David Youngberg
BSAD 210—Montgomery College

LECTURE 16: HYPOTHESES AND TYPES OF ERROR

I. Introduction
   a. Hypothesis testing is the bread and butter of statistics. By The Central Limit Theorem, we know that samples can be, by chance, unusually high or unusually low.
   b. When wondering if some treatment (launching an ad campaign, hiring a new worker, changing a production process) had an effect on something, you don't know if any difference you detect is due to random chance or is the result of a genuine change the population. That's what hypothesis testing is for.
   c. So let's begin with what statistician mean by "hypothesis."

II. Hypotheses
   a. *Null Hypothesis*—assertion corresponding to the default position, where there is no significant different, or where nothing is happening
      i. The null hypothesis captures the state where nothing makes a difference (even if the intuition is that it should).
      ii. Example: Income ***doesn't*** predict how much you spend
      iii. Example: An hour of exercise each day for a year ***won't*** cause people to lose weight.
   b. *Alternative Hypothesis*—assertion that claims there is a significant difference.
      i. Example: Income ***does*** predict how much you spend
      ii. Example: An hour of exercise each day for a year ***will*** cause people to lose weight.
   c. Alternative hypothesis can be either one-tailed or two-tailed.
      i. Example: More income predicts that you spend less (one-tailed), that you spend more (one-tailed), or that you spend either more or less (two-tailed).
      ii. Example: An hour of exercise each day for a year will cause people to lose weight (one-tailed), or their weight will change (two-tailed).
      iii. We usually focus on two-tailed tests.
   d. *Level of significance*—determines a cut-off point when the null hypothesis is rejected or failed to be rejected. The standard is 95% (or 1.96).

e. There are two types of mistakes you can make when working with a null hypothesis.
    i. *Type I error*—false positive; a non-match is declared a match—or when you reject the null hypothesis and you should fail to reject it
    ii. *Type II error*—false negative; a good match is not detected—or when you fail to reject the null hypothesis and you should reject it

Examples

Type I	*Type II*
Convicting an innocent person	Letting the guilty go free
Approving a damaging drug	Rejecting a beneficial drug
Befriending a jerk	Ignoring a nice person
Funding a poor investment	Passing on a good investment

f. Type I and Type II errors are equally undesirable, but Type II errors are insidious because they are harder to notice when they happen.
    i. *In general*, Type I errors are self correcting; Type II errors are not. But precisely because Type I errors are self correcting, the fact that one made an error at all is evident thus there is a tendency for people to commit Type II errors.

III. The Critical Values

a. Let's revisit that list from an earlier lecture. Notice the expansion.

Confidence	α	$z_{\alpha/2}$	z_{α}
90%	10%	1.645	1.280
95%	5%	1.960	1.645
99%	1%	2.576	2.330
99.9%	0.1%	3.291	3.090

b. The additional column is for when the significance level is concentrated on just one side of the distribution.
c. This brings up the difference between a one-tail hypothesis test and a two-tail hypothesis test.
    i. In a *one-tail hypothesis test* the alternative hypothesis is stated with a "<" or a ">" and the null hypothesis is stated with a "≥" or a "≤", as appropriate.

ii. In a *two-tail hypothesis test* the alternative hypothesis is stated with a "≠" and the null hypothesis is stated with a "=".

d. There is no difference in equation when you consider a one-tail or two-tail test. The only difference is the significance levels.
   i. You can use the t-distributions to tell you one- or two-tail values. Note in the table above, the one-tail z-score at 95% is identical to the two tail score at 90%. That's because in both cases the number of observations under <u>one</u> tail is 5%.

IV. One-Tail or Two?
   a. The question then becomes: when should you use one-tail or two-tails? This involves answering a different question: what do people care about?
   b. One-tail tests are best for claims of improvement, where doing worse is effectively the same as doing average; neither is impressive. One-tail tests also used to refute points of view (someone might say something is popular so the alternative is that it's not popular).
      i. <u>Examples</u>: Longer battery life; faster acceleration time; the popularity of gay marriage.
      ii. One-tail tests are great because you get to claim a higher confidence level with the same z-score. If, for example, your z-test is 1.7, you're significant at the 95% level but for a two-tailed test, you'd only be significant at the 90% level.
   c. Two-tail tests are best for when you're trying to detect unusualness in either direction. In other words, there's a "sweet spot" that the null hits but the alternative doesn't. The question is *not* if the value is more or less than the average but if the value is *different*.
      i. <u>Examples</u>: the accuracy of a machine putting ketchup in a bottle (could be putting too much or too little in); the length of time you must stand in line at a store (both too little or too much is noteworthy); if a new employee is unusually good or bad at the job.
      ii. Because a two-tailed test has a higher standard than a one-tailed test, it's what you use when you're not sure.

