Testing Statistical Hypotheses

Recall the study where we estimated the difference between mean systolic blood pressure levels of users of oral contraceptives and non-users, μ_x-μ_y.

Such studies are sometimes viewed as attempts to answer a question like "Does the mean blood pressure of OC users differ from that of non-users? (Are μ_x and μ_y different?)"

This speculation, that maybe there really is no difference (maybe oral contraceptive use has no effect on blood pressure), is a hypothesis

$$H_0 : \mu_X = \mu_Y \qquad or \qquad H_0 : \mu_X - \mu_Y = 0$$

The alternative to H_0 is another hypothesis,

$$H_1 : \mu_X \neq \mu_Y \qquad or \qquad H_1 : \mu_X - \mu_Y \neq 0$$

The alternative hypothesis says that the mean blood pressure of OC users really is different from that of non-users. (It doesn't say which mean is bigger, or how big the difference is, only that there is a difference.)

In this example, the parameter of interest, μ_x-μ_y , has a wide range of possible values, and μ_x-μ_y=0 , is special because it represents "no difference" (between OC users and non-users), or "no effect" (of OC use).

- H_0 is called the null hypothesis. It states that the parameter μ_x-μ_y equals the special value, zero.

- H_1 , the alternative hypothesis, consists of the complement of H_0. It states that the parameter does not equal zero.

In our example we had observations on 8 OC users and 21 non-users.

The immediate question is "What do these observations tell us?" or

"Is it right to say that these data represent strong evidence that there is indeed a difference between the mean blood pressure levels of OC users and non-users?" or

"Do these observations justify rejecting H_0?"

(Note that: we **cannot** directly answer the question "Which of these hypotheses is true?" because we only have a sample of the population to work with.)

We could answer the previous questions using our new confidence interval tools:

Construct a 95% confidence interval for μ_x-μ_y , and answer "Yes, our data show that there is a difference" if the value μ_x-μ_y=0 is not in the interval. That is, reject the hypothesis of no difference (H_0) if the value μ_x-μ_y=0 is not in the confidence interval.

But if the value μ_x-μ_y=0 is inside the 95% CI , our answer is "These observations do not allow us to reject H_0."

We are not saying that the observations are evidence supporting H_0. (Remember that our best estimate of μ_x-μ_y was 5.42, not 0.) But the data do not represent strong enough evidence against H_0 to justify its rejection.

This procedure is essentially a classical hypothesis test: When H_0 specifies the value of a parameter, (as it does in our example)

(i) Construct a $100(1-\alpha)\%$ CI for the parameter.

(ii) Reject H_0 if the hypothesized value is not in the interval.

Recall our earlier interpretation of the CI as:

> *"the values of the parameter that are consistent with the observations."*

What we're doing here is checking to see if the value specified by H_0 is "consistent with the observations". If it is not, then we reject H_0.

⇨ As always, there is a more mathematically direct approach that is not only more complicated, but also more technically involved and, in the end, provides the same answer. ☺

A More Direct Approach To Hypothesis Testing

In the blood pressure example, if the two variances are assumed to be equal then

$$T^* = \frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{\sqrt{S_P^2\left(\frac{1}{n} + \frac{1}{m}\right)}} \sim t^{n+m-2}$$

Under our Null hypothesis (the mean pressures of OC users and non-users are no different), H_0: $\mu_X - \mu_y = 0$, m=21, n=8 so

$$T^* = \frac{5.42 - (0)}{\sqrt{S_P^2\left(\frac{1}{8} + \frac{1}{21}\right)}} \sim t^{27}$$

has Student's t distribution with 27 df (under H_0). For an observed sample we can calculate the value of this statistic. (T^* is called the **"test statistic"**)

If it falls in one of the tails of the t-distribution (say either $T^* > 2.052$ or $T^* < -2.052$) then we reject H_0.

A popular explanation for the reasoning behind this procedure is as follows:

If H_0 is true then T^* will usually (95% of the time) fall between -2.052 and 2.052. If it falls outside this interval, then either

(a) H_0 is true, and we just happened to observe a relatively rare sample where T^* falls this far from its expected value, 0, or

(b) H_0 is false; the expected value of T^* is really not 0.

Values of T^* that are far from zero (large values of $|T^*|$) are <u>improbable</u> under H_0. If we observe such a value, we reject H_0.

