STAT 509 – Sections 4.2-4.3 – Hypothesis Testing

• CIs are possibly the most useful forms of inference because they give a range of "reasonable" values for a parameter.
• But sometimes we want to know whether one particular value for a parameter is "reasonable."
• In this case, a popular form of inference is the hypothesis test.

We use data to test a claim (about a parameter) called the null hypothesis.

Example 1: We claim the proportion of USC students who travel home for Christmas is 0.95.

Example 2: We assume a milk carton filling machine produces cartons with a mean weight of 260 g.
• Question: Is this true, or is the process overfilling the cartons on average?

• If the engineer finds reason to believe the mean weight is greater than 260, he/she would correct the process.

Engineer's Decision

	Correct process	**Leave alone**
Actual wt = 260	*Type I error*	*OK*
Actual wt > 260	*OK*	*Type II error*

Hypotheses and Types of Errors

- **Null hypothesis (denoted H_0) often represents “status quo”, “previous belief” or “no effect”.**
- **Alternative hypothesis (denoted H_a) is usually what we seek evidence for.**

NOTE: There are two types of wrong decisions in a hypothesis test:

(1) Type I error: We reject null hypothesis when H_0 is true.
(2) Type II error: We fail to reject the null hypothesis when the alternative hypothesis is true.

	Statistician’s Decision	
Truth	**Reject H_0**	**Fail to reject H_0**
H_0 is true	*Type I error*	*OK*
H_a is true	*OK*	*Type II error*

Let α = P(type I error), β = P(type II error)
- **The power of the test is then $1-\beta$.**

Power = Probability of rejecting H_0 when H_0 is false.

Idea: We will reject H_0 and conclude H_a if the data provide convincing evidence that H_a is true.

Evidence in the data is measured by a <u>test statistic</u>.

- **A test statistic measures how far away the corresponding sample statistic is from the parameter value(s) specified by H_0.**

- **If the sample statistic is extremely far from the value(s) in H_0, we say the test statistic falls in the "rejection region" and we reject H_0 in favor of H_a.**

<u>Example 2</u>: Our claim assumed the mean milk carton weight is no more than 260 g, but we seek evidence that the mean weight is actually greater than 260. We randomly sample 49 cartons and calculate the sample mean weight $\overline{Y}$. *Assuming we know* σ, let $Z = \frac{\bar{Y} - 260}{\sigma / \sqrt{n}}$ be our "test statistic" here.

Note: If this *Z* value is much bigger than zero, then we have evidence against H_0: $\mu = 260$ and in favor of H_a: $\mu > 260$.

- **Suppose we'll reject H_0 if $Z > 1.645$.**

- **If μ really is 260, then *Z* has a standard normal distribution. (Why?)**

Picture:

- **If we reject H_0 whenever $Z > 1.645$, what is the probability we reject H_0 when H_0 <u>really is true</u>?**

$P(Z > 1.645 \mid \mu = 260) =$

- **This is the probability of making a Type I error (rejecting H_0 when it is actually true).**

P(Type I error) = "level of significance" of the test (denoted α).

- **We don't want to make a Type I error very often, so we choose α to be small:**

- **The α we choose will determine our rejection region (determines how strong the sample evidence must be to reject H_0).**

- **In the previous example, if we choose $\alpha = .05$, then $Z > 1.645$ is our rejection region.**

Hypothesis Tests of the Population Mean

In practice, we don't know σ, so we don't use the *Z*-statistic for our tests about μ.

Use the t-statistic: $t = \dfrac{\bar{Y} - \mu_0}{s / \sqrt{n}}$**, where μ_0 is the value in the null hypothesis.**

• This has a t-distribution (with *n* – 1 d.f.) if H_0 is true (if μ really equals μ_0).

Example 2: Milk carton: $H_0: \mu = 260$
$H_a: \mu > 260$

Sample 49 cartons, get $\overline{Y} = 260.8$ grams and $s = 1.95$.
Let's set $\alpha = .05$.

