10 Hypothesis Testing

Complementary reading: Chapter 10 (WMS)

10.1 Introduction and review

PREVIEW: Classical statistical inference deals with making statements about population (model) parameters. The two main areas of statistical inference are **estimation** (point estimation and confidence intervals) and **hypothesis testing**. Point and interval estimation were discussed CH8-9 (WMS). This chapter deals with hypothesis testing.

Example 10.1. Actuarial data reveal that the claim amount for a "standard class" of policy holders, denoted by Y (measured in \$1000s), follows an exponential distribution with mean $\theta > 0$. Suppose that we adopt this model for Y and that we observe an iid sample of claims, denoted by $Y_1, Y_2, ..., Y_n$. Recall the following facts from STAT 512:

1. A sufficient statistic for θ is
$$T = \sum_{i=1}^{n} Y_i.$$

2. The maximum likelihood estimator (MLE) for θ is
$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

3. The minimum variance unbiased estimator (MVUE) for θ is $\overline{Y}$.

4. The quantity
$$Q = \frac{2T}{\theta} \sim \chi^2(2n),$$
and therefore is a pivot. This is an exact (finite sample) result; i.e., $Q \sim \chi^2(2n)$ exactly for all n.

5. The quantity
$$Z = \frac{\overline{Y} - \theta}{S/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1),$$

as $n \to \infty$, and therefore Z is a large sample pivot. This means that $Z \sim \mathcal{AN}(0,1)$, when n is large. The larger the n, the better the approximation.

INTERVAL ESTIMATION: We have at our disposal two pivots, namely,

$$Q = \frac{2T}{\theta} \sim \chi^2(2n)$$

and

$$Z = \frac{\overline{Y} - \theta}{S/\sqrt{n}} \sim \mathcal{AN}(0,1).$$

The (exact) confidence interval for θ arising from Q is

$$\left(\frac{2T}{\chi^2_{2n,\alpha/2}}, \frac{2T}{\chi^2_{2n,1-\alpha/2}} \right),$$

where $\chi^2_{2n,1-\alpha/2}$ and $\chi^2_{2n,\alpha/2}$ denote the lower and upper $\alpha/2$ quantiles of a $\chi^2(2n)$ distribution, respectively. The (approximate) confidence interval for θ arising from Z is

$$\overline{Y} \pm z_{\alpha/2} \left(\frac{S}{\sqrt{n}} \right),$$

where S is the sample standard deviation and $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the $\mathcal{N}(0,1)$ distribution.

SIMULATION EXERCISE: To compare the coverage probabilities of the exact and approximate intervals, we will use **Monte Carlo simulation**. In particular, we use R to generate $B = 10,000$ iid samples

$$Y_1, Y_2, ..., Y_n \sim \text{exponential}(\theta),$$

with $n = 10$.

- For each of the $B = 10,000$ samples, we will keep track of the values of T, $\overline{Y}$, and S. We will then compute the exact and approximate 95 percent confidence intervals for θ with each sample (that is, $1 - \alpha = 0.95$).
- Therefore, at the end of the simulation, we will have generated $B = 10,000$ exact intervals and $B = 10,000$ approximate intervals.
- We can then compute the proportion of the intervals (both exact and approximate) which contain θ. For purposes of illustration, we take $\theta = 10$. Because we are computing 95 percent confidence intervals, we would expect this proportion to be close to $1 - \alpha = 0.95$.
- We then repeat this simulation exercise for $n = 30$, $n = 100$, and $n = 1000$. Here are the results:

Interval	$n = 10$	$n = 30$	$n = 100$	$n = 1000$
Exact	0.953	0.949	0.952	0.951
Approximate	0.868	0.915	0.940	0.951

Table 10.1: Monte Carlo simulation. Coverage probabilities for exact and approximate 95 percent confidence intervals for an exponential mean θ, when $\theta = 10$.

DISCUSSION: As we can see, regardless of the sample size n, the exact interval produces a coverage probability that hovers around the nominal $1-\alpha = 0.95$ level, as expected. On the other hand, the coverage probability of the approximate interval is much lower than the nominal $1 - \alpha = 0.95$ level when n is small, although, as n increases, the coverage probability does get closer to the nominal level.

MORAL: We will discuss two types of statistical inference procedures: those that are **exact** and those that are **approximate**. Exact procedures are based on exact distributional results. Approximate procedures are typically based on large sample distributional results (e.g., Central Limit Theorem, Delta Method, Slutsky's Theorem, etc.).

- In some problems, exact inference may not be available or the exact distributional results needed may be so intractable that they are not helpful. In these instances, approximate procedures can be valuable.
- Approximate procedures are based on the (rather nonsensical) notion that the sample size $n \to \infty$. However, these procedures often do confer acceptable results for reasonably sized samples.

PREVIEW: Suppose your colleague claims that the mean claim amount θ for a new class of customers is larger than the mean amount for the standard class of customers, known to be θ_0. How can we determine (statistically) if there is evidence to support this claim? Here, it makes sense to think of two competing hypotheses:

$$H_0 : \theta = \theta_0$$

versus

$$H_a : \theta > \theta_0.$$

- H_0 says that the mean claim amount for the new class of customers, θ, is the same as the mean claim amount for the standard class, θ_0.
- H_a says that the mean claim amount for the new class of customers, θ, is larger than the mean claim amount for the standard class, θ_0, that is, your colleague's claim is correct.
- Based on a sample of claim amounts $Y_1, Y_2, ..., Y_n$ from the new class of customers, how should we formally decide between H_0 and H_a? This question can be answered by performing a hypothesis test.

