CHAPTER 10

ONE- AND TWO-SAMPLE TESTS OF HYPOTHESES

Confidence intervals represent the first of two kinds of inference that we study in this course. **Hypothesis testing**, or **test of significance** is the second common type of formal statistical inferences. It has a different goal than confidence intervals.

The big picture is that the test of hypothesis helps us to assess the evidence provided by the data in favor of some claim concerning a population.

Let us look at an example first.

EXAMPLE 10.1 (Tomato). Suppose that you are in charge of quality control in your food company. The company produces a type of packs of cherry tomatoes, labeled $1/2$ lb (227 g). To monitor the mean weight of the packs of tomatoes, you sample randomly four packs of cherry tomatoes, and find the following weights:

$$212 \text{ g}, 205 \text{ g}, 225 \text{ g}, 230 \text{ g}$$

The average weight from the four boxes is 218 g. You suspect that the mean weight is not half a pound (less than $1/2$ lb is illegal, and more than $1/2$ lb will result in profit loss of the company). Does the data provide evidence that the mean weight of tomato packs is not $1/2$ lb?

- If it does, the calibrating machine that sorts cherry tomatoes into packs needs revision.
- We cannot expect that boxes filled with tomatoes weigh exactly half a pound. The mean weight of cherry tomato boxes is expected to be 227 g.
- Suppose the weights of tomato boxes is normally distributed with standard deviation $\sigma = 5$ g.
- Does $\bar{x} = 218$ provide evidence that the mean weight μ of tomato boxes is not 227 g?
- The true value of mean weight μ is unknown but a decision about μ must be made. Is $\mu = 227$ or $\mu \neq 227$?

10.1 Statistical Hypotheses

A test of statistical significance tests a specific hypothesis using sample data to decide on the validity of the hypothesis.

Statistical Hypotheses

A *statistical hypothesis* is an assertion or conjecture concerning one or more populations.

In other words, we have some claim about the value of some population parameter and we would like to provide evidence that either supports or rejects this claim. We accomplish this by looking at sample data and seeing if they are representative of the claim.

We can **never** prove that a parameter has any particular value, so we try and reach our conclusions with a high probability of being correct.

The statements of claim we are testing are expressed as **two hypotheses**:

Null Hypothesis H_0

The statement being tested in a statistical test is called the *null hypothesis*, denoted H_0.

- The test is designed to assess and determine the strength of evidence in the data **against** the null hypothesis H_0.
- The null hypothesis H_0 is a clear and definite statement of "no change" or "no difference" about a parameter of the population, often being stated using the *equality sign*.

NOTE. Because we are interested in the value of a parameter for the whole population, we always express our statements of interest in terms of population parameters (μ, p or σ^2).

The null hypothesis of Tomato Example 10.1 is:

$$H_0 : \mu = 227$$

where μ is the true average weight of tomato packs.

NOTE. The conclusions do NOT involve a formal and literal "accept H_0."

One should reach at one of the two following conclusions:

- reject H_0 in favor of H_1 because of sufficient evidence in the data
- fail to reject H_0 because of insufficient evidence in the data.

Alternative Hypothesis H_1

The statement making the claim which we are trying to find evidence for supporting is called the *alternative hypothesis*, denoted H_1 or H_a.

The alternative hypothesis of Tomato Example 10.1 is:

$$H_1 : \mu \neq 227$$

NOTE. The alternative hypothesis could be **one-sided** or **two-sided**.

What determines the choice of a one-sided versus a two-sided test is what we know about the problem before we perform a test of statistical significance. It is important to make that choice before performing the test or else you could make a choice of "convenience" or fall in circular logic.

EXAMPLE 10.2 (Hepatitis B Vaccine). Henning *et al* (Amer. J. Pub. Health 1992) found that in a sample of 670 infants, 66 percent of them had completed the hepatitis B vaccine series. Can we conclude on the basis of these data that in the sampled population, more than 60 percent infants have completed the hepatitis B vaccine series? What are the null and alternative hypotheses?

EXAMPLE 10.3 (Material). An experiment was performed to compare the abrasive wear of two different laminated materials. Twelve pieces of material 1 were tested by exposing each piece to a machine measuring wear. Ten pieces of material 2 were similarly tested. In each case, the depth of wear was observed. The samples of material 1 gave an average (coded) wear of 81 units with a sample standard deviation of 4, while the samples of material 2 gave an average of 85 with a sample standard deviation of 5. Can we conclude that the abrasive wear of material 1 is less than that of material 2?