Rejection region:

Reject H_0 if *t* is bigger than 1.68.

Conclusion:

• We never accept H_0; we simply "fail to reject" H_0.

• This example is a <u>one-tailed test</u>, since the rejection region was in one tail of the t-distribution.

• Only very <u>large</u> values of *t* provided evidence against H_0 and for H_a.

Suppose we had sought evidence that the mean weight was less than 264 g. The hypotheses would have been:

$$H_0: \mu = 264$$
$$H_a: \mu < 264$$

• Now very small values of $t = \dfrac{\bar{Y} - \mu_0}{s/\sqrt{n}}$ would be evidence against H_0 and for H_a.

Rejection region would be in left tail:

Rules for one-tailed tests about population mean

H_0: $\mu = \mu_0$		H_0: $\mu = \mu_0$
H_a: $\mu < \mu_0$	or	H_a: $\mu > \mu_0$

Test statistic: $t = \dfrac{\bar{Y} - \mu_0}{s / \sqrt{n}}$

Rejection Region: $t < -t_\alpha$ $\qquad$ $t > t_\alpha$

(where t_α is based on n – 1 d.f.)

Rules for two-tailed tests about population mean

H_0: $\mu = \mu_0$

H_a: $\mu \neq \mu_0$

Test statistic: $t = \dfrac{\bar{Y} - \mu_0}{s / \sqrt{n}}$

Rejection Region: $t < -t_{\alpha/2}$ **or** $t > t_{\alpha/2}$ **(both tails)**

(where $t_{\alpha/2}$ is based on n – 1 d.f.)

Example 3: We want to test (using $\alpha = .05$) whether the mean width of a manufactured part differs from 100 cm. Let μ = mean width.
<u>Hypotheses</u>:

We sample 20 of these parts. Sample data: $\overline{Y} = 105.0$ cm, $s = 6.2$ cm.

<u>Assumptions of t-test (and CI) about μ</u>

• We assume the data come from a population that is approximately normal.
• If this is not true, our conclusions from the hypothesis test may not be accurate (and our true level of confidence for the CI may not be what we specify).
• How to check this assumption?

• The t-procedures are robust: If the data are "close" to normal, the t-test and t CIs will be quite reliable.
• If the sample size is large, the t-procedures will work well even if the data are somewhat far from normal.

P-value of a hypothesis test

Recall that the significance level α is the desired P(Type I error) that we specify before the test.

The P-value (or "observed significance level") of a test is the probability of observing as extreme (or more extreme) of a value of the test statistic than we did observe, if H_0 was in fact true.

The P-value gives us an indication of the strength of evidence against H_0 (and for H_a) in the sample.

This is a different (yet equivalent) way to decide whether to reject the null hypothesis:

- **A small p-value (less than α) = strong evidence against the null => Reject H_0**

- **A large p-value (greater than α) = little evidence against the null => Fail to reject H_0**

How do we calculate the P-value? It depends on the alternative hypothesis.

One-tailed tests

Alternative	**P-value**
H_a: “ < ”	**Area to the left of the test statistic value in the appropriate distribution (t or z).**
H_a: “ > ”	**Area to the right of the test statistic value in the appropriate distribution (t or z).**

Two-tailed test

Alternative	**P-value**
H_a: “ ≠ ”	**2 times the “tail area” outside the test statistic value in the appropriate distribution (t or z). Double the tail area to get the P-value!**

We generally use software to give us P-values for t-tests:

Example 2: Testing $\mu = 260$ vs. $\mu > 260$
Sample data: $n = 49$, $\overline{Y} = 260.8$, $s = 1.95$.

Picture:

```
> pt( (260.8-260)/(1.95/sqrt(49)), df = 48,
lower=F)
[1] 0.003029211
```

Example 3: Testing $\mu = 100$ vs. $\mu \neq 100$
Sample data: $n = 20$, $\overline{Y} = 105.0$, $s = 6.2$.