10.2 The elements of a hypothesis test

TERMINOLOGY: A **hypothesis test** is an inferential technique which pits two competing hypotheses versus each other. The goal is to decide which hypothesis is more supported by the observed data. The four parts of a hypothesis test are

1. the null hypothesis, H_0
2. the alternative hypothesis, H_a
3. the test statistic
4. the rejection region.

TERMINOLOGY: The **null hypothesis** H_0 states the value of the parameter to be tested. For example, if our colleague in Example 10.1 wants to compare the mean claim amount of the new class θ to the mean claim amount for the standard class (known to be $\theta_0 = 10$, say), then the null hypothesis would be

$$H_0 : \theta = 10.$$

In this course, we will usually take the null hypothesis to be **sharp**; that is, there is only one value of the parameter θ possible under H_0.

TERMINOLOGY: The **alternative hypothesis** H_a describes what values of θ we are interested in testing H_0 against. For example, if our colleague in Example 10.1 believed that the mean claim amount for the new class of customers was

- greater than $\theta_0 = 10$, s/he would use

$$H_a : \theta > 10.$$

- less than $\theta_0 = 10$, s/he would use

$$H_a : \theta < 10.$$

- different than $\theta_0 = 10$, s/he would use

$$H_a : \theta \neq 10.$$

The alternative hypothesis H_a is sometimes called the **researcher's hypothesis**, since it is often the hypothesis the researcher wants to conclude is supported by the data.

NOTE: The first two examples of H_a above are called **one-sided** alternatives. The last example is called a **two-sided** alternative. One-sided alternatives state pointedly which direction we are testing H_0 against. A two-sided alternative does not specify this.

TERMINOLOGY: A **test statistic** is a statistic that is used to test H_0 versus H_a. We make our decision by comparing the observed value of the test statistic to its sampling distribution under H_0.

- If the observed value of the test statistic is consistent with its sampling distribution under H_0, then this is not evidence for H_a.
- If the observed value of the test statistic is not consistent with its sampling distribution under H_0, and it is more consistent with the sampling distribution under H_a, then this is evidence for H_a.

TERMINOLOGY: The **rejection region**, denoted by RR, specifies the values of the test statistic for which H_0 is rejected. The rejection region is usually located in tails of the test statistic's sampling distribution computed under H_0. This is why we take H_0 to be sharp, namely, so that we can construct a single sampling distribution.

PREVAILING RULE: In any hypothesis test, if the test statistic falls in rejection region, then we reject H_0.

STATES OF NATURE: Table 10.2 summarizes the four possible outcomes from performing a hypothesis test.

	Decision: Reject H_0	Decision: Do not reject H_0
Truth: H_0	Type I Error	correct decision
Truth: H_a	correct decision	Type II Error

Table 10.2: States of nature in testing $H_0 : \theta = \theta_0$ versus $H_a : \theta \neq \theta_0$ (or any other H_a).

TERMINOLOGY: **Type I Error**: *Rejecting H_0 when H_0 is true.* The probability of Type I Error is denoted by α. Notationally,

$$\begin{aligned} \alpha = P(\text{Type I Error}) &= P(\text{Reject } H_0 | H_0 \text{ is true}) \\ &= P(\text{Reject } H_0 | \theta = \theta_0). \end{aligned}$$

The Type I Error probability α is also called the **significance level** for the test. We would like α to be small. It is common to choose this value up front.

TERMINOLOGY: **Type II Error**: *Not rejecting H_0 when H_a is true.* The probability of Type II Error is denoted by β. Notationally,

$$\beta = P(\text{Type II Error}) = P(\text{Do not reject } H_0 | H_a \text{ is true}).$$

REMARK: Obviously, $H_a : \theta \neq \theta_0$ (or any other H_a) can be true in many ways, so we can compute β for different values of θ under H_a. Specifically, the probability of Type II error, when $\theta = \theta_a \in H_a$, is

$$\beta = \beta(\theta_a) = P(\text{Do not reject } H_0 | \theta = \theta_a).$$

That is, this probability will be different for different values of $\theta_a \in H_a$. Ideally, we would like β to be small for all $\theta_a \in H_a$.

Example 10.2. Suppose that industrial accidents occur according to a Poisson process with mean $\theta = 20$ per site per year. New safety measures have been put in place to decrease the number of accidents at industrial sites all over the US. Suppose that after implementation of the new measures, we will observe the number of accidents for a sample of $n = 10$ sites. Denote these data by $Y_1, Y_2, ..., Y_{10}$. We are interested in testing

$$H_0 : \theta = 20$$

versus

$$H_a : \theta < 20.$$

To perform the test, suppose we use the test statistic

$$T = \sum_{i=1}^{10} Y_i$$

and the rejection region RR $= \{t : t \leq 175\}$.

QUESTIONS:

(a) What is the distribution of T when H_0 is true?

(b) What is $\alpha = P(\text{Type I Error})$ for this RR?

(c) Suppose that $\theta = 18$, that is, H_a is true. What is the probability of Type II Error when using this RR?

