

Hypothesis Testing

Xiuxia Du

Department of Bioinformatics & Genomics
University of North Carolina at Charlotte

`Xiuxia.Du@uncc.edu`

Fall 2009

Contents

1	Introduction	3
2	Hypothesis testing: A single population mean	8
2.1	Sampling from normally distributed populations: population variances known	9
2.2	Sampling from a normally distributed population: population variance unknown	15
2.3	Sampling from a population that is not normally distributed	19
2.4	The power of a test	20
2.4.1	One-sided test	21
2.4.2	Two-sided test	24
2.5	Sample size determination	24
2.5.1	One-sided test	24
2.5.2	Two-sided test	26
2.6	The relationship between hypothesis testing and confidence intervals . .	26
3	Hypothesis testing: the difference between two population means	28
3.1	t-test for the significance between the means of two independent samples	28
3.1.1	Sampling from normally distributed populations: population variances known	29
3.1.2	Sampling from normally distributed populations: population variances unknown	30
3.1.3	Sampling from populations that are not normally distributed . .	34
3.2	t-test for the significance of the difference between the means of two correlated samples	35

1 Introduction

We have discussed the first type of statistical inference, estimation. The other type, hypothesis testing, is the subject of this chapter. Actually, estimation and hypothesis testing are not as different as they are made to appear. As we will see in this chapter, one may use confidence intervals to arrive at the same conclusions that are reached by using the hypothesis testing procedures [1].

Definition 1 (Hypothesis). A hypothesis may be simply defined as a statement about one or more population.

The hypothesis is frequently concerned with the **parameters** of the populations about which the statement is made. A hospital administrator may hypothesize that the average length of stay of patients admitted to the hospital is 5 days; a physician may hypothesize that a certain drug will be effective in 90 percent of the cases for which it is used. By means of hypothesis testing one determines whether or not such statements are compatible with the available data.

Hypothesis testing steps

1. **Data.**
2. **Assumptions.** The same assumptions that are of importance in estimation are important in hypothesis testing. These include the normality of the population distribution, the independence of samples, etc.
3. **Hypothesis.**

There are two statistical hypotheses involved in hypothesis testing. The **null hypothesis** is the **hypothesis to be tested**. It is designated by the symbol H_0 . The null hypothesis is sometimes referred to as a **hypothesis of no difference**, since it is a statement of agreement with (or no difference from) conditions presumed to be true in the population of interest. In general, the null hypothesis is set up for the express purpose of being discredited. Consequently, the complement of the conclusion that the researcher is seeking to reach becomes the statement of the null hypothesis. In the testing process the null hypothesis is either rejected or not rejected. If the null hypothesis is not rejected, we will say that the data on which the test is based on do not provide sufficient evidence to cause rejection. If the testing procedure leads to rejection, we will say that the data at hand are not compatible with the null hypothesis, but are supportive of some other hypothesis. The **alternative hypothesis** is a statement of what we will believe is true if our sample data cause us to reject the null hypothesis. We designate the alternative hypothesis by the symbol H_1 .

Suppose, for example, that we want to answer the question: Can we conclude that a certain population mean is not 50? The null hypothesis is

$$H_0 : \mu = 50$$

and the alternative is

$$H_1 : \mu \neq 50$$

Suppose that we want to know if we can conclude that the population mean is greater than 50. Our hypotheses are

$$H_0 : \mu \leq 50, \quad H_1 : \mu > 50$$

If we want to know if we can conclude that the population mean is less than 50, the hypotheses are

$$H_0 : \mu \geq 50, \quad H_1 : \mu < 50$$

In summary, we may state the following rules of thumb for deciding what statement goes in the null hypothesis and what statement goes in the alternative hypothesis:

- What you hope or expect to be able to conclude as a result of the test usually should be placed in the alternative hypothesis.
- The null hypothesis should contain a statement of equality, either $=$, \leq , or \geq .
- The null hypothesis is the hypothesis that is tested.
- The null and alternative hypothesis are complementary. That is, the two together exhaust all possibilities regarding the value that the hypothesized parameter can assume.

A precaution It should be pointed out that neither hypothesis testing or statistical inference, in general, leads to the proof of a hypothesis; it merely indicates whether the hypothesis is supported or is not supported by the available data. When we fail to reject a null hypothesis, therefore, we do not say that it is true, but that it may be true. When we speak of accepting a null hypothesis, we have this limitation in mind and do not wish to convey the idea that accepting implies proof.

4. **Test statistic.** The test statistic is some statistic that may be computed from the data of the sample. As a rule, there are many possible values that the test statistic may assume, the particular value observed depending on the particular sample drawn. As we will see, the test statistic serves as a decision maker, since

the decision to reject or not to reject the null hypothesis depends on the magnitude of the test statistic. An example of a test statistic is the quantity

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

where μ_0 is a hypothesized value of a population mean. This test statistic is related to the statistic

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

with which we are already familiar.

General formula for test statistic

$$\text{test statistic} = \frac{\text{relevant statistic-hypothesized parameter}}{\text{standard error of the relevant statistic}}$$

5. **Distribution of test statistic.** It has been pointed out that the key to statistical inference is the sampling distribution. The distribution of the test statistic

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

for example, follows the standard normal distribution if the null hypothesis is true and the assumptions are met.

6. **Decision rule.** All possible values that the test statistic can assume are points on the horizontal axis of the graph of the distribution of the test statistic and are divided into two groups: one group constitutes what is known as the **rejection region** and the other group makes up the **nonrejection region**. The values of the test statistic forming the rejection region are those values that are less likely to occur if the null hypothesis is true, while the values making up the acceptance region are more likely to occur if the null hypothesis is true. *The decision rule tells us to reject the null hypothesis if the value of the test statistic that we compute from our sample is one of the values in the rejection region and to not reject the null hypothesis if the computed value of the test statistic is one of the values in the nonrejection region.*

Definition 2. The **rejection region** is the range of values of the test statistic for which H_0 is rejected.

Definition 3. The **nonrejection region** is the range of values of the test statistic for which H_0 is accepted.

Significance level The decision as to which values go into the rejection region and which ones go into the nonrejection region is made on the basis of the desired **level of significance**, designated by α . The term level of significance reflects the fact that hypothesis tests are sometimes called significance tests, and a computed value of the test statistic that falls in the rejection region is said to be **significant**. The level of significance α , specifies the area under the curve of the distribution of the test statistic that is above the values on the horizontal axis constituting the rejection region.

Definition 4. The level of significance α is the probability of rejecting a true null hypothesis.

Since to reject a true null hypothesis would constitute an error, it seems only reasonable that we should make the probability of rejecting a true null hypothesis small. We select a small value of α in order to make the probability of rejecting a true null hypothesis small. The more frequently encountered values of α are .01, .05, and .10.

If the results of a hypothesis test is to reject the null hypothesis based on the decision rule, we say that these results are **statistically significant**. However, a distinction should be made between scientific and statistical significance. The results of a study can be statistically significant but can still not be scientifically important. Conversely, some statistically nonsignificant results can be scientifically important, encouraging researchers to perform larger studies to confirm the direction of the findings and possibly reject H_0 with a larger sample size.

Types of errors The error committed when a true null hypothesis is rejected is called the **type I error**. The **type II error** is the error committed when a false null hypothesis is not rejected. The probability of committing a type II error is designated by β .

Definition 5. The probability of a type I error, α , is the significance level of a test.

Definition 6. The **power** of a test is defined as

$$1 - \beta = 1 - \text{probability of a type II error} = P[\text{rejecting } H_0 | H_1 \text{ is true}]$$

Whenever we reject a null hypothesis there is always the concomitant risk of committing a type I error, rejecting a true null hypothesis. Whenever we fail to reject a null hypothesis the risk of failing to reject a false null hypothesis is always present.