Picture:

P-value from R:

```
> 2*pt( (105-100)/(6.2/sqrt(20)), df = 19, lower=F)
[1] 0.001880167
```

Example based on raw data:

- **Testing whether the mean lifetime of a population of lightbulbs is less than 800 hours. Random sample of 15 bulbs' lifetimes: 769.3 730.3 737.9 794.9 791.5 827.2 885.7 775.0 779.4 703.2 791.5 764.4 870.2 798.6 789.7**

Hypotheses:

Checking normality assumption:

```
mydata <- c(769.3, 730.3, 737.9, 794.9, 791.5,
827.2, 885.7, 775.0, 779.4, 703.2, 791.5, 764.4,
870.2, 798.6, 789.7)

qqnorm(mydata)
```

Getting test statistic value and P-value in R:

```
t.test(mydata, mu=800, alternative="less")
```

Conclusion?

Note: In R, choices are: `alternative="less"` **,** `alternative="greater"` **, or** `alternative="two.sided"`

Practical and Statistical Significance

- A rejection of H_0 is called a "statistically significant" result.
- This simply means that we conclude that, say, μ is something bigger than 260.
- We are not concluding that it is much bigger than 260.
- So a result may be "statistically significant" without being "practically significant."
- Often we are able to reject H_0 when our sample size is large.

Example: Suppose we're testing

$$H_0: \mu = 260 \text{ vs. } H_a: \mu > 260$$

and suppose the true mean weight is 260.03. With a sample of size 5000 cartons, we will likely reject H_0.

Is this a statistically significant result?

Is this a practically significant result?

- A solution: Provide a CI for μ, so we can get an idea about the likely values for the true mean.

Relationship between a CI and a (two-sided) hypothesis test about μ:

- **A test of H_0: $\mu = m^*$ vs. H_a: $\mu \neq m^*$ will reject H_0 if and only if a corresponding CI for μ does not contain the number *m**.**

Example: A 95% CI for μ is (2.7, 5.5).

(1) At $\alpha = 0.05$, would we reject H_0: $\mu = 3$ in favor of H_a: $\mu \neq 3$?

(2) At $\alpha = 0.05$, would we reject H_0: $\mu = 2$ in favor of H_a: $\mu \neq 2$?

(3) At $\alpha = 0.10$, would we reject H_0: $\mu = 2$ in favor of H_a: $\mu \neq 2$?

(4) At $\alpha = 0.01$, would we reject H_0: $\mu = 3$ in favor of H_a: $\mu \neq 3$?

Power of a Hypothesis Test

• Recall the significance level α is our desired P(Type I error) = P(Reject H_0 | H_0 true)

The other type of error in hypothesis testing:
Type II error =

P(Type II error) = β

The power of a test is

• High power is desirable, but we have little control over it (different from α)

Calculating Power: The power of a test about μ depends on several things: α, *n*, σ, and the true μ.

Example 1: Suppose we test whether the true mean nicotine contents in a population of cigarettes is greater than 1.5 mg, using $\alpha = 0.01$.

H_0: **H_a:**

We take a random sample of 36 cigarettes. Suppose we know $\sigma = 0.20$ mg. Our test statistic is

We reject H_0 if:

- Now, suppose μ is actually 1.6 (implying that H_0 is false). Let's calculate the power of our test if $\mu = 1.6$:

This is just a normal probability problem!

- What if the true mean were 1.65?

Verify:

- The farther the true mean is into the "alternative region," the more likely we are to correctly reject H_0.

Recap: Steps of a hypothesis test:

(1) Determine the null and alternative hypotheses.
(2) Determine the appropriate test statistic and rejection region ("critical region").
(3) Collect data and calculate test statistic value.
(4) Determine whether test statistic value falls in the rejection region (or else find the P-value of the test).
(5) Draw conclusion and state it in English.