K300 (4392) Statistical Techniques (Fall 2007)

Lecture Note: Hypothesis Testing

Instructor: Hun Myoung Park

1. Hypothesis

A *hypothesis* is a specific conjecture (statement) about a characteristic of a population. This conjecture may or may not be true; we may not know exactly whether it is true or false forever. 1) *A hypothesis should be specific* enough to be falsifiable; otherwise, the hypothesis cannot be tested successfully. A bad example is "The war on Iraq may or may not produce a positive impact on the U.S. civil society." 2) A *hypothesis is a conjecture about a population (parameter), not about a sample (statistic)*. Therefore, $H_0 : \bar{x} = 0$ is not valid because we already know the sample mean $\bar{x}$ from a sample. 3) A *valid hypothesis should not be based on the sample to be used to test the hypothesis*. "Hmm… my sample mean is 5, so my null hypothesis is the population mean is 5." What? This is a tautology. Finally, 4) *A appropriate hypothesis needs to be interesting and challenging (informative)*. Some bad examples includes "Average personal income is zero."

The *null hypothesis*, denoted H_0, is a specific baseline statement to be tested and often (not always) takes such forms as "no effect" or "no difference." Why? Simply because it is easy to compute the statistic and interpret the output. The *alternative hypothesis* (or research hypothesis), denoted H_a, is the denial of the null hypothesis and tends to state "significant effect" or "significant difference." A hypothesis is either *two-tailed* (e.g., $H_0 : \mu = 0$) or *one-tailed* (e.g., $H_0 : \mu \geq 0$ or $H_0 : \mu \leq 0$).

2. Correct Decision and Error

When H_0 is true, in order word, H_a is false, you should not reject the null hypothesis; otherwise, you should reject the null hypothesis in favor of the alternative hypothesis. When H_0 is true, but you may mistakenly reject the null hypothesis. This is an erroneous decision, which is called the *Type I error*. If you do not reject the null hypothesis when H_0 is false, you makes another erroneous decision, which is called the *Type II error*.

	Do not reject H_0	*Reject H_0*
H_0 is true	Correct Decision 1-α: Confidence level	Type I Error Size of a test α: Significance level
H_0 is false	Type II Error β	Correct Decision 1-β: Power of a test

3. Component of Hypothesis Testing

A research design contains *specific models* (tests) to test the research hypothesis. Different models (tests) have different formulas to compute test statistics. Components of hypothesis testing include 1) Hypothesis to be tested, 2) Standardized effect size (test statistic): effect size and variation (variability), 3) Sample size (n) to be used when computing the degrees of freedom and standard error, and 4) Test size (significance level α) that is subjective criterion to evaluate the null hypothesis .

A *standardized effect size*, a test statistic (e.g., *z*, *t*, and *F* scores), is computed by combining the effect size and variation. One example is $z=(x-\mu)/\sigma$. An effect size in actual units of the response is the "degree to which the phenomenon exists" (Cohen 1988). Alternatively, an effect size is the deviation of the hypothesized value in the alternative hypothesis from the baseline in the null hypothesis. Variation (variability) is the standard deviation of the population. Cohen (1988) calls it the reliability of sample results. Therefore, *a standardized effect size can be viewed as unit effect size or the deviation from the baseline per variation*.

Sample size (N) is the number of observations (cases) in a sample. As N increases, the standardized effect size tends to increase because the standard error becomes smaller. In other word, it is more likely to reject the null hypothesis in favor of the alternative hypothesis when N becomes large.

The *test size* or *significance level* (α) is the probability of rejecting the null hypothesis that is true. This *size of a test* is the probability of Type I error or "the maximum tolerable probability of a Type I error" (Hildebrand 2005: 310). The Type I error occurs when a null hypothesis is rejected when it is true. This test size is denoted by α (*alpha*). The .05 level means that we are taking a risk of being wrong five times per 100 trials. In contrast, a stringent test size like .01 reports significant effects only when the effect size (deviation from the baseline) is large. Instead, the conclusion is more convincing (less risky). The 1-α is called the *confidence level*. You may also construct the (1- α) percent confidence interval as a criterion.

