8 章 仮説検定 (Hypothesis Testing)

東京大学

2014

Hypothesis Testing

In this chapter, we will discuss one of the most useful tools to write a research paper. Most research questions are so called "hypotheses" which we need to test. And one of the most useful ways to do this is testing these hypotehses statistically. There are many examples of "hypothesis".

For Example...

(1) Do the Democrats have a majority in a population of voters?

(2) Is there any difference between men's salaries and women's salaries?

(3) Does TV watching in childhood affect cognitive development measured by math and reading scores?

(4) Are there any effects of new drags or new surgical procedures? (in medical literature)

(5) Are there any effects of government programs or policies, such as those that subsidize training for disadvantaged workers? (in economic literature)

In this chapter, we will discuss how to conduct the statistical hypothesis testing formally.

We begin with the definition of a statistical hypothesis.

Definition

A **hypothesis** *is a statement about a population parameter.*

The goal of a hypothesis test is to decide, based on a sample from the population, which of two complementary hypothesis is true.

Definition

The two complementary hypothesis in a hypothesis testing problem are called the **null hypothesis** *and the* **alternative hypothesis**. *They are denoted by* $\mathbf{H_0}$ *and* $\mathbf{H_1}$, *respectively.*

In the examples, the **null hypotheses** may be
(1) $H_0 : \pi > 0.5$, where π is the population proportion of the Democrat voters.
(2) $H_0 : \mu_1 - \mu_2 = 0$, where μ_1 is the population mean of men's salaries and μ_2 is that of women's salaries.
(3) $H_0 : \mu_1 - \mu_2 = 0$, where μ_1 is the population mean of scores of those who watch TV in childhood and μ_2 is that of those who do not.
(4) $H_0 : \mu_1 - \mu_2 = 0$, where μ_1 is the population cure rate of those who receive the treatment and μ_2 is that of those who do not.
(5) $H_0 : \mu_1 - \mu_2 = 0$, where μ_1 is the population mean of wages of those who participate in the program and μ_2 is that of those who do not.

Hypothesis Testing using Confidence Intervals

How can we decide either to **accept** H_0 as true or to reject **reject** H_0 as false and decide H_1 is true?

Example 9-1: Sex Difference in Salaries
In a large American university, 10 men and 5 women professors were independently sampled in 1969, yielding the annual salaries given below (in thousand dollars)

Men (X_1)		Women (X_2)
13	20	9
11	14	12
19	17	8
15	14	10
22	15	16
$\bar{X} = 16$		$\bar{X} = 11$

Is there any difference between men's salaries (μ_1) and women's salaries (μ_2)? ($H_0 : \Delta \equiv \mu_1 - \mu_2 = 0$).

Once we construct the confidence interval, we can use it to test a hypothesis because, by definition, the confidence interval is the interval where the true value is bracketed with some probability, say 95%.

A confidence interval may be regarded as just the set of acceptable hypothesis.

Intuitively, any hypothesis that lies outside the confidence interval may be judged implausible –that is, can be **rejected**. On the other hand, any hypothesis that lies within the confidence interval may be judged plausible or **acceptable**.

In Example 9-1, the 95% confidence interval is

$$\left[(\bar{X}_1 - \bar{X}_2) - t_{.025} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},\ (\bar{X}_1 - \bar{X}_2) + t_{.025} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]$$

$$= \left[(16 - 11) - 2.16\sqrt{\frac{146}{13}}\sqrt{\frac{1}{10} + \frac{1}{5}},\ (16 - 11) + 2.16\sqrt{\frac{146}{13}}\sqrt{\frac{1}{10} + \frac{1}{5}}\right]$$

$$\approx [1.0,\ 9.0] \tag{1}$$

The claim $H_0 : \mu_1 - \mu_2 = 0$ seems implausible, because it falls outside the confidence interval.

Remark: As with the confidence interval, there is always some error probability of the decision. We need to decide the **error level** and it is common to choose 5% error level. If you use a 95% confidence interval, the hypothesis testing is with 5% error level.

In Example 9-1, we formally conclude that, **with a 5% error level, we can reject the hypothesis of no difference**.

When we reject the hypothesis of no difference with 5% level, we call the difference "**statistically discernible**" or "**statistically significant**" at 5% "**error level**" or "**significance level**".

Remark: Our formal statistical language must not obscure the important commonsense aspects of this problem, of course. Although we have shown (at the 5% error level) that men's and women's salaries are different, we have NOT shown that discrimination necessarily exists. There are many alternative explanations. For example, men may have more education than women, on average. What we really should do then is compare men and women of the same **qualification**. (This can be done using **regression** model.)

