8 Hypothesis Testing

8.1 Introduction

In Chapter 7, we've studied the estimation of parameters, point or interval estimates. The construction of CI relies on the sampling distribution of $\bar{X}$ using CLT.

In this chapter, we look at the problem of making decision about the population parameters, especially the population mean, μ based on sample information. For example, we wish to confirm if the mean of a particular population is 90 (a hypothetical value based on our belief and/or experience) or more than 90. This statistical inference problem is known as *hypothesis testing*. It is very important in statistics because it helps us to draw conclusions about the population parameters.

8.2 General Concepts (P.111-114)

First we look at some terms to be used in this section.

- **Statistical hypothesis** is a claim about a population parameter. This claim may or may not be true.
- **Null hypothesis:** A default or unverified statement about the population parameter is called null hypothesis. Many authors prefer to use the notation H_0 to represent a null hypothesis.
- **Alternative hypothesis:** A claim or statement about the population parameter that contradicts or against the null hypothesis. A useful notation is H_1.

Motivating example: Suppose that a university reports that the average UAI for 2012 entry was 90. However, an education agency claims that the university's report is incorrect and its average entry UAI was in fact greater than 90.

Note: It clear that the university's claim is a default statement and the agency wants to verify it. Therefore, it is the null hypothesis about the population mean, $H_0 : \mu = 90$, where $\mu =$ average UAI entry of all accepted students in 2012.

The alternative hypothesis is the claim from the education agency against the university's claim. In our example, $H_1 : \mu > 90$.

Clearly, there is a dispute between this two groups. How can we resolve this dispute or test this claim using a good statistical argument?

Method:

Take a random sample of students and calculate their average UAI. Draw your conclusion based on sample information and your statistics knowledge. The procedure to decide whether the sample data are *agreeable* or *consistent* with the null hypothesis is called *statistical hypothesis testing* or simply *hypothesis testing* or *test of significance*. It is clear now that the decision includes some randomness.

Possible outcomes in hypothesis testing:

- Decide H_0 is correct when in fact it is true. This is the correct decision.
- Decide H_0 is not correct when in fact it is true. This is an

incorrect decision or an error has been made in the process.

Another example: A suspected person in a court is assumed to be innocent (H_0: innocence) and a judge may decide that the suspected person is guilty only when there is strong evidence that argues again H_0 of innocence, in favor of H_1 of guilty. Every person including the judge may draw incorrect decision. But this error of drawing incorrect decision must be made as small as possible because the error of charging an innocent person guilty can be fatal. This error of *deciding* H_0 *as incorrect when* H_0 *is in fact true* is known as the *Type I error* in practice.

Definition: The probability of Type I error is called the *level of significance* for the test. The letter α is used to represent the level of significance in hypothesis testing. That is

P(Type I error) = P(Reject $H_0|H_0$ is true) = α.

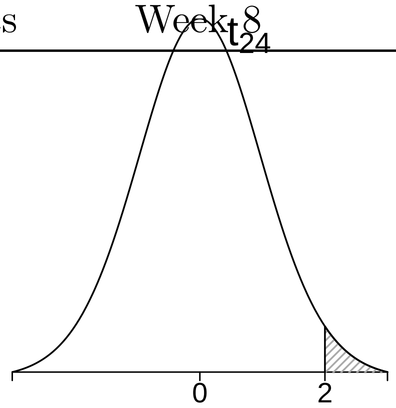
Back to the example:

Take a random sample of say 25 students and calculate the sample average UAI to be $\bar{x} = 93$ and sample s.d. to be $s = 7.5$. Assuming the distribution of X = individual UAI, is normal with certain variance σ^2, that is, $X \sim N(90, \sigma^2)$ under H_0, find the probability of getting the observed and more extreme values; that is, find $P(\overline{X} > 93)$.