We make α small, but we generally exercise no control over β , although we know that in most practical situations it is larger than α .

We never know whether we have committed one of these errors when we reject or fail to reject a null hypothesis, since the true state of affairs is unknown. If the testing procedure leads to rejection of the null hypothesis, we can take comfort from the fact that we made α small and, therefore, the probability of committing a type I error was small. If we fail to reject the null hypothesis, we do not know the concurrent risk of committing a type II error, since β is usually unknown but, as has been pointed out, we do know that, in most practical situation, it is larger than α .

Figure 1 shows for various conditions of a hypothesis test the possible actions that an investigator may take and the conditions under which each of the two types of errors will be made.

		condition of null hypothesis	
		TRUE	FALSE
possible action	fail to reject H_0	correct action	type II error
	reject H_0	type I error	correct action

Figure 1: Conditions under which type I and type II errors may be committed.

Note that regardless of what is done, an error is possible. Anytime H_0 is rejected, a type I error might occur; anytime H_0 is not rejected, a type II error might occur. There is no way to avoid this dilemma. The job of the statistician is to design methods for deciding whether or not to reject H_0 that keep the probabilities of making either error reasonably small.

7. **Calculation of test statistic.** From the data contained in the sample we compute a value of the test statistic and compare it with the rejection and nonrejection regions that have already been specified.
8. **Statistical decision.** The statistical decision consists of rejecting or of not rejecting the null hypothesis. It is rejected if the computed value of the test statistic falls in the rejection region, and it is not rejected if the computed value of the test statistic falls in the nonrejection region.
9. **Conclusion.** If H_0 is rejected, we conclude that H_1 is true. If H_0 is not rejected, we conclude that H_0 may be true.

10. **p values.** The p value is a number that tells us how unusual our sample results are, given that the null hypothesis is true. A p value indicating that the sample results are not likely to have occurred, if the null hypothesis is true, provides justification for doubting the truth of the null hypothesis.

Definition 7. A p value is the probability that the computed value of a test statistic is at least as extreme as a specified value of the test statistic when the null hypothesis is true. Thus, the p value is the smallest value of α for which we can reject a null hypothesis.

We emphasize that when the null hypothesis is not rejected one should not say that the null hypothesis is accepted. We should say that null hypothesis is **not rejected**. We avoid using the word *accept* in this case because we may have committed a type II error. Since, frequently, the probability of committing a type II error can be quite high, we do not wish to commit ourselves to accepting the null hypothesis.

It is worth pointing out that the hypothesis testing procedure entails deciding on the level of significance α before the data are gathered and the test statistic is evaluated. That is, it involves presetting α . There are several reasons for wanting to do this. It gives a clear-cut way of making a decision. Once α is set, the critical region for the test is fixed also. If the observed value of the test statistic falls into this region, we reject H_0 ; otherwise, we do not. There is no room for debate after the data are gathered. Hence there can be no charge that the statisticians are manipulating the results to suit themselves. In addition, if the consequences of making a type I error are very serious, then by presetting α we are able to specify **before the fact** exactly how large a risk we are willing to tolerate.

2 Hypothesis testing: A single population mean

In this section, we consider the testing of a hypothesis about a population mean under three different conditions: (1) when sampling is from a normally distributed population of values with known variance; (2) when sampling is from a normally distributed population with unknown variance, and (3) when sampling is from a population that is not normally distributed. Although the theory for conditions 1 and 2 depends on normally distributed populations, it is common practice to make use of the theory when relevant populations are only approximately normally distributed. This is satisfactory as long as the departure from normality is not drastic.

2.1 Sampling from normally distributed populations: population variances known

As we did in the chapter Estimation, we again emphasize that situations in which the variable of interest is normally distributed with a known variance are rare.

Example 1. Researchers are interested in the mean age of a certain population. Let us say that they are asking the following question: Can we conclude that the mean age of this population is different from 30 years? To answer this question, a random sample of 10 individuals were withdrawn from the population of interest and their ages were taken. The sample mean $\bar{x} = 27$. Assume that the population has a known variance of $\sigma^2 = 20$.

Solution:

To answer a hypothesis testing problem, we usually follow the following ten steps.

1. **Data.** The data available to us is the sample mean of the random sample of 10 individuals.
2. **Assumption.** It is assumed that the sample comes from a population whose ages are approximately normally distributed. The population variance is assumed to be 20.
3. **Hypothesis.** We are identifying with the alternative hypothesis the conclusion the researchers wish to reach, so that if the data permit rejection of the null hypothesis, the researchers' conclusion will carry more weight, since the accompanying probability of rejecting a true null hypothesis will be small. Our hypotheses are:

$$H_0 : \mu = 30, \quad H_1 : \mu \neq 30$$

4. **Test statistic.** Since we are testing a hypothesis about a population mean, since we assume that the population is normally distributed, and since the population variance is known, our test statistic is given as

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

5. **Distribution of test statistic.** Based on our knowledge of sampling distributions and the normal distribution, we know that the test statistic is normally distributed with a mean of 0 and a variance of 1, if H_0 is true. *There are many possible values of the test statistic that the present situation can generate; one for every possible sample of size 10 that can be drawn from the population. Since we draw only one sample, we have only one of these possible values on which to base a decision.*

6. Decision rule.

Now we need to ask ourselves what magnitude of the values of the test statistic will cause rejection of H_0 . If the null hypothesis is false, it may be so either because the population mean is less than 30 or because the population mean is greater than 30. Therefore, either sufficiently small values or sufficiently large values of the test statistic will cause rejection of the null hypothesis. We want these extreme values to constitute the rejection region. How extreme must a possible value of the test statistic be to qualify for the rejection region? The answer depends on the **level of significance** we choose, that is, the size of the probability of committing a type I error. Let us say that we want the probability of rejecting a true null hypothesis to be $\alpha = .05$. Since our rejection region is to consist of two parts, sufficiently small values and sufficiently large values of the test statistic, part of α will have to be associated with the large values and part with the small values. It seems reasonable that we should divide α equally and let $\alpha/2 = 0.025$ be associated with small values and $\alpha/2 = 0.025$ be associated with large values.

Critical values of test statistic For the standard normal distribution, we know that the area between -1.96 and 1.96 is .95. Thus, we may state the decision rule for this test as follows: *reject H_0 if the computed value of the test statistic is either ≥ 1.96 or ≤ -1.96 .*

7. Calculation of test statistic.

From our sample we compute

$$z = \frac{27 - 30}{\sqrt{20/10}} = -2.12$$

8. Statistical decision.

Abiding by the decision rule, we are able to reject the null hypothesis since -2.12 is in the rejection region. We can say that the computed value of the test statistic is significant at the .05 level.

9. Conclusion.

We conclude that μ is not equal to 30.

10. p values.

Instead of saying that an observed value of the test statistic is significant or is not significant, most writers in the research literature prefer to report the exact probability of getting a value as extreme as or more extreme than that observed if the null hypothesis is true.

In the present instance these writers would give the computed value of the test statistic along with the statement $p = .0340$. This statement means that the probability of getting a value as extreme as 2.12 in either direction, when the null hypothesis is true, is .0340. That is, when H_0 is true, $P[z \geq 2.12] = .0170$ and $P[z \leq -2.12] = .0170$.

The R command to obtain the probability of $P[z \leq -2.12]$ is `pnorm(-2.12, mean=0, sd=1)`.

