STA 570 — Spring 2012 — Dr. Charnigo

Lecture 5

Introduction to hypothesis testing

Motivation. A researcher who formulates a scientific hypothesis about a population would like to determine whether that hypothesis is supported by sample data. If she can formulate the hypothesis in terms of a population parameter (e.g., a mean μ, a variance σ^2, or a proportion p), then the hypothesis testing framework to be developed presently allows such a determination to be made objectively. For example, she might hypothesize that the population mean reduction in serum cholesterol after one month on a vegetarian diet is in excess of 10 points. Symbolically, such a hypothesis would be written as "$\mu > 10$".

Null and alternative hypotheses. In our hypothesis testing framework, we actually specify two competing hypotheses: a null hypothesis (often denoted H_0) and an alternative hypothesis (often denoted H_1 or H_a). The null hypothesis represents a default position that we will reluctantly retain in the absence of sufficient evidence to the contrary. The alternative hypothesis represents a new position (i.e., a scientific discovery) that we will gladly embrace if the sample data provide sufficient evidence in its favor. In any case, our decision to retain or reject the null hypothesis will be made objectively, on the basis of the sample data.

Asymmetry and court analogy. There is an inherent asymmetry in our hypothesis testing framework: the burden of proof, so to speak, is attached to the alternative hypothesis. Thus, since the researcher wishes to be convincing in advocating her scientific hypothesis, she usually chooses the alternative hypothesis to be her scientific hypothesis and chooses the null hypothesis to be its logical negation.

An analogy may be drawn to a criminal trial in the United States legal system. The defendant [null hypothesis] is suspected of being guilty [false], otherwise he would not be on trial [the study would not be conducted]. However, the defendant [null hypothesis] cannot be convicted [rejected] unless the evidence supporting conviction [rejection] is sufficient. If the evidence is not sufficient, then the defendant [null hypothesis] must be released [retained], even though the jury [researcher] may not truly believe that the defendant [null hypothesis] is innocent [true].

Just as the legal system labels a defendant who is not convicted as "not guilty" rather than "innocent", some researchers prefer to say "I fail to reject the null hypothesis" instead of "I accept the null hypothesis".[1]

Example (null and alternative hypotheses). Consider the researcher who hypothesizes that the population mean reduction in serum cholesterol after one month on a vegetarian diet is in excess of 10 points. To be convincing in advocating her scientific hypothesis, the researcher chooses the alternative hypothesis to be "$\mu > 10$" and the null to be "$\mu \leq 10$".

We could express these choices by saying that the researcher is testing $H_0 : \mu \leq 10$ against $H_1 : \mu > 10$. However, a common convention would be to abbreviate the null hypothesis from "$\mu \leq 10$" to "$\mu = 10$", in which case we would write that the researcher is testing $H_0 : \mu = 10$ against $H_1 : \mu > 10$. The reason for this is to facilitate probabilistic calculations.[2]

[1] I regard this mostly as an issue of semantics. Ultimately, a researcher is either successful or unsuccessful in advocating the alternative hypothesis; the wording used to describe an unsuccessful attempt is of secondary concern.

[2] For instance, if we ask about the distribution of $(\bar{X} - 10)/(\sigma/\sqrt{n})$ when the null hypothesis is true,

Type I error and significance level. Suppose that the null hypothesis is true. If you accept the null hypothesis, then you have made a correct decision. If you reject the null hypothesis in favor of the alternative, then you have made a Type I error (Definition 7.2). The probability of making a Type I error is typically denoted by α (Definition 7.4), and we refer to α as the significance level of the hypothesis test.

Type II error and power. Suppose that the alternative hypothesis is true. If you accept the alternative hypothesis, then you have made a correct decision. If you accept the null hypothesis, then you have made a Type II error (Definition 7.3). The probability of making a Type II error is typically denoted by β (Definition 7.5), and we refer to $1 - \beta$ as the power of the hypothesis test (Definition 7.6).

Significance level versus power. There is a tradeoff between significance level and power for a hypothesis test. You can be assured of never making a Type I error if you always accept the null hypothesis; however, the price of having $\alpha = 0$ is having $\beta = 1$ (i.e., no power). Also, you can be assured of never making a Type II error if you always reject the null hypothesis; however, the price of having $\beta = 0$ is having $\alpha = 1$. Often we fix α at 0.05 and then try to choose the sample size large enough so that β is not unreasonably large (e.g., not more than 0.20).