Solution: Since X is normal, $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$. Then the distribution of the *standardized difference* between sample mean $\bar{X}$ and hypothesized mean μ or called the *test statistics* is

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

Hence



$$P(\bar{X} \geq 93) = $$

$$= \underline{\hspace{4cm}}$$

$$= \underline{\hspace{4cm}}$$

This is the *observed* level of significance and it is a small probability. Therefore, either

1. the null hypothesis is correct but we observe a very rare event ($\approx 2.5\%$ chance), or
2. the null hypothesis is incorrect.

Notes:

1. Throughout this course, unless stated otherwise, we will assume that $\alpha = 0.05$ or 5%.
2. The above probability of getting the observed and more extreme values under the null hypothesis is less than 5% or very small. Therefore, it is very unlikely to observe a sample mean of 93 or higher if the true mean is 90 under H_0. This shows that the data are *inconsistent with* or *against* our assumption of $H_0 : \mu = 90$.

Conclusion: the data provides strong evidence against the null hypothesis, H_0.

Definition: The probability of getting the observed and more extreme values under the condition of null hypothesis is called the *P-value* or *observed significance level* of the test.

Remark: According to the previous argument,

1. small P-values (i.e., P-value $< \alpha$) argue against H_0. Similarly,

2. large P-values (i.e., P-value $> \alpha$) support H_0 and we say that the data are consistent with H_0.

Note: When we say the data are consistent with H_0, *it doesn't prove H_0 to be true* because the P-value is calculated assuming H_0 is true. What we can say is that there is not sufficient evident from the data to argue against H_0. In other words, we are *inconclusive* about H_0 based on the data.

Read pages 116-118.

Now we look at hypothesis tests using the t distribution.

8.3 One-sample t-test for the mean of a Normal population (P.114-119)

Recall the test discussed before for testing the UAI of 2012 entrants at a university. This test was based on a single sample and therefore such a test is called a *one-sample* test.

Hypotheses:
$H_0 : \mu = 90$ (null hypothesis) against
$H_1 : \mu > 90$ (alternative hypothesis)

Note that this alternative hypothesis is one-sided and the test is said to be *one-sided* as we test for values only on one side of 90. Also since n *is small* and σ *is unknown*, we use the t distribution and this kind of test is called a *t-test*.

8.4 Steps in Hypothesis Testing: t-test

The steps of any hypothesis test are very similar. Since this is the first test we study, we will list them in detail for the t-test:

1. Identify the null and alternative hypotheses.

2. Write down the test statistic, $t = \dfrac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$ and evaluate it using the sample mean, $\bar{x}$, the sample standard deviation, s and the value of μ assuming H_0 is true.

3. Write down a statement for the P-value to calculate the extreme probability under H_0.

4. Draw your conclusion based on the P-value in step 3:
   a large P-value ($> \alpha$): the data are consistent with H_0;
   a small P-value ($< \alpha$): the data provide evidence against H_0.

Remark: All statistical tests are based on *test statistics* which have different formulas, depending on parameters to be tested and assumptions on data.

Note: Some authors replace the decision

"data are consistent with H_0" by "*accept H_0*" and
"there is evidence against H_0" by "*reject H_0*".

We prefer to use the former expressions.

Example 1: A city health department wishes to determine the mean bacteria count per unit volume of water at a lake. The regulation says that the bacteria count per unit volume of water should be less than 200 for safety use. To test the claim that the water is safe, a researcher collected 10 water samples of unit volume and found the bacteria count to be:

$$175,\ 190,\ 215,\ 198,\ 184,$$
$$207,\ 210,\ 193,\ 196,\ 180$$

Do these data indicate that there is no cause for concern? Assume that the measurements constitute a sample from a normal population.

Solution: Let μ be the current population mean bacteria count per unit volume.

1. Hypotheses: Clearly, the null hypothesis is ____________.

'No cause for concern if the bacteria count is less than 200' and therefore, the alternative hypothesis is ____________.