In this example we chose 2.052 as the **<u>critical value</u>** for our test, and used the rule

"Reject H_0 if $|T^*|$ exceeds 2.052"

This rule will sometimes lead us to reject H_0 when it is in fact true. It is easy to calculate the probability of making this error:

$$P\left(\left|T^*\right| > 2.052 \mid H_0\right) = 1 - P\left(-2.052 < T^* < 2.052\right) = 0.05$$

By choosing this critical value, we control the error probability. If we use a larger critical value, say t = 2.771, then we reduce the probability that we will reject H_0 when it is really true:

$$P\left(\left|T^*\right| > 2.771 \mid H_0\right) = 1 - P\left(-2.771 < T^* < 2.771\right) = 0.01$$

This is an example of another general scheme for hypothesis testing: Instead of constructing a confidence interval, choose an observable "test statistic" (T^*) for which

(i) the probability distribution is known (at least approximately) when H_0 is true, and

(ii) the larger the value of the test statistic, the stronger the evidence against H_0.

Then we choose some critical value, and reject H_0 if the observed value of the test statistic exceeds the critical value.

In our OC example the test statistic was

$$|T^*| = \frac{|5.42 - (0)|}{\sqrt{S_P^2\left(\frac{1}{8}+\frac{1}{21}\right)}} \sim t^{27}$$

When H_0 is true, T^* has a Student's t distribution with 27 df.

Since the distribution of the test statistic when H_0 is true is known, we can choose the critical value to fix the error probability (reject H_0 when H_0 is really true) at whatever value we like.

Popular choices are 0.05 and 0.01 (1/20 and 1/100). This error probability is often called the "significance level" of the test and often represent by

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ true})$$

(Remember, we use the fact that H_0 is true to construct the test statistic.)

Testing a Hypothesis About a Single Mean

Assume that the data have normal probability distribution, variance known. The Null hypothesis is H_0: $\mu = \mu_0$. The Test statistic:

$$T^* = \frac{\sqrt{n}\left|\overline{X}_n - \mu_0\right|}{\sigma}$$

Notice the form of T^*?

$$T^* = \frac{\overline{X}_n - E\left[X \mid H_0\right]}{\sqrt{Var\left[\overline{X}_n\right]}}$$

This is an appropriate test statistic for this hypothesis, because

(i) its distribution when H_0 is true is known (standard normal), and

(ii) the larger the value of the test statistic, the farther the sample mean, $\overline{X}_n$, is from μ_0 , and it seems appropriate to say that the farther $\overline{X}_n$ is from μ_0, the stronger the evidence that μ_0 is not the expected value of $\overline{X}_n$.

If we choose the critical value 1.96, then we have

$$\alpha = P(reject\ H_0 \ | H_0) = P\left(\frac{\sqrt{n}|\overline{X}_n - \mu_0|}{\sigma} > 1.96 \,|\, H_0\right)$$
$$= P(|Z| > 1.96) = 0.05$$

This gives a test of H_0 at the 5% significance level.

To test H_0 at the 1% significance level, we must use a larger critical value, 2.576.

$$\alpha = P(reject\ H_0 \ | H_0) = P\left(\frac{\sqrt{n}|\overline{X}_n - \mu_0|}{\sigma} > 2.576 \,|\, H_0\right)$$
$$= P(|Z| > 2.576) = 0.01$$

To reject H_0 at the 1% level requires stronger evidence (a more extreme value of the test statistic) than is required to reject at the 5% level.

- At the 5% level we reject the hypothesis that H_0: μ =E[X]= μ_0 if $\overline{X}_n$ falls further than 1.96 SE's away from the hypothesized value, μ_0.

- At the 1% level we only reject if $\overline{X}_n$ falls further than 2.576 SE's away from μ_0.

Example

We know (Pagano and Gauvreau) that the distribution of serum cholesterol in adult males in the U.S. has a mean of 211 mg/100ml and a standard deviation of 46 mg/100ml.

We want to learn about men who are hypertensive smokers. Specifically, is the distribution of serum cholesterol among such men any different from that in the general population?

Let's suppose the standard deviation is the same as in the general population (46 mg/100ml), and test whether the mean, μ, is the same or not.

$$H_0: \mu = 211 \text{ mg/100ml}$$

$$H_1: \mu \neq 211 \text{ mg/100ml}$$

Our test statistic will be

$$T^* = \frac{\sqrt{n}\left|\overline{X}_n - 211\right|}{46}$$

which, under H_0, has a standard normal distribution.

Thus to test at the 5% level, we will reject H_0 if the absolute value of the test statistic exceeds 1.96.

This means that we will reject unless

$$-1.96 < \frac{\sqrt{n}\left(\overline{X}_n - 211\right)}{46} < 1.96$$

which implies that we will reject unless the 95% CI contains the hypothesized value, 211:

$$\overline{X}_n - 1.96\frac{46}{\sqrt{n}} < 211 < \overline{X}_n + 1.96\frac{46}{\sqrt{n}} \qquad ???$$

In Pagano's example, the sample size is n = 12 and the sample mean is 217, so the test statistic's value is

$$T^* = \frac{\sqrt{12}\left|217 - 211\right|}{46} = 0.45$$

Since this is much less than the critical value, 1.96, we do not come close to rejecting H_0.