Example 9-2: Sex Difference in Salaries, but Weaker Data
Suppose the confidence interval (1) had been based on a smaller sample and, consequently, had been vaguer. Specifically, suppose we calculated the 95% confidence interval to be:

$$[5 - 8,\ 5 + 8] = [-3,\ 13]. \tag{2}$$

Are the following interpretations (at the 5% error level) true or false?
a. Since the hypothesis $\Delta = 0$ falls within the interval (2), it can not be rejected.
b. The true (population) difference may well be 0. That is, the population of men's salaries may be the same as the women's on average. The difference in sample means ($\bar{X}_1 - \bar{X}_2 = 5$) may represent only random fluctuation, and therefore cannot be used to show that a real difference exists in the population means.
c. The plausible population differences with in the interval (2) include both negative and positive values; that is, we cannot even decide whether men's salaries on the whole are better or worse than women's.
d. In (2), we see that the sampling allowance (± 8) overwhelms the estimated difference (5). Whenever there is this much sampling "fog", we call the difference **statistically indiscernible** or **statistically insignificant**.

Type I error and Type II error

- Type I error: Rejecting H_0 when it is true, with its probability of course being α, the error level of the test.
- Type II error: Accepting H_0 when it is false.
- If we make α small, the risk of Type I error decreases, while the risk of Type II error increases.

p-Value

- After a hypothesis test is done, the conclusions must be reported in some statistically meaningful way.
- One method of reporting the results of a hypothesis test is to report the error level (we sometimes call it "**size**"), α, of the test used and the decision to reject H_0 or accept H_0.
- The error level (or size) of the test carries important information.
- If α is **small**, the decision to reject H_0 is fairly convincing, but if α is **large**, the decision to reject H_0 is not very convincing because the test has a large probability of incorrectly making that decision.
- Another way of reporting the results of a hypothesis test is to report the value of a certain kind of test statistic called **p-value**.

Example 9-3: New TV Tube Production make difference?
A standard manufacturing process has produced millions of TV tubes, with a mean life $\mu = 1200$ hours and a standard deviation $\sigma = 300$ hours. A new process produces a sample of 100 tubes, with an average $\bar{X} = 1265$. Although this sample makes the new process look better, is this just a sampling fluke? Is it possible that the new process is really no better (and worse) than the old, and we have just turned up a misleading sample?
To give this problem more structure, we state the null hypothesis: the new process would produce a population that is no different from the old–that is, $H_0 : \mu = 1200$. The alternative hypothesis can be written as $H_1 : \neq 1200$.

How consistent is the sample $\bar{X} = 1265$ with the null hypothesis H_0? Specifically, **if the null hypothesis H_0 is true, what is the probability that $\bar{X}$ would be as extreme as 1265?** It is this probability that we call **p-value**.

Solution:
If H_0 is true, by the Normal Approximation Rule, the distribution of $\bar{X}$ is normal, with mean $\mu_0 = 1200$, and standard error$= \sigma/\sqrt{n} = 300/\sqrt{100} = 30$. We use these to standardize the observed value $\bar{X} = 1265$:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{1265 - 1200}{30} = 2.17$$

Thus,

$$Pr(|\bar{X}| \geq 1265|H_0) = Pr(|Z| \geq 2.17) = .03.$$

What does this mean in plain English? If H_0 were true, there would be only 3% probability of observing $\bar{X}$ as large or small as 1265. This 3% is therefore called the **p-value** for H_0 (or more specifically, the two-sided p-value, matching the two-sided alternative hypothesis).
The p-value summarizes very clearly how much agreement there is between the data and H_0.

Using Student's t statistic

We have seen how $\bar{X}$ was standardized so that the standard normal table could be used. The key statistic we evaluated was

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{X} - \mu_0}{\text{exact SE}}$$

Usually σ is **unknown**, and has to be estimated with the sample standard deviation s. Then the statistic is called **t-statistic** instead of Z:

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{\bar{X} - \mu_0}{\text{estimated SE}}.$$

Since $\bar{X}$ fluctuates around μ_0, Z fluctuates around 0. Similarly the t-statistic fluctuates around 0 –but with **wider** variability, as already noted in Chapter 7.

When H_0 is true, the t-statistic follows the student's t-distribution. So we can calculate the p-value using the t-table .

There are many t-distributions, one for each sample size; hence, one for each d.f. (degrees of freedom).

The same recommendation as in Chapter 7 applies here:

As the sample size grows larger than 100, say, the t-distribution becomes very close to the standard normal distribution. (For example, as we read down t-Table, for d.f. $= 120$ $t_{.025} = 1.98$, while $z_{.025} = 1.96$, which is an excellent approximation.)