Example 10.3. Suppose that $Y_1, Y_2, ..., Y_{25}$ is an iid sample of $n = 25$ observations from a $\mathcal{N}(\theta, \sigma_0^2)$ distribution, where $\sigma_0^2 = 100$ is known. We would like to test

$$H_0 : \theta = 75$$

versus

$$H_a : \theta > 75.$$

To perform the test, suppose we use the test statistic

$$\overline{Y} = \frac{1}{25}\sum_{i=1}^{25} Y_i$$

and the rejection region $\text{RR} = \{\overline{y} : \overline{y} > k\}$, where k is a constant.

QUESTIONS:

(a) What is the distribution of $\overline{Y}$ when H_0 is true?

(b) Find the value of k that provides a level $\alpha = 0.10$ test.

(c) Suppose that $\theta = 80$, that is, H_a is true. What is the probability of Type II Error when using this RR?

Example 10.4. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid Bernoulli(p) sample, where $n = 100$. We would like to test

$$H_0 : p = 0.10$$

versus

$$H_a : p < 0.10.$$

To perform the test, suppose we use the test statistic

$$T = \sum_{i=1}^{100} Y_i$$

and the rejection region RR $= \{t : t \leq k\}$, where k is a constant.

QUESTIONS:

(a) What is the distribution of T when H_0 is true?

(b) Is it possible to find an exact level $\alpha = 0.05$ rejection region?

(c) With $k = 5$, find the probability of Type II Error when $p = 0.05$.

10.3 Common large sample tests

REMARK: The term "large sample" is used to describe hypothesis tests that are constructed using asymptotic (large sample) theory, so the following tests are approximate for "large" sample sizes. We present large sample hypothesis tests for

1. one population mean μ
2. one population proportion p
3. the difference of two population means $\mu_1 - \mu_2$
4. the difference of two population proportions $p_1 - p_2$.

TEST STATISTIC: In each of these situations, we will use a point estimator $\widehat{\theta}$ which satisfies

$$Z = \frac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}} \xrightarrow{d} \mathcal{N}(0,1),$$

as $n \to \infty$. Recall that

$$\sigma_{\widehat{\theta}} = \sqrt{V(\widehat{\theta})}$$

denotes the **standard error** of $\widehat{\theta}$. In most cases, the estimated standard error $\widehat{\sigma}_{\widehat{\theta}}$ must be used in place of $\sigma_{\widehat{\theta}}$. The estimated standard error $\widehat{\sigma}_{\widehat{\theta}}$ is simply a point estimator for the true standard error $\sigma_{\widehat{\theta}}$. In fact, if

$$\frac{\sigma_{\widehat{\theta}}}{\widehat{\sigma}_{\widehat{\theta}}} \xrightarrow{p} 1,$$

as $n \to \infty$, then

$$Z^* = \frac{\widehat{\theta} - \theta}{\widehat{\sigma}_{\widehat{\theta}}} \xrightarrow{d} \mathcal{N}(0,1),$$

as $n \to \infty$, by Slutsky's Theorem.

TWO-SIDED TEST: Suppose that we would like to test

$$H_0 : \theta = \theta_0$$

versus

$$H_a : \theta \neq \theta_0.$$

This is called a **two-sided test** because H_a does not specify a direction indicating departure from H_0. Therefore, large values of

$$Z = \frac{\widehat{\theta} - \theta_0}{\sigma_{\widehat{\theta}}},$$

in either direction, are evidence against H_0. Note that, for n large, $Z \sim \mathcal{AN}(0, 1)$ when $H_0 : \theta = \theta_0$ is true. Therefore,

$$\text{RR} = \{z : |z| > z_{\alpha/2}\}$$

is an approximate level α rejection region. That is, we will reject H_0 whenever $Z > z_{\alpha/2}$ or $Z < -z_{\alpha/2}$. For example, if $\alpha = 0.05$, then $z_{\alpha/2} = z_{0.025} = 1.96$.

ONE-SIDED TESTS: Suppose that we would like to test

$$H_0 : \theta = \theta_0$$

versus

$$H_a : \theta > \theta_0.$$

This is called a **one-sided test** because H_a indicates a specific direction indicating a departure from H_0. In this case, only large values of

$$Z = \frac{\widehat{\theta} - \theta_0}{\sigma_{\widehat{\theta}}},$$

are evidence against H_0. Therefore,

$$\text{RR} = \{z : z > z_\alpha\}$$

is an approximate level α rejection region. That is, we will reject H_0 whenever $Z > z_\alpha$. For example, if $\alpha = 0.05$, then $z_\alpha = z_{0.05} = 1.65$. By an analogous argument, the one-sided test of

$$H_0 : \theta = \theta_0$$

versus

$$H_a : \theta < \theta_0$$

can be performed using

$$\text{RR} = \{z : z < -z_\alpha\}$$

as an approximate level α rejection region.

10.3.1 One population mean

SITUATION: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from a distribution with mean μ and variance σ^2 and that interest lies in testing

$$H_0 : \mu = \mu_0$$

versus

$$H_a : \mu \neq \mu_0$$

(or any other H_a). In this situation, we identify

$$\begin{aligned} \theta &= \mu \\ \widehat{\theta} &= \overline{Y} \\ \sigma_{\widehat{\theta}} &= \frac{\sigma}{\sqrt{n}} \\ \widehat{\sigma}_{\widehat{\theta}} &= \frac{S}{\sqrt{n}}, \end{aligned}$$

where S denotes the sample standard deviation. Therefore, if σ^2 is known, we use

$$Z = \frac{\overline{Y} - \mu_0}{\sigma/\sqrt{n}}.$$

Otherwise, we use

$$Z = \frac{\overline{Y} - \mu_0}{S/\sqrt{n}}.$$

Both statistics have large sample $\mathcal{N}(0, 1)$ distributions when $H_0 : \mu = \mu_0$ is true.

