MATH 243: HYPOTHESIS TESTING

N. CHRISTOPHER PHILLIPS

Procedures

Exams to be returned in discussion sections.

Grading corrections due in writing at the beginning of Wednesday's lecture.

I am away for the next three lectures (and office hours are cancelled).
Wednesday: David Levin.
Friday: David Levin.
Monday: Dev Sinha.

Midterm 1 solutions to be posted later this week.

Examples of questions

Example 1: We know US women of ages 20–29 years have mean height 64 inches. Are women in this age group in Eugene taller on average?

Example 2: Do women in this age group in Eugene differ, on average, from those in the general population?

Example 3: Are female intercollegiate soccer players taller than average?

A warning

As we will see, we will not detect very small differences. Indeed, it is almost certainly true that the mean height of Eugene women of ages 20–29 years is not the same as the mean height of US women of ages 20–29 years. However, if the difference is small enough, we will not detect it.

Example 4: Does being left handed affect IQ?

Date: 5 May 2008.

Example 5: Suppose we know the mean weight of spiral-horned snorkacks. Are crumple-horned snorkacks, on average, smaller (lighter)?

Example 6: You suspect that the Consolidated Generic Unhealthy Foods Corporation is, on average, underfilling its 12 ounce bags of potato chips.

Example 7: You have bought a large quantity of 7 mm ball bearings from Wang's Widgets Inc., and you are preparing to sue Wang's Widgets Inc. because, you claim, the diameter of the balls in the ball bearings is not as advertised.

The basic idea

We choose, for example, a simple random sample of balls in 7 mm ball bearings from Wang's Widgets Inc., and find their mean diameter $\overline{x}$ (in mm). If $\overline{x}$ is far enough away from 7, then we interpret this as evidence that the true mean diameter μ is different from 7.

Suppose, for example, that the standard deviation of the diameters of the balls is known to be 0.1 (in mm), and that we chose a simple random sample of size 100.

Suppose we got $\overline{x} = 7.03$. The sampling standard deviation is $\sigma/\sqrt{n} = 0.1/\sqrt{100} = 0.01$. If the true mean were 7, it would be unlikely that we got a sample mean as far away as $\overline{x} = 7.03$. Indeed, the probability that the sample mean $\overline{x}$ is in (6.997, 7.003) is (by the Rule of Thumb) about 0.997, leaving only a very small probability of about 0.003 that $\overline{x} \leq 6.97$ or $\overline{x} \geq 7.03$.

We interpret this as strong evidence that the true mean μ was not, after all, equal to 7.

Suppose instead we got $\overline{x} = 7.003$. The sampling standard deviation is $\sigma/\sqrt{n} = 0.1/\sqrt{100} = 0.01$. If the true mean were 7, our $\overline{x}$ would be only 0.3 sampling standard deviations away from the true mean. Using Table A (details omitted), I found that the probability that $\overline{x}$ is that far, or farther, from μ is about 0.7642. Accordingly, this observed $\overline{x}$ would not be unreasonable is the true mean were 7. We don't have good evidence on which to base a lawsuit against Wang's Widgets Inc.

Caution: We *don't* conclude that the true mean is 7! (Maybe we were extremely lucky with our sample, and the true mean is equal to

7.003.) All we conclude is that we didn't find good evidence that the true mean isn't 7.

The formal setup: Hypotheses

We operate in terms of two *hypotheses*.

The *null hypothesis* H_0 is the statement that there is no difference, no change, etc., as appropriate. You are seeking evidence against this hypothesis.

The *alternative hypothesis* H_a is the statement that there is a difference. There two forms: one sided and two sided.

Both hypotheses make assertions about the *population*.

Formulating hypotheses

The hypotheses should be formulated *before* designing the experiment, survey, or other procedure. They *must* be formulated before examining the data. A P-value (the appropriate numerical measure in a hypothesis test–see below) formulated with the data in mind is suspect. (See Chapter 16.)

Example 1 revisited

We know US women of ages 20–29 years have mean height 64 inches.

