Chapter 10

Hypothesis Testing

We previously examined how the parameters for a probability distribution can be estimated using a random sample and maximum likelihood (Chapter 8), as then showed how confidence intervals provide a measure of the reliability of these estimates (Chapter 9). In hypothesis testing, the subject of this chapter, we examine the consistency of observed data sets with a null hypothesis, commonly a statement about the parameter values within a statistical model. We conduct a statistical test of this null hypothesis, with the result being a decision to accept or reject the null hypothesis based on the magnitude of a quantity called a P value. Small values of P indicate a test result inconsistent with the null hypothesis, suggesting it might be false and some alternative hypothesis more valid. In the following, we discuss the different components and steps of hypothesis testing.

10.1 The null and alternative hypotheses

As an example of hypothesis testing, suppose that we rear n tilapia on a commercial diet, and want to compare their body size with ones reared using a natural diet. Fish reared on natural food are already known to have a weight of 500 g at a certain age, and weight is normally distributed. We could test whether the fish reared on the commercial diet have the same mean weight as ones reared on natural food (500 g) using the **null hypothesis** that $\mu = 500$ g, where μ is the mean parameter for the normal distribution. This can be written as $H_0 : \mu = 500$ g, where H_0 stands for null hypothesis. Null hypotheses of this type can be written more generally as $H_0 : \mu = \mu_0$, where

μ_0 is the hypothesized mean of the distribution. For the tilapia problem, we would have $\mu_0 = 500$ g.

An **alternative hypothesis** for this example is that the mean weight of tilapia on commercial diet is different from 500 g. This can be written as $H_1 : \mu \neq 500$ g, where H_1 stands for the alternative hypothesis. Alternative hypotheses of this type are written generally as $H_1 : \mu \neq \mu_0$. We may also be interested in particular values of the alternative mean, such as $H_1 : \mu = 490$ g or $H_1 : \mu = 530$ g, or more generally $H_1 : \mu = \mu_1$.

10.2 Test statistics

A test statistic is a quantity that measures the consistency of the observed data with the null hypothesis. Test statistics are usually chosen so that large values occur when the data are inconsistent with H_0. What would be a suitable test statistic for the tilapia problem, using $H_0 : \mu = \mu_0$ as the null hypothesis? Suppose we rear n fish on the commercial diet, and then calculate the sample mean $\bar{Y}$ of their weights. The statistic $\bar{Y}$ is an estimator of the true mean μ for this statistical population, which may or may not be equal to the μ_0 under the null hypothesis. A value of $\bar{Y}$ substantially greater than μ_0, or smaller than μ_0, would be inconsistent with H_0. This suggests using the quantity $\bar{Y} - \mu_0$ as the test statistic for the problem. What about the other parameter for the normal distribution, σ^2 or σ? For simplicity, we will assume that it is a known quantity, although this is rare in practice. We could then use the test statistic

$$Z_s = \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} \tag{10.1}$$

to test $H_0 : \mu = \mu_0$ (Bickel & Doksum 1977). We use this statistic because it has a standard normal distribution under H_0 ($Z_s \sim N(0, 1)$, see Chapter 9) which makes it straightforward to employ the test. Note that Z_s becomes large (positive or negative) if the sample mean $\bar{Y}$ differs greatly from μ_0. In general, tests based on the standard normal distribution are called Z tests.

10.3 Acceptance and rejection regions – Type I error

Given a suitable test statistic, how large must it be before we decide the data are inconsistent with H_0? This is determined by finding an interval that defines an **acceptance region** for the test, and its complement, called the **rejection** or **critical region** (Bickel & Doksum 1977). We then accept H_0 if the test statistic falls within the acceptance region, and reject H_0 if it falls outside or lies on its boundary. The boundaries of the acceptance and rejection regions are determined by setting the probability of a Type I error. **A Type I error is defined as the test rejecting H_0 when H_0 is true. The probability of committing a Type I error is called the Type I error rate, usually denoted with the symbol α.** It is common practice to set $\alpha = 0.05$, meaning there is a 1 in 20 chance that the test will reject H_0 even when it is true. It follows that the probability of the test accepting H_0 if it is true is $1 - \alpha$. For $\alpha = 0.05$, we have $1 - \alpha = 1 - 0.05 = 0.95$.

The acceptance region is determined as follows. Suppose that $H_0 : \mu = \mu_0$ is true. Because the test statistic $Z_s \sim N(0, 1)$ under H_0, the following is a true statement:

$$P[-c_\alpha < Z_s < c_\alpha] = P[-c_\alpha < \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} < c_\alpha] = 1 - \alpha. \tag{10.2}$$

The quantity c_α would be chosen using Table Z to satisfy this equation (for details see Chapter 9). The interval $(-c_\alpha, c_\alpha)$ is the acceptance region of a test with a Type I error rate of α. Under H_0, the test statistic Z_s would lie within this interval with probability $1 - \alpha$ and outside this region with probability α, which is the required Type I error rate. The rejection region would be the complement of the acceptance region, i.e., all values on the boundary or outside of $(-c_\alpha, c_\alpha)$.

For example, with $\alpha = 0.05$ we find that $c_{0.05} = 1.96$, and so we would accept H_0 if Z_s lies within $(-1.96, 1.96)$ and reject H_0 if it lies outside this interval or exactly on the boundary (see Fig. 10.1). The acceptance region for this test can also be expressed using absolute values - we would accept H_0 if $|Z_s| < 1.96$ and reject it if $|Z_s| \geq 1.96$.

The acceptance region becomes larger (and the rejection region smaller) for smaller α values. For $\alpha = 0.01$, we find that $c_{0.01} = 2.576$ and so the acceptance region is $(-2.576, 2.576)$ (Fig. 10.2). Using absolute values, we

would accept H_0 if $|Z_s| < 2.576$ and reject it otherwise. Using a smaller value of α indicates we are more concerned about making a Type I error. For $\alpha = 0.01$ there is only a 1 in 100 chance we would reject H_0 if H_0 were true, but this also reduces the power of the test (see below) to detect whether H_0 is false.

The acceptance and rejection regions we just developed are for a **two-tailed test**, which tests the null hypothesis $H_0 : \mu = \mu_0$ with $H_1 : \mu \neq \mu_0$ the alternative hypothesis. This test statistic will reject H_0 for either large and small values of the test statistic Z_s, which occurs when $\bar{Y}$ is greater than μ_0 or less than μ_0. We will later examine the behavior of **one-tailed tests**, where the null is $H_0 : \mu = \mu_0$ while the alternative is of the form $H_1 : \mu > \mu_0$, or $H_1 : \mu < \mu_0$. Note that the two alternative hypotheses here specify that μ is either greater or less than μ_0. One-tailed tests are designed to reject H_0 in only one direction.

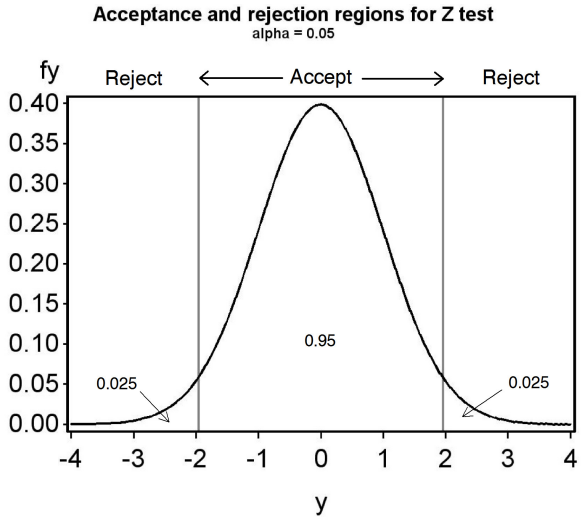


Figure 10.1: Acceptance and rejection regions for a one-sample Z test, $\alpha = 0.05$. Also shown is the distribution of Z_s under H_0.

