Hypothesis Testing

In these notes we will propose methods of testing scientific hypotheses. Studies and experiments allow one to verify scientific hypotheses. Before collecting data, we consider two hypotheses about the probability distribution of a random variable: the *null hypothesis* and the *alternative hypothesis*.

In these notes, we will consider hypotheses about the mean of a population. We will assume that we collect a random sample of size n from a population of mean μ and standard deviation σ.

Formulating hypotheses about μ

Oftentimes there is an established theory that suggests that the population mean μ should be a certain value, μ_0 (we call this the *null hypothesis*, H_0). On the other hand, it is also possible that this theory does not apply in our situation and that μ is not equal to μ_0 (we call this the *alternative hypothesis* H_1). So to summarize, we would like to test

$$H_0 \ : \ \mu = \mu_0 \quad \text{against} \quad H_1 \ : \ \mu \neq \mu_0,$$

where μ_0 is a real number.

Remarks:

- H_0 is called the null hypothesis.
- H_1 is called the alternative hypothesis.
- $H_1 \ : \ \mu \neq \mu_0$ is a **two-sided alternative**
- Sometimes our scientific hypothesis can be expressed in terms of an inequality. For example, we might hope that a certain modification to a machine will increase the mean output of this machine, and we would like to test whether making this modification actually helps. In this situation, we will use a one-sided alternative hypothesis.
- If our alternative hypothesis is that $\mu > \mu_0$, we call this a **right-sided alternative**

$$H_0 \ : \ \mu = \mu_0 \quad \text{against} \quad H_1 \ : \ \mu > \mu_0.$$

- If our alternative hypothesis is that $\mu < \mu_0$, we call this a **left-sided alternative** :

$$H_0 \ : \ \mu = \mu_0 \quad \text{against} \quad H_1 \ : \ \mu < \mu_0.$$

- After formulating the hypothesis, we conduct an experiment or study and evaluate our evidence for H_1 against H_0. If the evidence is significant enough, then we reject H_0 in favor of H_1. If the evidence against H_0 is not significant enough, then we do not reject H_0.

- **Types of errors:** There is always a risk that the conclusion of our test is false.
  - If we reject H_0 when H_0 is true, we say we committed an error of **type I**.
  - If we fail to reject H_0 when H_1 is true, we say we committed an error of **type II**.

Example 60a: A company makes 6-meter tubes. We sampled $n = 10$ tubes at random with sample mean $\overline{x} = 5.7$ meters and sample standard deviation $s = 1.2$ meters. Formulate hypotheses to test if the population of these tubes is actually 6 meters.

solution: $H_0 : \mu = 6\text{m}$ against $H_1 : \mu \neq 6\text{m}$

Example 61a: We study the yield of a chemical process. We would like the yield to be greater than 0.9. We will use this process only if we have evidence that the mean yield is greater than 0.9.

(i) Formulate hypotheses to test whether we should keep using this process.

solution: $H_0 : \mu = 0.9$ against $H_1 : \mu > 0.9$

(ii) What does it mean for us to have a type 1 error in this situation? A type 2 error?

solution: A type 1 error means we will end up using the chemical process even though it has a bad yield. A type 2 error means that we will fail to use the chemical process even though it has a good yield.

Example 62a: We study the yield of a chemical process. We would like the yield to be at least 0.9. If we have proof that the mean yield μ is less than 0.9, then we must make improvements to the chemical process.

(i) Formulate hypotheses to test whether we should make improvements to our chemical process.

solution: $H_0 : \mu = 0.9$ against $H_1 : \mu < 0.9$

(ii) What does it mean for us to have a type 1 error in this situation? A type 2 error?

solution: A type 1 error means that we will end up making improvements even though these improvements are unnecessary. A type 2 error means that we will not make improvements even though improvements are needed.

Example 63a: Suppose that a manufacturer declares that the mean lifetime of their tires is $60,000$ kilometers. A group of consumers suspects that the company is lying and that the lifetime might actually be less than $60,000$ kilometers. To file a complaint, the consumer group needs evidence that the company is lying. Formulate a null hypothesis and alternative hypothesis to test the hypothesis of this consumer group.

solution: $H_0 : \mu = 60,000\text{km}$ against $H_1 : \mu < 60,000\text{km}$

Evaluating the evidence against a null hypothesis: confidence intervals and critical regions

We will see that there are several equivalent approaches for testing hypotheses. Using the **confidence interval**, we will obtain a test based on a **critical region** for a statistic T.

Case i: Consider a **two-sided test**.

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_1 : \mu \neq \mu_0,$$

where μ_0 is a real number.