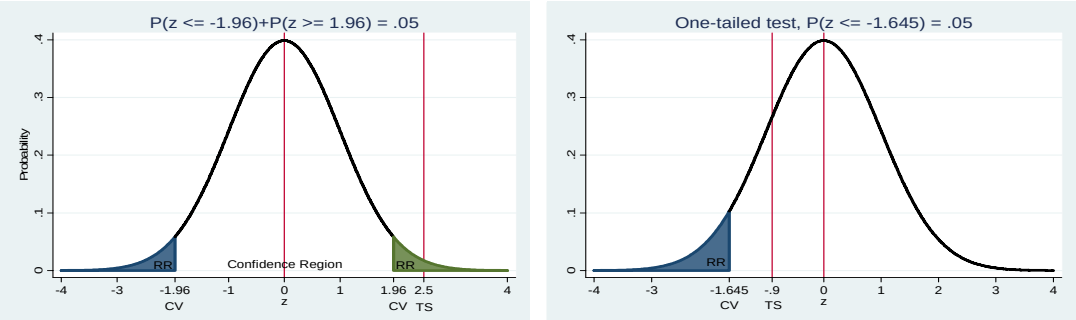
The *power of the test* (1–ß) is the probability that it will correctly lead to the rejection of a false null hypothesis (Greene 2000). Note that ß (*beta*) is the Type II error. The statistical power is the ability of a test to detect an effect, if the effect actually exists (High 2000). When the significance level increases, the power of a test decreases; they have a trade-off relationship.

4. Subjective Criteria: Critical Value and Rejection Region

Once a test size, a subjective criterion, is determined (the .10, .05, and .01 levels are conventionally used), the critical value (CV) needs to be found. A critical value is a cut point that makes the probability of occurring from the value to the (positive and/or negative) infinity the significance level. In order word, the probability that a random variable *x* is greater (or smaller) than or equal to the critical value is the significance level or the size of a test. In a two-tailed test, $P(-\infty \le z \le -z_{\alpha/2}) + P(z_{\alpha/2} \le z \le \infty)=\alpha$, where $z_{\alpha/2}$ is the critical value for α significance level. The critical values of z at the .05 and .01 level are 1.96 and 2.58 in a two-tailed test, respectively (see the left figure below).

The *rejection region (or critical region)* is the area encompassed by the critical value, the (positive and/or negative) infinity, and the probability distribution function curve, and x axis (see the shaded areas below). Why is it called the rejection region (RR)? We reject the null hypothesis if a test statistic falls into this region. This area represents the extent that you are willing to take a risk of making wrong conclusion when the null hypothesis is true. This willingness is noting but the significance level α discussed above.

The significance level determines the critical value and rejection region (critical region) of a hypothesis test. The critical value distinguishes between the rejection region and the confidence region or "probable range." The rejection region visually represents the significance level (test size) on the probability distribution (e.g., the standard normal, *t*, *F*, and Chi-squared distributions). Therefore, one implies the other.



5. Objective Criteria: Test Statistic and Its P-value

A test statistic (or test value) is computed from the sample to evaluation the null hypothesis. The information from sample is summarized in this test statistic. *F*, *z*, and *t* values are common examples of test statistics. For example, ±2.5 and -.9 in the above figures are test statistics.

The *p-value* is the probability that you get even more unlike sample than test statistic. Technically speaking, the p-value is the area underneath the probability distribution curve from the test statistic to the (positive or negative) infinity. For instance, the p-value in the left figure above is the sum of shaded areas from $-\infty$ to -2.5 and from 2.5 to ∞. In the right figure, the p-value is the area from $-\infty$ to -.9 since this is the one-tailed test.

Like the critical value and rejection region, the test statistic and its p-value are both sides of a coin. They both depend on the model and probability distribution used. For example, if you get a *t* value, you need to take a look at the *t* distribution table to get the p-value.

6. Z versus t Distribution

The textbook says if population variance σ^2 is known or the sample size is larger than 30, you may compute a *z* score to use the standard normal distribution. If σ^2 is not known and N is smaller than 30, the t distribution is used instead. This distinction reflects the fact that the t distribution is approximated to the normal distribution when N is large. In fact, when N is greater than, say, 100, the standard normal and t distribution become similar, producing the same probability, $P(z) \approx P(t)$. But if N is between 30 and 100, probably the t distribution may give you a more accurate (conservative) answer although most textbooks including ours do not report such probabilities. But I do not care much in K300.