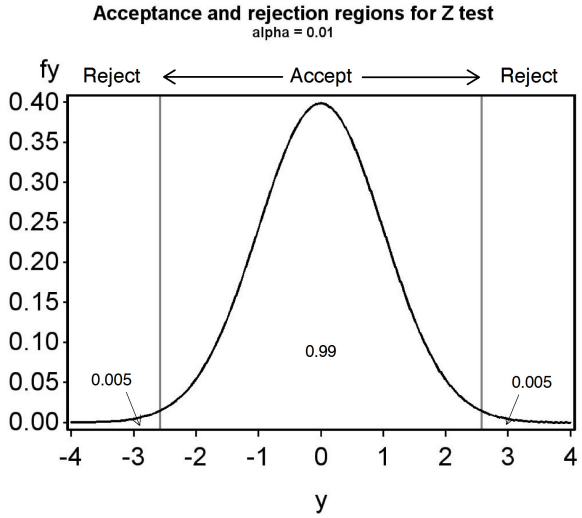


Figure 10.2: Acceptance and rejection regions for a one-sample Z test, $\alpha = 0.01$. Also shown is the distribution of Z_s under H_0.

10.3.1 One-sample Z test - sample calculation

We will now do an example of this test, known as a one-sample Z test. Recall the tilapia diet example, where it is known that fish reared on natural food have a mean weight of 500 g. We rear $n = 10$ fish on a commercial diet, and want to compare the weight of fish on the commercial diet with ones reared on natural food. In particular, we want to test $H_0 : \mu = 500$ g. We find that $\bar{Y} = 495$ g for the fish reared on the commercial diet, and already know that $\sigma^2 = 49$ g^2, so $\sigma = 7$ g. Because $\bar{Y} = 495$ g is less than 500 g, it already appears that the commercial diet produces smaller fish than natural food, but a statistical test is still needed to provide convincing evidence against H_0. For the test statistic, we have

$$Z_s = \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} = \frac{495 - 500}{7/\sqrt{10}} = \frac{-5}{2.214} = -2.258 \tag{10.3}$$

For a Type I error rate of $\alpha = 0.05$, the acceptance region for Z_s is $(-1.96, 1.96)$. $Z = -2.258$ lies outside this interval, so we would reject H_0 at the $\alpha = 0.05$ level. For $\alpha = 0.01$ the acceptance region is $(-2.576, 2.576)$. Because Z_s lies within this interval, we would accept H_0 at this α level. Thus, the decision to accept or reject H_0 depends on both the test statistic value and the value of α.

10.4 P values

As noted above, the value of α can affect whether we accept or reject H_0. Rather than force a particular α on the analyst, the test results can also be presented in the form of a P value. **A P value is defined as the smallest value of α for which one can just reject** H_0 (Bickel & Doksum 1977). It is calculated by finding an α such that the test statistic Z_s is equal to c_α.

Recall from Chapter 9 that c_α is defined so that the following equation is true:

$$P[Z < c_\alpha] = 1 - \alpha/2. \tag{10.4}$$

To find the P value for the tilapia example, we substitute the test statistic value Z_s for c_α in the above equation, ignoring the fact that Z_s is negative. We have

$$P[Z < Z_s] = P[Z < 2.258] = 1 - \alpha/2. \tag{10.5}$$

From Table Z, we see that $P[Z < 2.258] \approx 0.9881$. We then solve the equation

$$0.9881 = 1 - \alpha/2 \tag{10.6}$$

for α to obtain the P value. We have $\alpha = 2(1 - 0.9881) = 0.0238$. This is the P value for the test, reported as $P = 0.0238$. Given the P value, the analyst or other interested parties can decide for themselves whether to reject or accept H_0.

A P value can also be thought of as the probability of obtaining a test statistic equal to or more extreme than the observed one, under the null hypothesis. We can see this from a graph of the acceptance and rejection regions for the tilapia example, where $Z_s = -2.258$ and $P = 0.0238$ (Fig. 10.3). The probabilities outside the acceptance region correspond to $P[Z_s \leq -2.258]$ and $P[Z_s \geq 2.258]$, which are the probabilities of observing values of Z_s equal to or more extreme than the observed value of $Z_s = -2.258$. The two definitions of a P value are equivalent.

A P value is also a measure of the consistency of the observed data with the null hypothesis. If the P value is large, say $P > 0.05$, then the observed data generated a test statistic value that is fairly likely under the null hypothesis. On the other hand, if P is small then the observed data has generated a test statistic that is unlikely under the null hypothesis. This suggests the observed data are inconsistent with the null hypothesis, and the null may be false.

There are specific phrases generally used to describe the significance of a statistical test result. If a test yields $P \leq 0.05$, it is described as being **significant**, while if $P \leq 0.01$ it is **highly significantly**. If $P > 0.05$ the test is described as **nonsignificant**. The tilapia example with $P = 0.0238$ would be described as significant because $0.0238 < 0.05$, but not highly significant.

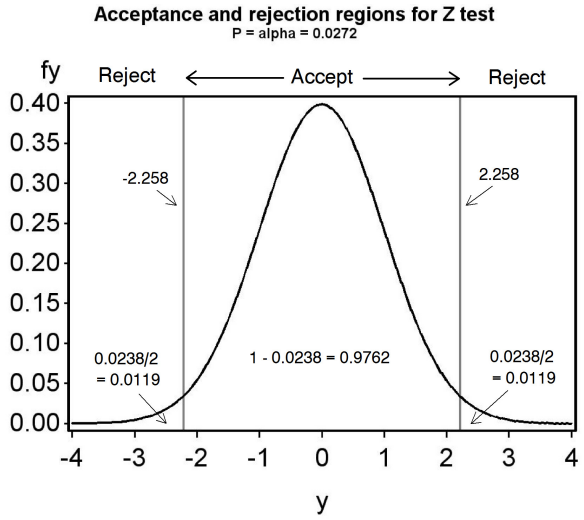


Figure 10.3: Acceptance-rejection region for a one-sample Z test, exact $P = 0.0238$

10.5 Type II error and power

Suppose now that H_0 is actually false and some alternative hypothesis H_1 is true. **A Type II error is defined as failing to reject H_0 when H_0 is false**. The probability of committing a Type II error is called the Type II error rate, usually denoted by the symbol β. It follows that the probability of the test rejecting H_0 if it is false is $1 - \beta$, and this quantity is called the **power** of the test (Bickel & Doksum 1977). High power values indicate the test is capable of detecting departures from the null hypothesis.

The power and Type II error rate of a statistical test depends on the sample size n of the test, the standard deviation of the observations σ, the Type I error rate α, and the particular alternative hypothesis chosen. An analyst interested in determining the power of a test will fix some of these values, often α and σ, and then examine how changes in n and the alternative hypothesis affect power. This procedure is called a **power analysis**. A power value of 0.8 is believed to be adequate in most situations (Cohen 1988). This implies that a statistical test will reject H_0 when it is false 80% of the time.

It is relatively easy to calculate the power for a one-sample Z test, using the distribution of Z_s under H_1. Suppose that we choose $\alpha = 0.05$, so that the acceptance region is the interval $(-1.96, 1.96)$, and that the alternative hypothesis is $H_1 : \mu = \mu_1$ for some μ_1. Under $H_0 : \mu = \mu_0$ the test statistic has a standard normal distribution, implying $Z_s \sim N(0, 1)$, but what is its distribution under H_1? Using the expected value and variance rules in Chapter 7, one can show that

$$E[Z_s] = \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} = \phi \tag{10.7}$$

and also that $Var[Z_s] = 1$. So, Z_s has the same variance under both H_1 and H_0, but the mean under H_1 is equal to ϕ, not zero as under H_0. It follows that under H_1 the test statistic $Z_s \sim N(\phi, 1)$. The probability of rejecting H_0 when H_1 is true, the power of the test, is the probability that Z_s lies outside the interval $(-1.96, 1.96)$, or

$$\text{power} = P[Z_s \leq -1.96] + P[Z_s \geq 1.96]. \tag{10.8}$$

The Type II error rate β can be calculated as $1-$power, or directly by finding

$$\beta = P[-1.96 < Z_s < 1.96] \tag{10.9}$$

when H_1 is true.