We want to know if women in Eugene of ages 20–29 years are on average taller than those in the general population. Let μ be the mean height of women of ages 20–29 years in Eugene.

H_0: $\mu = 64$.
H_a: $\mu > 64$.

This is a one sided hypothesis test.

Example 2 revisited

Change the question slightly.

We want to know if the mean height of women in Eugene of ages 20–29 years differs from the mean height of those in the general population. Again, let μ be the mean height of women of ages 20–29 years in Eugene.

H_0: $\mu = 64$.
H_a: $\mu \neq 64$.

This is a two sided hypothesis test. Note the change in H_a (but H_0 is the same).

Caution: It is very unlikely that the mean height of women of ages 20–29 years in Eugene is exactly 64 inches. Thus, the null hypothesis is unlikely to be exactly true. However, we don't reject it unless we have what we believe to be strong evidence against it.

Example 3 revisited
Are female intercollegiate soccer players taller than average?

We compare with the mean height of US women of ages 20–29 years, given as 64 inches.

Let μ be the mean height of female intercollegiate soccer players.

H_0: $\mu = 64$.
H_a: $\mu > 64$.

Example 4 revisited
Does being left handed affect IQ?

We compare with the mean IQ of left handed people with the mean IQ of the general population, which is 100.

Let μ be the mean IQ of left handed people.

H_0: $\mu = 100$.

H_a: $\mu \neq 100$.

Example 5 revisited

Suppose we know the mean weight of spiral-horned snorkacks. Are crumple-horned snorkacks, on average, smaller (lighter)?

To be specific, suppose that spiral-horned snorkacks are known to have a mean mass of 60 kg.

Let μ be the mean mass, in kg, of crumple-horned snorkacks.

H_0: $\mu = 60$.
H_a: $\mu < 60$.

Example 6 revisited

You suspect that the Consolidated Generic Unhealthy Foods Corporation is, on average, underfilling its 12 ounce bags of potato chips.

We compare the mean weight of the contents of bags of potato chips with 12 ounces, which is what the weight is supposed to be.

Let μ be the mean weight, in ounces, of the contents of 12 ounce bags of potato chips produced by the Consolidated Generic Unhealthy Foods Corporation.

H_0: $\mu = 12$.
H_a: $\mu < 12$.

Example 7 revisited

You have bought a large quantity of 7 mm ball bearings from Wang's Widgets Inc., and you are preparing to sue Wang's Widgets Inc. because, you claim, the diameter of the balls in the ball bearings is not as advertised.

We compare the mean diameter of the balls in the ball bearings with 7 mm, which is what the diameter is supposed to be.

Let μ be the mean diameter, in mm, of the balls in 7 mm ball bearings from Wang's Widgets Inc.

H_0: $\mu = 7$.
H_a: $\mu \neq 7$.

The test statistic

We calculate a *test statistic*. It is a number which compares an estimate of the parameter obtained from our sample with the value the null hypothesis says that parameter is supposed to have.

In this situation, we use the z-score of the sample mean $\overline{x}$, based on the sampling distribution that would hold if the null hypothesis were true.

(Every computation is done on the assumption that the null hypothesis is true.)

So our test statistic here is

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}.$$

Example 7 (continued)

Let μ be the mean diameter, in mm, of the balls in 7 mm ball bearings from Wang's Widgets Inc. Suppose the standard deviation of their diameters is known to be 0.1 (in mm).
H_0: $\mu = 7$.
H_a: $\mu \neq 7$.

We choose a simple random sample of size 100, and get $\overline{x} = 7.03$.

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} = \frac{7.03 - 7}{0.1/\sqrt{100}} = 3.$$

Note: We are assuming H_0 is true. Therefore we take $\mu = 7$ in this calculation, because that is what H_0 says μ is.

We choose another simple random sample of size 100, and get $\overline{x} = 6.98$.

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} = \frac{6.98 - 7}{0.1/\sqrt{100}} = -2.$$

P-value
The *P-value*, written P, of our test is the probability that, *if the null hypothesis is true*, then the value of our test statistic is as extreme, or more extreme, than was actually observed.

