Stat 401, section 9.1 z Tests Comparing Two Population Means

notes by Tim Pilachowski

In chapter 9, we turn to comparing two groups with each other.

One scenario would be comparing genders, geography, levels of education, etc. That is, the same parameter is being measured in two different populations ("independent samples" – sections 9.1 and 9.2).

Another scenario would involve two groups from the same population, measuring different treatments or comparing a treatment group with a control group (either "independent samples" or "paired data" – section 9.3).

We'll be relying on groundwork done in Lecture 3.6b (Change of Variables) and used in Chapter 5 (Joint Probability Distributions).

In section 5.3, we outlined necessary requirements for a random sample of size n taken from a population. We now expand these to apply to two random samples taken from two populations.

basic assumptions:
Random variables $X_1, X_2, X_3, \ldots, X_m$, are from a population distribution with mean μ_1 and variance $(\sigma_1)^2$.
Random variables $Y_1, Y_2, Y_3, \ldots, Y_n$, are from a population distribution with mean μ_2 and variance $(\sigma_2)^2$.
The X and Y samples are independent of one another.

Note that it is not necessary that the two sample sizes are equal. In other words, it is possible that $m \neq n$.

When comparing two populations, we'll be considering the parameter $\mu_1 - \mu_2$. When we have two independent randomly-chosen large samples, we'll use the two sample means and two sample standard deviations (or variances) in our point estimates of expected value and standard error.

Proposition: $\overline{X} - \overline{Y}$ is an unbiased estimator of $\mu_1 - \mu_2$. The standard deviation of $\overline{X} - \overline{Y}$ is $\sigma_{\overline{X}-\overline{Y}} = \sqrt{\frac{(\sigma_1)^2}{m} + \frac{(\sigma_2)^2}{n}}$. (For the definition of "unbiased", see Lecture 6.1.)

notes on the proofs

$$E\left(\overline{X} - \overline{Y}\right) = E\left(\overline{X}\right) - E\left(\overline{Y}\right)$$
$$= \mu_1 - \mu_2$$

$$V\left(\overline{X} - \overline{Y}\right) = V\left(\overline{X}\right) + V\left(\overline{Y}\right)$$
$$= \frac{(\sigma_1)^2}{m} + \frac{(\sigma_2)^2}{n}$$

$$\sigma_{\overline{X}-\overline{Y}} = \sqrt{\frac{(\sigma_1)^2}{m} + \frac{(\sigma_2)^2}{n}}$$

Previous work in chapters 5 and 6 provides us with important and useful information. If both population distributions are normal, then the sampling distributions for both $\overline{X}$ and $\overline{Y}$ are normal.

Furthermore, independence of the two samples implies that $\overline{X}$ and $\overline{Y}$ are independent of one another.

Thus, the difference $\overline{X} - \overline{Y}$ will be normally distributed, and we can consider this a large sample case.

The processes for finding a confidence interval and conducting a hypothesis test will proceed along much the same lines as did the single large sample cases in chapters 7 and 8.

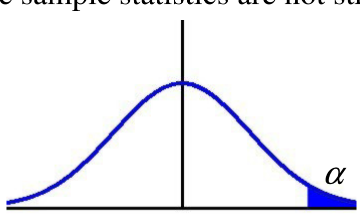
Example A – confidence interval: In doing a survey of textbook costs, Rodney found that at Whatsammatta U., in a sample of 42, the mean cost was \$63.42 with a known (from prior studies) σ = \$8.13. At Matriarch College (affectionately known as U. Mama) the mean textbook cost was \$67.19 in a sample of 35 with a known σ = \$7.29. Construct a 95% confidence interval for the difference between the mean costs of textbooks at the two schools. *answer*: $\approx$ $(-7.22, -0.32)$

For a hypothesis test, the initial assumption ("prior belief claim") is called the **null hypothesis**, and is denoted by H_0.

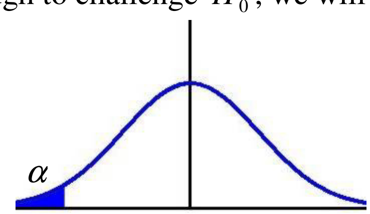
The challenge to the null hypothesis is called the **alternate hypothesis** ("assertion that is contradictory to H_0") and is denoted by H_a.

If the sample statistics present strong enough evidence that H_0 is false, we will "reject H_0".

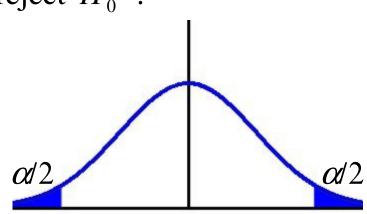
If the sample statistics are not strong enough to challenge H_0, we will "fail to reject H_0".



$$H_0: \mu_1 - \mu_2 = \Delta_0$$
$$H_a: \mu_1 - \mu_2 > \Delta_0$$



$$H_0: \mu_1 - \mu_2 = \Delta_0$$
$$H_a: \mu_1 - \mu_2 < \Delta_0$$



$$H_0: \mu_1 - \mu_2 = \Delta_0$$
$$H_a: \mu_1 - \mu_2 \neq \Delta_0$$

The large sample case test statistic is $z = \dfrac{(\bar{x} - \bar{y}) - \Delta_0}{\sqrt{\dfrac{(\sigma_1)^2}{m} + \dfrac{(\sigma_2)^2}{n}}}$.