Fig. 10.4 shows the power and Type II error for the tilapia example with $H_0 : \mu = 500$ vs. a particular alternative hypothesis, $H_1 : \mu = 495$. We assume $\sigma = 7$ as before, with $n = 10$ and $\alpha = 0.05$. For this alternative hypothesis, we have

$$\phi = \frac{\mu_1 - \mu_0}{\sigma/\sqrt{10}} = \frac{(495 - 500)}{7/\sqrt{10}} = -2.26. \tag{10.10}$$

Thus, under H_1 we have $Z_s \sim N(-2.26, 1)$, and this distribution is shown as well as the distribution of Z_s under H_0 and the acceptance and rejection regions for the test. The power is the area Z_s under H_1 outside the acceptance region, while β is the area in the region.

What happens to the power as we vary μ_1? Suppose now that $H_1 : \mu_1 = 490$ is the alternative hypothesis. As we can see from Fig. 10.5, in this case the power is substantially higher and β is lower. Fig. 10.6 shows how power changes as we vary μ_1 across a range of values. Power is quite high (nearly 1) for μ_1 far from μ_0, but approaches a minimum value of α for μ_1 near μ_0. The minimum power is α, not zero, because the test will reject H_0 even if it is true ($\mu_1 = \mu_0$) at this rate.

Power is also affected by sample size. If we use $H_1 : \mu = 495$ and increase the sample size from $n = 10$ to $n = 20$, this also increases the power (Fig. 10.7). However, an increase in the standard deviation from $\sigma = 7$ to $\sigma = 10$ lowers the power (Fig. 10.8).

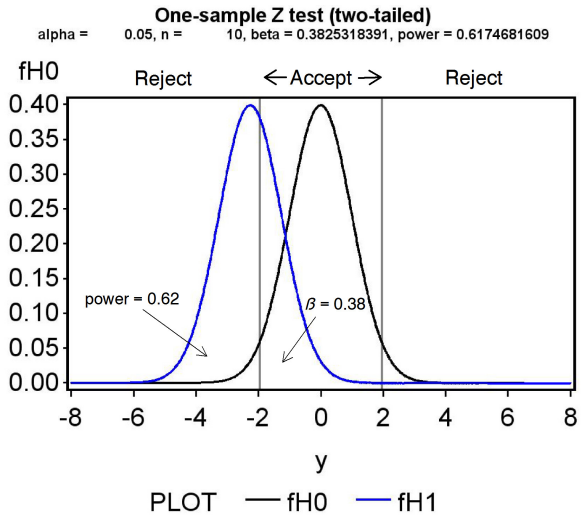


Figure 10.4: Distribution of Z_s under $H_1 : \mu = 495$, with $\sigma = 7, n = 10$ ($\phi = -2.26$). Almost all of the power occurs to the left of the acceptance region, but there is also a small amount to the right. Also shown is the distribution of Z_s under H_0.

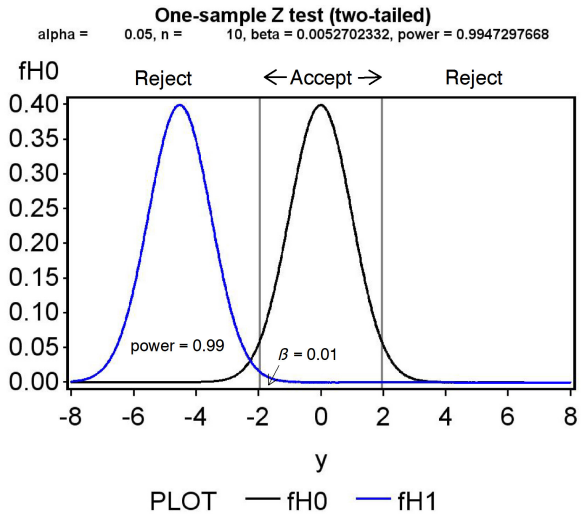


Figure 10.5: Distribution of Z_s under $H_1 : \mu = 490$, with $\sigma = 7, n = 10$ ($\phi = -4.52$).

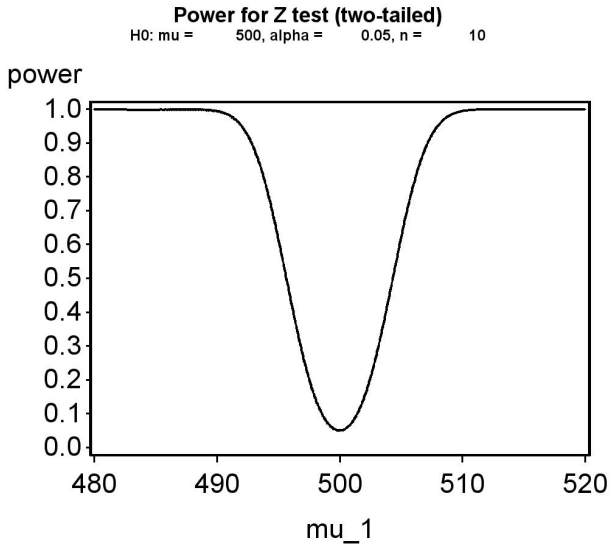


Figure 10.6: Power across a range of μ_1 values, for $H_0 : \mu = 500$, $\sigma = 7$, and $n = 10$

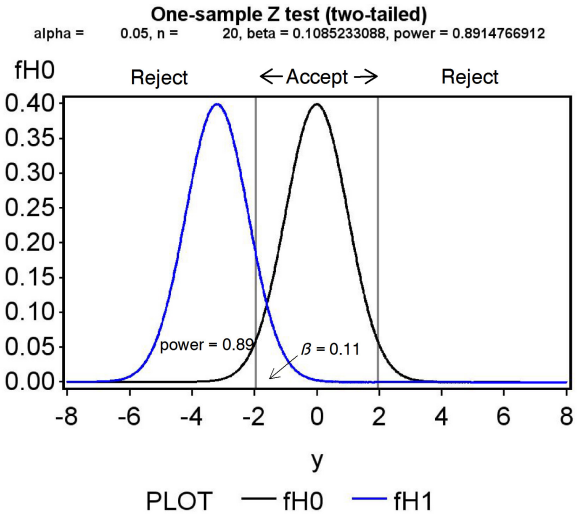


Figure 10.7: Distribution of Z_s under $H_1 : \mu = 495$, with $\sigma = 7, n = 20$ ($\phi = -3.19$).

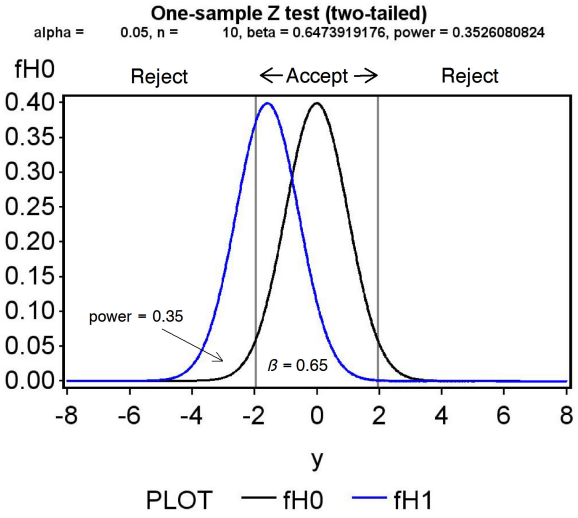


Figure 10.8: Distribution of Z_s under $H_1 : \mu = 495$, with $\sigma = 10, n = 10$ ($\phi = -1.58$).

