SAN JOSÉ STATE UNIVERSITY

College of Social Work
S. W. 242
Spring 2008
Edward Cohen

Week 5: February 22, 2008
Testing Hypotheses and Statistical Significance

"Statistics means never having to say you're certain!"
Anon.

- Normal distribution
- Z score
- Skewed distributions
- Inferential analysis
- Sampling error
- Null hypothesis
- Two tailed hypothesis
- One-tailed hypothesis
- Alpha (or rejection) level
- Type I error
- Type II error

I. Continuation from last week's notes:

A. Measures of Dispersion—Big Question: Why are large samples better than smaller ones?

Consider the "Descriptives" data results from last week's exercise (dataset from Ch. 6, Kirkpatrick & Feeney)

Statistics

		STUDENT	GENDER	SCORE
N	Valid	30	30	30
	Missing	0	0	0
Mean		15.5000	1.5000	78.7333
Median		15.5000	1.5000	82.5000
Mode		1.00[a]	1.00[a]	89.00
Std. Deviation		8.80341	.50855	14.22221
Variance		77.50000	.25862	202.27126
Range		29.00	1.00	65.00
Minimum		1.00	1.00	33.00
Maximum		30.00	2.00	98.00

[a] Multiple modes exist. The smallest value is shown

1. **Range –** (the largest value minus the smallest value + 1)

Calculate the Score range from the data set: (Maximum-Minimum +1) = (98-33) +1 = 66

2. **Variance** – the average of the squared deviations from the mean

3. **Standard deviation --** the square root of the variance

Calculate the variance and standard deviation for the Score variable (Ch. 6 Kirkpatrick & Feeney):

Student	Score	Subtract mean	Deviation from mean	Squared deviation from mean
1	87	-78.33	8.67	75.17
2	53	-78.33	-25.33	641.61
3	92	-78.33	13.67	186.87
4	70	-78.33	-8.33	69.39
5	78	-78.33	-0.33	0.11
Do this with all 30 cases	.	.	.	.
.	.	.	.	.
.	.	.	.	.
30	.	.	.	.
Variance: Sum of all 30 squared deviations divided by *n* (or 30)				202.21
Standard Deviation: Square root of squared deviations				14.22

Fascinating! properties of the standard deviation:

1. The SD is in the same units as the original measure (Score points)

2. For the SD of Score, we say that “one standard deviation is equivalent to 14.22 score points”

3. Different samples of classes taking this test could possibly have the same mean, but different standard deviations. What does this say about means, in general? What does this say about sample size, in general?

4. A *population* can have a standard deviation. Since the population mean and SD are rarely known in social research, we *infer* it from samples. Then, the question is, does the sample mean represent the mean of the population from which the sample was chosen? What’s the role of the SD in answering this question?

5. ***One more highly interesting fact:*** the Weinbach/Grinnell book computes the variance with a denominator of *n*. SPSS (and many other statistics books) use *n-1,* a smaller denominator (larger Variance) hence a more “conservative” approximation of the variance in the population. There’s a theory behind this. Ask me sometime. OK, since you insist: *since the sum of the (non-squared) deviations from the mean = 0 (try it), then you can determine the last one once we know the others. We really only need to average 29 in the example, not 30. So using the 30th in the calculation is superfluous.* ***However, for our purposes we will stick with the Weinbach/Grinnell (n).***

Note: will you have to learn formulas for the test or papers? No.

II. Normal Distributions

A. A <u>normal distribution</u> (or normal curve) is a bell-shaped curve (a type of frequency polygon) generated from the plotting of values from *interval* and *ratio* level variables. Some properties are:

1. Bell-shaped

2. Symmetrical

3. Mode, median, and mean all occur at the highest point and in the center of the distribution

4. The curve goes to infinity and never touches the x-axis. This accounts for extreme +,- values

*** Can you draw this distribution?

B. What does curve frequency polygon tell us about the distribution of the variable?

1. Whether the variable is normally distributed, also known as symmetric (see above, #3), or **skewed**

    a) Positively skewed—the right tail is longer than the left one

    b) Negatively skewed—the left tail is longer than the right one

2. What proportion of values occur in any given distance from the mean

C. If a dataset is normally distributed in the population, then we can "standardize" the original raw data into **z scores**.

1. Calculating a z score: [(raw score – mean) divided by the SD], e.g. from our practice dataset (87-78.33/14.22) = 0.61 . This is in standard deviation units (.61 of a standard deviation).

***Using Table 4.3 (Weinbach & Grinnell, p. 70), what is the area under the normal curve between this z score and the mean? Another question: what is the percentile of this z score? (Hint—use the area to the left of the mean)

2. Several datasets of different scales and units of measures can be transformed into z scores and compared to each other.

3. Understanding the area under the normal curve is important to understand hypothesis testing, in general

III. Statistical Significance and Hypothesis Testing—What can we *infer* from the sample about the population?

A. **Inference** – using a statistic (that is, a summary of the sample) to draw tentative conclusions about the population from which the sample was drawn, and to make a probability statement (such as p < .05) about our confidence in those conclusions

B. What can influence this confidence?

1. Rival hypotheses (remember that from 240?)

2. Measurement problems (remember measurement error?)

These two influences are the result of research design.

3. **Sampling bias, or sampling error** -- Chance contends that no matter how well a research study is designed, any relationship between variables within the sample could be just a "fluke" or random occurrence. Sampling error is the natural tendency of any sample to differ from the population from which it was drawn. Sampling error is especially likely to occur in small samples, even when rival hypotheses and measurement error are taken care of.

