Statistical Inference and Hypothesis Testing in Data Science Applications

DTSA 5003 offered on Coursera

by the University of Colorado, Boulder

Instructor: J.N. Corcoran

Simple versus Simple

In this document, we are only going to look at "simple versus simple" hypothesis tests such as

$$H_0 : \mu = 3 \qquad \text{versus} \qquad H_1 : \mu = 5.$$

and maybe something a little more abstract looking like

$$H_0 : \mu = \mu_0 \qquad \text{versus} \qquad H_1 : \mu = \mu_1$$

where μ_0 and μ_1 are numbers that are fixed and known. Even if we are not interested in simple versus simple problems, what we do here will be a building block towards dealing with more complicated hypotheses.

Level of Significance

When constructing a hypothesis test (coming up with a suitable rejection rule for H_0), we want to manage the probabilities of making one or both of Type I and Type II errors. In this document, we will look at the Type I error only.

Definition

For a simple null hypothesis, the **level of significance** or **size** of a test is defined to be the probability of making a Type I error. It is denoted by an α.

$$\begin{aligned} \alpha &= P(\text{ Type I Error }) \\ &= P(\text{ Reject } H_0 \text{ when it's true }) \end{aligned}$$

Given an acceptable level of Type I error probability, we can construct a test.

Four Steps to a Hypothesis Test

Suppose that we have a random sample $X_1, X_2, \ldots, X_n$ from the normal distribution with unknown mean μ and a known variance σ^2. Further suppose that we wish to test the hypotheses

$$H_0 : \mu = 3 \qquad \text{versus} \qquad H_1 : \mu = 5$$

at a given level of significance α.

There are four steps to follow.

Step One: Choose a statistic on which to base the test.

This step might have been written as "Choose an estimator on which to base the test." but it is more important to choose something we can work with in the following steps. Because the hypothesis test we are doing is about a population mean, we will choose the "common sense estimator" $\widehat{\mu} = \overline{X}$ which is the sample mean. However, in the steps to follow, we will need to compute probabilities involving the chosen statistic. If we don't know its distribution then we should probably choose something else to work with. Knowing what to do here in Step One comes with practice!

Step Two: Give the "form" of the test.

Remember that we are assuming that H_0 is true and looking at the data to see if we find evidence for H_1. For this problem, we have to decide between two values for the mean of the underlying distribution. How would the statistic $\overline{X}$ behave if H_1 were true? H_1 postulates that the true mean is larger than it is under H_0. We might see evidence of this if the statistic $\overline{X}$ is "large".

The form of the test is to reject H_0, in favor of H_1 if

$$\overline{X} > c$$

for some c to be determined.

Step Three: Find c.

Here is where the level of significance comes in. We have that

$$\begin{aligned} \alpha &= P(\text{ Type I Error }) \\ &= P(\text{ Reject } H_0 \text{ when it's true }) \end{aligned}$$

Since our rejection rule has the form "reject H_0 when $\overline{X} > c$", this becomes

$$\alpha = P(\,\overline{X} > c \text{ when } \mu = 3\,).$$

Under the assumption that $\mu = 3$, we have that

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(3, \sigma^2)$$

where σ^2 is known. This means that

$$\overline{X} \sim N\left(3, \frac{\sigma^2}{n}\right)$$

and so we know how to compute this probability. (And we do this below!)

Step Four: Give the conclusion.

Now that we have found the value of c we will pull it all together and state the rejection rule with c plugged in.

An Example

Let $X_1, X_2, \ldots, X_{10} \overset{iid}{\sim} N(\mu, 4)$. Find a test of

$$H_0 : \mu = 3 \qquad \text{versus} \qquad H_1 : \mu = 5$$

at level of significance $\alpha = 0.05$.

Step One: Choose a statistic on which to base the test.

Since this is a test about a population mean, we will try to use the sample mean $\overline{X}$.

Step Two: Give the "form" of the test.

If the alternative hypothesis is true, it will be reflected in the data by the sample mean being "large". The form of the test is to

$$\text{"Reject } H_0 \text{, in favor of } H_1 \text{ if } \overline{X} > c."$$

for some c to be determined.

Step Three: Find c.

$$
\begin{aligned}
0.05 &= P(\text{ Type I Error }) \\
&= P(\text{ Reject } H_0 \text{ when it's true }) \\
&= P(\overline{X} > c \text{ when } \mu = 3\,) \\
&= P\left(\tfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} > \tfrac{c-3}{2/\sqrt{10}} \text{ , when } \mu = 3\right) \\
&= P\left(Z > \tfrac{c-3}{2/\sqrt{10}}\right)
\end{aligned}
$$

where $Z \sim N(0,1)$.

Note that we have dropped the "$\mu = 3$". The information is no longer of any use to us. It was there to give us information about $\overline{X}$ but the distribution of Z does not depend on knowing whether or not μ is 3.

What number will Z be above with probability 0.05? The answer is the critical value $z_{0.05} \approx 1.645$. Thus, we must have that

$$\frac{c-3}{2/\sqrt{10}} = 1.645$$

which gives us that $c \approx 4.04$.

Step Four: Conclusion

"We reject H_0, in favor of H_1 if $\overline{X} > 4.04$."