Table 10.1: Effects on power and the Type II error rate β of changes in various parameters. The arrows indicate if a particular quantity increases or decreases.

Parameter	Direction	ϕ	power	β
$\lvert\mu_1 - \mu_0\rvert$	$\uparrow$	$\uparrow$	$\uparrow$	$\downarrow$
n	$\uparrow$	$\uparrow$	$\uparrow$	$\downarrow$
σ	$\uparrow$	$\downarrow$	$\downarrow$	$\uparrow$
α	$\uparrow$	no change	$\uparrow$	$\downarrow$

All of these effects on power can be understood through their influence on ϕ. Any change in a parameter value that makes ϕ larger increases power and reduces β, because it shifts the distribution of Z_s under H_1 away from the acceptance and into the rejection region. Thus, large differences between μ_1 and μ_0, large n, and small σ will all increase power because they increase ϕ. Conversely, close values of μ_1 and μ_0, small n, and large σ would all reduce power. Table 10.1 summarizes how the different parameter values influence ϕ, power, and the Type II error rate β. Also shown is the effect of the Type I error rate α on power. If an investigator can accept a larger value of α, so that Type I errors are more common, this reduces the acceptance and increases the rejection region size, and thus increases power.

Note that a sufficiently large value of n can generate a large value of ϕ, even when μ_1 and μ_0 are close or σ is large. Thus, large sample sizes can yield adequate power even when the data are noisy, or the two means are close in value. This basically arises from the inverse relationship between the variance of $\bar{Y}$ and n, i.e., $Var[\bar{Y}] = \sigma^2/n$, which is incorporated in the test statistic Z_s (see Eqn. 10.1).

10.6 Summary table

A common way of summarizing the different outcomes in hypothesis testing is the table below. The null hypothesis H_0 can be either true or false. If H_0 is true, then the test may accept H_0 and make a correct decision, or reject it and make a Type I error, with a Type I error rate of α. If H_0 is false, then the test may accept H_0 and make a Type II error with an error rate of β, or reject it and make a correct decision.

Table 10.2: Table summarizing the different outcomes in hypothesis testing, with the corresponding Type I (α) and Type II (β) error rates.

	Accept H_0	Reject H_0
H_0 true	Correct 1-α	Type I error α
H_0 false	Type II error β	Correct 1-β = power

10.7 One-sample t test

In the preceding, we used the test statistic Z_s to test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$, for the case where σ^2 or σ was known. Although this simplifies the statistics, in most cases we will need to estimate σ^2 and σ from the data using the sample variance s^2 and standard deviation s. We then use the test statistic

$$T_s = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} \tag{10.11}$$

to conduct the test (Bickel & Doksum 1977). T_s has a t distribution with $n-1$ degrees of freedom under H_0 (see Chapter 9). The following is therefore a true statement:

$$P[-c_{\alpha,n-1} < T_s < c_{\alpha,n-1}] = P[-c_{\alpha,n-1} < \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} < c_{\alpha,n-1}] = 1 - \alpha. \tag{10.12}$$

The quantity $c_{\alpha,n-1}$ would be chosen using Table T, using the entry for $2(1-p)$ corresponding to α and the appropriate degrees of freedom (see Chapter 9). The interval $(-c_{\alpha,n-1}, c_{\alpha,n-1})$ is the acceptance region of a test with a Type I error rate of α, while the rejection region is its complement.

For example, with $\alpha = 0.05$ and $n = 10$, we have $c_{0.05,9} = 2.262$. We would therefore accept H_0 if T_s lies within $(-2.262, 2.262)$, and reject it if T_s lies outside this interval (see Fig. 10.9). Using absolute values, we would accept H_0 if $|T_s| < 2.262$ and reject it otherwise. For $\alpha = 0.01$ and $n = 10$, we have $c_{0.01,9} = 3.250$, and would accept H_0 if T_s lies within $(-3.250, 3.250)$ and reject it otherwise.

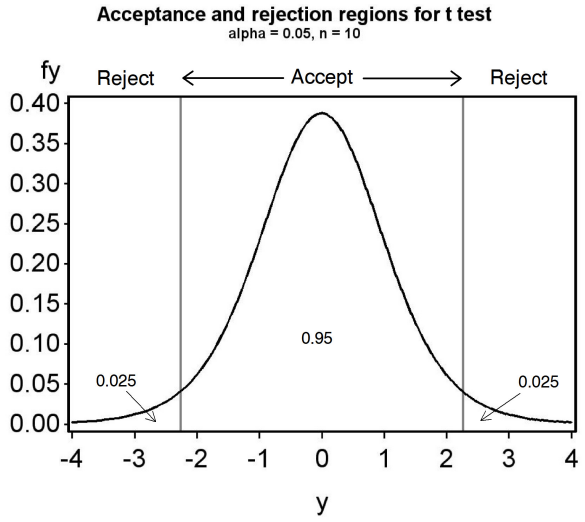


Figure 10.9: Acceptance and rejection regions for a one-sample t test, $\alpha = 0.05, n = 10$. The distribution shown is for the t distribution with $n - 1 = 9$ degrees of freedom.

10.7.1 One-sample t test - sample calculation

Recall the tilapia example, and suppose that $\bar{Y} = 493$ g and $s^2 = 48.2$ g^2, so that $s = 6.94$ g, with $n = 10$. We wish to test $H_0 : \mu = 500$ g vs. $H_1 : \mu \neq 500$ g. For the test statistic, we have

$$T_s = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} = \frac{493 - 500}{6.94/\sqrt{10}} = \frac{-7}{2.19} = -3.196 \tag{10.13}$$

For $\alpha = 0.05$, the acceptance region for T_s is $(-2.262, 2.262)$ with $n - 1 = 10 - 1 = 9$ degrees of freedom (Fig. 10.9). $T_s = -3.196$ lies outside this interval, so we would reject H_0 at the $\alpha = 0.05$ level. For $\alpha = 0.01$ the acceptance region is $(-3.250, 3.250)$. Because T_s lies within this interval, we would accept H_0 at this α level. We can also determine a P value for this test using Table T. The P value is found by scanning along the row in the table corresponding to 9 degrees of freedom, looking for two values that bracket T_s while ignoring its sign. We see that the values 2.821 and 3.250 bracket $T_s = -3.196$. Looking at the values for $2(1 - p)$, which correspond to α, this implies that $0.010 < P < 0.020$. This is the best accuracy that can be