Recall that the p value for a test may be defined also as the smallest value of α for which the null hypothesis can be rejected. In this example, our p value is .0340, we know that we could have chosen an α value as small as .0340 and still have rejected the null hypothesis. If we had chosen an α smaller than .0340, we would not have been able to reject the null hypothesis. A general rule worth remembering, then, is this: *if $p \leq \alpha$, we reject the null hypothesis; if $p \geq \alpha$, we do not reject the null hypothesis.*

Testing H_0 by means of a confidence interval In Example 1, we used a hypothesis testing procedure to test $H_0 : \mu = 30$ against the alternative $H_1 : \mu \neq 30$. We were able to reject H_0 because the computed value of the test statistic fell in the rejection region. Let us see how we might have arrived at this same conclusion by using a $100(1 - \alpha)$ percent confidence interval. The 95 percent confidence interval for μ is

$$\begin{aligned} &27 \pm 1.96\sqrt{20/10} \\ &27 \pm 1.96(1.414) \\ &27 \pm 2.7714 \\ &[24.2286, 29.7714] \end{aligned}$$

Since this interval does not include 30, we say 30 is not a candidate for the mean we are estimating, and therefore, μ is not equal to 30 and H_0 is rejected. This is the same conclusion reached by means of the hypothesis testing procedure.

If the hypothesized parameter, 30, had been within the 95 percent confidence interval, we would have said that H_0 is not rejected at the .05 level of significance. In general, *when testing a null hypothesis by means of a two-sided confidence interval, we reject H_0 at the α level of significance if the hypothesized parameter is not contained within the $100(1 - \alpha)$ percent confidence interval. If the hypothesized parameter is contained within the interval, H_0 cannot be rejected at the α level of significance.*

One-sided hypothesis tests The hypothesis test in Example 1 is an example of a **two-sided test**, so called because the rejection region is split between the two sides or tails of the distribution of the test statistic. A hypothesis test may be **one-sided**, in which case all the rejection region is in one or the other tail of the distribution. Whether a one-sided or a two-sided test is used depends on the nature of the question being asked by the researcher.

If both large and small values will cause rejection of the null hypothesis, a two-sided test is indicated. When either sufficiently small values only or sufficiently large values

only will cause rejection of the null hypothesis, a one-sided test is indicated.

In summary, a hypothesis on μ can take one of three general forms. With μ_0 denoting the null value of the mean, these are:

- $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$, two-tailed test
- $H_0 : \mu \leq \mu_0$, $H_1 : \mu > \mu_0$, right-tailed test
- $H_0 : \mu \geq \mu_0$, $H_1 : \mu < \mu_0$, left-tailed test

It is worth pointing out that $\mu = \mu_0, \mu \neq \mu_0, \mu \leq \mu_0, \mu \geq \mu_0$ hold in a statistical sense, that is:

1. $\mu = \mu_0$: μ and μ_0 will not significantly differ, i.e. μ will be equal to μ_0 within the limits of random variability.
2. $\mu \neq \mu_0$: μ and μ_0 will significantly differ (there will be a difference between μ and μ_0 that goes beyond what could be expected on the basis of mere random variability).
3. $\mu \geq \mu_0$: μ will be significantly greater than μ_0 (μ will be greater than μ_0 in a degree that goes beyond what could be expected on the basis of mere random variability).
4. $\mu \leq \mu_0$: μ will be significantly smaller than μ_0 (μ will be smaller than μ_0 in a degree that goes beyond what could be expected on the basis of mere random variability).

There is one general statement to keep in mind when you test a hypothesis on any parameter: *To test a hypothesis on a parameter θ , you must find a statistic whose probability distribution is known at least approximately under the assumption that $\theta = \theta_0$.* This statistic will serve as a test statistic.

Example 2. Refer to Example 1. Suppose, instead of asking if they could conclude that $\mu \neq 30$, the researchers had asked: can we conclude that $\mu < 30$? To this question, we would reply that they can so conclude if they can reject the null hypothesis that $\mu \geq 30$.

Solution: Let us go through the ten-step procedure to reach a decision based on a one-sided test.

1. **Data.**

2. **Assumptions.**

3. **Hypothesis.**

$$H_0 : \mu \geq 30, \quad H_1 : \mu < 30$$

The inequality in the null hypothesis implies that the null hypothesis consists of an infinite number of hypotheses. The test will be made only at the point of equality, since it can be shown that if H_0 is rejected when the test is made at the point of equality, it would be rejected if the test were done for any other value of μ indicated in the null hypothesis.

4. **Test statistic.**

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

5. **Distribution of test statistic.**

6. **Decision rule.** Let us again use $\alpha = .05$. To determine where to place the rejection region, let us ask ourselves what magnitude of values would cause rejection of the null hypothesis. If we look at the hypothesis, we see that sufficiently small values would cause rejection and that large values would tend to reinforce the null hypothesis. We will want our rejection region to be where the small values are – at the lower tail of the distribution. This time, since we have a one-sided test, all of α will go in the one tail of the distribution. We can find the value of z to the left of which lies .05 of the area under the standard normal curve is -1.645. The R command to get z is `qnorm(.05, mean=0, sd=1)`. Our decision rule tells us to reject H_0 if the computed value of the test statistic is ≤ -1.645 .

7. **Calculation of test statistic.**

$$z = \frac{27 - 30}{\sqrt{20/10}}$$

8. **Statistical decision.** We are able to reject the null hypothesis since $-2.12 < -1.645$

9. **Conclusion.** We conclude that the population mean is smaller than 30 and act accordingly.

10. **p values.** The p value for this test is .0170 since $P[z \leq -2.12] = .0170$.

If the researcher's question has been, "can we conclude that the mean is greater than 30?," following the above ten-step procedure would have led to a one-sided test with all the rejection region at the upper tail of the distribution of the test statistic and a critical value of +1.645.

Next, we summarize the three cases of hypothesis test when the test statistic is

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

- The following hypothesis testing

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0$$

is a two-tailed hypothesis test. The regions for rejecting and accepting the null hypothesis are

$$\text{Rejection:} \quad Z > Z_{1-\alpha/2} \text{ or } Z < Z_{\alpha/2}$$

$$\text{Acceptance:} \quad Z_{\alpha/2} < Z < Z_{1-\alpha/2}$$

if the required significance level of the test is α . The corresponding probability of rejecting the null hypothesis when the null hypothesis is true, i.e. the probability of committing a type I error, is

$$\begin{aligned} P[\text{reject } H_0 | H_0 \text{ true}] &= P[Z > Z_{1-\alpha/2} \text{ or } Z < Z_{\alpha/2} | H_0 \text{ true}] \\ &= P[Z > Z_{1-\alpha/2} | H_0 \text{ true}] + P[Z < Z_{\alpha/2} | H_0 \text{ true}] = \alpha \end{aligned}$$

The p value can be calculated by

$$p \text{ value} = \begin{cases} 2 \times (1 - P[Z \leq z]) & \text{if } z > 0 \\ 2 \times P[Z \leq z] & \text{if } z < 0 \end{cases}$$

where z is the calculated test statistic.

- The following hypothesis test

$$H_0 : \mu \geq \mu_0, \quad H_1 : \mu < \mu_0$$

is a left-tailed hypothesis test. The regions for rejecting and accepting the null hypothesis are

$$\text{Rejection:} \quad Z < Z_{\alpha}$$

$$\text{Acceptance:} \quad Z > Z_{\alpha}$$

if the required significance level of the test is α . The corresponding probability of rejecting the null hypothesis when the null hypothesis is true, i.e. the probability of committing a type I error, is

$$P[\text{reject } H_0 | H_0 \text{ true}] = P[Z < Z_{\alpha} | H_0 \text{ true}] = \alpha \quad (1)$$

The p value can be calculated by

$$p \text{ value} = P[Z \leq z]$$

- The following hypothesis test

$$H_0 : \mu \leq \mu_0, \quad H_1 : \mu > \mu_0$$

is a right-tailed hypothesis test. The regions for rejecting and accepting the null hypothesis are

$$\begin{array}{ll} \text{Rejection:} & Z > Z_{1-\alpha} \\ \text{Acceptance:} & Z < Z_{1-\alpha} \end{array}$$

if the required significance level of the test is α . The corresponding probability of rejecting the null hypothesis when the null hypothesis is true, i.e. the probability of committing a type I error, is

$$P[\text{reject } H_0 | H_0 \text{ true}] = P[Z > Z_{1-\alpha} | H_0 \text{ true}] = \alpha \quad (2)$$

