CLARK UNIVERSITY

Hypothesis tests
Math 218, Mathematical Statistics

D Joyce, Spring 2016

Introduction to hypothesis tests. As stated in our text, it is no an exaggeration to say that, for better or worse, hypothesis testing is the most widely used statistical toolin practice. Unfortunately, it's also one of the most misunderstood and misused of tools.

In this introduction to hypothesis tests, we'll only consider hypothesis tests concerning the population mean μ, but in later chapters we'll look at hypothesis tests that concern other parameters such as σ^2.

For a hypothesis test, we assume the population distribution comes from a known family of distributions, but an unknown mean μ. We have under consideration two hypotheses concerning the value of μ. One hypothesis, H_0, is called the *null* hypothesis, the other, H_1, is called the *alternative hypothesis*. A test is designed to determine whether to reject or not reject H_0 at some prespecified confidence level. (In practice, these tests are designed to show that the null hypothesis is false.) After the test is performed, there are two possible results, either the data strongly contradict H_0, in which case we reject H_0 and accept H_1, or the data are consistent with H_0, in which case we don't reject H_0. In the second case, not rejecting H_0 does not mean we accept H_0 or reject H_1 as the data may not be strong enough for those conclusions.

There are different forms for these hypotheses. Here are four of them. In each, μ_0 is some specified constant.

- Single population. The mean is μ_0.

  H_0: $\mu = \mu_0$; H_1: $\mu \neq \mu_0$.

  This form will require a two-sided test.

- Single population. The mean is at most μ_0.

  H_0: $\mu \leq \mu_0$; H_1: $\mu > \mu_0$.

  This form will require a one-sided test.

  Of course, there's an analogous one-sided test to see if the mean is at least μ_0.

- Two populations. The means of the two populations are the same.

  H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 \neq \mu_2$.

  This form will require a two-sided test.

- Two populations. The mean of the first population is less than or equal to the mean of the second population.

  H_0: $\mu_1 \leq \mu_2$; H_1: $\mu_1 > \mu_2$.

  This form will require a one-sided test.

We'll look at a couple of examples in the text.

The form of the hypothesis test for the mean is usually to evaluate the sample mean $\overline{X}$ and determine whether $\overline{X}$ falls in a *rejection region* or its complement, an *acceptance region*. The boundaries between these two regions are called *critical constants*. In a one-sided test, there is only one critical constant and each of the regions is a half-infinite interval. In a two-sided test, there are two constants, the acceptance region is the interval between them, and the rejection region is the union of two half-infinite intervals.

A test for a fair coin. Let's design a test for a fair coin. We want a test that will reject or not reject the null hypothesis H_0 that the coin is fair, and let's choose the confidence level to be 95%. The alternative hypothesis H_1 is that the coin is not fair.

So, H_0 is that p, which is μ, equals $\frac{1}{2}$, while H_1 is that it doesn't. Now, we know that for large n,

$$P(\overline{X} - 1/\sqrt{n} \leq p \leq \overline{X} + 1/\sqrt{n}) \geq 0.95.$$

We developed this probability above when we were looking at confidence intervals. We found a 95% confidence interval for p was the interval

$$[\overline{X} - 1/\sqrt{n}, \overline{X} + 1/\sqrt{n}].$$

Hypothesis tests are directly related to confidence intervals, and we can turn this confidence interval into this hypothesis test:

> Reject the null hypothesis H_0 that $p = \frac{1}{2}$ in favor of the alternative hypothesis H_1 if $\frac{1}{2} \notin [\overline{X} - 1/\sqrt{n}, \overline{X} + 1/\sqrt{n}]$.

In other words, we conclude, at the 95% confidence level, that the coin is unfair if $\overline{X}$ is further from $\frac{1}{2}$ than $\frac{1}{\sqrt{n}}$.

How big does n have to be to make this conclusion? You get to decide, but n shouldn't be too small, or it won't be possible to conclude the coin is unfair. For instance, if you take $n = 4$, the test says never says that the coin is unfair, since $\frac{1}{2}$ is never further from $\frac{1}{2}$ than $1/\sqrt{4}$. But suppose you let $n = 100$. Then you could say the coin is unfair if $\overline{X}$ lies outside the interval $[0.4, 0.6]$. That's a pretty big interval, but with $n = 10000$, you could say that the coin is unfair if $\overline{X}$ lies outside the interval $[0.49, 0.51]$. Even so, at confidence level 95%, you'd be wrong 5% of the time.

Type I and Type II errors, α-risks and β-risks. Errors in hypothesis tests have these two types.