***Why would sampling error be more likely to occur in small samples?

4. A statistic and its associated probability (*p* value) will quantify the extent of sampling error. It's up to the researcher to interpret whether rival hypotheses or measurement error are likely culprits based on limitations of the research design, or if it's mainly an issue about sampling.

5. What's a "*p* value" (*p* = .05)? *"We can say with a degree of certainty that the relationship found between these variables can only happen by chance 5% of the time."*

6. This probability statement only can confirm or disconfirm a *null hypothesis*…

C. Hypothesis testing—what are we testing, anyway?

1. **Null Hypothesis** – a statement that there is no relationship between two variables of interest, e.g. "there is no relationship between gender of social workers and whether or not they are satisfied with their jobs" OR "there is no difference in job satisfaction between female and male social workers"

2. This doesn't state *why* there is no relationship, however it is based on sampling theory, e.g. the statistical procedure you are using cannot find any evidence to the contrary, due to factors such as high variance and SD in the sample. Another way of saying the null hypothesis: "with this sample we cannot infer that there is a relationship between these two variables *in the population*."

3. What’s the opposite of Null Hypothesis? Answer: the research hypothesis. There are two types--

a) Two-tailed (non-directional) hypothesis: “Gender of medical social workers is related to job satisfaction levels” OR “there is a difference in job satisfaction between female and male social workers” (both ways of saying it are correct)

b) One-tailed (directional) hypothesis: Female medical social workers will have higher levels of job satisfaction than male social workers

4. Type I and Type II Errors. Errors in drawing conclusions about relationships.

		Our Decision (Based on sample)	
		Reject Null Hypothesis	Accept Null Hypothesis
TRULY in the Real World (Population)	Null hypothesis is false (The alternative hypothesis is true and there is truly an effect)	No error	Type II error (We have wrongly rejected the alternative hypothesis)
	Null hypothesis is true (The alternative hypothesis is false and there is truly no effect)	Type I Error (We have wrongly accepted the alternative hypothesis)	No error

*** If you were conservative in evaluating practice, an intervention, or the effectiveness of a program, would you rather make a Type I or a Type II error?

a) A Type II error is preferred if being conservative.

b) For example, if you do a clinical trial for Drug X, the null hypothesis is that Drug X is not related to improvement. If Drug X were TRULY EFFECTIVE but you made a Type II error, you would be accepting the null hypothesis and rejecting the alternative hypothesis, thus delaying or preventing Drug X from use.

c) BUT, if Drug X were TRULY INEFFECTIVE and you made a Type I error, you would be rejecting the null hypothesis and accepting the alternative hypothesis, thus approving an ineffective (or dangerous) drug for use!

5. Statistical Significance

a) <u>p-value</u> (*p*) is the mathematical probability that a relationship between variables found within a sample may have been produced by chance or error.

b) <u>rejection level</u> (α , alpha level, or significance level) is the point that we are willing to accept a Type I error and say with a degree of certainty that we can reject the null hypothesis and say there is a significant relationship between our variables of interest. Let's discuss $\alpha = .05$ vs $\alpha = .01$ vs $\alpha = .001$

Example:

If we find from our statistical analyses that female social workers are more satisfied than male social workers with a p-value = .04, what does this mean?

If we find from our statistical analyses that female social workers are more satisfied than male social workers with a p-value = .15, what does this mean?

How do these examples relate to our hypothesis testing? **<u>When the p-value is less than the alpha level, we reject the null hypothesis.</u>**

a) How is the alpha level determined? In social sciences research, use of $\alpha = .01$ or .05 is customary. For very large datasets, criteria might be more stringent, e.g. $\alpha = .01$. For clinical drug trials, when lives may be at stake, the alpha cutoff might be quite low (.001)—scientists want to be sure not to make a Type I error. If you were ill and waiting for the results of an experimental medication trial, and the Null was rejected at $\alpha = .05$, would you favor taking the medicine for your illness instead of another medicine that has already been proven at $\alpha < .001$?

b) In social sciences research, we tend to have the opposite problem—our samples are often too small and we have trouble statistically rejecting the Null when, in fact, the intervention actually does have an effect (i.e. we make too many Type II errors)!

c) Statistically significant versus meaningful findings: Is every finding that is mathematically significant useful, practical, or relevant? Numbers must be interpreted in context and with other corroborating evidence (experience, triangulation, scientific rigor, common sense…) in mind.

Self-esteem example: We want to evaluate whether our new intervention elevates self-esteem. We use the Cohen Self Esteem Scale (CSES) to measure esteem (scores from 10-50, with higher scores meaning higher self-esteem). We use a classic experimental design. At pretest, we find that scores measuring self-esteem are comparable between the control and experimental groups, both at 15 (randomization did its job). At posttest the experimental group has a mean score of 25 and the control group a score of 19. The statistic used (we'll get to that in a few weeks) rejected the Null Hypothesis (that there is no posttest difference between the experimental and control group, $p<.05$). In other words, the probability that this difference would be by chance alone is less than 5 out of 100. However, Cohen says that according to his validation of the CSES, the change from 15 to 25 in the experimental group is not clinically meaningful. If you were reporting these results, how would you handle this in your Results and Discussion sections?

Class Process Questionnaire -- ScWk 242 Section: __3
__4

Date of Class: _________

Instructions—near the end of class or at anytime after class please complete *any* of these questions and return to the Instructor. **This is *not required* for grade or class participation.** Do not put your name on this sheet.

1. At what point in this class did you feel most engaged?

2. At what point in this class did feel the least engaged?

3. What action or discussion by anyone in the room did you find most helpful?

4. What action or discussion by anyone in the room did you find most confusing?

5. What surprised you most about class today?

6. Is there a topic that we should definitely get back to in the next weeks? Or, any particular types of teaching techniques you would recommend?