The p value can be calculated by

$$p \text{ value} = P[Z \geq z]$$

Tests of hypotheses on μ are actually conducted by testing $H_0 : \mu = \mu_0$ against one of the alternatives $\mu > \mu_0$, $\mu < \mu_0$, or $\mu \neq \mu_0$. This is because values of the test statistic that lead us to reject μ_0 and to conclude that $\mu > \mu_0$ will also lead us to reject any value less than μ_0 ; values of the test statistic that lead us to reject μ_0 and to conclude that $\mu < \mu_0$ will also lead us to reject any value greater than μ_0 . For this reason, many statisticians prefer to express the three forms of hypotheses as

- $H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0$, two-tailed test
- $H_0 : \mu = \mu_0, \quad H_1 : \mu > \mu_0$, right-tailed test
- $H_0 : \mu = \mu_0, \quad H_1 : \mu < \mu_0$, left-tailed test

This emphasizes the fact that when performing a hypothesis test on μ , α and p value are computed assuming that $\mu = \mu_0$.

2.2 Sampling from a normally distributed population: population variance unknown

As we have already noted, the population variance is usually unknown in actual situations involving statistical inference about a population mean. When sampling is from

an approximately normal population with an unknown variance, the test statistic for testing $H_0 : \mu = \mu_0$ is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

which, when H_0 is true, is distributed as Student's t with $n - 1$ degrees of freedom.

Example 3. Subjects with medical collateral ligament (MCL) and anterior cruciate ligament (ACL) tears. Between February 1995 and December 1997, 17 consecutive patients with combined acute ACL and grade III MCL injuries were treated by the same physician at the research center. One of the variables of interest was the length of time in days between the occurrence of the injury and the first magnetic resonance imaging (MRI). The number of days for the 17 subjects are:

14, 9, 18, 26, 12, 0, 10, 4, 8, 21, 28, 24, 24, 2, 3, 14, 9

We wish to know if we can conclude that the mean number of days between injury and initial MRI is not 15 days in a population presumed to be represented by these sample data.

Solution: We will be able to conclude that the mean number of days for the population is not 15 if we can reject the null hypothesis that the population mean is equal to 15.

1. **Data.**

2. **Assumptions.** The 17 subjects constitute a simple random sample from a population of similar subjects. We assume that the number of days until MRI in this population is approximately normally distributed.

3. **Hypothesis.**

$$H_0 : \mu = 15, \quad H_1 : \mu \neq 15$$

4. **Test statistic.**

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

5. **Distribution of test statistic.** Student's t with $n - 1 = 16$ degrees of freedom if H_0 is true.

6. **Decision rule.** Let $\alpha = .05$. Since we have a two-sided test, we put $\alpha/2 = .025$ in each tail of the distribution of our test statistic. The t values to the right and left of which .025 of the area lies are 2.1199 and -2.1199. These values can be obtained by the R command `qt(.025, df=16)`. Thus, if the computed test statistic $t \geq 2.1199$ or $t \leq -2.1199$, we will reject H_0 .

7. **Calculation of test statistic.** From our sample data we compute a sample mean of 13.2941 and a sample standard deviation of 8.88654. Thus, our test statistic is

$$t = \frac{13.2941 - 15}{8.88654/\sqrt{17}} = -.791$$

8. **Statistical decision.** Do not reject H_0 since $-.791$ falls in the nonrejection region.

9. **Conclusion.** Our conclusion, based on these data, is that the mean of the population from which the sample came may be 15.

10. **p value.**

$$\begin{aligned} P[t \leq -.791 \cap t \geq .791] &= P[t \leq -.791] + P[t \geq -.791] \\ &= 2P[t \leq -.791] = 2(.22) = .44 \end{aligned}$$

where .22 can be obtained using the R command `pt(-.791, df=16)`

If in this example, the hypothesis had been

$$H_0 : \mu \geq 15, \quad H_1 : \mu < 15$$

the testing procedure would have led to a one-sided test with all the rejection region at the lower tail of the distribution, and if the hypothesis had been

$$H_0 : \mu \leq 15, \quad H_1 : \mu > 15$$

we would have had a one-sided test with all the rejection region at the upper tail of the distribution.

Next, we summarize the three cases of hypothesis test when the test statistic is

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

- The hypothesis testing

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0$$

is a two-tailed hypothesis test. The regions for rejecting and accepting the null hypothesis are

$$\begin{aligned} \text{Rejection:} & \quad t_{n-1} > t_{(n-1, 1-\alpha/2)} \text{ or } t_{n-1} < t_{(n-1, \alpha/2)} \\ \text{Acceptance:} & \quad t_{(n-1, \alpha/2)} < t_{n-1} < t_{(n-1, 1-\alpha/2)} \end{aligned}$$

if the required significance level of the test is α . The corresponding probability of rejecting the null hypothesis when the null hypothesis is true, i.e. the probability of committing a type I error, is

$$\begin{aligned} P[\text{reject } H_0 | H_0 \text{ true}] &= P[t_{n-1} > t_{(n-1, 1-\alpha/2)} \text{ or } t_{n-1} < t_{(n-1, \alpha/2)} | H_0 \text{ true}] \\ &= P[t_{n-1} > t_{(n-1, 1-\alpha/2)} | H_0 \text{ true}] + P[t_{n-1} < t_{(n-1, \alpha/2)} | H_0 \text{ true}] = \alpha \end{aligned}$$

The p value can be calculated by

$$p \text{ value} = \begin{cases} 2 \times (1 - P[t_{n-1} \leq t_0]) & \text{if } t_0 > 0 \\ 2 \times P[t_{n-1} \leq t_0] & \text{if } t_0 < 0 \end{cases}$$

where t_0 is the computed test statistic.

- The following hypothesis test

$$H_0 : \mu \geq \mu_0, \quad H_1 : \mu < \mu_0$$

is a left-tailed hypothesis test. The regions for rejecting and accepting the null hypothesis are

$$\begin{aligned} \text{Rejection:} & \quad t_{n-1} < t_{(n-1, \alpha)} \\ \text{Acceptance:} & \quad t_{n-1} > t_{(n-1, \alpha)} \end{aligned}$$

if the required significance level of the test is α . The corresponding probability of rejecting the null hypothesis when the null hypothesis is true, i.e. the probability of committing a type I error, is

$$P[\text{reject } H_0 | H_0 \text{ true}] = P[t_{n-1} < t_{(n-1, \alpha)} | H_0 \text{ true}] = \alpha \quad (3)$$

The p value can be calculated by

$$p \text{ value} = P[t_{n-1} \leq t_0]$$

- The following hypothesis test

$$H_0 : \mu \leq \mu_0, \quad H_1 : \mu > \mu_0$$

is a right-tailed hypothesis test. The regions for rejecting and accepting the null hypothesis are

$$\begin{aligned} \text{Rejection:} & \quad t_{n-1} > t_{(n-1, 1-\alpha)} \\ \text{Acceptance:} & \quad t_{n-1} < t_{(n-1, 1-\alpha)} \end{aligned}$$

if the required significance level of the test is α . The corresponding probability of rejecting the null hypothesis when the null hypothesis is true, i.e. the probability of committing a type I error, is

$$P[\text{reject } H_0 | H_0 \text{ true}] = P[t_{n-1} > t_{(n-1, 1-\alpha)} | H_0 \text{ true}] = \alpha \quad (4)$$

The p value can be calculated by

$$p \text{ value} = P[t_{n-1} \geq t_0]$$

2.3 Sampling from a population that is not normally distributed

If, as is frequently the case, the sample on which we base our hypothesis test about a population mean comes from a population that is not normally distributed, we may, if our sample is large (≥ 30), take advantage of the central limit theorem and use

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

as the test statistic. If the population standard deviation is not known, the usual practice is to use the sample standard deviation as an estimate. The test statistic for testing $H_0 : \mu = \mu_0$, then, is

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

which, when H_0 is true, is distributed approximately as the standard normal distribution if n is large. The rationale for using s to replace σ is that the large sample, necessary for the central limit theorem to apply, will yield a sample standard deviation that closely approximates σ .

