatment (dependent on changes in the independent variable)
- Control condition vs. experimental condition (examples)

Hypothesis Testing

- As we have discussed in past classes, we use inferential statistics to make inferences about a population from a sample
- Hypothesis testing is an inferential procedure that uses sample data to evaluate a hypothesis about a population parameter
  - The first hypothesis test we will learn combines our knowledge of z-scores, probability, and the distribution of sample means…
  - Other hypothesis tests build on this basic structure.

Logic of Hypothesis Testing

1. State a hypothesis about a population (what we believe is true about a population parameter, ex. mean)
2. Based on hypothesis, we then predict what a random sample of that population should look like (sample should be close to population but have some error)
3. We take a random sample from the population of interest
4. We compare our sample to our prediction.
   - If the sample data is similar to our prediction about the population, we conclude that we made a reasonable hypothesis.
   - If our sample data is not similar to prediction, we conclude that our hypothesis is wrong.

Hypothesis Testing

- We will first use hypothesis testing to determine if a certain treatment has an effect on a given sample
  - In this case, we know what the population is like *before treatment* (population parameters are known)
  - We want to know what the population would look like if we gave the treatment to the entire population (same or different?)
- We assume that a given treatment will add a constant to each score.
  - Mean changes, but SD and distribution shape remains the same.
- We administer our treatment to a random sample of the population so that we can generalize our results to everyone in the population
  - If our treatment has a (big) effect on the sample, we *infer* that the treatment would have an effect on everyone in the population

Hypothesis Testing

- With hypothesis testing, we must follow a standard series of operations that all researchers follow
- One of the foundations of the scientific method is that research must be replicable.
- The standardized procedure allows others to know how their results were obtained and gives them the ability to replicate the study using the same procedures

Testing a Hypothesis

- Example – We've developed an amazing (but safe, nonaddictive, etc.) study pill
- STEP 1: State the hypothesis
  - The *null* hypothesis – the treatment (pill) has no effect
    - H_0: $\mu = 0$
  - And the alternative hypothesis – the treatment does have an effect (does not indicate direction of effect yet)
    - $H_{1:}$ $\mu \neq 0$

Testing a Hypothesis

- STEP 2: Set the criteria for decision
  - We determine what sample means are consistent with the null hypothesis
  - If the null hypothesis is true, we would expect our sample mean to be relatively close to our population mean
  - We must make a "cut-off" that tells us what sample means we are likely to get if the null hypothesis is true and what sample means we are unlikely to get if the null hypothesis is true

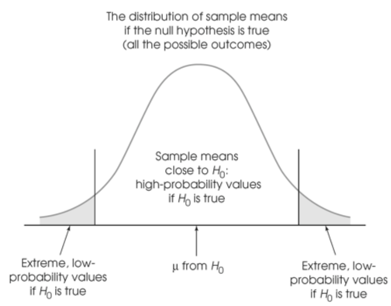

Figure 8-2 (p. 182)
The set of potential samples is divided into those that are likely to be obtained and those that are very unlikely if the null hypothesis is true.

Alpha Level

- To determine boundaries that separate high probability from low probability samples, use alpha level (level of significance)
  - Commonly use $\alpha = .05$, or $\alpha = .01$
- **The alpha level or level of significance** – a probability value that is used to define the very unlikely sample outcomes if the null hypothesis is true

Critical Region

- **The critical region** – the area beyond the alpha level that is composed of extreme sample values that are very unlikely to be obtained if the null hypothesis is true
- Boundaries of critical region determined by alpha level
- If the sample data fall in the critical region, we reject the null hypothesis
  - *Convincing evidence* that the treatment really does have an effect.