NOTE. Note again that hypotheses always refer to some population or model, not to a particular outcome. For this reason, we must state H_0 and H_1 in terms of *population* parameters. So, $\bar{x} > 100$, $\bar{x}_1 - \bar{x}_2 < 1.2$, $\hat{p} \neq 0.75$, $\sigma^2 < 0.4$ or $\sigma_1^2/\sigma_2^2 \neq 1$ would NOT be the correct statistical hypothesis.

The most well-known illustration for hypothesis testing is on the predicament encountered in a jury trial, where the null and alternative hypotheses are:

H_0: defendant is innocent
H_1: defendant is guilty.

Would "failure to reject H_0" imply innocence? Do you think the jury should "accept H_0" or "fail to reject H_0"?

NOTE. **Keep in mind** that one should reach at one of the two following conclusions:

- reject H_0 in favor of H_1 because of sufficient evidence in the data
- fail to reject H_0 because of insufficient evidence in the data.

So, your conclusions do NOT involve a formal and literal "accept H_0."

10.2 Testing a Hypothesis

Test Statistics

A *test statistic* estimates the parameter that appears in the hypotheses.

- When H_0 *is true*, we expect the estimate to take a value *near* the parameter value specified by H_0.
- Values of the estimate far from the parameter value specified by H_0 give evidence *against* H_0.
- The alternative hypothesis determines which direction count against H_0.

The test statistic of Tomato Example 10.1 can be found to be

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{218 - 227}{5/\sqrt{4}} = -3.6$$

The test statistic assesses how far the estimate is from the parameter. The standardized estimate is usually

adopted. In many common situations the test statistic is of form

$$\frac{\text{estimate - hypothesized value}}{\text{standard deviation of the estimate}}$$

For example, the statistic for testing $H_0: \mu = \mu_0$ is

$$\begin{cases} z_0 = \dfrac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}, & \text{when } \sigma \text{ is known} \\ t_0 = \dfrac{\bar{x}-\mu_0}{s/\sqrt{n}}, & \text{when } \sigma \text{ is unknown.} \end{cases}$$

Significance Level α

The probability that the test statistic will fall in the critical region when the null hypothesis is actually true. It is also the probability of making the mistake of rejecting the null hypothesis when it is true. Common choices of α is 0.01, 0.05, 0.10, with 0.05 being most common.

NOTE. If the level of significance α is NOT given, a rule of thumb is to take $\alpha = 0.05$.

Critical Region (Rejection Region)

The set of all values of the test statistic that cause us to reject the null hypothesis. They are the extreme regions bounded by the critical values.

For example, again, suppose that we test

- $H_0: \mu = \mu_0$ v.s. $H_1: \mu \neq \mu_0$: The critical region is in the two extreme regions (both tails).

$$\left|\text{test statistic value}\right| > z_{\alpha/2} \text{ or } t_{\alpha/2}$$

- $H_0: \mu = \mu_0$ v.s. $H_1: \mu < \mu_0$: The critical region is in the extreme left region (left tail).

$$\text{test statistic value} < -z_\alpha \text{ or } -t_\alpha$$

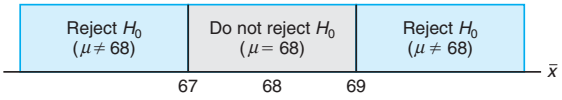
- $H_0: \mu = \mu_0$ v.s. $H_1: \mu > \mu_0$: The critical region is in the extreme right region (right tail).

$$\text{test statistic value} > z_\alpha \text{ or } t_\alpha$$

Decision Criterion using Critical Regions

- If the test statistic falls within the critical region, we reject H_0 at a level of significance α.
- If the test statistic does not fall within the critical region, we fail to (do not) reject H_0 at an α level.

For testing $H_0: \mu = 68$ v.s. $H_1: \mu \neq 68$,



A summary for the Critical/Rejection Region Approach

1. State the null and alternative hypotheses.
2. Choose an appropriate test statistic and establish the critical region based on α.
3. Reject H_0 if the computed test statistic is in the critical region. Otherwise, do not reject.
4. Draw scientific or engineering conclusions.