Example 4. A study was conducted to determine how symptom recognition and perception influence clinical presentation as a function of race. Researchers characterized symptoms and care-seeking behavior in African-American patients with chest pain seen in the emergency department. One of the presenting vital signs was systolic blood pressure. Among 157 African-American men, the mean systolic blood pressure was 146 mm Hg with a standard deviation of 27. We wish to know if, on the basis of these data, we may conclude that the mean systolic blood pressure for a population of African-American men is greater than 140.

Solution:

1. **Data.**

2. **Assumptions.** The data constitute a simple random sample from a population of African-American men who report to an emergency department with symptoms similar to those in the sample. We are unwilling to assume that systolic blood pressure values are normally distributed in such a population.

3. **Hypotheses.**

$$H_0 : \mu \leq 140, \quad H_1 : \mu > 140$$

4. **Test statistic.**

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

5. **Distribution of test statistic.** Because of the central limit theorem, the test statistic is approximately normally distributed with $\mu = 0$ if H_0 is true.

6. **Decision rule.** Let $\alpha = .05$. The critical value of the test statistic is 1.645.

7. **Calculation of test statistic.**

$$z = \frac{146 - 140}{27/\sqrt{157}} = 2.78$$

8. **Statistical decision.** Reject H_0 since $2.78 > 1.645$.

9. **Conclusion.** Conclude that the mean systolic blood pressure for the sampled population is ≥ 140 .

10. **p value.** The p value is $P[z \geq 2.78] = 1 - P[z \leq 2.78] = .0027$.

2.4 The power of a test

The power of a test tells us how likely it is that a statistically significant difference will be detected based on a finite sample size n , if the alternative hypothesis is true - that is, if the true mean μ differs from the mean under the null hypothesis (μ_0). If the power is too low, then there is little chance of finding a significant difference and nonsignificant results are likely even if real differences exist between the true mean μ of the group being studied and the null mean μ_0 . An inadequate sample size is usually the cause of low power to detect a scientifically meaningful difference [2].

2.4.1 One-sided test

We consider the three cases when test statistic z is used.

- Left-tailed test: $H_0 : \mu = \mu_0, H_1 : \mu = \mu_1 < \mu_0$

$$\begin{aligned}
 \text{power} &= 1 - \beta \\
 &= P[\text{reject } H_0 | H_0 \text{ false}] \\
 &= P[Z < Z_\alpha | \mu = \mu_1] \\
 &= P\left[\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < Z_\alpha | \mu = \mu_1\right] \\
 &= P\left[\bar{X} < \mu_0 + Z_\alpha \frac{\sigma}{\sqrt{n}} | \mu = \mu_1\right]
 \end{aligned}$$

We know that under $H_1, \bar{X} \sim N(\mu_1, \sigma^2/n)$. Hence, on standardization of limits,

$$\begin{aligned}
 \text{power} &= \Phi\left(\frac{\mu_0 + Z_\alpha \sigma/\sqrt{n} - \mu_1}{\sigma/\sqrt{n}}\right) \\
 &= \Phi\left(Z_\alpha + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}\right)
 \end{aligned}$$

- Right-tailed test: $H_0 : \mu = \mu_0, H_1 : \mu = \mu_1 > \mu_0$

$$\begin{aligned}
 \text{power} &= P\left[\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > Z_{1-\alpha} | \mu = \mu_1\right] \\
 &= 1 - P\left[\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < Z_{1-\alpha} | \mu = \mu_1\right] \\
 &= 1 - \Phi\left(\frac{\mu_0 + Z_{1-\alpha} \sigma/\sqrt{n} - \mu_1}{\sigma/\sqrt{n}}\right) \\
 &= 1 - \Phi\left(Z_{1-\alpha} + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}\right)
 \end{aligned}$$

Using the relationship $\Phi(-x) = 1 - \Phi(x)$ and $Z_\alpha = -Z_{1-\alpha}$, this expression can be rewritten as

$$\text{power} = \Phi\left(-Z_{1-\alpha} - \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}\right) = \Phi\left(Z_\alpha + \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right) \text{ if } \mu_1 > \mu_0$$

- One-tailed alternative test: $H_0 : \mu = \mu_0, H_1 : \mu = \mu_1$

$$\text{power} = \Phi\left(Z_\alpha + \frac{|\mu_0 - \mu_1|}{\sigma/\sqrt{n}}\right) = \Phi\left(-Z_{1-\alpha} + \frac{|\mu_0 - \mu_1|}{\sigma/\sqrt{n}}\right)$$

Example 5. A current area of research interest is the familial aggregation of cardiovascular risk factors in general and lipid levels in particular. Suppose the average cholesterol level in children is 175 mg/dL. A group of men who have died from heart disease within the past year are identified, and the cholesterol levels of their offspring are measured. Suppose the mean cholesterol level of 10 children whose fathers died from heart disease is 200 mg/dL and the sample standard deviation is 50 mg/dL. Test the hypothesis that the mean cholesterol level is higher in this group than in the general population with a 5% level of significance. Compute the power of the test for the cholesterol data with an alternative mean of 190 mg/dL, a null mean of 175 mg/dL, and a standard deviation of 50 mg/dL.

Solution: The hypotheses are

$$H_0 : \mu = 175, \quad H_1 : \mu > 175$$

The test statistic is

$$t = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{200 - 175}{50/\sqrt{10}} = 1.58$$

The critical value is $t_{9,.95} = 1.833$. Because $1.833 > 1.5$, it follows that we accept H_0 at the 5% level of significance.

We can also use the p -value method to arrive at the same conclusion. The p -value can be computed as:

$$p = P[t_9 > 1.58] = 1 - P[t_9 \leq 1.58] = .074$$

Because $p > .05$, we conclude that our results are not statistically significant and the null hypothesis is accepted.

We have $\mu_0 = 175, \mu_1 = 190, \alpha = .05, \sigma = 50, n = 10$. Thus

$$\begin{aligned} \text{power} &= \Phi\left(1.645 + \frac{190 - 175}{50/\sqrt{10}}\right) \\ &= \Phi(-0.696) = 1 - \Phi(0.696) = .243 \end{aligned}$$

Therefore, the chance of finding a significant difference in this case is only 24%.

Example 6. Suppose we want to test the hypothesis that mothers with low socioeconomic status (SES) deliver babies whose birth weights are lower than normal. To test this hypothesis, a list is obtained of birth weights from 100 consecutive, full-term, live-born deliveries from the maternity ward of a hospital in a low-SES area. The mean birth weights (\bar{x}) is found to be 115 oz with a sample standard deviation (s) of 24 oz. Suppose we know from nationwide surveys based on millions of deliveries that the mean

birth weight in the United States is 120 oz. Can we actually say the underlying mean birth weight from this hospital is lower than the national average with a significance level $\alpha = .05$? Compute the power of the test for the birth weight data with an alternative mean of 115 oz, assuming the true standard deviation = 24 oz.

Solution: The hypotheses are

$$H_0 : \mu = \mu_0 = 120, \quad H_1 : \mu = \mu_1 < 120$$

We compute the test statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{115 - 120}{24/\sqrt{100}} = -2.08$$

For $\alpha = .05$, the critical value = $t_{99,.05} = -1.66$. Because $-2.08 < -1.66$, we can reject H_0 at the significance level of .05. Since the sample size is quite large, we can use test statistic z instead of t .