---

we will be assuming that $\mu = 10$ rather than allowing μ to be some unspecified number no larger than 10. Describing the distribution of $(\bar{X} - 10)/(\sigma/\sqrt{n})$ in the former scenario is relatively easy, but describing its distribution in the latter scenario is rather difficult.

Hypothesis test for a population mean (one-sided)

Introduction. Consider testing $H_0 : \mu = \mu_0$ against the "one-sided" alternative $H_1 : \mu > \mu_0$. Since $\bar{x}$ is an estimate of μ, we will be inclined to reject H_0 in favor of H_1 if $\bar{x}$ is much larger than μ_0. The question is, how much larger than μ_0 must $\bar{x}$ be?

Preliminary formulation. Suppose that the null hypothesis is true. If the population is normal, then $\bar{X}$ is normal with mean μ_0 and variance σ^2/n. If the population is not normal but $n \geq 30$, we may apply the Central Limit Theorem to conclude that $\bar{X}$ is approximately normal with mean μ_0 and variance σ^2/n. For brevity we will suppress the word "approximately" in what follows.

Define

$$Z := \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}},$$

which is standard normal since $\bar{X}$ is normal with mean μ_0 and standard deviation $\sigma/\sqrt{n}$. Note that $P(Z > z_{1-\alpha}) = \alpha$. Hence, if we want a test with significance level α such that large $\bar{x}$ leads to rejection of the null hypothesis, we can agree to reject the null hypothesis if

$$z := \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha}.$$

In this context, both random Z and numerical z are commonly referred to as a "test statistic", while $z_{1-\alpha}$ is called a "critical value".

Practical formulation. In practice σ is unknown. If $n \geq 200$, we may simply replace σ by s. If $n < 30$ and the population is normal, we may replace σ by s if we also replace Z, z, and $z_{1-\alpha}$ by T, t, and $t_{n-1,1-\alpha}$ (Equation 7.6).

If $30 \leq n < 200$, there are differing opinions about what to do. Some people favor using T, t, and $t_{n-1,1-\alpha}$. Others prefer to use Z, z, and $z_{1-\alpha}$. My opinion is that the former strategy is better if the population is normal, while the latter strategy is better if the population cannot be assumed normal.

Example (practical formulation). Taking $\alpha = 0.05$ and assuming normality, let us test $H_0 : \mu = 10$ against $H_1 : \mu > 10$ for the serum cholesterol data (Table 2.13, page 39). Having dealt with this data set many times before, we know that $\bar{x} = 19.54$, $s^2 = 282.4$, $s = \sqrt{282.4} = 16.80$, and $n = 24$. Noting that $\mu_0 = 10$, we compute

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{19.54 - 10}{16.80/\sqrt{24}} = 2.782.$$

Since $t_{23,0.95} = 1.714$, we reject the null hypothesis and conclude that the population mean reduction in serum cholesterol after one month on a vegetarian diet is greater than 10 points.

The p-value. The p-value is the probability of getting a result at least as extreme as what we actually obtained, calculated supposing that the null hypothesis were true (Definition 7.14). The p-value may also be characterized as the smallest significance level at which we could have rejected the null hypothesis (Definition 7.13). Hence, a small p-value is supportive of the alternative hypothesis.

Your textbook author offers guidelines for assessing p-values in Equations 7.4 and 7.5. A result is called statistically significant if the p-value is less than 0.05 (since 0.05 is the usual choice for α). In the present context, a formula for the p-value is $P(Z > z)$ or $P(T > t)$, calculated supposing that the null hypothesis were true.

Example (the p-value). Continuing from our previous example, the p-value would be $P(T > 2.782)$, where T has a T distribution on 23 degrees of freedom. The p-value turns out to be 0.0053, but Table 5 in the back of your book does not provide a mechanism for you to recover such a p-value. Even so, you could note that the p-value must be less than 0.01 (because 2.782 is greater than the 2.500 that would be required for rejection of the null hypothesis if the significance level were 0.01) but greater than 0.005 (because 2.782 is less than the 2.807 that would be required for rejection of the null hypothesis if the significance level were 0.005).