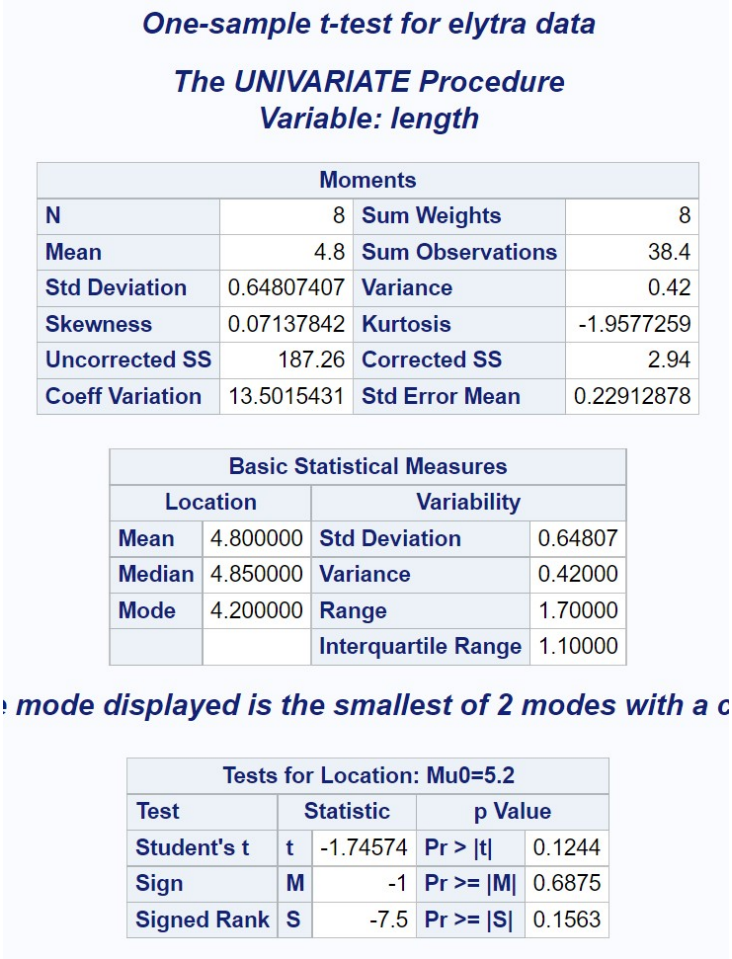
accomplished using Table T, and to obtain an exact P value would require the use of SAS.

10.7.2 Hypothesis testing - SAS demo

To illustrate hypothesis testing using SAS, we will use a subset ($n = 8$) of the elytra data for the insect predator *Thanasimus dubius* (see Chapter 3). These observations are from a study that used an artificial diet to rear the insects, and we would like to compare their size to wild individuals. Suppose that wild predators have an elytral length of 5.2 mm. This suggests testing $H_0 : \mu = 5.2$ mm vs. $H_1 : \mu \neq 5.2$ mm. We can conduct a one-sample t test for this null hypothesis using `proc univariate`, by adding the option `mu0=5.2` as an option. See SAS program and Fig. 10.10 below. The test statistic T_s and its P value are listed on one line at the bottom of the output. We see that $T_s \approx -1.75$ for this test. What is its P value? The notation Pr $>$ |t| in the output is shorthand for the $P[T_s < -1.75] + P[T_s > 1.75]$, the P value for this two-tailed test. We thus have $P = 0.1244$, a non-significant test result because $P > 0.05$. The degrees of freedom for the test are not reported by SAS, but are equal to $n - 1 = 8 - 1 = 7$. A sentence reporting this test result in a scientific journal would be something like 'A one-sample t test comparing the elytra length of individuals reared on artificial diet vs. wild individuals was non-significant ($t_7 = -1.746, P = 0.1244$).' Note that the degrees of freedom are reported as a subscript on the test statistic.

SAS Program

```
* one-sample_t_test.sas;
title 'One-sample t-test for elytra data';
data elytra;
    input sex $ length;
    datalines;
F   5.2
F   4.2
F   5.7
F   5.4
F   4.0
F   4.5
F   5.2
F   4.2
;
run;
* Generate t test and plots;
proc univariate mu0=5.2 data=elytra;
    var length;
    histogram length / vscale=count normal;
    qqplot length / normal;
run;
quit;
```



One-sample t-test for elytra data

The UNIVARIATE Procedure
Variable: length

Moments			
N	8	Sum Weights	8
Mean	4.8	Sum Observations	38.4
Std Deviation	0.64807407	Variance	0.42
Skewness	0.07137842	Kurtosis	-1.9577259
Uncorrected SS	187.26	Corrected SS	2.94
Coeff Variation	13.5015431	Std Error Mean	0.22912878

Basic Statistical Measures			
Location		Variability	
Mean	4.800000	Std Deviation	0.64807
Median	4.850000	Variance	0.42000
Mode	4.200000	Range	1.70000
		Interquartile Range	1.10000

mode displayed is the smallest of 2 modes with a c

Tests for Location: Mu0=5.2				
Test	Statistic		p Value	
Student's t	t	-1.74574	Pr > |t|	0.1244
Sign	M	-1	Pr >= |M|	0.6875
Signed Rank	S	-7.5	Pr >= |S|	0.1563

Figure 10.10: `one-sample_t_test.sas - proc univariate`

10.7.3 Power analysis for one-sample t tests - SAS demo

A power analysis can be used to determine an adequate sample size n for a one-sample t test, as well as many other statistical tests. To conduct a power analysis, you need to specify a null and alternative hypothesis, a Type I error rate α, and have some estimate of the standard deviation σ of the population in question. The analysis then calculates the power for a range of n values. **The idea is to choose a value of n that gives power close to 0.8, often regarded as an adequate level of power (Cohen 1988).** The power analysis for a one-sample t test involves the same quantity

$$\phi = \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} \tag{10.14}$$

as for the one-sample Z test, and its power is influenced by the same factors (see Table 10.1). The power calculation involves the **non-central t distribution** with a non-centrality parameter of ϕ. One subtle difference is that acceptance and rejection regions for the t test depends on n through the degrees of freedom, unlike the Z test. Larger values of n lead to smaller values of $c_{\alpha,n-1}$, shrinking the acceptance region and affecting the power calculation in this way.

Returning to the elytra example, suppose we want to test if the length of predators reared on an artificial diet differs from wild individuals, which have a length of 5.2 mm. This implies $H_0 : \mu = 5.2$ mm. For biological reasons, we are interested in detecting an decrease or increase in length of approximately 10% on the artificial diet, about 0.5 mm. This suggests an alternative hypothesis of the form $H_1 : \mu = 5.2 - 0.5 = 4.7$ mm (or $H_1 : \mu = 5.2 + 0.5 = 5.7$ mm). How many predators need to be reared on artificial diet to give a power of at least 0.8? Assume we already have an estimate of σ from another study, say $s = 0.6$ mm, and let $\alpha = 0.05$.

We can use `proc power` to find the sample size n that gives this power (SAS Institute Inc. 2018). See program plus Fig. 10.11 and 10.12 below. We first specify a one-sample t test using the `onesamplemeans` option, followed by values for μ under H_0 (`nullmean = 5.2`), σ (`stddev = 0.6`), and μ under H_1 (`mean = 4.7`). The default value of α is 0.05. We then specify a range of sample sizes (n) for which we want the power to be calculated, using the option `ntotal = 2 to 20 by 1`. This finds the power for $n = 2, 3, \ldots, 20$. The `power = .` option tells SAS solve for power (there are other possibilities, like finding n for a given power value). The option `plot x=n` generates a plot of

power vs. n. We see that a sample size of $n = 14$ gives power > 0.8 for this scenario.

While power increases rapidly for small sample sizes, there are diminishing returns once the power exceeds about 0.8. In other words, obtaining higher power values requires many more observations.