The power of the test is

$$\begin{aligned} \text{power} &= \Phi \left(z_{.05} + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} \right) \\ &= \Phi \left(-1.645 + \frac{120 - 115}{24/\sqrt{100}} \right) = \Phi(.438) = .669 \end{aligned}$$

Therefore, there is about a 67% chance of detecting a significant difference using a 5% significance level with this sample size.

Example 7. A new drug in the class of calcium-channel blockers is to be tested for the treatment of patients with unstable angina, a severe type of angina. The effect of this drug will have on heart rate is unknown. Suppose 20 patients are to be studied and the change in heart rate after 48 hours is known to have a standard deviation of 10 beats per minute. What power would such a study have of detecting a significant difference in heart rate over 48 hours if it is hypothesized that the true mean change in heart rate from baseline to 48 hours could be either a mean increase or a decrease of 5 beats per minute?

Solution: We have $\sigma = 10$, $|\mu_0 - \mu_1| = 5$, $\alpha = .05$, $n = 20$ and

$$\text{power} = \Phi \left(-z_{1-.05/2} + \frac{5}{10\sqrt{20}} \right) = \Phi(-1.96 + 2.236) = .61$$

Thus this study would have a 61% chance of detecting a significant difference.

Factors affecting the power

1. If the significance level is made smaller (α decreases), Z_{alpha} decreases and hence the power decreases.
2. If the alternative mean is shifted away from the null mean ($|\mu_0 - \mu_1|$ decreases), then the power increases.
3. If the standard deviation of the distribution of individual observations increases (σ increases), then the power decreases.
4. If the sample size increases (n increases), then the power increases.

2.4.2 Two-sided test

Two-tailed test: $H_0 : \mu = \mu_0$, $H_1 : \mu = \mu_1 \neq \mu_0$. Assume the underlying distribution is normal and the population variance is known, the power is given exactly by

$$\begin{aligned}
 \text{power} &= P \left[\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < z_{\alpha/2} \mid \mu = \mu_1 \right] + P \left[\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha/2} \mid \mu = \mu_1 \right] \\
 &= P \left[\bar{X} < \mu_0 + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \mid \mu = \mu_1 \right] + P \left[\bar{X} > \mu_0 + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \mid \mu = \mu_1 \right] \\
 &= \Phi \left(\frac{\mu_0 + z_{\alpha/2} \sigma / \sqrt{n} - \mu_1}{\sigma / \sqrt{n}} \right) + 1 - \Phi \left(\frac{\mu_0 + z_{1-\alpha/2} \sigma / \sqrt{n} - \mu_1}{\sigma / \sqrt{n}} \right) \\
 &= \Phi \left(-z_{1-\alpha/2} + \frac{\mu_0 - \mu_1}{\sigma / \sqrt{n}} \right) + \Phi \left(-z_{1-\alpha/2} + \frac{\mu_1 - \mu_0}{\sigma / \sqrt{n}} \right)
 \end{aligned}$$

This equation is more tedious to use than is usually necessary. Specifically, if $\mu_1 < \mu_0$, then the second term is usually negligible relative to the first term. However, if $\mu_1 > \mu_0$, then the first term is usually negligible relative to the second term. Therefore, we can approximate the power formula as follows:

$$\text{power} \approx \Phi \left(-z_{1-\alpha/2} + \frac{|\mu_0 - \mu_1|}{\sigma / \sqrt{n}} \right)$$

2.5 Sample size determination

2.5.1 One-sided test

For planning purposes, we frequently need some idea of an appropriate sample size for investigation before a study actually begins. One possible result of making these calculations is finding out that the appropriate sample size is far beyond the financial means of the investigator(s) and thus abandoning the proposed investigation. Obviously, reaching this conclusion before a study starts is much better than after it is in progress [2].

The problem of determining sample size can be summarized as follows: Given that a one-sided significance test will be conducted at level α and that the true alternative mean is expected to be μ_1 , what sample size is needed to be able to detect a significant difference with probability $1 - \beta$?

For a left-sided test, H_0 will be rejected if $\bar{x} < \mu_0 + z_\alpha\sigma/\sqrt{n}$. Hence, the area to the left of $\mu_0 + z_\alpha\sigma/\sqrt{n}$ under the rightmost curve is α . However, we also want the area to the left of $\mu_0 + z_\alpha\sigma/\sqrt{n}$ under the leftmost curve, which represents the power, to be $1 - \beta$. These requirements will be met if n is made sufficiently large, because the variance of each curve σ^2/n will decrease as n increases and thus the curves will separate. We know that the power formula is

$$\text{power} = \Phi\left(z_\alpha + \frac{|\mu_0 - \mu_1|}{\sigma/\sqrt{n}}\right) = 1 - \beta$$

We want to solve for n in terms of $\alpha, \beta, |\mu_0 - \mu_1|$, and σ . To accomplish this, recall that

$$\Phi(z_{1-\beta}) = 1 - \beta$$

and therefore

$$z_\alpha + \frac{|\mu_0 - \mu_1|}{\sigma/\sqrt{n}} = z_{1-\beta}$$

Subtracting z_α from both sides of the equation and multiply by $\sigma/|\mu_0 - \mu_1|$ to obtain

$$\sqrt{n} = \frac{(z_\alpha + z_{1-\beta})\sigma}{|\mu_0 - \mu_1|}$$

Replace $-z_\alpha$ by $z_{1-\alpha}$, and square both sides of the equation to obtain

$$n = \frac{(z_{1-\alpha} + z_{1-\beta})^2\sigma^2}{(\mu_0 - \mu_1)^2}$$

Similarly, if we were to test the hypotheses

$$H_0 : \mu = \mu_0, \quad H_1 : \mu = \mu_1 > \mu_0$$

using a significance level of α and a power of $1 - \beta$, the same sample-size formula would hold. This procedure can be summarized as follows:

Theorem 1. Suppose we wish to test

$$H_0 : \mu = \mu_0, \quad H_1 : \mu = \mu_1$$

where the data are normally distributed with mean μ and known variance σ^2 . The **sample size** needed to conduct a one-sided test with significance level α and probability of detecting a significant difference = $1 - \beta$ is

$$n = \frac{(z_{1-\alpha} + z_{1-\beta})^2\sigma^2}{(\mu_0 - \mu_1)^2}$$

Notice that the sample size is very sensitive to the alternative mean chosen. The sample size is inversely proportional to $(\mu_0 - \mu_1)^2$. Thus, if the absolute value of the distance between the null and alternative means is halved, then the sample size needed is 4 times as large.

Example 8. Consider the birth weight example. Suppose that $\mu_0 = 120$ oz, $\mu_1 = 115$ oz, $\sigma = 24$, $\alpha = .05$, $1 - \beta = .80$, and we use a one-sided test. Compute the appropriate sample size needed to conduct the test.

Factors affecting the sample size

1. The sample size increases as σ^2 increases.
2. The sample size increases as the significance level is made smaller.
3. The sample size increases as the required power increases.
4. The sample size decreases as the absolute value of the distance between the null and alternative means $|\mu_0 - \mu_1|$ increases.

2.5.2 Two-sided test

Suppose we wish to test $H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_1$, where the data are normally distributed with mean μ and known variance σ^2 . The sample size needed to conduct a two-sided test with significance level α and power $1 - \beta$ is

$$n = \frac{\sigma^2(z_{1-\beta} + z_{1-\alpha/2})^2}{(\mu_0 - \mu_1)^2}$$

2.6 The relationship between hypothesis testing and confidence intervals

Theorem 2 (The relationship between hypothesis testing and confidence intervals (two-sided case)). Suppose we are testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. H_0 is rejected with a two-sided level α test if and only if the two-sided $100(1 - \alpha)$ percent confidence interval for μ **does not** contain μ_0 . H_0 is accepted with a two-sided level α test if and only if the two-sided $100(1 - \alpha)$ percent confidence interval for μ **does** contain μ_0 .