10.3.2 One population proportion

SITUATION: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid Bernoulli(p) sample and that interest lies in testing

$$H_0 : p = p_0$$

versus

$$H_a : p \neq p_0$$

(or any other H_a). In this situation, we identify

$$\begin{aligned} \theta &= p \\ \widehat{\theta} &= \widehat{p} \\ \sigma_{\widehat{\theta}} &= \sqrt{\frac{p(1-p)}{n}} \\ \widehat{\sigma}_{\widehat{\theta}} &= \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}. \end{aligned}$$

To perform this test, there are two candidate test statistics. The first is

$$Z_W = \frac{\widehat{p} - p_0}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}},$$

which arises from the theory we have just developed. A second test statistic is

$$Z_S = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

The test statistic Z_S uses the standard error

$$\widehat{\sigma}_{\widehat{p}} = \sqrt{\frac{p_0(1-p_0)}{n}}$$

which is the correct standard error when $H_0 : p = p_0$ is true. For theoretical reasons, Z_W is called a **Wald statistic** and Z_S is called a **score statistic**. Both have large sample $\mathcal{N}(0,1)$ distributions when $H_0 : p = p_0$ is true. The score statistic Z_S is known to have better properties in small (i.e., finite) samples; i.e., it possesses a true significance level which is often closer to the nominal level α. The Wald statistic is often liberal, possessing a true significance level larger than the nominal level.

10.3.3 Difference of two population means

SITUATION: Suppose that we have two **independent** samples; i.e.,

Sample 1: $Y_{11}, Y_{12}, ..., Y_{1n_1}$ are iid with mean μ_1 and variance σ_1^2

Sample 2: $Y_{21}, Y_{22}, ..., Y_{2n_2}$ are iid with mean μ_2 and variance σ_2^2,

and that interest lies in testing

$$H_0 : \mu_1 - \mu_2 = d_0$$

versus

$$H_a : \mu_1 - \mu_2 \neq d_0$$

(or any other H_a), where d_0 is a known constant. Note that taking $d_0 = 0$ allows one to test the equality of μ_1 and μ_2. In this situation, we identify

$$\begin{aligned}
\theta &= \mu_1 - \mu_2 \\
\widehat{\theta} &= \overline{Y}_{1+} - \overline{Y}_{2+} \\
\sigma_{\widehat{\theta}} &= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \\
\widehat{\sigma}_{\widehat{\theta}} &= \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}.
\end{aligned}$$

If σ_1^2 and σ_2^2 are both known (which would be unlikely), then we would use

$$Z = \frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

Otherwise, we use

$$Z = \frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - d_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}.$$

Both statistics have large sample $\mathcal{N}(0, 1)$ distributions when $H_0 : \mu_1 - \mu_2 = d_0$ is true.

10.3.4 Difference of two population proportions

SITUATION: Suppose that we have two **independent** samples; i.e.,

$$\begin{aligned} &\text{Sample 1:} \quad Y_{11}, Y_{12}, ..., Y_{1n_1} \text{ are iid Bernoulli}(p_1) \\ &\text{Sample 2:} \quad Y_{21}, Y_{22}, ..., Y_{2n_2} \text{ are iid Bernoulli}(p_2), \end{aligned}$$

and that interest lies in testing

$$H_0 : p_1 - p_2 = d_0$$

versus

$$H_a : p_1 - p_2 \neq d_0$$

(or any other H_a), where d_0 is a known constant. Note that taking $d_0 = 0$ allows one to test the equality of p_1 and p_2. In this situation, we identify

$$\begin{aligned} \theta &= p_1 - p_2 \\ \widehat{\theta} &= \widehat{p}_1 - \widehat{p}_2 \\ \sigma_{\widehat{\theta}} &= \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \\ \widehat{\sigma}_{\widehat{\theta}} &= \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}. \end{aligned}$$

The Wald statistic is

$$Z_W = \frac{(\widehat{p}_1 - \widehat{p}_2) - d_0}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}}.$$

A score statistic is available when $d_0 = 0$. It is given by

$$Z_S = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

where

$$\widehat{p} = \frac{n_1\widehat{p}_1 + n_2\widehat{p}_2}{n_1 + n_2}$$

is the **pooled sample proportion**, as it estimates the common $p_1 = p_2 = p$ under H_0. Both statistics have large sample $\mathcal{N}(0, 1)$ distributions when $H_0 : p_1 - p_2 = d_0$ is true. As in the one-sample problem, the score statistic performs better in small samples.

10.4 Sample size calculations

IMPORTANCE: We now address the problem of sample size determination, restricting attention to one-sample settings. We assume that the estimator $\widehat{\theta}$ satisfies

$$Z = \frac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}} \sim \mathcal{AN}(0, 1),$$

for large n, where $\sigma_{\widehat{\theta}}$ is the standard error of $\widehat{\theta}$. Recall that $\widehat{\theta}$, and consequently its standard error $\sigma_{\widehat{\theta}}$, depends on n, the sample size. We focus on the one-sided test

$$H_0 : \theta = \theta_0$$

versus

$$H_a : \theta > \theta_0,$$

that employs the level α rejection region

$$\text{RR} = \{z : z > z_\alpha\} \Longleftrightarrow \{\boldsymbol{y} : \widehat{\theta} > k\},$$

where k is chosen so that $P_{H_0}(\widehat{\theta} > k) \equiv P(\widehat{\theta} > k | H_0 \text{ is true}) = \alpha$.