SAS Program

```
* One-sample_t_test_power2.sas;
title 'Power analysis for one-sample t test';
proc power;
    onesamplemeans
        nullmean = 5.2
        stddev = 0.6
        mean = 4.7
        ntotal = 2 to 20 by 1
        power = . ;
    plot x=n;
run;
quit;
```

Power analysis for one-sample t test

The POWER Procedure
One-Sample t Test for Mean

Fixed Scenario Elements	
Distribution	Normal
Method	Exact
Null Mean	5.2
Mean	4.7
Standard Deviation	0.6
Number of Sides	2
Alpha	0.05

Computed Power		
Index	N Total	Power
1	2	0.081
2	3	0.142
3	4	0.218
4	5	0.300
5	6	0.381
6	7	0.457
7	8	0.528
8	9	0.593
9	10	0.651
10	11	0.703
11	12	0.748
12	13	0.788
13	14	0.822
14	15	0.851
15	16	0.876
16	17	0.897
17	18	0.915
18	19	0.930
19	20	0.942

Figure 10.11: `one-sample_t_test_power2.sas - proc power`

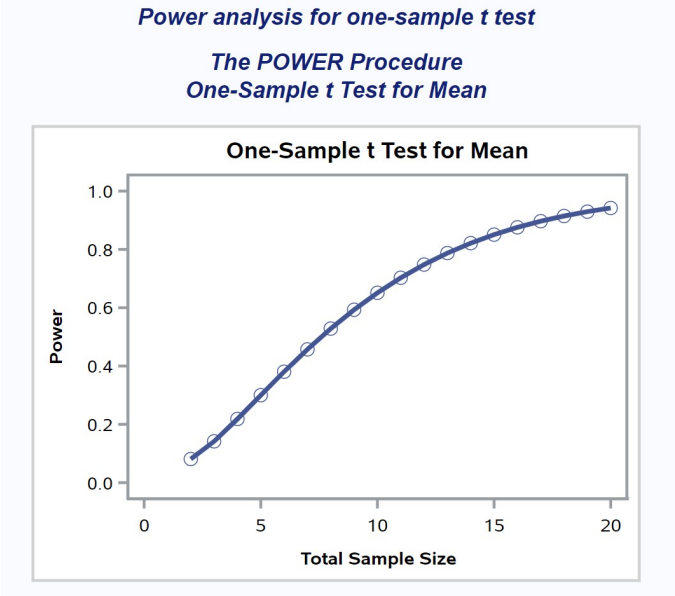


Figure 10.12: `one-sample_t_test_power2.sas - proc power`

10.8 One-tailed t test

The tests we have examined so far are known as two-tailed tests. They are called this because the test statistic Z_s or T_s can detect departures from $H_0 : \mu = \mu_0$ in both directions, for $H_1 : \mu > \mu_0$ and $H_1 : \mu < \mu_0$, although the alternative for these tests is usually written more compactly as $H_1 : \mu \neq \mu_0$. We will now examine one-tailed tests, which have the same null hypothesis but the alternative is one direction or the other.

Suppose we are interested in testing $H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$. We can use the same test statistic as before, namely

$$T_s = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}}. \tag{10.15}$$

If H_1 is true, we would expect to see $\bar{Y}$ values larger than μ_0, and so T_s would be positive. We would reject H_0 if T_s was sufficiently positive, with the acceptance and rejection regions determined as before by controlling the Type I error rate. Therefore, if the Type I error rate is α we want to determine a constant $c'_{\alpha,n-1}$ such that the following statement is true:

$$P[T_s < c'_{\alpha,n-1}] = 1 - \alpha \tag{10.16}$$

The quantity $c'_{\alpha,n-1}$ would be chosen using Table T, using the entry for p corresponding to $1 - \alpha$. We would accept H_0 if $T_s < c'_{\alpha,n-1}$ and reject it if $T_s \geq c'_{\alpha,n-1}$.

For example, with $\alpha = 0.05$ so that $p = 0.95$, and $n = 10$ (degrees of freedom $= n - 1 = 10 - 1 = 9$), we have $c'_{0.05,9} = 1.833$. We would therefore accept H_0 if $T_s < 1.833$ and reject it if $T_s \geq 1.833$ (see Fig. 10.13). For $\alpha = 0.01$ and $n = 10$, we have $c'_{0.01,9} = 2.822$, and would accept H_0 if $T_s < 2.822$ and reject it otherwise.

If we now wish to test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu < \mu_0$, we would use the same test statistic as above. However, if H_1 is true we would expect $\bar{Y}$ to be smaller than μ_0, and so T_s would be negative. To determine the acceptance and rejection regions we would find $c'_{\alpha,n-1}$ in the same way as above, except we would use its negative. We would accept H_0 if $T_s > -c'_{\alpha,n-1}$ and reject it if $T_s \leq -c'_{\alpha,n-1}$. For example, if $\alpha = 0.05$ and $n = 10$, we would accept H_0 if $T_s > -1.833$ and reject it if $T_s \leq -1.833$ (Fig. 10.14). For $\alpha = 0.01$, we would accept H_0 if $T_s > -2.822$ and reject it otherwise.

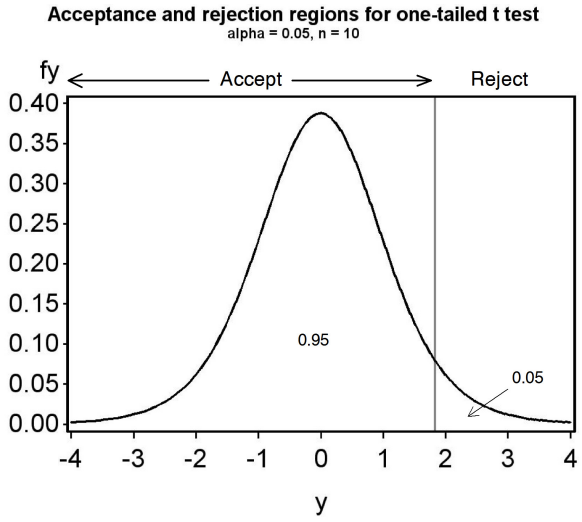


Figure 10.13: Acceptance and rejection regions for one-tailed t test, $H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$, for $\alpha = 0.05$ and $n = 10$.

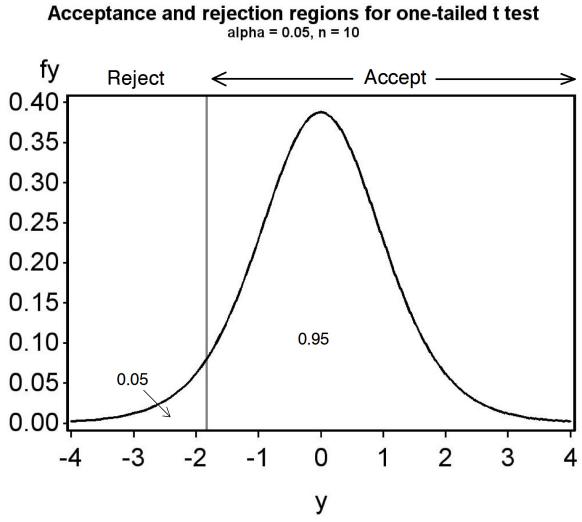


Figure 10.14: Acceptance and rejection regions for a one-tailed t test, $H_0 : \mu = \mu_0$ vs. $H_1 : \mu < \mu_0$, for $\alpha = 0.05$ and $n = 10$.

10.8.1 One-tailed t test - sample calculation

Recall the tilapia example, with $\bar{Y} = 493$ g, $s^2 = 48.2$ g^2, $s = 6.94$ g, and $n = 10$. Suppose we are only interested in detecting diets that produce fish of lower weight than natural food, implying we wish to test $H_0 : \mu = 500$ g vs. $H_1 : \mu < 500$ g. The test statistic value is again

$$T_s = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} = \frac{493 - 500}{6.94/\sqrt{10}} = \frac{-7}{2.19} = -3.196 \tag{10.17}$$

For $\alpha = 0.05$ and $n - 1 = 10 - 1 = 9$ degrees of freedom, we have $-c'_{0.05,9} = -1.833$. Because $T_s = -3.916 < -1.833$, we would reject H_0 at the $\alpha = 0.05$ level. For $\alpha = 0.01$, we have $-c'_{0.01,9} = -2.821$, and again $T_s = -3.196 < -2.821$. Thus, we can also reject H_0 at the $\alpha = 0.01$ level. We could continue this process with successively smaller values of α by scanning the row corresponding to 9 degrees of freedom in Table T, but cannot reject H_0 for smaller ones. Therefore, we have $P < 0.01$ for this test.