2. Test statistic: Since the population is normal and the sd is not known, we employ the t-test with $(n-1)$ df using

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}.$$

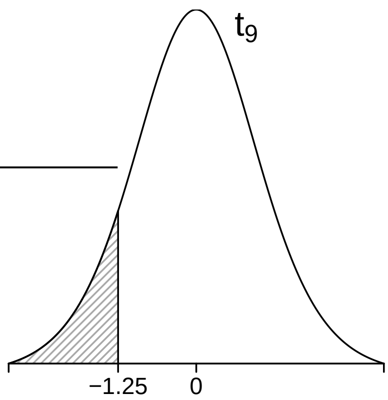
Now calculate the sample mean, $\bar{x} = 194.8$ and sd, $s = 13.14$ from the data and calculate t using these sample values under the null hypothesis H_0. The observed test statistic is

$$t_{obs} = \frac{\qquad}{\qquad\qquad}$$

3. P-value: Since the df = 9, the P-value is

$P(\bar{X} \leq 194.8) =$ ______________________________

from the t-table with 9 df.



4. Conclusion: ________________________. In other words, we don't have sufficient evidence to show that the true mean level of bacteria is within the safety level of below 200.

Example 2: It is believed that the mean plasma aluminium level for population of healthy infants is 4.13 micro gram/l and this level increases for infants receiving antacid containing aluminium. To test this claim, a random sample of 11 infants receiving antacid containing aluminium is examined and gave a mean of $\bar{x} = 5.20$ and $s = 1.13$ (both are in micro gram/l). Test this claim assuming normality.

Solution: Let μ be the current mean plasma aluminium level for population of infants.

1. Hypotheses: By the default, μ should be 4.13, then the null hypothesis is

____

"It is believed that the mean plasma aluminium level for infants receiving antacid containing aluminium is higher" and so the corresponding alternative hypothesis is __________.

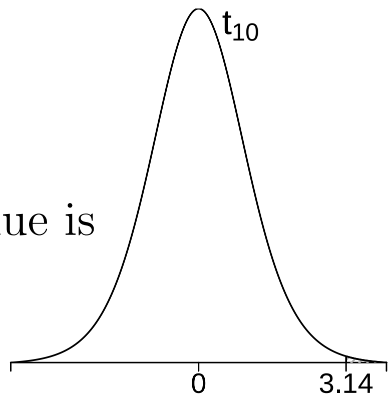
2. Test statistic: Since the population is normal and the sd is not known, we employ the t-test with $(n-1)$ df, using

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}.$$

Now use the given sample mean and sd to calculate t under the null hypothesis H_0. That is,

$$t_{obs} =$$

____



3. *P*-value: Since the df = 10, the P-value is

____

from the t-table with 10 df.

4. Conclusion: ____.
In other words, the true mean plasma aluminium level for infants receiving antacid containing aluminium is higher than the normal infants.

Note: Clearly understand this procedure for calculating the associated P-value according to the alternative hypothesis, H_1.

8.5 Two-Sided Alternatives (P.120-121)

In our previous examples on hypothesis testing, the alternative hypotheses considered only one direction (with $>$ or $<$). For example

in (1) $H_1 : \mu < 200$ and
in (2) $H_1 : \mu > 4.13$.

These H_1 look for either larger or smaller values of μ. Therefore it is reasonable call these tests as one-sided.

However, in many practical problems we need to look for a specific (exact) value for the parameter, μ. We cannot accept either larger or smaller values than this specific value. In other words, we will not accept any claim if the value is *different* (even slightly) from the specified value.

Example 3: Suppose that a manufacturing company of surgical tubes claims that its production meets the specified average diameter of 8.5 mm. A medical research team believes that the manufacturer's claim is incorrect. To test this claim they took a random sample of 16 tubes and found that the sample mean, $\bar{x} = 8.3$ and $s = 0.8$ (both are in mm). Set up the null hypothesis H_0 and the alternative hypothesis H_1 in this problem.

Solution: Let μ be the population mean diameter of all such potential surgical tubes.