EXAMPLE 10.4 (Health). A health advocacy group suspects that cigarette manufacturers sell cigarettes with a nicotine content higher than what they advertise (1.4 mg per cigarette) in order to better addict consumers to their products and maintain revenues. The health advocacy group took a sample random sample of 12 cigarettes and found that the sample mean was 1.6 mg with a standard deviation of 0.3 mg. Suppose that the measurement of nicotine in cigarettes is normally distributed. Is there any evidence for the suspicion of the health advocacy group at $\alpha = 5\%$? State the hypotheses, calculate the value of the test statistic, write the rejection region and draw the statistical conclusion.

Type I Error, Type II Error and Power

As shown in the following diagram, there are four situations we may encounter upon making a conclusion for a statistical test of hypothesis:

	H_0 **is true**	H_0 **is false**
Do not reject H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

Obviously, we want to reject H_0 when it is false and we want to fail to reject H_0 when it is true.

Type I Error

Rejection of the null hypothesis when it is true is called a *type I error*.

We say we have made a **Type I Error** if we **reject** H_0 when H_0 is **true**.

$$\begin{aligned} \alpha &= \mathrm{P}(\text{type I error}) \\ &= \mathrm{P}(\text{reject } H_0 | H_0 \text{ true}) \end{aligned}$$

It is also called the **size of the test**.

NOTE. This is just our regular **level of significance** we have been using all along.

We control the probability of Type I Error directly through our specification of the significance level α. For example, when $\alpha = 0.05$, we have only a 5% chance of incorrectly rejecting the null hypothesis when it is true.

Type II Error

Nonrejection of the null hypothesis when it is false is called a *type II error*.

We say we have made a **Type II Error** if we **fail to reject H_0** when H_1 is **true**.

$$\begin{aligned}\beta &= P(\text{type II error}) \\ &= P(\text{fail to reject } H_0 | H_1 \text{ true})\end{aligned}$$

Power

The *power* of a test is the probability of rejecting H0 given that a specific alternative is true.

$$\begin{aligned}\text{Power} &= 1 - \beta \\ &= P(\text{reject } H_0 | H_1 \text{ true})\end{aligned}$$

NOTE. The power of a test of significance depends on μ_1, the specific value of the mean under the alternative hypothesis.

EXAMPLE 10.5. Suppose we wanted to conduct a hypothesis test to determine whether or not a coin is fair, i.e., $H_0 : p = 0.5$ vs. $H_1 : p \neq 0.5$, where p is the probability of the coin landing on heads. We will flip the coin ten times and count X, the number of heads we observe. We will reject H_0 if $X \leq 2$ or if $X \geq 8$.

(a) What is the probability of a Type I Error?

(b) What is the power of the test if $p = 0.3$?

EXAMPLE 10.6. A manufacturer has developed a new fishing line, which the company claims has a mean breaking strength of 15 kilograms with a standard deviation of 0.5 kilogram. To test the hypothesis that $\mu = 15$ kilograms against the alternative that $\mu < 15$ kilograms, a random sample of 50 lines will be tested. The critical region is defined to be $\bar{x} < 14.9$.

(a) Calculate the probability of committing a Type I Error.

(b) Calculate the power for each of the alternatives $\mu = 14.9$, $\mu = 14.8$ and $\mu = 14.7$ kilograms, respectively.

(c) Compare, graphically, the results in part (b) and discuss your findings.

Power Relationships with α, β and n

- **Raising** the sample size has greatly **increased the power** of the test.
- For a fixed level of significance α, an increase in the sample size n will result in a **decrease** in β (and thus an **increase** in power.)
- For a fixed sample size n, **reducing** the probability of a Type I Error α will result in an **increase** of the probability of a Type II Error β (and thus a **decrease** in power).
- The **further away** the specified value of the true mean μ_1 is from μ_0, the **higher** the power.

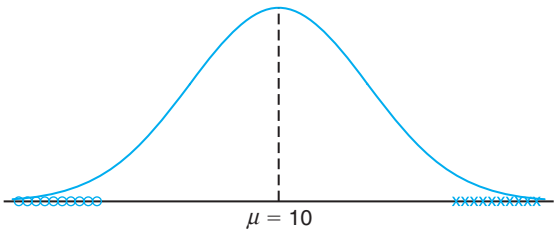
10.3 The Use of P -Values

P-value

The probability, computed assuming that H_0 is true, that the test statistic would take a value as extreme or more extreme than the actual observed value is called the *P-value* of the test.

In other words, a *P-value* is the lowest level (of significance) at which the observed value of the test statistic is significant.