SETTING: Our goal is to determine the sample size n that confers a specified Type II Error probability β. However, β is a function θ, so we must specify a particular value of θ to consider. Because the alternative hypothesis is of the form $H_a : \theta > \theta_0$, we are interested in a value $\theta_a > \theta_0$; i.e.,

$$\theta_a = \theta_0 + \Delta,$$

where $\Delta > 0$ is the **practically important difference** that we wish to detect.

IMPORTANT: To derive a general formula for the sample size in a particular problem, we exploit the following two facts:

- when H_0 is true, our level α rejection region RR $= \{z : z > z_\alpha\}$ implies that
$$\frac{k - \theta_0}{\sigma_{\widehat{\theta}}} = z_\alpha.$$
- when H_a is true and $\theta_a = \theta_0 + \Delta$, then for a specified value of β, it follows that
$$\frac{k - \theta_a}{\sigma_{\widehat{\theta}}} = -z_\beta;$$
see Figure 10.5, pp 508 (WMS). These two formulae provide the basis for calculating the necessary sample size n. When a two-sided alternative $H_a : \theta \neq \theta_0$ is specified, the only change is that we replace z_α with $z_{\alpha/2}$.

POPULATION MEAN: For the one-sample test regarding a population mean, that is, of $H_0 : \mu = \mu_0$ versus $H_a : \mu > \mu_0$, we have
$$\frac{k - \mu_0}{\sigma/\sqrt{n}} = z_\alpha.$$

When $\mu_a = \mu_0 + \Delta$, then for a specified value of β, we have
$$\frac{k - \mu_a}{\sigma/\sqrt{n}} = -z_\beta.$$

Solving these two equations simultaneously for n gives
$$n = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{\Delta^2},$$

where $\Delta = \mu_a - \mu_0$. Note that the population variance σ^2 must be specified in advance. In practice, we must provide a "guess" or an estimate of its value. This guess may be available from preliminary studies or from other historical information.

Example 10.5. A marine biologist, interested in the distribution of the size of a particular type of anchovy, would like to test

$$H_0 : \mu = 20$$

versus

$$H_a : \mu > 20,$$

where μ denotes the mean anchovy length (measured in cm). She would like to perform this test using $\alpha = 0.05$. Furthermore, when $\mu = \mu_a = 22$, she would like the probability of Type II Error to be only $\beta = 0.1$. What sample size should she use? Based on previous studies, a guess of $\sigma \approx 2.5$ is provided.

POPULATION PROPORTION: For the one-sample test regarding a population proportion, that is, $H_0 : p = p_0$ versus $H_a : p > p_0$, it follows that

$$\frac{k - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = z_\alpha.$$

When $p_a = p_0 + \Delta$, then for a specified value of β, we have

$$\frac{k - p_a}{\sqrt{\frac{p_a(1-p_a)}{n}}} = -z_\beta.$$

Eliminating the common k in these two equations and solving for n produces

$$n = \left[\frac{z_\alpha \sqrt{p_0(1-p_0)} + z_\beta \sqrt{p_a(1-p_a)}}{\Delta} \right]^2,$$

where $\Delta = p_a - p_0$.

Example 10.6. Researchers are planning a Phase III clinical trial to determine the probability of response, p, to a new drug treatment. It is believed that the standard treatment produces a positive response in 35 percent of the population. To determine if the new treatment increases the probability of response, the researchers would like to test, at the $\alpha = 0.05$ level,

$$H_0 : p = 0.35$$

versus

$$H_a : p > 0.35.$$

In addition, they would like to detect a "clinically important" increase in the response probability to $p = p_a = 0.40$ with probability 0.80 (so that the Type II Error probability $\beta = 0.20$). The clinically important difference $\Delta = p_a - p_0 = 0.05$ is a value that represents "a practically important increase" for the manufacturers of the new drug. What is the minimum sample size that should be used in the Phase III trial?

10.5 Confidence intervals and hypothesis tests

REVELATION: There is an elegant duality between confidence intervals and hypothesis tests. In a profound sense, they are essentially the same thing, as we now illustrate. Suppose that we have a point estimator, say, $\widehat{\theta}$, which satisfies

$$Z = \frac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}} \sim \mathcal{N}(0, 1).$$

Using Z as a pivot, it follows that

$$\widehat{\theta} \pm z_{\alpha/2}\sigma_{\widehat{\theta}}$$

is a $100(1-\alpha)$ percent confidence interval for θ.

REMARK: In what follows, we assume that $\sigma_{\widehat{\theta}}$ does not depend on θ (although the following conclusions hold even if it does). If $\sigma_{\widehat{\theta}}$ depends on other nuisance parameters, without loss, we assume that these parameters are known.