Suppose we had wanted to test $H_0 : \mu = 500$ g vs. $H_1 : \mu > 500$ g using the same data and test statistic value, namely $T_s = -3.196$. The scenario here could be that we want a commercial diet that actually increases the weight of tilapia over natural food, and are not interested in ones that yield lower weights. In this case, for $\alpha = 0.05$ we would not reject H_0, because $T_s = -3.196 < 1.833$. The test was non-significant, with $P > 0.05$.

10.8.2 One-tailed t test - SAS demo

Recall the elytra length example, where we tested $H_0 : \mu = 5.2$ mm vs. $H_1 : \mu \neq 5.2$ mm using SAS (Fig. 10.10). While there is no option for one-tailed tests in `proc univariate`, we can reinterpret the output and so derive a P value for a one-tailed test.

Suppose we want to test $H_0 : \mu = 5.2$ mm vs. $H_1 : \mu < 5.2$ mm. This implies we want to test whether predators reared on artificial diet are smaller than those reared on natural food, which have a length of 5.2 mm. This would be reasonable if we want to detect diets that are deficient in some manner. If H_1 were true we would expect to see a negative value of T_s, because $\bar{Y}$ would likely be smaller than μ_0. This is what occurred in the SAS output, because $\bar{Y} = 4.8 < 5.2$ mm and $T_s = -1.75$. The one-tailed P value in this case is simply half the two-tailed P value, or P(one-tailed) = P(two-tailed)/2= 0.1244/2 = 0.0622. This is because the two-tailed test

gives the P value for both tails (see Fig. 10.9), but for this one-tailed test we only need the probability for the left tail of the t distribution (Fig. 10.14).

Now suppose we want to test $H_0 : \mu = 5.2$ mm vs. $H_1 : \mu > 5.2$ mm. This implies we want to test whether predators reared on artificial diet are larger than those reared on natural food. If H_1 were true we would expect to see a positive value of T_s, because $\bar{Y}$ would likely be greater than μ_0. This is not what occurred in the SAS output, because $\bar{Y} = 4.8 < 5.2$ mm and $T_s = -1.75$. The P value should therefore be large in this case, and in fact the one-tailed P value is $1 - P(\text{two-tailed})/2 = 1 - 0.1244/2 = 0.9378$. This is the probability for the right tail of the t distribution, which is large because T_s is negative.

We can distill the above procedures to a simple rule that will convert the SAS two-tailed P value to the appropriate one-tailed one. Assume $H_0 : \mu = \mu_0$ is the null hypothesis. **If the test statistic favors the alternative hypothesis, then the one-tailed P value is P(two-tailed)/2, otherwise it is $1 - P$(two-tailed)/2.** For example, if we have $H_1 : \mu > \mu_0$ and $T_s > 0$, the test statistic favors H_1 and the P value is P(two-tailed)/2. This procedure also works for tests calculated by hand. You first find the P value for the two-tailed test, then convert it to a one-tailed P value using the same rule.

10.8.3 One-tailed tests - a warning

As discussed above, the P value for a one-tailed test may sometimes be half the two-tailed P value. This makes it tempting to employ a one-tailed test after a two-tailed test yields a nonsignificant result. However, the proper procedure is to determine whether a one-tailed alternative hypothesis and test is appropriate for the situation **before** conducting the test. For example, artificial diets for insects are unlikely to yield larger insects than natural diets, and so it seems reasonable to use an alternative hypothesis of the form $H_1 : \mu < \mu_0$, where μ_0 is the size of insects reared on natural foods. This choice of an alternative hypothesis can be justified based on prior knowledge of the system.

10.9 Confidence intervals and tests

Confidence intervals are typically used as measures of the accuracy or reliability of parameter estimates, but can also be used for hypothesis testing. Why might you do this? There are cases where the statistical software only provides confidence intervals for a parameter, but a test can still be developed using these intervals. Also, a publication may only provide confidence intervals for a parameter, but the reader can still conduct a test if required using these intervals. Some statisticians argue that this makes confidence intervals more useful than hypothesis testing, because they also provide information on the magnitude of a population parameter, and how reliably it is estimated (see Yaccoz 1991).

We will now demonstrate how a confidence interval for μ is equivalent to a one-sample t test. Recall that a $100(1-\alpha)\%$ confidence interval for μ has the form

$$\left(\bar{Y} - c_{\alpha,n-1}\frac{s}{\sqrt{n}}, \bar{Y} + c_{\alpha,n-1}\frac{s}{\sqrt{n}}\right) \tag{10.18}$$

(see Chapter 9). Suppose that we want to test $H_0 : \mu = \mu_0$. If we accept H_0 when this confidence interval includes μ_0, and reject it if the interval does not include μ_0, this is an α level test of H_0, equivalent to running a one-sample t test.

To see this connection, note that we would accept H_0 if μ_0 was inside this interval, or

$$\bar{Y} - c_{\alpha,n-1}\frac{s}{\sqrt{n}} < \mu_0 < \bar{Y} + c_{\alpha,n-1}\frac{s}{\sqrt{n}}. \tag{10.19}$$

Rearranging these inequalities, we see it is equivalent to saying

$$-c_{\alpha,n-1} < \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} < c_{\alpha,n-1}, \tag{10.20}$$

or

$$-c_{\alpha,n-1} < T_s < c_{\alpha,n-1}, \tag{10.21}$$

where T_s is the test statistic for a one-sample t test. We would reject H_0 if T_s falls outside this interval. Note that this acceptance region is exactly the same as for the t test with Type I error rate of α, which is of the form $(-c_{\alpha,n-1}, c_{\alpha,n-1})$. Thus, the test based on a $100(1-\alpha)\%$ confidence interval is equivalent to an α level test. In particular, a 95% confidence interval is equivalent to an $\alpha = 0.05$ test.

Conversely, it is often possible to reverse this process and obtain a confidence interval from a statistical test. The procedure is called 'inverting the test' (Bickel & Doksum 1977).

10.10 Likelihood ratio tests

We saw earlier how statisticians use the concept of maximum likelihood to estimate population parameters (Chapter 8). The maximum likelihood method begins by constructing a likelihood function based on the distribution of the data (Poisson, normal, etc.) and the observed data. We then maximize the likelihood as a function of the parameters of the distribution (μ, σ^2, etc). The values of the parameters that maximize the likelihood are the maximum likelihood estimates of the parameters. The likelihood function is not a fixed quantity but instead varies with the observed data, so that different data sets yield different estimates of the population parameters. Maximum likelihood estimators have desirable statistical properties and in many cases yield estimators that seem reasonable (like using $\bar{Y}$ to estimate μ).

Likelihood methods can also be used to develop statistical tests called **likelihood ratio tests**. These tests also have desirable statistical properties and in many cases are identical to classical statistical tests. Likelihood methods thus provide a theoretical framework for many statistical problems, including parameter estimation, confidence intervals, and hypothesis testing. The main drawback of these methods is that one must be willing to specify the distribution of the data, be it Poisson, binomial, normal, or more exotic distributions.

10.10.1 Example of a likelihood ratio test

We will now develop a likelihood ratio test that leads to the familiar one-sample t test (Mood et al. 1974) . We suppose that the data are normally distributed and we wish to test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$. A random sample with n observations has been obtained.

We can think of H_0 and H_1 as two different statistical models for the data. Under H_0, the data are assumed to be normally distributed with $\mu = \mu_0$, but can have any value of σ^2 because this parameter is left unspecified. Under H_1, the data are permitted to have any value of μ and σ^2.

The first step in constructing a likelihood ratio test is to find the maximum likelihood estimates of the parameters for each of these two statistical models. We have already dealt with this problem for the model specified by H_1 – this is just maximum likelihood estimation of μ and σ^2 for the normal distribution. The same methods can be used to estimate σ^2 under H_0, but we will not go into the details.