- Those sample outcomes that fall in the critical region are nearly impossible to get unless there is something (treatment of some kind) affecting them
- Psychologists typically set the alpha level at .05 (meaning that they are separating the middle 95% of the distribution from the outer 5%)
- Because we are splitting up 5% between the two tails of the distribution, we would look in the unit normal table under 2.5% probability in each tail (z = 1.96)

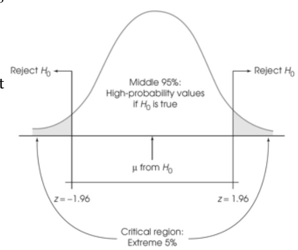

Figure 8-3 (p. 195)
The critical region (very unlikely outcomes) for $\alpha = .05$.

- So if our mean had a z-score of 1.96 or above, we would reject the null hypothesis
- We could also set the alpha level at 1% or .01 What z-score would we have to meet or exceed to reject the null at the .01 level?

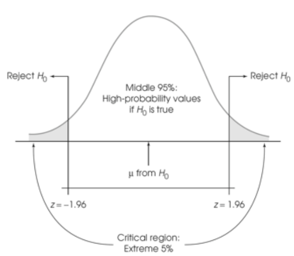

Figure 8-3 (p. 183)
The critical region (very unlikely outcomes) for $\alpha = .05$.

Testing a Hypothesis

- STEP 3: Collect the data and compute the *test statistic* (this is done after we state the hypotheses and establish the criteria)
- We would, in this case, compute the z-score for the sample mean and determine where it falls on the distribution in relation to the population mean

z-score formula reminder

$$z = \frac{M - \mu}{\sigma_M}$$

$$Z = \frac{\text{sample mean – hypothesized population mean}}{\text{Standard error between } M \text{ and } \mu}$$

$$\sigma_M = \frac{\sigma}{\sqrt{n}}$$

Testing a Hypothesis

- STEP 4: Make a decision
  - Two possible decisions and both are stated in terms of the null hypothesis
    - Reject the null hypothesis (sample data fall in the critical region)
    - Fail to reject the null hypothesis (sample data are reasonably close to the null hypothesis – not in critical region)
  - Can *never* say that we have proven something to be true in science
    - To prove true, have to show in all cases
    - To prove false, have to show in one case

Factors that influence hypothesis testing…

1. The size of the difference between the sample mean and the original population mean.
2. The variability of the scores
3. The sample size.

Testing a Hypothesis

- Because we are making decisions in terms of probability, we state that it is very unlikely to draw a sample from a population that falls in the tail beyond an alpha of .05
- We will, however, sometimes draw the wrong conclusion…

Error

- Two kinds of error in hypothesis testing:
- **Type I error** – reject a true null hypothesis
  - In other words, reject the null hypothesis when in fact the treatment has no effect.
- Type I errors are probably considered to be the most serious because you are reporting that your treatment had an effect when it didn't.

Error

- **Type II error** – fail to reject a false null hypothesis
  - In other words, your treatment has an effect but you say it doesn't
- This is less serious than Type I error
  - (it depends on the situation however. If you discover a drug that cures cancer but you claim that it has no effect, your drug will not be given to the public)

Error

- The probability of making a Type I error = alpha level
  - $\alpha = .05$, means 5% chance of rejecting a true null
- The probability of making a Type II error = beta (β)
  - Depends on a variety of factors – not easy to determine.

Making a decision

		Actual situation: No effect, H_0 true	Actual situation: Effect exists, H_0 false
Experimenter's decision	Reject H_0	Type I error	Decision correct
	Retain H_0	Decision correct	Type II error

Table 8-1 (p. 201)
Possible outcomes of a statistical decision.

Minimizing Error

- We select our alpha level to minimize the risk of Type I error
  - Again, typically set $\alpha = .05$, means 5% chance of committing Type I error
- Alpha levels of .01 or even .001 are obviously preferred (less chance of committing Type I error), but they are much harder to obtain
  - Would need a large treatment effect before sample data would reach critical boundaries
  - What would happen to Type II error?