Use a $(1-\alpha)100\%$ confidence interval to estimate μ and compare this estimate to the value μ_0 of the null hypothesis. If the population is normal and σ is unknown, the confidence interval for μ is

$$\text{C.I.} = [\overline{x} - t_{\alpha/2;n-1}\frac{s}{\sqrt{n}}, \overline{x} + t_{\alpha/2;n-1}\frac{s}{\sqrt{n}}].$$

Hypothesis test based on a Confidence Interval: If $\mu_0 \notin C.I.$, then we are confident that $\mu \neq \mu_0$. So, we will say that we have significant evidence against H_0 in favor of H_1. Otherwise, we say that the evidence against H_0 is not significant.

Remarks:

- We say that α is the **significance level** of the test.
- If H_0 is true, then μ_0 will be inside the $C.I.$ with probability $(1-\alpha)100\%$. So we will end up rejecting H_0 (in error) with probability α when H_0 is true. (this is an error of type 1)
- The significance level is

  $$\alpha = P[\text{reject } H_0 \mid H_0 \text{ is true}].$$

  So, α is our risk of commiting an error of type 1.

Construction of a critical region. If $\mu_0 \notin$ CI, this means that

$$\mu_0 < \overline{x} - t_{\alpha/2;n-1}\frac{s}{\sqrt{n}} \quad \text{or} \quad \mu_0 > \overline{x} + t_{\alpha/2;n-1}\frac{s}{\sqrt{n}}.$$

This is equivalent to the following ineqaulity

$$|t_0| > t_{\alpha/2;n-1}, \quad \text{where} \quad t_0 = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}.$$

Defining the Hypothesis test using the critical region:
Let

$$t_0 = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}.$$

We reject H_0 in favor of H_1 if $|t_0| > t_{\alpha/2;n-1}$.

Remarks:

- Observe that t_0 is a measure of the difference between our estimation of μ (which is $\overline{x}$) and the value of μ under H_0 (which is μ_0). As t_0 moves away from 0, our evidence against H_0 becomes stronger.
- The statistic and the critical region we will use depends on the conditions of the applications.
  - If $n \geq 40$ with σ unknown, then the test statistic is

    $$z_0 = \frac{\overline{x} - \mu_0}{s/\sqrt{n}},$$

    and we reject H_0 in favor of H_1 if $|z_0| > z_{\alpha/2}$.
  - If the population is normal with σ unknown (but with n possibly smaller than 40) then the test statistic is

    $$t_0 = \frac{\overline{x} - \mu_0}{s/\sqrt{n}},$$

    and we reject H_0 in favor of H_1 if $|t_0| > t_{\alpha/2;n-1}$.

Example 60b: A company manufactures 6-meter tubes. We selected a random sample of $n = 10$ of these tubes with sample mean $\overline{x} = 5.7$ meters and sample standard deviation $s = 1.2$ meters. Do we have significant evidence that the mean is *not* 6 meters? Assume that the length of each tube is normally distributed. (use a significance level of 5% for the test)

(a) Test this hypothesis using a confidence interval.

solution: We need to test
$H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$,
where $\mu = 6$m

$\alpha = 0.05$, $n = 10$, so we can use the t-quantile table to look up the following quantile:

$$t_{\alpha/2,n-1} = t_{0.025,9} = 2.262$$

so the confidence interval is

$$\text{C.I.} = \overline{x} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} = 5.7 \pm 2.262\frac{1.2}{\sqrt{10}} = [4.84, 6.56]$$

so $\mu_0 = 6\text{m} \in$ C.I. is inside the confidence interval, so No, we do not have sufficient evidence that $\mu \neq 6\text{m}$. Therefore, we accept the null hypothesis $H_0 : \mu = \mu_0$ by default.

(b) Test this hypothesis again using a t or z statistic.

solution: the t-statistic is

$$t_0 = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{5.6 - 6}{1.2/\sqrt{10}} = -0.791$$

so

$$|t_0| = 0.791 \leq 2.262 = t_{0.025,9} = t_{\alpha/2;n-1}$$

Therefore, since $|t_0| \leq t_{\alpha/2;n-1}$, we do not have sufficient evidence that $\mu \neq 6\text{m}$. Therefore, we accept the null hypothesis $H_0 : \mu = \mu_0$ by default.

Remark : We can define "one-sided hypothesis tests" in a similar way as we defined two-sided hypothesis tests, except that we use confidence bounds in place of two-sided confidence intervals. (although we do not do so in these notes).

The p value: a measure of significance

Suppose that we use a significance level of $\alpha = 5\%$ (so the probability of commiting a type 1 error is 5%). We can determine if our evidence against H_0 in favor of H_1 is significant by using a confidence interval or a critical region. A natural question is then to ask if our evidence will still be significant at 2% significance level, at 1% significance level, at 0.5% significance level, and so on. The p-value is a measure that allows us to answer this question, without having to try many different significance levels.

Definition: The p value is the smallest significance level for which we reject H_0.