Example A – hypothesis test: In doing a survey of textbook costs, Rodney found that at Whatsammatta U., in a sample of 42, the mean cost was \$63.42 with known σ = \$8.13. At Matriarch College (affectionately known as U. Mama) the mean textbook cost was \$67.19 in a sample of 35 with σ = \$7.29. Test the hypothesis (α = 0.05) that textbooks cost less at Whatsammatta U. than they do at U. Mama. *answer*: reject H_0, P-value = 0.0162

As before, the CLT assures us that, if sample sizes are sufficiently large, even if we do not know the population variances, we can construct confidence intervals and conduct hypothesis tests of the difference between means as a large-sample scenario. The rule of thumb is that, if both m and n are greater than 40, then sample variances can be used as point estimates for population variances for both confidence intervals and hypothesis tests.

Example B: In doing a survey of hours spent watching educational television per month, Regina found that in Pleasantville, in a sample of 41 households, the mean was 60 hours with a sample variance of 64. In Nicetown the mean was 63 hours in a sample of 50 households with a sample variance of 81. Is there a statistically significant difference between the two towns' educational television viewing habits? (α = 0.05)
answer: fail to reject H_0, P-value = 0.0930

Two Final Notes: 1) As we have done for recent sections, you are not responsible for knowing "β and the Choice of Sample Size".

2) Next, a short side jaunt to consider "Using a Comparison to Identify Causality". (Read your text, also.)

Back in chapter 5 we encountered correlation – a relationship between two non-independent random variables. At the time, we noted that while the numeric value of the correlation coefficient indicated the strength or weakness of the connection, the correlation may or may not be statistically significant. (We'll conduct the hypothesis test in Chapter 12.)

Also, we noted that, even when correlation tests to be significant, "correlation" is not the same thing as "cause". The existence of a statistical relationship does not confirm direction – X may cause Y, Y may cause X, or there may be outside factors that affect both X and Y in a similar fashion.

So how do statisticians go about supporting a claim of "cause and effect"? One approach is comparison of populations.

For example, in a Stat 400 class in a given semester, it is observed [observational study] that Mr. Jones' students did not do as well as Ms. Smith's students. Can we conclude that Ms. Smith is a more effective teacher than Mr. Jones? [cause: method of teaching, effect: student performance] Probably not – there are too many uncontrolled variables: student background and preparation, student choice of section based on time of day and/or reputation of teacher, one teacher whose requirements are easier than the others, etc.

However, if external variables are controlled [by random selection of students into sections, matching of student background characteristics between sections, use of the same tests in both sections, results replicated over a number of semesters, control group vs. treatment group, etc.], statisticians can build a case for "cause and effect". In an analogy to a court case, statisticians demonstrate a "preponderance of the evidence" to support their conclusion.

A case in point: Smoking cigarettes and lung disease. As early as the 1920s, it was generally known that cigarette smoking and lung cancer were correlated, but tobacco companies were able to claim that cause and effect had not been demonstrated. It was not until the 1950s that controlled studies were designed and conducted, and replicated, providing significant statistical evidence suggesting that cigarette smoking caused lung cancer and other diseases. The Royal College of Physicians in Britain appointed a committee to investigate the relationship between smoking and health, and the committee's report, issued on March 7, 1962, clearly indicted cigarette smoking as a cause of lung cancer and bronchitis, and indicated that it probably contributed to cardiovascular disease as well.

In the U.S., the Surgeon General's Advisory Committee on Smoking and Health released a similar report on January 11, 1964: *Smoking and Health: Report of the Advisory Committee to the Surgeon General of the United States*. In this report, the committee concluded that lung cancer and chronic bronchitis are causally related to cigarette smoking. The report also stated that there was suggestive evidence, if not definite proof, for a causative role of smoking in other illnesses such as emphysema, cardiovascular disease, and various types of cancer. This report was directly responsible for the passage of the Cigarette Labeling and Advertising Act of 1965, which mandated the familiar Surgeon General's health warnings on cigarette packages.

Since that time, much work has been done to identify biomedical links between cigarette smoking and various diseases, and to provide evidence that secondhand smoke can be almost as dangerous as smoking itself. As of 2016, the practice of "vaping" has not been around long enough to provide evidence of long-term effects, but preliminary observations indicate some of the same health risks.

Sources:

U.S. Department of Health and Human Services. *How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease: A Report of the Surgeon General.* Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, 2010.
http://www.ncbi.nlm.nih.gov/books/NBK53017/

U.S. Department of Health and Human Services. *The health consequences of smoking – 50 years of progress: a report of the Surgeon General*. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, 2014.
http://www.surgeongeneral.gov/library/reports/50-years-of-progress/full-report.pdf