7. Procedures of Hypothesis Testing

In general, there are five steps for hypothesis testing. The textbook explains a little bit different way.

1) State a null and alternative *hypothesis* (one-tailed or two-tailed test)

2) Determine a *test size* (*significance level*) and find the critical value and/or rejection region. You need to pay attention to whether a test is one-tailed or two-tailed to get the right critical value and rejection region.
3) Compute a *test statistic* (test value) and its *p-value* or construct the confidence interval. Collect all necessary information (e.g., N, μ, σ, and s) and compute statistics using formula dictated by the model (e.g., z-test and t-test).
4) Reject or do not reject the null hypothesis by comparing the subjective criterion in 2) and the objective test statistic calculated in 3)
5) Draw a conclusion and interpret substantively.

There are three different approaches for hypothesis testing. Each approach has both advantages and disadvantages. The important fact is that *three approaches conduct the same hypothesis test and thus give you the same conclusion*. If one of these approaches gives you a different answer, you must be off the track.

8. Classical Approach: Test Statistic versus Critical Value

The classical approach asks, "*Is the sample mean one that would likely to occur if the null hypothesis is true?*" This test statistic approach examines how far the sample statistic is away from the hypothesized value (i.e., population mean). This is *point estimation*. If the sample statistic is far away from the hypothesized value, we may suspect that the hypothesized value is not true. If the observance of the sample statistic is likely (or probable), do not reject the null hypothesis; if unlikely, reject the null hypothesis (our conjecture may be wrong).

$$P(\mu_{null} - t_{\alpha/2} \times s_{\bar{x}} < \bar{x} < \mu_{null} + t_{\alpha/2} \times s_{\bar{x}}) = 1 - \alpha$$

The "likely" and "unlikely" mean whether the sample mean is far away from the hypothesized mean. If a (objective) test statistic is smaller than the (subjective) critical value or if the test statistic is not farther away from the hypothesized mean than the critical value, it is more likely to get such a sample mean than you thought if the hypothesized value hypothesis is true. Since you have a probable sample mean, you may not change your conjecture (the null hypothesis); your conjecture appears correct. In the right figure above, the test statistic -.9 is closer to the 0 (mean) than the critical value -1.645. Thus, the sample mean is not unlikely at the .05 significance level.

If a test statistic is larger than the critical value (opposite in the left-hand side), it is less likely than you thought to get such a sample mean. The sample mean is farther away than you thought from the hypothesized mean (the null hypothesis). Therefore, you do have some reason to reject the null hypothesis. Your conjecture may be wrong. In the left figure above, the test statistic -2.5 and 2.5 are farther away from the mean than their critical value counterparts (-1.96 and 1.96). Therefore, the sample mean you got is not likely at the .05 significance level. So you need to reject your conjecture in the null hypothesis.

9. P-value Approach: P-value versus the Significance Level α

This approach asks, "*What is the probability that we get a more unlikely (extreme or odd) test statistic if the null hypothesis is true?*" What does "a p-value is smaller than the

significance level" mean? If a p-value is smaller than the significance level or falls into the rejection region, you have a smaller risk of being wrong than your subjective criterion (significance level). Thus, you become more confident to reject the null hypothesis in favor of the alternative (research) hypothesis because it is less risky than you thought to reject the null hypothesis. Keep in mind that the .05 significance level, for example, means you are willing to take a risk of making a wrong conclusion by that amount. When you reject the null hypothesis, there is still 5 percent chance that your conclusion is wrong (rejecting the true null hypothesis).

Put it differently, a small p-value, say .003, indicates that your sample mean is vary far away from the hypothesized mean. You just have only .3 percent chance to observe sample means less than the one you get if the hypothesized mean is true. Obviously, this is an extremely unlikely event. Therefore, it is less risky than you thought to reject the null hypothesis or conclude that the hypothesized mean is wrong.

In the left figure above, the p-value for the test statistic 2.5 is .0124, which is much smaller than the .05 significance level. Therefore you reject the null hypothesis. In the right figure, the p-value for the one-tailed test is .1841, which is greater than .05; do not reject the null hypothesis.