- This is a way of assessing the "believability" of the null hypothesis given the evidence provided by a random sample.
- The P-value is used to measure the strength of the evidence in the data against H_0.
- The *smaller* the P-value, the *stronger* the evidence against H_0.
- The P-value is calculated assuming that the *null hypothesis* H_0 *is true*.
- The pre-fix significance level is not required for calculating the P-values.



For example, again, the P-values can be calculated as follows for one sample mean testing problem.

To test a population mean μ when σ is known,

- $2\mathrm{P}(Z \geq |z_0|)$ for $\mathrm{H}_1 : \mu \neq \mu_0$
- $\mathrm{P}(Z \geq z_0)$ for $\mathrm{H}_1 : \mu > \mu_0$
- $\mathrm{P}(Z \leq z_0)$ for $\mathrm{H}_1 : \mu < \mu_0$

where $z_0 = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ and $Z \sim \mathrm{N}(0,\ 1)$.

To test a population mean μ when σ is unknown,

- $2\mathrm{P}(T \geq |t_0|)$ for $\mathrm{H}_1 : \mu \neq \mu_0$
- $\mathrm{P}(T \geq t_0)$ for $\mathrm{H}_1 : \mu > \mu_0$
- $\mathrm{P}(T \leq t_0)$ for $\mathrm{H}_1 : \mu < \mu_0$

where $t_0 = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$ and $T \sim \mathrm{t}(n-1)$.

NOTE. The inequality sign for P-value is *consistent* with that in the one-sided alternative hypothesis.

Decision Criterion using P-values

- If the P-value $\leq \alpha$, we reject H_0 at the level of significance α.
- If the P-value $> \alpha$, we fail to (do not) reject H_0 at the level of significance α.

A summary for the P-value Approach

1. State the null and alternative hypotheses.
2. Choose an appropriate test statistic and compute the P -value based on the computed value of the test statistic.
3. Reject H_0 if the P-value is smaller than α. Otherwise, do not reject.
4. Draw scientific or engineering conclusions.

EXAMPLE 10.7. Refer to Example Health (10.4) using the P-value approach at a level of significance $\alpha = 5\%$.

Actually, we can also use confidence intervals to help us make decisions.

Confidence Interval Approach

- If the hypothesized parameter θ_0 falls **outside** the $100(1-\alpha)\%$ confidence interval, we **reject H_0**.
- If the hypothesized value θ_0 fall **inside** the $100(1-\alpha)\%$ confidence interval, we **fail to reject H_0**.

NOTE. This approach and the rejection region approach are the same in nature. The confidence area is the "**opposite**" of the rejection region.

EXAMPLE 10.8. Bottles of antacid tablets include a printed label claiming that they contain 1000 mg of calcium carbonate. In a simple random sample of 21 bottles of tablets made by the Medassist Pharmaceutical Company, the amounts of calcium carbonate are measured and the average amount of calcium carbonate is found to be 985 mg. Suppose that the amounts of calcium carbonate in bottles of antacid tablets produced by the company follows a normal distribution with a standard deviation of 50 mg. Is there sufficient evidence to support the claim that consumers are being cheated at the significance level of $\alpha = 0.01$. Use the confidence interval approach.

NOTE. Which is better?

- The **critical region** method and the **confidence interval** method give a black and white answer: Reject or do not reject H_0. But it also estimates a range of likely values for the true population parameters.
- A **P-value** quantifies how strong the evidence is against the H_0. But if you reject H_0, it does not provide any information about the true population parameter.

10.4 Single Sample: Testing Mean

We have seen the procedures of the examples in previous sections. Let's pay attention to the requirements and look at more examples.

EXAMPLE 10.9. A pharmaceutical company makes a therapy cream (for face and hand) bottles that contain a mean amount of therapy cream of 650 *ml* per bottle as indicated on the label. To monitor its quality, the company random selected 25 bottles from the production line and the sample mean amount of cream was 640 *ml* per bottle. Assume that the amount of cream follows a normal distribution with a standard deviation of 4 *ml*. Is there evidence at 0.01 level of significance to conclude that the population mean amount of cream is not 650 *ml* per bottle?

EXAMPLE 10.10. *Do middle-aged male executives have higher average blood pressure than the general population?* The National Center for Health Statistics reports that the mean systolic blood pressure for males 35 to 44 years of age is 128. The medical director of a company looks at the medical records of 28 company executives in this age group and finds that the mean systolic blood

pressure in this sample is 126.07 and the standard deviation is 15. Is this evidence that executive blood pressures are higher than the national average? Assume normality.