V. Simple tests of hypothesis
   a. Terminology
      i. The calculated value is the result of your calculation. You take the absolute value when comparing.
      ii. The critical value is value you compare the calculated value to.
      iii. Sometimes they are called scores instead of values.

b. The basics idea is that you'll calculate a z-value and compare the absolute value to a critical value.
    i. If the absolute value of your calculated score is greater than the critical value, you reject the null hypothesis. The difference is probably not due to chance.
        1. It can <u>still</u> be due to chance. We could be committing Type I error: 95% isn't 100. The phrase "statistically significant" comes from this idea that passing the threshold means the results are interesting, even if nothing's been "proven."
    ii. If the absolute value of your calculated score is less than the critical value, you fail to reject the null hypothesis.
        1. Why "fail to reject" rather than "accept?" Because that sounds like we're saying the null hypothesis is true; it may not be. In other words, we could be committing Type II error.
    iii. Note that as α decreases, the critical value increases. Thus, if you reject the null hypothesis at a particular α, you should reject the null at a higher α.
c. What should matter when determining if you should reject or fail to reject the null hypothesis?
    i. How different the sample average is from the population average. If the difference is really large, then the sample mean is far from the center of the distribution and thus the likelihood this sample is different by chance falls.
    ii. If values for the variable tend to fall in a narrow range, it should be less likely (all things other being equal) the less likely the null hypothesis is correct. The lower the standard deviation, the lower the likelihood that the sample mean is different by chance.
    iii. If you have a lot of observations, you know more about the population you sampled from. The bigger the sample, the less likely the sample mean is different by chance.
d. The larger the absolute value of the z-score, the greater the chance that you'll reject the null hypothesis.

VI. Known σ
a. If you know the population standard deviation:

$$z_{\bar{x}} = \left| \frac{\bar{x} - \mu_{H_0}}{\sigma/\sqrt{n}} \right|$$

i. Where $z_{x\text{-bar}}$ is the z-test statistics;
ii. x-bar is the sample mean;
iii. μ_{H0} is the mean of the sample distribution, which is assumed to be true for the null hypothesis;
iv. σ is the population standard deviation; and
v. n is the sample size.
vi. Note that we take the absolute value; this only for purposes of comparing to critical values.

b. Example: You want to know if your new sports drink improves athletic performance over what people normally do. Suppose the average athlete can run a mile in exactly 15 minutes with a standard deviation of exactly 2 minutes. You give 100 athletes your sports drink and time their run. Your sample mean is exactly 14 minutes.
   i. Your null hypothesis is that it does nothing to reduce the time it takes to run a mile: x-bar $\geq \mu$
   ii. Your alternative hypothesis is that it reduces the time it takes to run a mile: x-bar $< \mu$.
   iii. Here's what your equation should look like:

$$z_{\bar{x}} = \left|\frac{\bar{x} - \mu_{H_0}}{\sigma/\sqrt{n}}\right| = \left|\frac{14 - 15}{2/\sqrt{100}}\right| = \left|\frac{-1}{2/10}\right| = \left|\frac{-1}{0.2}\right| = 5.00$$

   iv. Note that because this is a one-tailed test, our critical z-scores are 1.645 (95%); 2.330 (99%); and 3.090 (99.9%). Because the absolute value of -5 is greater than 3.090, we have evidence—strong evidence—that our sports drink improves performance.

c. You might be tempted to say that we've "proved" our sports drink make a positive difference. We haven't "proved" anything. Because there's always a chance of luck, statisticians state that we have evidence for something. "Proof" is not an option.

d. The example also highlights the difference between statistical significance and practical significance.
   i. Yes, our sports drink really does improve performance (statistical significance).
   ii. But is a minute really that big of a deal? If the answer is no, then it's not really interesting and people might not be willing to spend the money on it (practical significance).

VII. Unknown σ

a. You don't know the population standard deviation:

$$t_{\bar{x}} = \left| \frac{\bar{x} - \mu_{H_0}}{s/\sqrt{n}} \right|$$

i. Where $t_{x\text{-bar}}$ is the z-test statistics; and
ii. s is the standard deviation of the sample.

VIII. Proportion

a. You are using a proportion:

$$z = \left| \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \right|$$

i. Where π is the estimate of the population's proportion; and
ii. p is the sample estimate.