10. Modern Approach: Hypothesized Value versus the Confidence Interval

This approach asks, "*Is the hypothesized value in the null hypothesis the value (parameter) that we would estimate?*" You need to construct the (1- α) percent confidence interval using the critical value and sample statistics (i.e., standard error). This approach uses *interval estimation*. The null hypothesis expects that the hypothesized value exists somewhere in the confidence interval. Unlikely the classical approach, the modern approach focuses on the location of the hypothesized value μ_{null}.

$$P(\bar{x} - t_{\alpha/2} \times s_{\bar{x}} < \mu_{null} < \bar{x} + t_{\alpha/2} \times s_{\bar{x}}) = 1 - \alpha$$

The (1- α) percent confidence interval means that you are (1- α) percent sure that the population mean (not sample mean) exists in the (1- α) percent confidence interval. If the hypothesized value falls into the confidence interval, you may not reject the null hypothesis and conclude that "Yes, my conjecture is right."

In contrast, the presence of the hypothesized value out of the confidence interval indicates you got an odd observation (sample mean) that went beyond your prediction (confidence interval). You may not be confident that your hypothesized value is likely. Therefore, there must be something wrong in your conjecture (the null hypothesis). In other word, the true parameter does not appear to be the hypothesized value.

11. Comparison of Three Approaches

Which one is better than the others? It depends on researchers' purposes. The (classical) test statistic approach is easy to understand. The modern approach (confidence interval approach) does not require computing the test statistic and p-value. The p-value approach reports correct information about making a wrong conclusion. In general, p-value approach appears more useful than the other two approaches. Most statistical software

packages tend to focus on the p-value approach although they produce the confidence interval as well. The following table summarizes these three approaches.

	Test Statistic	**P-Value**	**Confidence Interval**
Key	*Sample statistic*	*Amount of risk taking*	*Hypothesized value*
1	Determine H_0 and H_a	Determine H_0 and H_a	Determine H_0 and H_a
2	Determine the test size α Find the critical value	Determine the test size α	Determine the test size α or 1-α (confidence level) Find the critical value
3	Compute the test statistic	Compute the test statistic Compute the p-value	Construct the (1- α)% confidence interval
4	Reject H_0 if TS > CV	Reject H_0 if p-value < α (rejection region)	Reject H_0 if the hypothesized value is beyond CI
5	Substantive interpretation	Substantive interpretation	Substantive interpretation

* TS (test statistic), CV (critical value), and CI (confidence interval)

12. Why Standard Error?

What is the standard error? Why do we have to use the standard error instead of the sample standard deviation when evaluating sample mean? The standard error is the standard deviation of the sample mean. Since we are talking about the distribution of sample mean (as opposed to the distribution of data points of a random variable), we have to use the standard deviation of the sample mean. Please pay special attention to the difference between $s_{\bar{x}}$ (standard error) and s_x (sample standard deviation).

Sample standard deviation $s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$ and standard error $s_{\bar{x}} = \frac{s_x}{\sqrt{n}}$

13. What Is Wrong with "Statistically Significant?"

Researchers often state, on the basis of a hypothesis test, that something is "statistically significant" or "statistically insignificant." The "significant" may be misleading since the word implies that something is important, large, and meaningful. As a result, some may want to say like "we confirm the null hypothesis" or "we have enough evidence to prove that the null hypothesis is true." However, we never know exactly if a null hypothesis is true or not. As shown in the above, we just evaluate the deviation of sample mean (test statistic) from the hypothesized mean (our subject criterion). Therefore, many statisticians prefer "statistically discernable" to "statistically significant" in order to emphasize the nature of hypothesis testing. Only the degree of deviation matters.

14. Good and Bad Expressions

a. At 5 percent significance level (X)→ at the .05 significance level (O)
b. At .01 significant level (X)→ at the .01 (significance) level (O)
c. The .05 confidence interval (X)→ the 95 percent confidence interval (O)
d. Accept the null hypothesis (X)→ Do not reject the null hypothesis (O)
e. Do not confirm the null hypothesis (X)→ Reject the null hypothesis (O)
f. Have sufficient evidence to prove H_0 (X) → Do not reject H_0 at the .05 level (O)
g. We cannot believe that H_0 is true (X) → Reject H_0 at the .01 level (O)