This process can be illustrated by plotting the likelihood function as a function of μ and σ^2. To make things more concrete, we show the likelihood function for a data set with three data points ($Y_1 = 4.5$, $Y_2 = 5.3$, and $Y_3 = 5.4$). Also shown is a possible null hypothesis for these data, such as $H_0 : \mu = 4.7$. See Fig. 10.15 below.

The maximum likelihood estimates of μ and σ^2 under H_1 are the values of μ and σ^2 found at the peak of the likelihood function. However, the maximum likelihood estimate of σ^2 under H_0 occurs at a different location. Because μ is fixed at 4.7 under H_0, σ^2 is only free to vary along the vertical line shown in the figure. The maximum likelihood estimate of σ^2 under H_0 is the value of σ^2 that maximizes the likelihood along this line.

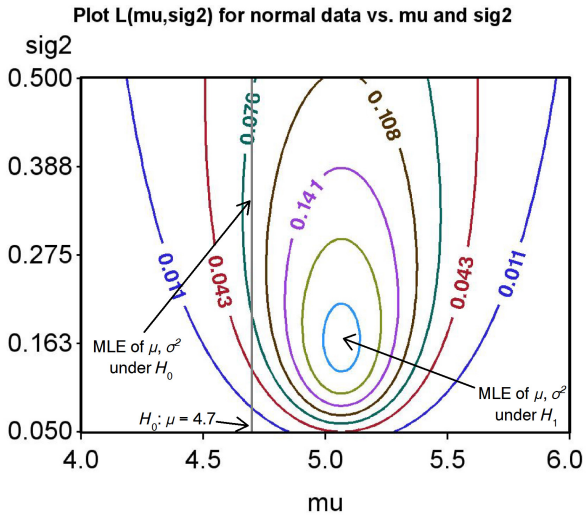


Figure 10.15: Likelihood ratio test for $H_0 : \mu = 4.7$

We are now ready to construct the likelihood ratio test statistic. Let L_{H_0} be the maximum height of the likelihood surface under H_0, which occurs at the maximum likelihood estimate of σ^2 under H_0. Similarly, let let L_{H_1} be

the maximum height under H_1, which occurs at the estimates of μ and σ^2 under H_1. The test statistic λ is just the ratio of these two quantities:

$$\lambda = \frac{L_{H_0}}{L_{H_1}}. \tag{10.22}$$

How does this statistic behave? If H_0 is true, the peak of the likelihood function will often be near the vertical line, and the height of the likelihood function will be similar at the two locations. This implies a value of $\lambda \approx 1$ because $L_{H_0} \approx L_{H_1}$. If H_0 is false and H_1 true, however, we would expect to see $L_{H_0} < L_{H_1}$ and so $\lambda < 1$. We would therefore reject H_0 for sufficiently small values of λ.

More formally, we reject H_0 if $\lambda < c$ and accept H_0 otherwise. The value of c is determined using the Type I error rate α and the distribution of λ under H_0.

An alternate form of the test uses $-2\ln(\lambda)$ rather than λ itself, and rejects H_0 for values of $-2\ln(\lambda) > d$, where d is a constant that controls the Type I error rate. This form of the test rejects for large values of the test statistic, similar to other tests we have developed. Note that

$$-2\ln(\lambda) = 2\ln(L_{H_1}) - 2\ln(L_{H_0}) \tag{10.23}$$

by the properties of logarithms, and is a positive quantity. SAS provides values of the likelihood function in this format for some statistical procedures, and these can be used to construct likelihood ratio tests.

How is the likelihood ratio test related to a t test? It can be shown mathematically that the value of the test statistic

$$T_s = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} \tag{10.24}$$

is directly proportional to $-2\ln(\lambda)$, the likelihood ratio test statistic (Mood et al. 1974). Figure 10.16 plots the value of $-2\ln(\lambda)$ vs. T_s for a scenario matching our example data set. We observe there is a one-to-one correspondence between the two test statistics. When such a correspondence occurs between two test statistics, the tests are considered to be statistically equivalent. We will later see that many statistical tests are in fact likelihood ratio tests. These include tests in analysis of variance, regression, and methods for categorical data such as χ^2 tests.

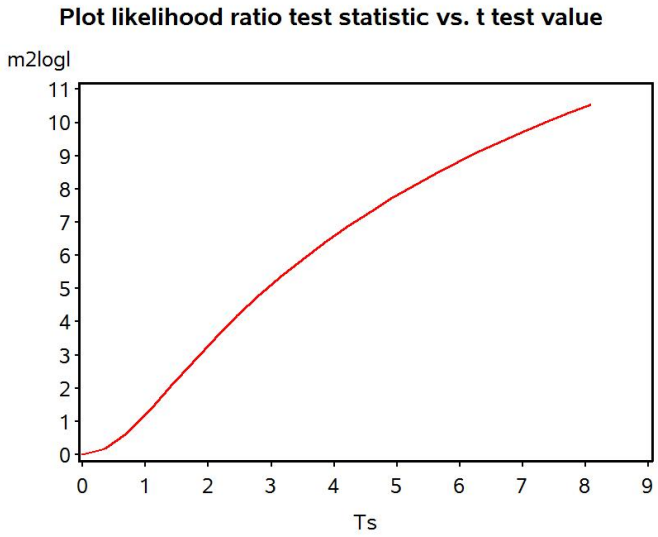


Figure 10.16: Likelihood ratio vs. t test statistics.

10.11 References

Bickel, P. J. & Doksum, K. A. (1977) *Mathematical Statistics: Basic Ideas and Selected Topics*. Holden-Day, Inc., San Francisco, CA.

Mood, A. M., Graybill, F. A. & Boes, D. C. (1974) *Introduction to the Theory of Statistics*. McGraw-Hill, Inc., New York, NY.

SAS Institute Inc. (2018) *SAS/STAT 15.1 Users Guide* SAS Institute Inc., Cary, NC

Yaccoz, N. G. (1991) Use, overuse, and misuse of significance tests in evolutionary biology and ecology. *Bulletin of the Ecological Society of America* 72: 106-111.

10.12 Problems

1. A company that rears beneficial insects produces lacewings (Chrysopidae: Neuroptera) whose mean length is 10 mm. A new method of rearing is being tested and the company wants to determine if the new method changes lacewing length. A sample of 10 insects is collected for the new method, yielding the following lengths:

```
10.3  14.1  11.5  9.9  12.6  9.7  11.0  9.5  12.4  13.5
```

   (a) Test whether the lacewings produced using the new method have the same length as before ($H_0 : \mu = 10$ vs. $H_1 : \mu \neq 10$), using a two-tailed test and Table T. Provide a P value and discuss the significance of the test. Show your calculations.

   (b) Suppose the company is only interested in rearing methods that yield larger lacewing lengths, because bigger is better with beneficial insects. Test $H_0 : \mu = 10$ vs. $H_1 : \mu > 10$. Provide a P value and discuss the significance of the test.

   (c) Use SAS and `proc univariate` to carry out the same two tests. What are the exact P values for these tests? Attach your SAS program and printout.

2. A study is done to measure the concentration of a particular chemical (ppm) in drinking water, with samples taken at eight locations. The samples were analyzed and the following results obtained:

```
23 20 24 20 23 24 21 22
```

   (a) Test whether the concentration of the chemical is significantly different from 20 ppm, the level set by the EPA, using a two-tailed test and Table T. Provide a P value and discuss the significance of the test. Show your calculations.

   (b) The EPA actually requires that the concentration of the chemical be equal to or below 20 ppm. Test whether the chemical concentration exceeds this level using a one-tailed test and Table T. In particular, test $H_0 : \mu = 20$ vs. $H_1 : \mu > 20$. Provide a P value and discuss the significance of the test.

(c) Use SAS and `proc univariate` to carry out the same two tests. What are the exact P values for these tests? Attach your SAS program and printout.