Reversing the direction. If we are testing $H_0 : \mu = \mu_0$ against $H_1 : \mu < \mu_0$, then having $\bar{x}$ much less than μ_0 is what will persuade us to reject H_0 in favor of H_1. We will reject H_0 if $z < -z_{1-\alpha}$ or $t < -t_{n-1,1-\alpha}$ (Equation 7.2). In this case, the p-value is $P(Z \leq z)$ or $P(T \leq t)$.

Hypothesis test for a population mean (two-sided)

Introduction. Consider testing $H_0 : \mu = \mu_0$ against the "two-sided" alternative $H_1 : \mu \neq \mu_0$. If $\bar{x}$ is much larger or much smaller than μ_0, we will be persuaded to reject H_0 in favor of H_1.

Preliminary formulation. Suppose that the null hypothesis is true. If the population is normal, then $\bar{X}$ is normal with mean μ_0 and variance σ^2/n. If the population is not normal but $n \geq 30$, we may apply the Central Limit Theorem to conclude that $\bar{X}$ is approximately normal with mean μ_0 and variance σ^2/n. For brevity we will suppress the word "approximately" in what follows.

Define

$$Z := \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}},$$

which is standard normal since $\bar{X}$ is normal with mean μ_0 and standard deviation $\sigma/\sqrt{n}$. Note that $P(|Z| > z_{1-\alpha/2}) = \alpha$. Hence, if we want a test with significance level α such that either very large $\bar{x}$ or very small $\bar{x}$ leads to rejection of the null hypothesis, we can agree to reject the null hypothesis if $|z| > z_{1-\alpha/2}$, where

$$z := \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}.$$

Practical formulation. In practice σ is unknown. If $n \geq 200$, we may simply replace σ by s. If $n < 30$ and the population is normal, we may replace σ by s if we also replace Z, z, and $z_{1-\alpha/2}$ by T, t, and $t_{n-1,1-\alpha/2}$ (Equation 7.10). If $30 \leq n < 200$, there are differing opinions about what to do. The p-value is $2P(T \leq t)$ or $2P(Z \leq z)$ if t or z is negative. If t or z is positive, then the p-value is $2P(T > t)$ or $2P(Z > z)$.

Power of a hypothesis test

Introduction. The significance level α reflects how strictly we guard against making a Type I error. We must also be concerned with β, the probability of making a Type II error. Let us discuss how to compute β. Our discussion will take place in the context of hypothesis tests concerning a population mean (Equations 7.19 to 7.21).

Power for a one-sided test. Suppose that we are testing $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$ and (temporarily) that we can treat σ as known. Suppose, moreover, that $\mu = \mu_1 > \mu_0$. In this case, the distribution of $\bar{X}$ is centered at μ_1 rather than at μ_0 (Figure 7.6). Consequently,

$$Z := \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

is not standard normal. Rather,

$$\frac{\bar{X} - \mu_1}{\sigma/\sqrt{n}} = \frac{\bar{X} - \mu_0 + \mu_0 - \mu_1}{\sigma/\sqrt{n}} = Z + \frac{(\mu_0 - \mu_1)}{\sigma/\sqrt{n}}$$

is standard normal. Hence,

$$\beta = P(Z \leq z_{1-\alpha}) = P\left(Z + \frac{(\mu_0 - \mu_1)}{\sigma/\sqrt{n}} \leq z_{1-\alpha} + \frac{(\mu_0 - \mu_1)}{\sigma/\sqrt{n}}\right) = \Phi\left(z_{1-\alpha} + \frac{(\mu_0 - \mu_1)}{\sigma/\sqrt{n}}\right).$$

Thus, by symmetry of the standard normal distribution, the power is

$$1 - \beta = \Phi\left(-z_{1-\alpha} + \frac{(\mu_1 - \mu_0)}{\sigma/\sqrt{n}}\right).$$

In practice we must decide what to use for μ_1 and σ in the formula above. This decision can be based on data from a pilot study or on scientific knowledge.[3]

Example (power for a one-sided test). Suppose that the mean reduction in diastolic blood pressure associated with a new antihypertensive agent ("Hypertension", page 287) is $\mu_1 = 4.8$ and that the standard deviation is

[3] You may wonder why the textbook author does not present a formula with $-t_{n-1,1-\alpha}$ instead of $-z_{1-\alpha}$ and the cumulative distribution function for a T random variable on $(n-1)$ degrees of freedom instead of the standard normal cumulative distribution function. I have seen such a modification suggested, but there is little difference in the answer unless n is extremely small. Moreover, the entire computation is rather crude anyway since we do not really know μ and σ.