So in practice when σ is unknown, the t-table needs to be used only for small samples ($n < 100$).

Of course, the t-statistic can be easily generalized to cover any of the other situations encountered in Chapter 7:

$$t = \frac{\text{estimate} - \text{null hypothesis}}{\text{estimated SE}}.$$

Example 9-4: Sex Difference in Salaries Once Again
For the mean difference in men's and women's salaries (in thousands of dollars, annually), we calculated $\bar{X}_1 - \bar{X}_2 = 5.0$, $SE = 1.84$, and d.f.$= 13$. Since the null hypothesis is $H_0 : \Delta = \mu_1 - \mu_2 = 0$, the t-statistic is

$$t = \frac{\text{estimate} - \text{null hypothesis}}{\text{estimated SE}} = \frac{5.0 - 0}{1.84} = 2.72.$$

Since d.f.$= 13$, we scan along the thirteenth row of the t-table, and find that the the value of t-statistic, 2.72, lies beyond $t_{.01} = 2.65$. This means that the tail probability is smaller than .01, that is,

$$\text{p-value} < .01.$$

Since the p-value is a measure of the credibility of H_0, such a low value leads us to conclude that H_0 is implausible.

One-sided test

When there is one sided claim to be made, such as, “more than”, “less than”, “better than”, “worse than”, “at least”, then the one-sided test is more appropriate.

Example 9-3’: TV Tube Production Improved?
A standard manufacturing process has produced millions of TV tubes, with a mean life $\mu = 1200$ hours and a standard deviation $\sigma = 300$ hours. A new process, recommended by the engineering department as better, produces a sample of 100 tubes, with an average $\bar{X} = 1265$. Although this sample makes the new process look better, is this just a sampling fluke? Is it possible that the new process is really no better than the old, and we have just turned up a misleading sample?
To give this problem more structure, we state the null hypothesis: the new process would produce a population that is no different from the old –that is, $H_0 : \mu = 1200$. The claim of the engineering department that the new process is better can be written as $H_1 : \mu > 1200$.

Solution:
If H_0 is true, by the Normal Approximation Rule, the distribution of $\bar{X}$ is normal, with mean $\mu_0 = 1200$, and standard error$= \sigma/\sqrt{n} = 300/\sqrt{100} = 30$. We use these to standardize the observed value $\bar{X} = 1265$:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{1265 - 1200}{30} = 2.17$$

From the standard normal table, we find that the the value of 2.17 lies beyond $z_{.05} = 1.64$. This means that we can reject H_0 with 5% level.

p-Value (one-sided)

In Example 9-3', the one-sided p-value is

$$Pr(\bar{X} \geq 1265|H_0) = Pr(Z \geq 2.17) = .015,$$

that is, if the new process were no better (that is, if H_0 were true), there would be only 1.5% probability of observing $\bar{X}$ as large as 1265. So, we can conclude that the data provided very little support for H_0.

Proportion

Example 9-5: Suppose General Electric regularly receives shipments of cooling units to install in its refrigerators and over the past 18 months, only 2% of these units have been substandard. When the supplier switches production to a new plant however, General Electric is concerned that quality may have deteriorated. Therefore they randomly sample 500 units of the next shipment, and find that 21 are substandard.

a. Calculate the 95% confidence interval for the proportion of substandard units in the whole shipment.

b. Calculate the p-value for the null hypothesis that quality remains unchanged ($\pi = .02$).

Solution:

a. The sample proportion of substandards is $P = 21/500 = .042$. Around it we construct the familiar 95% confidence interval for π, the population proportion:

$$
\begin{aligned}
&[P - 1.96\sqrt{\frac{P(1-P)}{n}},\ P + 1.96\sqrt{\frac{P(1-P)}{n}}] \\
&= [.042 - 1.96\sqrt{\frac{.042(.958)}{500}},\ .042 + 1.96\sqrt{\frac{.042(.958)}{500}}] \\
&= [.042 - 1.96(.0090),\ .042 + 1.96(.0090)] \\
&= [4.2\% - 1.8\%,\ 4.2\% + 1.8\%].
\end{aligned}
$$

b. Since the null hypothesis is $H_0 : \pi = .02$, the t-statistic is

$$
t = \frac{\text{estimate} - \text{null hypothesis}}{\text{estimated SE}} = \frac{.042 - .02}{.0090} = 2.45.
$$

Since $n = 500$ is large, we can look up the tail area above 2.45 in the normal table: p-value $= .007$. That is, the sampling data shows little credibility for the null hypothesis that the new shipment has maintained the same quality.