Hypothesis Tests as Decisions

Sometimes, we may be forced to make a decision based on a hypothesis test: For instance, should we ban a food additive based on results of an experiment looking for an increase in cancer in people eating it?

We might calculate a P-value for H_0 (no effect on cancer) versus H_a (an increase in cancer), and *reject* H_0 if the P-value is less than some level α (eg, 0.05).

We can make two kinds of errors:

Type I: We reject H_0 when it is true.

Type II: We don't reject H_0 when it is false.

The probability of a Type I error will be α. The probability of a Type II error will depend on how big the effect is.

The Power of a Hypothesis Test

The *power* of a test is the probability of rejecting H_0 (at some level α) when H_0 is false, and a specific alternative is true instead.

For example, if we are doing a test with

$$H_0: \quad \mu = 0$$
$$H_a: \quad \mu \neq 0$$

we might ask what the power is if $\mu = 5$.

The power is one minus the probability of a Type II error if μ has that value.

Why do we care about power?

- We want to design our experiment to have high power for an interesting alternative.
- If the power is low, the logic of rejecting H_0 when the P-value is small may not be compelling — a test statistic as big as that observed might be just as unlikely under H_a as under H_0.

Getting Hypothesis Tests From Confidence Intervals

One way to do a level-α hypothesis test for

$$H_0: \quad \mu = d$$
$$H_a: \quad \mu \neq d$$

is to compute a $1-\alpha$ confidence interval for μ and reject if it does not contain d.

Why does this work? Let the $1-\alpha$ confidence interval be (low, high). We know that if H_0 is true,

$$P(\text{low} \leq d \leq \text{high}) \;=\; 1-\alpha$$

Hence

$$P(d < \text{low or } d > \text{high}) \;=\; \alpha$$

So if H_0 is true, we reject with probability α, as we should for a level-α test.

Getting Confidence Intervals From Hypothesis Tests

One way to construct a level $1-\alpha$ confidence interval for a parameter (say μ) is to find the interval of values for μ that we would reject at level α in a two-sided test. (This set of values won't always be an interval, but it often is).

That is, we make the confidence interval be (low, high), with low and high chosen so that for any d between low and high (but for no others), the hypothesis test

$$H_0: \quad \mu = d$$
$$H_a: \quad \mu \neq d$$

does not reject at level α, for the particular data that we observed.

Why does this work? If d is the true value of μ, the test of $H_0: \mu = d$ will reject a fraction α of the time. So this d will be in the confidence interval a fraction $1-\alpha$ of the time, as desired.

Confidence Interval or Hypothesis Test?

Confidence intervals and hypothesis tests are closely related. Which one should we use?

A hypothesis test for $H_0 : \mu = d$ may be sensible when d has some special significance:

- If we think μ may be equal to d, but we have no reason to think μ is any other particular value.
- If we are especially interested in whether $\mu = d$, or $\mu < d$, or $\mu > d$.

You should **not** test the hypothesis that $\mu = d$ if d is some arbitrarily chosen value. Compute a confidence interval for μ instead.

Some Sensible Two-Sided Tests

Use a two-sided hypothesis test when a special value for the parameter exists and departures from that value either way are important:

- To check a new scale for systematic errors, we weight a 10kg weight several times. We do a hypothesis test of the form
$$H_0 : \quad \mu = 10$$
$$H_a : \quad \mu \neq 10$$
- To test whether a coin is fair, we flip it many times, recording how often it lands heads. We do a hypothesis test of the form
$$H_0 : \quad \pi = 1/2$$
$$H_a : \quad \pi \neq 1/2$$
- To see whether chimpanzees tend to prefer one hand, we record the number of times each chimpanzee in a sample reaches for a banana with its right hand minus the number of times it reaches with its left hand, for 100 bananas. We do a test for the mean of this difference, of the form
$$H_0 : \quad \mu = 0$$
$$H_a : \quad \mu \neq 0$$

A Sensible One-Sided Test

You should usually use a two-sided test, because usually differences in either direction are both possible and interesting.

But one-sided tests are sometimes appropriate:

> Gold has a density of 19.8 g/cm^2, which is denser than any other common material (and all cheaper materials). A shipment of coins that are supposed to be pure gold is suspected to instead consist of fake coins that aren't made of pure gold. Several of these coins are weighed, from which their density is computed, with some amount of error. We want to use this data to determine whether or not the coins are fake.

Here, a population mean density of $\mu = 19.8$ has special significance, and we expect that any alternative value for μ will be less than this. It is appropriate to do a test of the form

$$H_0 : \quad \mu = 19.8$$
$$H_a : \quad \mu < 19.8$$

Some Hypothesis Tests That Aren't Very Sensible

Unfortunately, many examples of hypothesis tests in the text by Kitchens aren't sensible:

Example 8.1: A conjecture is made that the mean starting salary for computer science graduates is \$30,000 per year. You believe it is less than \$30,000. Formulate null and alternative hypotheses to evaluate the claim.

Exercise 8.5: A psychologist suspects that more than 10% of the adult population is illiterate... State the null and alternative hypotheses to evaluate the psychologist's claim.

Exercise 8.26: The government believes that no more than 25% of all college students would favor reducing the penalties for the use of marijuana... Set up null and alternative hypotheses...

A confidence interval should be used in these examples, because the null hypotheses are chosen arbitrarily — eg, the 10% illiteracy rate is not based on any real information, and a rate of 10% is of no particular interest.