(The 95% CI is (191, 243), and the hypothesized value, 211, is well within the interval.)

If we take a much larger sample of hypertensive smokers, say 250 instead of 12, and find the same value for the sample mean, $\overline{X}_{250} = 217$, the test statistic will be

$$T^* = \frac{\sqrt{250}\left|217 - 211\right|}{46} = 2.06$$

Since this is greater than the critical value, 1.96, the result would now be "Reject H_0 at the 5% level."

⇨ *When the observations lead to rejection of* H_0 *at a specified level, such as 5%, they are said to be "statistically significant at the 5% level." (When they don't, they are "not statistically significant.")*

We saw that a sample of size 12 with mean 217 is not statistically significant for testing H_0: μ=211 at the 5% level (σ = 46 known).

A sample of 250 with the same mean, 217, <u>is</u> statistically significant at the 5% level
(but not at the 1% level, since the test statistic's value, 2.06, is less than the 1% critical value, 2.576.)

The 95% CI is 217 ± 1.96(46/$\sqrt{250}$), or
(211.3 , 222.7),
which excludes the hypothesized value, 211, and

The 99% CI is 217 ± 2.576(46/$\sqrt{250}$), or
(209.5 , 224.5)
which <u>includes</u> that value.

Textbooks often stress the importance of selecting the significance level (choosing the value for the error probability, ɑ) before `looking' at the data. We will see later why they stress this.

In practice it is very common not to select a fixed ɑ at all. Instead, what is often done is that for the value of the test statistic that is actually observed, the probability of observing a value that large or larger (under H_0) is calculated and reported. This quantity is called the **p-value**. It is a measure of how extreme the test statistic is, and it is used as a measure of how strong the evidence against H_0 is.

We've been considering a test procedure that chooses a test statistic, T^*, and a fixed error probability, α. Then it finds the critical value, say t_α, for which $P(T^*> t_\alpha| H_0)=\alpha$. This procedure rejects H_0 if the observed value of the test statistic, say $T_{obs}{}^*$, exceeds the critical value, t_α.

⇨ This is equivalent to finding the p-value, $P(T^*> T_{obs}{}^*| H_0)$, and then rejecting H_0 if this probability is less than α.

For the hypertensive smokers' cholesterol levels (when n = 250) the test statistic was

$$T* = \frac{\sqrt{250}\left|\overline{X}_{250} - 211\right|}{46}$$

and the observed value of the test statistic was

$$T_{obs}^{*} = \frac{\sqrt{250}\left|217 - 211\right|}{46} = 2.06$$

The probability (under H_0) of observing a value of the test statistic this extreme is

$$\text{P}(\text{T}^{*} > \text{T}_{\text{obs}}{}^{*}|\text{H}_0) = P\left(\frac{\sqrt{250}\left|\overline{X}_{250} - 211\right|}{46} > 2.06 \mid H_0\right) = P\left(\left|Z\right| > 2.06\right) = 0.039$$

Since this is greater than 0.01, but smaller than 0.05, it shows that our observations represent evidence strong enough to justify rejection at the 5% significance level, but not strong enough for the 1% level.

A note on the P-value

The p-value $P(T^{*} > T_{obs}^{*}|H_0)$, is something very familiar. It is just a probability statement about the sample average.

$P(T^{*} > T_{obs}^{*}|H_0)=$

$$
\begin{aligned}
&= P\left(\frac{\sqrt{250}\left|\overline{X}_{250} - 211\right|}{46} > 2.06 \mid H_0 \right) \\
&= P\left(\overline{X}_n > 211 + 2.06\frac{46}{\sqrt{250}} \right) \\
&= P\left(\overline{X}_n > 217 \right) \\
&= 0.039
\end{aligned}
$$

Which says that the probability of observing a sample mean greater than or equal to the 217 that was observed is only 0.039 if the true mean is really 211 (the null hypothesis is true).

Two Types of Errors

We have seen that for testing the hypothesis that the mean of a normal probability distribution has a specified value, H_0: $\mu = \mu_0$ (against the alternative hypothesis that the mean is not μ_0), we reject H_0 at significance level $\alpha = 0.05$ if

$$T^* = \frac{\sqrt{n}\left|\overline{X}_n - \mu_0\right|}{\sigma} > 1.96$$

This procedure can lead us to reject the hypothesis H_0 when it is really true. The probability of this error is 0.05, and is often called the **Type I error**, significance level α, or the size of the test .

But what if the hypothesis H_0 is false and we fail to reject H_0? This error is called the **Type II error** and represent by the symbol β.

⇨ Type I error : α=P(reject H_0| H_0 true)

⇨ Type II error: β=P(fail to reject H_0| H_0 false)

⇨ Power: 1-β= P(reject H_0| H_0 false)
(Loose definition – we'll be more specific later)