HYPOTHESIS TEST: The two-sided level α hypothesis test

$$H_0 : \theta = \theta_0$$

versus

$$H_a : \theta \neq \theta_0$$

employs the rejection region

$$\text{RR} = \{z : |z| > z_{\alpha/2}\}$$

which means that H_0 is not rejected when

$$-z_{\alpha/2} < \frac{\widehat{\theta} - \theta_0}{\sigma_{\widehat{\theta}}} < z_{\alpha/2}.$$

However, algebraically, the last inequality can be rewritten as

$$\widehat{\theta} - z_{\alpha/2}\sigma_{\widehat{\theta}} < \theta_0 < \widehat{\theta} + z_{\alpha/2}\sigma_{\widehat{\theta}},$$

which we recognize as the set of all θ_0 that fall between the $100(1-\alpha)$ percent confidence interval limits.

PUNCHLINE: The hypothesis $H_0 : \theta = \theta_0$ is not rejected in favor of $H_a : \theta \neq \theta_0$, at significance level α, whenever θ_0 is contained in the $100(1-\alpha)$ percent confidence interval for θ. If θ_0 is not contained in the $100(1-\alpha)$ percent confidence interval for θ, then this is the same as rejecting H_0 at level α.

10.6 Probability values (p-values)

REMARK: When performing a hypothesis test, simply saying that we "reject H_0" or that we "do not reject H_0" is somewhat uninformative. A probability value (p-value) provides a numerical measure of how much evidence we have against H_0.

TERMINOLOGY: The **probability value** for a hypothesis test specifies the smallest value of α for which H_0 is rejected. Thus, if the probability value is less than (or equal to) α, we reject H_0. If the probability value is greater than α, we do not reject H_0.

REMARK: Probability values are computed in a manner consistent with the alternative hypothesis H_a. Since the probability value is viewed as a measure of how much evidence we have against H_0, *it is always computed under the assumption that H_0 is true.*

Example 10.7. Suppose that $Y_1, Y_2, ..., Y_{100}$ is an iid $\mathcal{N}(\mu, \sigma_0^2)$ sample, where $\sigma_0^2 = 100$ is known, and that we want to test

$$H_0 : \mu = 75$$

versus

$$H_a : \mu > 75.$$

Suppose that the sample mean is $\overline{y} = 76.42$, and, thus, the one sample z statistic is

$$z = \frac{\overline{y} - \mu_0}{\sigma_0/\sqrt{n}} = \frac{76.42 - 75}{10/\sqrt{100}} = 1.42.$$

Since our alternative is one-sided, we would use the rejection region $\text{RR} = \{z : z > z_\alpha\}$, where z_α denotes the upper α quantile of the standard normal distribution.

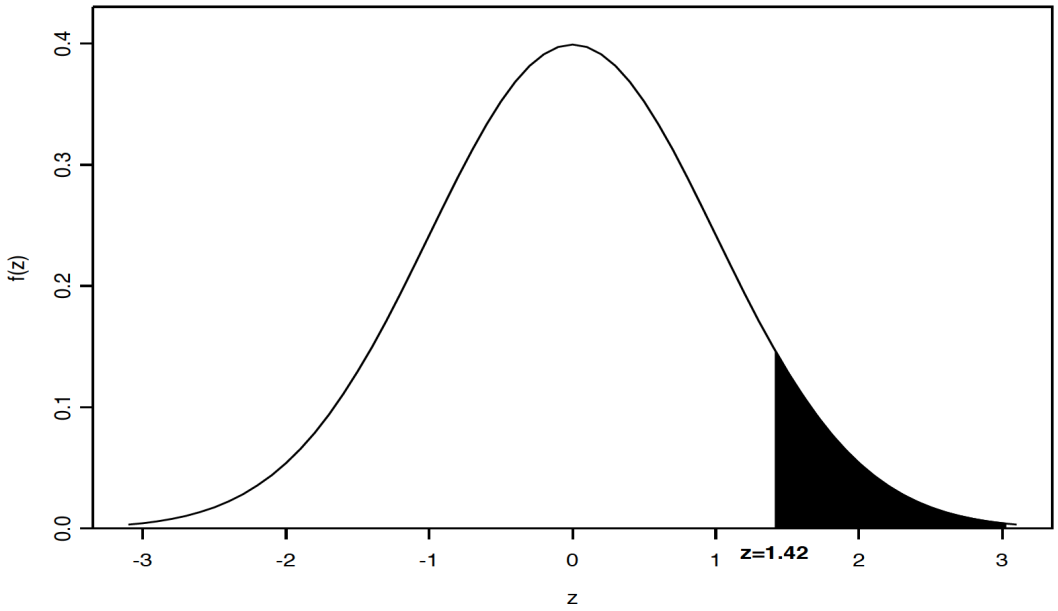

Figure 10.1: $\mathcal{N}(0,1)$ density with one-sided probability value $P(Z > 1.42) = 0.0778$.