A type I error occurs when the null hypothesis holds, but we reject it. Hypothesis tests are designed to control for type I errors. For instance, if the test is designed at the 95% confidence level, then when the null hypothesis actually holds, then 95% of the time we won't reject it, but $\alpha = 5\%$ of the time we will, so make this type I error 5% of the time. The probability of a type I error is denoted α, and it's sometimes called *level of significance* or the *α-risk*. To reduce the α-risk, just design the test to a higher confidence level.

A type II error occurs when the null hypothesis does not hold, but we don't reject it. Typically, we can't tell how often type II errors occur, because the frequency depends on the unknown parameter, and in the worst case they can occur 95% of the time when the null hypotheses does not hold.

Let's take an example to see this more clearly. Suppose the population distribution is a Bernoulli distribution with unknown parameter p. Last time we saw how to construct a test for a fair coin. The null hypothesis H_0 was that $p = \frac{1}{2}$, while the alternative hypothesis H_1 was that $p \neq \frac{1}{2}$. The hypothesis test said

> Reject the null hypothesis H_0 that $p = \frac{1}{2}$ in favor of the alternative hypothesis H_1 if $\frac{1}{2} \notin [\overline{X} - 1/\sqrt{n}, \overline{X} + 1/\sqrt{n}]$.

In other words, we conclude, at the 95% confidence level, that the coin is unfair if $\overline{X}$ is further from $\frac{1}{2}$ than $1/\sqrt{n}$.

To pin down this example, let's take $n = 10000$. Then we will reject H_0 if $|\overline{X} - \frac{1}{2}| > 0.01$.

Now, if $p = \frac{1}{2}$, then $P(|\overline{X} - \frac{1}{2}| > 0.01)$ is 0.05, leading to a type I error.

But if $p \neq \frac{1}{2}$, then the probability of a type II error,

$$\beta = P\left(|\overline{X} - \tfrac{1}{2}| \leq 0.01\right)$$

depends on what the value of p is. That is, the β-risk is actually a function that depends on the parameter. For instance, if p is very close to $\frac{1}{2}$, then this probability of a type II error will be very close to 0.95. But if p is near 0 or 1, this probability will be nearly 0.

Let's compute the probability of a type II error if $p = 0.49$. The distribution of $\overline{X}$ is almost normal with a mean of $p = 0.49$ and a variance of

$$\sigma^2_{\overline{X}} = \frac{pq}{n} = \tfrac{1}{10000} \cdot 0.49 \cdot 0.51 \approx \tfrac{1}{40000},$$

so a standard deviation of $\sigma_{\overline{X}} \approx \frac{1}{200} = 0.005$. Thus, standardizing the condition, we have

$$\begin{aligned} & P(0.49 \leq \overline{X} \leq 0.51) \\ = \; & P\left(\frac{0.49 - 0.49}{0.005} \leq \frac{\overline{X} - 0.49}{0.005} \leq \frac{0.51 - 0.49}{0.005}\right) \\ = \; & P(0 \leq Z \leq 4) \approx 0.5. \end{aligned}$$

In other words, a coin whose probability is heads is 0.49 will pass this fairness test half the time giving 50% type II errors.

This function β that depends on the unknown parameter θ, or rather the function that gives the probability that H_0 will not be rejected, is called the *operating characteristic function* of the test, and 1 minus it, that is the function that gives the probability that H_0 will be rejected, is called the *power function* $\pi(\theta)$ of the test. From either one, you can read off the α-risk and the β-risks for various values of θ.

The observed level of significance, called the P-value. Sometimes a hypothesis test just barely ends up rejecting or accepting H_0, and sometimes it clearly rejects or accepts H_0. The *observed level of significance*, or *P-value*, is a way of recording the information of near and far hits and misses. The value P is the smallest level of α for which H_0 is accepted; any lower and H_0 would be rejected.

Suppose we do a 95% confidence level test (so the level of significance is $\alpha = 0.05$). If H_0 is just barely accepted, then the observed level of significance is $P = 0.05$ or slightly larger. But if H_0 is just barely rejected, then P is just slightly smaller than 0.05. If the test accepts H_0 without question, then the P-value is higher, perhaps much higher than 0.05, while if the test clearly rejects H_0, then the P-value is smaller than 0.05, perhaps nearly 0.

P-values are fairly easy to compute. The fair-coin test is a special case of a two-sided hypothesis test on μ. We'll look at this in more detail in section 7.1. The test statistic for such tests is $z = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$ when σ is known, or, when n is large and σ not known, it's $z = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$, where s is the sample variance. This z has a standard normal distribution. The null hypothesis is $H_0 : \mu = \mu_0$, and the alternative hypothesis is $H_1 : \mu \neq \mu_0$. The P-value is the probability $P(Z \leq z \,|\, H_0)$, which is $2(1 - \Phi(|z|))$, which can be looked up in the standard normal table.

Math 218 Home Page at
`http://math.clarku.edu/~djoyce/ma218/`