Recall that the two-sided $100(1 - \alpha)$ percent confidence interval for μ is

$$\bar{x} \pm t_{n-1, 1-\alpha/2} \frac{s}{\sqrt{n}}$$

Suppose we reject H_0 at level α . Then either $t < -t_{n-1,1-\alpha/2}$ or $t > t_{n-1,1-\alpha/2}$. Suppose that

$$\begin{aligned} t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} &< -t_{n-1,1-\alpha/2} \\ \Rightarrow \bar{x} &< \mu_0 - t_{n-1,1-\alpha/2} \frac{s}{\sqrt{n}} \\ \Rightarrow \mu_0 &> \bar{x} + t_{n-1,1-\alpha/2} \frac{s}{\sqrt{n}} = l_2 \end{aligned}$$

Similarly, if $t > t_{n-1,1-\alpha/2}$, then

$$\begin{aligned} \bar{x} - \mu_0 &> t_{n-1,1-\alpha/2} \frac{s}{\sqrt{n}} \\ \Rightarrow \mu_0 &< \bar{x} - t_{n-1,1-\alpha/2} \frac{s}{\sqrt{n}} = l_1 \end{aligned}$$

Thus, if we reject H_0 at level α using a two-sided test, then either $\mu_0 < l_1$ or $\mu_0 > l_2$; that is, μ_0 fall outside the two-sided $100(1 - \alpha)$ percent confidence interval for μ . Similarly, it can be shown that if we accept H_0 at level α using a two-sided test, then μ_0 must fall within the two-sided $100(1 - \alpha)$ percent confidence interval for μ (or, $l_1 \leq \mu_0 \leq l_2$).

Hence, this relationship is the rationale for using confidence intervals in the Chapter of Estimation to decide on the reasonableness of specific values for the parameter μ . If any specific proposed value μ_0 did not fall in the two-sided $100(1 - \alpha)$ percent CI for μ , then we stated that it was an unlikely value for the parameter μ . Equivalently, we could have tested the hypothesis $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ and rejected H_0 at significance level α .

Here is another way of expressing this relationship:

Theorem 3. The two-sided $100(1 - \alpha)$ percent CI for μ contains all values μ_0 such that we accept H_0 using a two-sided test with significance level α , where the hypotheses are $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. Conversely, the $100(1 - \alpha)$ percent CI **does not** contain any value μ_0 for which we can reject H_0 , using a two-sided test with significance level α , where $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$.

Example 9. Suppose we want to compare fasting serum-cholesterol levels among recent immigrants to the United States with typical levels found in the general U.S. population. Suppose we assume cholesterol levels in women aged 21-40 in the U.S. are approximately normally distributed with mean 190 mg/dL. It is unknown whether cholesterol levels among recent immigrants are higher or lower than those in the general U.S. population. Let's assume that levels among recent female immigrants are normally distributed with unknown mean μ . Hence, we wish to test the null hypothesis $H_0 : \mu = \mu_0$ versus the alternative hypothesis $H_1 : \mu \neq \mu_0$. Blood tests are performed on 100 female immigrants

aged 21-40 and the mean level \bar{x} is 181.52 mg/dL with standard deviation = 40 mg/dL. What can we conclude on the basis of this evidence.

We compute the test statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{181.52 - 190}{40/\sqrt{100}} = -2.12$$

For a two-sided test with $\alpha = .05$, the critical values are $t_{99,0.025} = -1.98$ and $t_{99,0.975} = 1.98$.

Because $t = -2.12 < -1.98$, it follows that we can reject H_0 at the 5% level of significance. We conclude that the mean cholesterol level of recent immigrants is significantly different from the mean for the general U.S. population.

Consider $\bar{x} = 181.52$ mg/dL, $s = 40$ mg/dL, and $n = 100$. The two-sided 95% confidence interval for μ is given by

$$\begin{aligned} & \left[\bar{x} - t_{99,0.975} \frac{s}{\sqrt{n}}, \bar{x} + t_{99,0.975} \frac{s}{\sqrt{n}} \right] \\ & = \left[181.52 - 1.984 \frac{40}{10}, 181.52 + 1.984 \frac{40}{10} \right] \\ & = [173.58, 189.46] \end{aligned}$$

This CI contains all values for μ_0 for which we accept $H_0 : \mu = \mu_0$ and does not contain any value μ_0 for which we could reject H_0 at the 5% level. Specifically, the 95% CI does not contain $\mu_0 = 190$, which corresponds to the decision from the hypothesis test where we were able to reject $H_0 : \mu = 190$ at the 5% level of significance.

3 Hypothesis testing: the difference between two population means

3.1 t-test for the significance between the means of two independent samples

Hypothesis testing involving the difference between two population means is more frequently employed to determine whether or not it is reasonable to conclude that the two population means are unequal. In such cases, one or the other of the following hypotheses may be formulated:

1. $H_0 : \mu_1 - \mu_2 = 0, \quad H_1 : \mu_1 - \mu_2 \neq 0$
2. $H_0 : \mu_1 - \mu_2 \geq 0, \quad H_1 : \mu_1 - \mu_2 < 0$

$$3. H_0 : \mu_1 - \mu_2 \leq 0, \quad H_1 : \mu_1 - \mu_2 > 0$$

As was done in the previous section, hypothesis testing involving the difference between two population means will be discussed in three different contexts: (1) when sampling is from normally distributed populations with known population variances, (2) when sampling is from normally distributed populations with unknown population variances, and (3) when sampling is from populations that are not normally distributed.

3.1.1 Sampling from normally distributed populations: population variances known

When each of the two independent simple random samples has been drawn from a normally distributed population with a known variance, the test statistic for testing the null hypothesis of equal population means is

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

where the subscript 0 indicates that the difference is a hypothesized parameter. When H_0 is true, the test statistic is distributed as the standard normal.

Example 10. Researchers wish to know if the data they have collected provide sufficient evidence to indicate a difference in mean serum uric acid levels between normal individuals and individuals with Down's syndrome. The data consist of serum uric acid readings on 12 individuals with Down's syndrome and 15 normal individuals. The means are $\bar{x}_1 = 4.5mg/100ml$ and $\bar{x}_2 = 3.4mg/100ml$. Assume that both the normal and the Down's syndrome populations are normally distributed with a variance equal to 1.5 for the former and 1 for the latter.

Solution:

1. **Data.**
2. **Assumptions.**
3. **Hypotheses.**

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 - \mu_2 \neq 0$$

An alternative way of stating the hypotheses is as follows:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

4. **Test statistic.**

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

5. **Distribution of test statistic.**

6. **Decision rule.** Let $\alpha = .05$. The critical value of z are ± 1.96 . Reject H_0 unless $-1.96 < z < 1.96$.

7. **Calculation of test statistic.**

$$z = \frac{(4.5 - 3.4) - 0}{\sqrt{1/12 + 1.5/15}} = \frac{1.1}{.4282} = 2.57$$

8. **Statistical decision.** Reject H_0 since $2.57 > 1.96$.

9. **Conclusion.** Conclude that, on the basis of the data, there is an indication the two population means are not equal.

10. **p value.** p value = $2 * P[z \leq -2.57] = .012$.

A 95 percent confidence interval for $\mu_1 - \mu_2$. In the previous chapter, the 95 percent confidence interval for $\mu_1 - \mu_2$, computed from the same data, was found to be $[-.26, 1.94]$. Since this interval does not include 0, we say that 0 is not a candidate for the difference between population means, and we conclude that the difference is not zero. Thus we arrive at the same conclusion by means of a confidence interval.

3.1.2 Sampling from normally distributed populations: population variances unknown

As we have learned, when the population variances are unknown, two possibilities exist. The two population variances may be equal or they may be unequal.