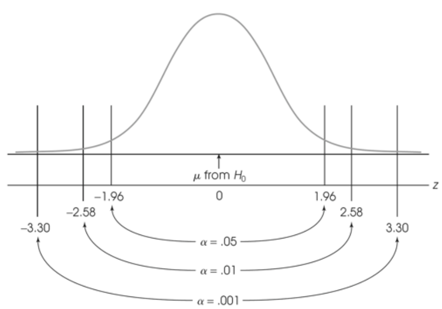

Figure 8-4 (p. 191)
The locations of the critical region boundaries for three different levels of significance: $\alpha = .05$, $\alpha = .01$, and $\alpha = .001$.

Example

- Back to our miracle study drug

1. State hypothesis
2. Set decision criteria
3. Collect data, compute test statistic
4. Make decision
5. APA style reporting of results

Assumptions of hypothesis testing using z-scores

- Random sampling
- Independent observations (occurrence of first event has no effect on the probability of the second event)
- Value of standard deviation is unchanged by treatment
- Normal sampling distribution

Effect Size

- Rejecting the null hypothesis (determining that there is a significant treatment effect), does not tell us anything about how large the effect is
  - *Significant* does not equal *large* or *important*
  - Remember, as sample size increases, standard error decreases
  - As standard error decreases, z-score increases, meaning more likely to reject null hypothesis
  - See example 8.5 in book

Effect Size

- Cohen's *d* is easiest way to measure effect size
  - Cohen's $d = \frac{\text{mean difference}}{\text{standard deviation}}$
- Effect size of 0 - .1 = small effect
- Effect size of .2 - .8 = medium effect
- Effect size greater than .8 = large effect

Same mean difference, very different effect size!

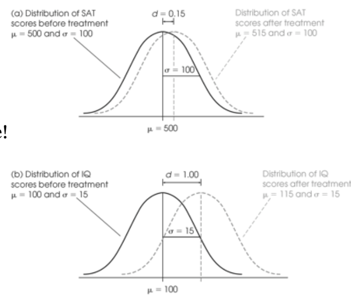

Figure 8-8 (p. 199)
The appearance of 15-point treatment effect in two different situation. In part (a), the standard deviation is $\sigma = 100$ and the 15-point effect is relatively small. In part (b), the standard deviation is $\sigma = 15$ and the 15-point effect is relatively large. Cohen's *d* uses the standard deviation to help measure effect size.

Power

- Power – the probability that the test will correctly reject a false null hypothesis
  - reject the null hypothesis if the treatment has an effect
  - Usually calculated *before* researchers conduct a study
- Influenced by effect size, sample size, alpha level, etc.
- Read about this in your book and get a general idea but don't worry too much about it now

One-tailed Hypothesis Tests

- If we have a prediction about the direction of the treatment effect, we will use a one-tailed hypothesis test
  - (we will specify either an increase or a decrease in the mean)
- The critical region is located entirely in one tail
  - We incorporate the directional prediction into the first step of the hypothesis test
  - In the second step, we locate the critical region in one tail

One-tailed Hypothesis Tests

- Example
  - Study drug
  - H_0: studying is not increased
  - H_1: studying is increased
  - Critical region only in one tail
- We will reject the null hypothesis more frequently in a one-tailed test because the critical region is larger
  - (entire 5% in one tail as opposed to being split between two tails)

One-tailed vs. Two-tailed

- Two-tailed more stringent
  - But may fail to pick up on real effect
- One-tailed more sensitive
  - But may commit Type I error
  - Or, may miss effect entirely
- Book says use one-tail with directional hypothesis
- Psychological community generally uses two-tail

Class Example

- A researcher is evaluating the effectiveness of a new physical education program for elementary school children. The program is designed to reduce competition and increase individual self-esteem. A sample of $n = 16$ children is selected and the children are placed in the new program. After 3 months, each child is given a standardized self-esteem test. For the general population of elementary school students, the scores on the self-esteem test form a normal distribution with $\mu = 40$ and $\sigma = 8$.
  - a. If the researcher obtains a sample mean of $M = 42$, is this enough evidence to conclude that the program has a significant effect?
  - b. If the sample mean is $M = 44$, is this enough to demonstrate a significant effect?