Suppose you got a P-value of 0.01, and your competitor got a P-value of 0.04, both on the same hypothesis test. Which of you has better evidence against the null hypothesis?

You have better evidence against the null hypothesis.

If the null hypothesis were true, your result would be more unlikely than your competitor's.

Recall from first day
From the first lecture, in a situation in which $P < 0.05$:

"We did an experiment .... If it were **not** true that students taking Jennifer's SAT preparation course get higher SAT scores than students taking no preparation course, then the observed outcome of the experiment would be very unlikely. In fact, it would have a probability of less than $\frac{1}{20}$ of occurring."

Example 7 (continued) (two-sided hypothesis test)
Let μ be the mean diameter, in mm, of the balls in 7 mm ball bearings from Wang's Widgets Inc. Suppose the standard deviation of their diameters is known to be 0.1 (in mm).
H_0: $\mu = 7$.
H_a: $\mu \neq 7$.
We chose a simple random sample of size 100, and got $\overline{x} = 7.03$.

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} = \frac{7.03 - 7}{0.1/\sqrt{100}} = 3.$$

P is the probability that $z \geq 3$ or $z \leq -3$. That is, P is the probability that, if $\mu = 7$, $\sigma = 0.1$, and the sample size is 100, then a simple random sample gives $\overline{x} \geq 7.03$ or $\overline{x} \leq 6.97$.

According to the Rule of Thumb, the probability that $-3 < z < 3$ is about 0.997, leaving only a very small probability of about 0.003 that $z \geq 3$ or $z \leq -3$. So $P \approx 0.003$.

Restating the reasoning

Let μ be the mean diameter, in mm, of the balls in 7 mm ball bearings from Wang's Widgets Inc. We assumed that $\mu = 7$, choose a simple random sample of balls from the ball bearings, and computed a test statistic z.

If the z we got were impossible, we could say that our assumption, namely $\mu = 7$, must have been wrong. We would conclude that μ is definitely not equal to 7.

We *cannot* say that the z we got is impossible. We can say that it is very unlikely. This outcome does not *prove* that our assumption $\mu = 7$ must have been wrong. But it gives evidence that this assumption was wrong.

The more unlikely our observed value of z is, the stronger the evidence. How strong is the evidence really? This issue is tricky, and will be discussed later.

Example 7 (continued) (two-sided hypothesis test)

Let μ be the mean diameter, in mm, of the balls in 7 mm ball bearings from Wang's Widgets Inc. Suppose the standard deviation of their diameters is known to be 0.1 (in mm).

H_0: $\mu = 7$.

H_a: $\mu \neq 7$.

Our competitor chose a simple random sample of size 100, and got $\overline{x} = 6.98$.

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} = \frac{6.98 - 7}{0.1/\sqrt{100}} = -2.$$

P is the probability that $z \leq -2$ or $z \geq 2$. That is, P is the probability that, if $\mu = 7$, $\sigma = 0.1$, and the sample size is 100, then a simple random sample gives $\overline{x} \leq 6.98$ or $\overline{x} \geq 7.02$.

According to Table A (details omitted), the probability that $-2 < z < 2$ is about 0.9544, leaving only a small probability of about 0.0456 that $z \geq 2$ or $z \leq -2$. So $P \approx 0.0456$.

The company's lawyers chose a simple random sample of size 100, and got $\overline{x} = 7.003$.

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} = \frac{7.003 - 7}{0.1/\sqrt{100}} = 0.3.$$

P is the probability that $z \geq 0.3$ or $z \leq -0.3$. That is, That is, P is the probability that, if $\mu = 7$, $\sigma = 0.1$, and the sample size is 100, then a simple random sample gives $\overline{x} \geq 7.003$ or $\overline{x} \leq 6.997$.

Using Table A (details omitted), one finds that the probability that $-0.3 < z < 0.3$ is about 0.2321, leaving only a quite large probability of about 0.7642 that $z \geq 0.3$ or $z \leq -0.3$. So $P \approx 0.7642$.