Normal population with σ unknown : The test statistic is

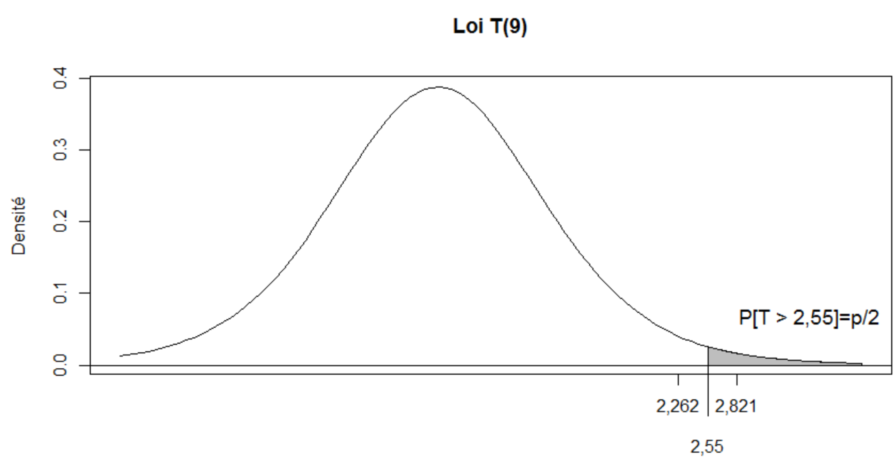
$$t_0 = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}, \quad \text{when } H_0 \ : \ \mu = \mu_0.$$

- **Two-sided alternative ($H_1 : \mu \neq \mu_0$) :** The critical region of t_0 is

$$|t_0| > t_{\alpha/2;n-1}.$$

  As α becomes smaller, $t_{\alpha/2;n-1}$ becomes larger and eventually the evidence will no longer be significant.

  For example, take $n = 10$, $t_0 = -2.73$, then $|t_0| = 2.73$. If we take $\alpha = 0.05$, then $t_{\alpha/2;n-1} = t_{0.025;9} = 2.262$ (so we reject H_0). But if we take $\alpha = 0.02$, then $t_{\alpha/2,n-1} = t_{0.01,9} = 2.821$ (we do not reject H_0). So we went too far with $\alpha = 0.02$.



We want to find the significance level α, which we denote by p, such that

$$t_{p/2;n-1} = |t_0|.$$

But this means that the area to the right of $|t_0|$ under the density of $T(n-1)$ is $p/2$, in other words $P(T > |t_0|) = p/2$, where $T \sim T(n-1)$. If we solve for p, we get

$$p = 2\,P(T > |t_0|), \quad \text{where} \;\; T \sim T(n-1) \;\; \text{and } t_0 = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}.$$

Remarks :

- **Hypothesis test (described in terms of p-values):** If we want to use a significance level of α, then we reject H_0 if the p-value satisfies $p < \alpha$.
- If the population is not necessarily normal, but $n \geq 40$, then the p-value for the two-sided hypothesis test is

$$p = 2\,P(Z > |z_0|), \quad \text{where} \;\; Z \sim N(0,1) \;\; \text{and } z_0 = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}.$$

Example 60c: A company makes 6-meter tubes. We have randomly sampled $n = 10$ of these tubes, with sample mean $\overline{x} = 5.7$ meters and sample standard deviation $s = 1.2$ meters. Is there significant evidence that the mean is not 6 meters? (Assume that the length is normally distributed. Calculate the p-value and use this to test whether there is significant evidence that the mean is not 6 meters. Use a significance level of 5% for the test.)

solution: $\alpha = 0.05$, $\nu = n - 1 = 9$, $\overline{x} = 5.7$, $s = 1.2$

$$t_0 = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{5.6 - 6}{1.2/\sqrt{10}} = -0.791$$

Now,

$$p = 2P(T > |t_0|) \qquad \text{where } T \sim T(n-1)$$

so

$$p = 2P(T > |t_0|) = 2P(T > 0.791) = 2[1 - P(T \leq 0.791)]$$
$$\approx 2[1 - 0.75] = 2 \times 0.25 = 0.5$$

So

$$0.5 = p > \alpha = 0.05$$

Since $p > \alpha$, we do not have sufficient evidence that $\mu \neq 6$m. Therefore, we accept the null hypothesis $H_0 : \mu = 6$m by default.

(**note:** In this example we looked up the p-value $P(T \leq 0.791) \approx 0.75$ using the T-table "backwards". However, in real life the p-value for the t-distribution is computed more precisely using a computer, so **we will give you any t-distribution p-values required on the exam** (you will not have to look up any p-values for the t-distribution using the t-table.))

Remark : We can define p-values in a similar way for one-sided hypothesis testing as for two-sided hypothesis testing (although we do not do so in these notes).