α	Test statistic	Rejection region	Reject H_0?
$\alpha = 0.05$	$z = 1.42$	$\{z : z > 1.65\}$	no
$\alpha = 0.10$	$z = 1.42$	$\{z : z > 1.28\}$	yes

From the table, we note that

$$1.28 = z_{0.10} < z = 1.42 < z_{0.05} = 1.65.$$

Therefore, the probability value is somewhere between 0.05 and 0.10. In fact, observing that our alternative is one-sided, we see that

$$\text{p-value} = P(Z > 1.42) = 0.0778$$

(see Figure 10.1). Therefore, if $\alpha < 0.0778$, we would not reject H_0. On the other hand, if $\alpha \geq 0.0778$, we would reject H_0. Remember, the probability value is the "borderline" value of α for which H_0 is rejected. □

Example 10.8. It has been suggested that less than 20 percent of all individuals who sign up for an extended gym membership continue to use the gym regularly six months

after joining. Suppose that Y denotes the number of members who use a certain gym regularly (i.e., at least 3 times per week on average) six months after joining, to be observed from a sample of $n = 50$ members. Assume that $Y \sim b(50, p)$ and that we are to test

$$H_0 : p = 0.20$$

versus

$$H_a : p < 0.20.$$

If $Y = y = 6$, the exact probability value is

$$\begin{aligned} \text{p-value} &= P(Y \leq 6) \\ &= \sum_{y=0}^{6} \underbrace{\binom{50}{y}(0.20)^y(1-0.20)^{50-y}}_{b(50,0.20)\text{ pmf}} \approx 0.1034, \end{aligned}$$

computed using the `pbinom(6,50,0.20)` command in R. This is somewhat strong evidence against H_0, although it is "not enough" at the standard $\alpha = 0.05$ level of significance. Instead of using the exact probability value, we could also compute the approximate probability value as

$$\begin{aligned} \text{p-value} &= P(\widehat{p} < 0.12) \\ &\approx P\left(Z < \frac{0.12 - 0.20}{\sqrt{\frac{0.20(1-0.20)}{50}}}\right) \\ &= P(Z < -1.41) = 0.0793. \end{aligned}$$

As you can see, there is a mild discrepancy here in the exact and approximate probability values. Approximate results should always be interpreted with caution. □

REMARK: In a profound sense, a probability value, say, P, is really a random variable. This should be obvious since it depends on a test statistic, which, in turn, is computed from a sample of random variables $Y_1, Y_2, ..., Y_n$. In the light of this, it seems logical to think about the distribution of P. If the test statistic has a continuous distribution, then **when H_0 is true**, the probability value $P \sim \mathcal{U}(0, 1)$. This is a theoretical result which would be proven in a more advanced course.

10.7 Small sample hypothesis tests using the t distribution

GOAL: We now focus on small sample hypothesis tests for

- a single population mean μ
- the difference of two population means $\mu_1 - \mu_2$.

In the one-sample problem, we know that when $H_0 : \mu = \mu_0$ is true,

$$Z = \frac{\overline{Y} - \mu_0}{S/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1),$$

as $n \to \infty$, by the Central Limit Theorem and Slutsky's Theorem. Therefore, Z can be used as a large sample test statistic to test $H_0 : \mu = \mu_0$. However, the large sample $\mathcal{N}(0,1)$ distribution may be inaccurate when n is small. This occurs when the underlying distribution is highly skewed and/or when S is not a good estimator of σ.

10.7.1 One-sample test

SETTING: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid $\mathcal{N}(\mu, \sigma^2)$ sample, where both parameters μ and σ^2 are unknown, and that we want to test

$$H_0 : \mu = \mu_0$$

versus

$$H_a : \mu \neq \mu_0$$

(or any other H_a). When $H_0 : \mu = \mu_0$ is true, the **one-sample t-statistic**

$$t = \frac{\overline{Y} - \mu_0}{S/\sqrt{n}} \sim t(n-1).$$

Therefore, to perform a level α (two-sided) test, we use the rejection region

$$\text{RR} = \{t : |t| > t_{n-1,\alpha/2}\}.$$

Probability values are also computed from the $t(n-1)$ distribution. One-sided tests use a suitably-adjusted rejection region.

Table 10.3: Crab temperature data. These observations are modeled as $n = 8$ iid realizations from a $\mathcal{N}(\mu, \sigma^2)$ distribution.

25.8	24.6	26.1	24.9	25.1	25.3	24.0	24.5

Example 10.9. A researcher observes a sample of $n = 8$ crabs and records the body temperature of each (in degrees C); see Table 10.3. She models these observations as an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution. She would like to test, at level $\alpha = 0.05$,

$$H_0 : \mu = 25.4$$

versus

$$H_a : \mu < 25.4.$$

The level $\alpha = 0.05$ rejection region is

$$\text{RR} = \{t : t < -t_{7,0.05} = -1.895\}.$$

From the data in Table 10.3, we compute $\overline{y} = 25.0$ and $s = 0.69$; thus, the value of the one-sample t-statistic is

$$t = \frac{\overline{y} - \mu_0}{s/\sqrt{n}} = \frac{25.0 - 25.4}{0.69/\sqrt{8}} = -1.64.$$

Therefore, we do not have sufficient evidence to reject H_0 at the $\alpha = 0.05$ level since our test statistic t does not fall in RR. Equivalently, the probability value is

$$\text{p-value} = P[t(7) \leq -1.64] \approx 0.073,$$

which is not smaller than $\alpha = 0.05$. I used the R command `pt(-1.64,7)` to compute this probability. □

10.7.2 Two-sample test

SITUATION: Suppose that we have two **independent** samples; i.e.,

Sample 1: $Y_{11}, Y_{12}, ..., Y_{1n_1}$ are iid with mean μ_1 and variance σ_1^2

Sample 2: $Y_{21}, Y_{22}, ..., Y_{2n_2}$ are iid with mean μ_2 and variance σ_2^2,

and that interest lies in testing

$$H_0 : \mu_1 - \mu_2 = d_0$$

versus

$$H_a : \mu_1 - \mu_2 \neq d_0$$

(or any other H_a), where d_0 is a known constant. When the population variances are equal; that is, when $\sigma_1^2 = \sigma_2^2$, we know that

$$\frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2),$$

where

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

is the pooled sample variance. Therefore, to perform a level α (two-sided) test, we use the test statistic

$$t = \frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - d_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

and the rejection region

$$\text{RR} = \{t : |t| > t_{n_1+n_2-2,\alpha/2}\}.$$

Probability values are also computed from the $t(n_1 + n_2 - 2)$ distribution. One-sided tests use a suitably-adjusted rejection region.