Population variances equal When the population variances are unknown, but assumed to be equal, we know that it is appropriate to pool the sample variances by means of the following formula:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

When each of two independent simple random samples has been drawn from a normally distributed population and the two populations have equal but unknown variances, the test statistic for testing $H_0 : \mu_1 = \mu_2$ is given by

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

which, when H_0 is true, is distributed as Student's t with $n_1 + n_2 - 2$ degrees of freedom.

Example 11. A study was conducted to investigate wheelchair maneuvering in individuals with lower-level spinal cord injury (SCI) and healthy controls (C). Subjects used a modified wheelchair to incorporate a rigid seat surface to facilitate the specified experimental measurements. Interface pressure measurement was recorded by using a high-resolution pressure-sensitive mat with a spatial resolution of four sensors per square centimeter taped on the rigid seat support. During static sitting conditions, average pressures were recorded under the ischial tuberosities (the bottom part of the pelvic bones). The data for measurements of the left ischial tuberosity (in mm Hg) for the SCI and control groups are shown in Table 1. We wish to know if we may conclude, on the basis of these data, that, in general, healthy subjects exhibit lower pressure than SCI subjects.

Table 1: Pressures (mm Hg) under the pelvis during static conditions.

Control	131	115	124	131	122	117	88	114	150	169
SCI	60	150	130	180	163	130	121	119	130	148

Solution:

1. **Data.**
2. **Assumptions.** The data constitute two independent simple random samples of pressure measurements, one sample from a population of control subjects and the other sample from a population with lower level of spinal cord injury. We shall assume that the pressure measurements in both populations are approximately normally distributed. The population variances are unknown but are assumed equal.
3. **Hypotheses.** $H_0 : \mu_C \geq \mu_{SCI}$, $H_1 : \mu_C < \mu_{SCI}$
4. **Test statistic.**

5. **Distribution of the test statistic.** When the null hypothesis is true, the test statistic follows Student's t distribution with $n_1 + n_2 - 2$ degrees of freedom.
6. **Decision rule.** Let $\alpha = .05$. The critical value of t is -1.7341 . Reject H_0 unless $t_{\text{computed}} > -1.7341$.
7. **Calculation of test statistic.** From the sample data, we compute

$$\bar{x}_C = 126.1, s_C = 21.8, \bar{x}_{SCI} = 133.1, s_{SCI} = 32.2$$

Next, we pool the sample variances to obtain

$$s_p^2 = \frac{9(21.8)^2 + 9(32.2)^2}{9 + 9} = 756.04$$

We now compute

$$t = \frac{(126.1 - 133.1) - 0}{\sqrt{\frac{756.04}{10} + \frac{756.04}{10}}} = -.569$$

8. **Statistical decision.** We fail to reject H_0 , since $-1.7341 < -.569$; that is, $-.569$ falls in the nonrejection region.
9. **Conclusion.** On the basis of these data, we cannot conclude that the population mean pressure is less for healthy subjects than for SCI subjects.
10. **p value.** The p value is $.288$.

Population variances unequal When two independent simple random samples have been drawn from normally distributed populations with unknown and unequal variances, the test statistic for testing $H_0 : \mu_1 = \mu_2$ is

$$t' = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The critical value of t' for an α level of significance and a two-sided test is approximately

$$t'_{1-(\alpha/2)} = \frac{w_1 t_1 + w_2 t_2}{w_1 w_2}$$

where $w_1 = s_1^2/n_1, w_2 = s_2^2/n_2, t_1 = t_{1-(\alpha/2)}$ for $n_1 - 1$ degrees of freedom, and $t_2 = t_{1-(\alpha/2)}$ for $n_2 - 1$ degrees of freedom. The critical value of t' for a one-sided test is

found by computing $t'_{1-\alpha}$, using $t_1 = t_{1-\alpha}$ for $n_1 - 1$ degrees of freedom and $t_2 = t_{1-\alpha}$ for $n_2 - 1$ degrees of freedom.

For a two-sided test, reject H_0 if the computed value of t' is either greater than or equal to the critical value or less than or equal to the negative of that value.

For a one-sided test with the rejection region in the right tail of the sampling distribution, reject H_0 if the computed t' is equal to or greater than the critical t' . For a one-sided test with a left-tail rejection region, reject H_0 if the computed value of t' is equal to or smaller than the negative of the critical t' .

Example 12. Researchers examined subjects with hypertension and healthy control subjects. One of the variables of interest was the aortic stiffness index. Measures of this variable were calculated from the aortic diameter evaluated by M-mode echocardiography and blood pressure measured by a sphygmomanometer. Generally, physicians wish to reduce aortic stiffness. In the 15 patients with hypertension (group 1), the mean aortic stiffness index was 19.16 with a standard deviation of 5.29. In the 30 control subjects (group 2), the mean aortic stiffness index was 9.53 with a standard deviation of 2.69. We wish to determine if the two populations represented by these samples differ with respect to mean aortic stiffness index.

Solution:

1. **Data.** The sample sizes, means, and sample standard deviations are:

$$n_1 = 15, \bar{x}_1 = 19.16, s_1 = 5.29$$

$$n_2 = 30, \bar{x}_2 = 9.53, s_2 = 2.69$$

2. **Assumptions.** The data constitute two independent random samples, one from a population of subjects with hypertension and the other from a control population. We assume that aortic stiffness values are approximately normally distributed in both populations. The population variances are unknown and unequal.

3. **Hypotheses.**

$$H_0 : \mu_1 - \mu_2 = 0, \quad H_1 : \mu_1 - \mu_2 \neq 0$$

4. **Test statistic.**

5. **Distribution of test statistic.** The statistic given by Equation 3.1.2 does NOT follow Student's t distribution. We, therefore, obtain its critical values by Equation 3.1.2.

6. **Decision rule.** Let $\alpha = .05$. Before computing t' , we calculate $w_1 = (5.29)^2/15 = 1.8656$ and $w_2 = (2.69)^2/30 = .2412$. $t_{14,.975} = 2.1448$ (R command `qt(.975,`

df=14)) and $t_{29,.975} = 2.0452$ (R command `qt(.975, df=29)`). By Equation 3.1.2, we compute

$$t' = \frac{1.8656(2.1448) + .2412(2.0452)}{1.8656 + .2412} = 2.133$$

Our decision rule, then, is reject H_0 if the computed t is either ≥ 2.133 or ≤ -2.133 .

7. **Calculation of test statistic.** By equation 3.1.2 we compute

$$t' = \frac{(19.16 - 9.53) - 0}{\sqrt{\frac{(5.29)^2}{15} + \frac{(2.69)^2}{30}}} = 6.63$$

8. **Statistical decision.** Since $6.63 > 2.133$, we reject H_0 .

9. **Conclusion.** On the basis of these results we conclude that the two population means are different.

10. **p value.**

3.1.3 Sampling from populations that are not normally distributed

When sampling is from populations that are not normally distributed, the results of the central limit theorem may be employed if sample sizes are large (say, ≥ 30). This will allow the use of normal theory since the distribution of the difference between sample means will be approximately normal. When each of two large independent simple random samples has been drawn from a population that is not normally distributed, the test statistic for testing $H_0 : \mu_1 = \mu_2$ is

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

which, when H_0 is true, follows the standard normal distribution. If the population variances are known, they are used; but if they are unknown, as is the usual case, the sample variances, which are necessarily based on large samples, are used as estimates. Sample variances are NOT pooled, since equality of population variances is not a necessary assumption when the z statistic is used.

3.2 t-test for the significance of the difference between the means of two correlated samples

References

- [1] Wayne W. Daniel. *Biostatistics: A foundation for analysis in the health sciences*. Wiley, 2009.
- [2] Bernard Rosner. *Fundamentals of biostatistics*. Thomson Brooks/Cole, 2006.