10.5 Two Samples: Testing Difference in Means

Assumptions

Similar to the interval estimation for difference in two population means, we require the random samples from each of the populations that are normally distributed (if not, the sample sizes must be large).

Hypotheses

$$\begin{aligned} &H_0 : \mu_1 - \mu_2 = d_0 \\ &H_1 : \mu_1 - \mu_2 \neq d_0 \qquad (\text{or } < \text{ or } >), \end{aligned}$$

where d_0 is the hypothesized value for difference in population means.

Test Statistic for $H_0 : \mu - \mu_2 = d_0$

$$z_0 = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \text{ for } \sigma_1^2 \text{ and } \sigma_2^2 \text{ known}$$

$$t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \text{ for } \sigma_1^2 \text{ and } \sigma_2^2 \text{ unknown \& Equal}$$

$$t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \text{ for } \sigma_1^2 \text{ and } \sigma_2^2 \text{ unknown \& Unequal}$$

Rejection Region

$$\begin{aligned} &|Z| > z_{\alpha/2} \text{ or } |T| > t_{\alpha/2}, && \text{for } H_1 : \mu_1 - \mu_2 \neq d_0 \\ &Z > z_\alpha \text{ or } T > t_\alpha, && \text{for } H_1 : \mu_1 - \mu_2 > d_0 \\ &Z < -z_\alpha \text{ or } T < -t_\alpha, && \text{for } H_1 : \mu_1 - \mu_2 < d_0 \end{aligned}$$

P-value

$$\begin{aligned} &2P(Z \geq |z_0|) \text{ or } 2P(T \geq |t_0|), && \text{for } H_1 : \mu_1 - \mu_2 \neq d_0 \\ &P(Z \geq z_0) \text{ or } P(T \geq t_0), && \text{for } H_1 : \mu_1 - \mu_2 > d_0 \\ &P(Z \leq z_0) \text{ or } P(T \leq t_0), && \text{for } H_1 : \mu_1 - \mu_2 < d_0 \end{aligned}$$

Conclusion

We reject H_0 if P-value $< \alpha$, or t_0/z_0 falls in the RR.

We do NOT reject H_0 if P-value $\geq \alpha$, or t_0/z_0 does not fall in the RR.

EXAMPLE 10.11. Refer to Example Material (10.3). Conduct a hypothesis test at a level of significance 5%. Assume the populations to be approximately normal with unknown-but-equal variances.

EXAMPLE 10.12. We take a random sample of five 10 year-old boys and four 10 year-old girls and measure their heights. The sample means are 55.7 inches and 54.1 inches, respectively . Suppose we know that heights of 10 year-old boys follow a normal distribution with standard deviation 2.9 inches, and that heights of 10 year-old girls follow a normal distribution with standard deviation 2.6 inches. Conduct a hypothesis test, at a level of significance 10% to determine whether the mean height of 10 year-old boys is greater than the mean height for the 10 year-old girls. Assume normality.

EXAMPLE 10.13. In a memory experiment, the volunteers are randomly divided into two groups. Group 1 (10 volunteers) will be given a list of words to study and Group 2 (9 volunteers) will listen to a recording of the same list of words. After half an hour, the subjects will be asked to write down all words they can remember. The scores for each of the subjects are shown below:

Group 1 (read):	52	71	43	39	42
	63	50	48	47	35
Group 2 (listen):	30	28	53	44	31
	32	27	54	43	

Would you be convinced that "read" is better than "listen"? Conduct a test of significance to answer it. Choose $\alpha = 1\%$ and assume scores follow normal distributions for both groups.

EXAMPLE 10.14. An engineer suspects that a power plant in the city is increasing the air pollution in the vicinity of the plant. A sample of measurements of carbon monoxide (in mg/m^3 of air) in the vicinity of the plant is obtained for comparison with another sample obtained from other locations in the city. The data are as follows:

Near plant:	40	16	44	47	26	64	47
	35	53	45	52	31	38	45
	44	29					
Other locations:	34	28	36	43	35	36	38
	39	32	25	37	40	44	37

Concentrations near the plant and at other locations in the city are both assumed to follow normal distributions. Run a hypothesis test for the suspicion of the engineer. Assume unequal variances.