$\sigma = 9.0$. If we intend to test $H_0 : \mu = 0$ against $H_1 : \mu > 0$ at significance level $\alpha = 0.05$, we may wonder what the power is with $n = 24$. We have $-z_{1-\alpha} = -z_{0.95} = -1.645$, so the power is

$$\Phi\left(-1.645 + \frac{(4.8 - 0)}{9.0/\sqrt{24}}\right) = \Phi(0.968) = 0.833.$$

Reversing the direction. Suppose that we are testing $H_0 : \mu = \mu_0$ against $H_1 : \mu < \mu_0$. Suppose, moreover, that $\mu = \mu_1 < \mu_0$. In this case (Figure 7.5),

$$\beta = 1 - \Phi\left(-z_{1-\alpha} + \frac{(\mu_0 - \mu_1)\sqrt{n}}{\sigma}\right) \quad \text{and} \quad 1 - \beta = \Phi\left(-z_{1-\alpha} + \frac{(\mu_0 - \mu_1)\sqrt{n}}{\sigma}\right).$$

Power for a two-sided test. Suppose that we are testing $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$. Suppose, moreover, that $\mu = \mu_1 \neq \mu_0$. In this case (Figure 7.8),

$$\beta = 1 - \Phi\left(-z_{1-\alpha/2} + \frac{(\mu_0 - \mu_1)\sqrt{n}}{\sigma}\right) - \Phi\left(-z_{1-\alpha/2} + \frac{(\mu_1 - \mu_0)\sqrt{n}}{\sigma}\right),$$

so that the power is

$$1 - \beta = \Phi\left(-z_{1-\alpha/2} + \frac{(\mu_0 - \mu_1)\sqrt{n}}{\sigma}\right) + \Phi\left(-z_{1-\alpha/2} + \frac{(\mu_1 - \mu_0)\sqrt{n}}{\sigma}\right).$$

If $\mu_0 < \mu_1$, then the first summand will be close to 0. If $\mu_0 > \mu_1$, then the second summand will be close to 0. Either way, an approximate formula for the power is

$$\Phi\left(-z_{1-\alpha/2} + \frac{|\mu_0 - \mu_1|\sqrt{n}}{\sigma}\right).$$

Choosing the sample size

Introduction. We have discussed how to calculate power for a given sample size. We can also choose the sample size to attain a desired power. Let us discuss how this is done in the context of hypothesis tests concerning a population mean (Equations 7.26 through 7.28).

Sample size for a one-sided test. If we are testing $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$, then the formula for power can be algebraically rearranged as

$$n = \frac{\sigma^2(z_{1-\beta} + z_{1-\alpha})^2}{(\mu_0 - \mu_1)^2}.$$

We always round the result up to the next positive integer. The same formula applies if we are testing $H_0 : \mu = \mu_0$ against $H_1 : \mu < \mu_0$.[4]

Example (sample size for a one-sided test). Continuing the previous example, suppose that we want to know what sample size will yield 90% power. Noting that $z_{1-\alpha} = z_{0.95} = 1.645$ and $z_{1-\beta} = z_{0.90} = 1.282$, we obtain

$$n = \frac{9.0^2(1.282 + 1.645)^2}{(0 - 4.8)^2} = 30.12 \approx 31.$$

Sample size for a two-sided test. If we are testing $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$, then

$$n = \frac{\sigma^2(z_{1-\beta} + z_{1-\alpha/2})^2}{(\mu_0 - \mu_1)^2}.$$

[4] One could make an argument for replacing $z_{1-\beta}$ and $z_{1-\alpha}$ by quantiles of a T distribution. The main problem here is that we wouldn't know which T distribution we should use (i.e., how many degrees of freedom there should be) since we are solving for n. There are ways to get around this problem, but attempting to do so doesn't seem worthwhile for the reasons given earlier.