1. Hypotheses: We need to investigate the manufacturer's claim against the medical research team. Therefore, the null hypothesis is ____________.

Since any large or small tubes cannot be used in a surgery, the alternative hypothesis is ___________.

2. Test statistic: Since the population is normal and the sd is not known, we employ the t-test with $(n-1)$ df using

$$t = \frac{\overline{X} - \mu}{S/\sqrt{n}}.$$

Use the given sample mean and sd to calculate t under the null hypothesis H_0. That is,

$$t_{obs} =$$

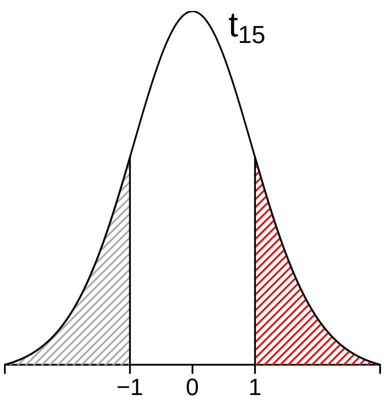
____

Note that the sample mean is below the null hypothetical value of 8.5 and therefore expect a negative value for t_{obs}.

3. ***P*-value:** Note that the df = 15. Since we need to allow both the large and small values to argue against H_0, the P-value is calculated by

____

which is very large from the t table with 15 df.



Note: we multiply $P(t_{15} \leq -1.0)$ by 2 to accommodate both large and small values for this two-sided alternative.

Remark: In general, the P-value for a two-sided t-test is:

$$P\text{-value} = 2\text{P}(T_{n-1} \geq |t_{obs}|)$$

4. Conclusion: ______________________________. In other words, the true mean diameter of the tube is not significantly different from 8.5mm or the data agree with the manufacturer's claim.

8.6 Relationship between Hypothesis Testing and CI's (P.122-125)

Suppose we are testing

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu \neq \mu_0$$

with a two-sided alternative. Then finding evidence against H_0 at significance level α is equivalent to μ_0 to be *outside* the $(1-\alpha)100\%$ CI for μ, that is,

$$\text{against} \quad H_0 : \mu = \mu_0 \quad \Longleftrightarrow \quad \mu_0 \ \text{lies outside} \ \text{CI for} \ \mu$$

$$\text{or, consistent with} \quad H_0 : \mu = \mu_0 \quad \Longleftrightarrow \quad \mu_0 \ \text{lies inside} \ \text{CI for} \ \mu$$

Remarks:

- Make sure that the α used for the hypothesis test is matching the confidence level of the CI. For example, we can not draw a conclusion for testing hypothesis at significance level $\alpha = 0.05$, based on a 98% CI.

- The duality between hypothesis testing and CI is valid only for two-sided alternatives. We do not study one-sided CI in this unit, therefore we can not apply the duality for one-sided alternative hypothesis.

- The same relationship between hypothesis testing and CI holds for any test. That is, we can apply it for tests other than the one-sample t-test for μ.

Example 4: Suppose that we want to know how many hours of TV per day (on average) a university student watches. We take a sample and find that the 99% CI for the true mean μ is (3.0, 3.4).

Assume that the common belief is that the mean number of hours of TV is 3.5 per day. Hence, $H_0 : \mu = 3.5$. We disagree with this claim, so $H_1 : \mu \neq 3.5$. Use the 99% CI to test the hypotheses.

Solution: Because the hypothesised value falls outside the CI ($3.5 \notin (3.0, 3.4)$), we can say that our data support the alternative hypothesis H_1 at the 1% significance level ($\alpha = 0.01$).

On the other hand, if the null hypothesis were $H_0 : \mu = 3.1$ with $H_1 : \mu \neq 3.1$, then we don't have enough evidence against the null hypothesis here for $\alpha = 0.01$, because 3.1 lies within the 99% CI.

Exercises: Book P.134 Q1-2, Q3 Should be repeat Q2(a), Q4-5.