REMARK: When $\sigma_1^2 \neq \sigma_2^2$; that is, when the population variances are **not** equal, we can use the modified t-statistic given by

$$t^* = \frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - d_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}.$$

Under H_0, the distribution of this modified t-statistic is approximated by a $t(\nu)$ distribution, where the degrees of freedom

$$\nu \approx \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}.$$

10.8 Hypothesis tests for variances

10.8.1 One-sample test

SETTING: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid $\mathcal{N}(\mu, \sigma^2)$ sample, where both parameters are unknown, and that interest lies in testing

$$H_0 : \sigma^2 = \sigma_0^2$$

versus

$$H_a : \sigma^2 \neq \sigma_0^2,$$

(or any other H_a), where σ_0^2 is a specified value. When H_0 is true; i.e., when $\sigma^2 = \sigma_0^2$, the statistic

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1).$$

Therefore, a level α (two-sided) rejection region is

$$\text{RR} = \{\chi^2 : \chi^2 < \chi^2_{n-1,1-\alpha/2} \text{ or } \chi^2 > \chi^2_{n-1,\alpha/2}\}.$$

Probability values are also computed from the $\chi^2(n-1)$ distribution. One-sided tests use a suitably-adjusted rejection region.

10.8.2 Two-sample test

SETTING: Suppose that we have two **independent** samples:

$$\begin{aligned} \text{Sample 1}: &\quad Y_{11}, Y_{12}, ..., Y_{1n_1} \sim \text{iid } \mathcal{N}(\mu_1, \sigma_1^2) \\ \text{Sample 2}: &\quad Y_{21}, Y_{22}, ..., Y_{2n_2} \sim \text{iid } \mathcal{N}(\mu_2, \sigma_2^2), \end{aligned}$$

and that interest lies in testing

$$H_0 : \sigma_1^2 = \sigma_2^2$$

versus

$$H_a : \sigma_1^2 \neq \sigma_2^2,$$

(or any other H_a). Recall from Chapter 7 (WMS) that, in general,

$$F = \frac{\frac{(n_1-1)S_1^2}{\sigma_1^2}/(n_1-1)}{\frac{(n_2-1)S_2^2}{\sigma_2^2}/(n_2-1)} \sim F(n_1-1, n_2-1).$$

However, note that when $H_0 : \sigma_1^2 = \sigma_2^2$ is true, F reduces algebraically to

$$F = \frac{S_1^2}{S_2^2} \sim F(n_1-1, n_2-1).$$

Therefore, a level α (two-sided) rejection region is

$$\text{RR} = \{F : F < F_{n_1-1,n_2-1,1-\alpha/2} \text{ or } F > F_{n_1-1,n_2-1,\alpha/2}\}.$$

Probability values are also computed from the $F(n_1-1, n_2-1)$ distribution. One-sided tests use a suitably-adjusted rejection region.

10.9 Power, the Neyman-Pearson Lemma, and UMP tests

10.9.1 Power

TERMINOLOGY: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from $f_Y(y;\theta)$ and that we use a level α rejection region to test $H_0 : \theta = \theta_0$ versus a suitable alternative. The **power function** of the test, denoted by $K(\theta)$, is given by

$$K(\theta) = P(\text{Reject } H_0|\theta).$$

That is, the power function gives the probability of rejecting H_0 as a function of θ.

- If $\theta = \theta_0$, that is H_0 is true, then $K(\theta_0) = \alpha$, the probability of Type I Error.
- For values of θ that are "close" to θ_0, one would expect the power to be smaller, than, say, when θ is far away from θ_0. This makes sense intuitively; namely, it is more difficult to detect a small departure from H_0 (i.e., to reject H_0) than it is to detect a large departure from H_0.
- The shape of the power function always depends on the alternative hypothesis.

NOTE: If θ_a is a value of θ in the alternative space; that is, if $\theta_a \in H_a$, then

$$K(\theta_a) = 1 - \beta(\theta_a).$$

Proof. This follows directly from the complement rule; that is,

$$\begin{aligned} K(\theta_a) &= P(\text{Reject } H_0|\theta = \theta_a) \\ &= 1 - P(\text{Do not reject } H_0|\theta = \theta_a) = 1 - \beta(\theta_a). \quad \square \end{aligned}$$

Example 10.10. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid $\mathcal{N}(\theta, \sigma_0^2)$ sample, where σ_0^2 is known, and that we would like to test

$$H_0 : \theta = \theta_0$$

versus

$$H_a : \theta > \theta_0.$$

Suppose that we use the level α rejection region $\text{RR} = \{z : z > z_\alpha\}$, where

$$Z = \frac{\overline{Y} - \theta_0}{\sigma_0/\sqrt{n}}$$

and z_α denotes the upper α quantile of the standard normal distribution. The power function for the test, for $\theta > \theta_0$, is given by

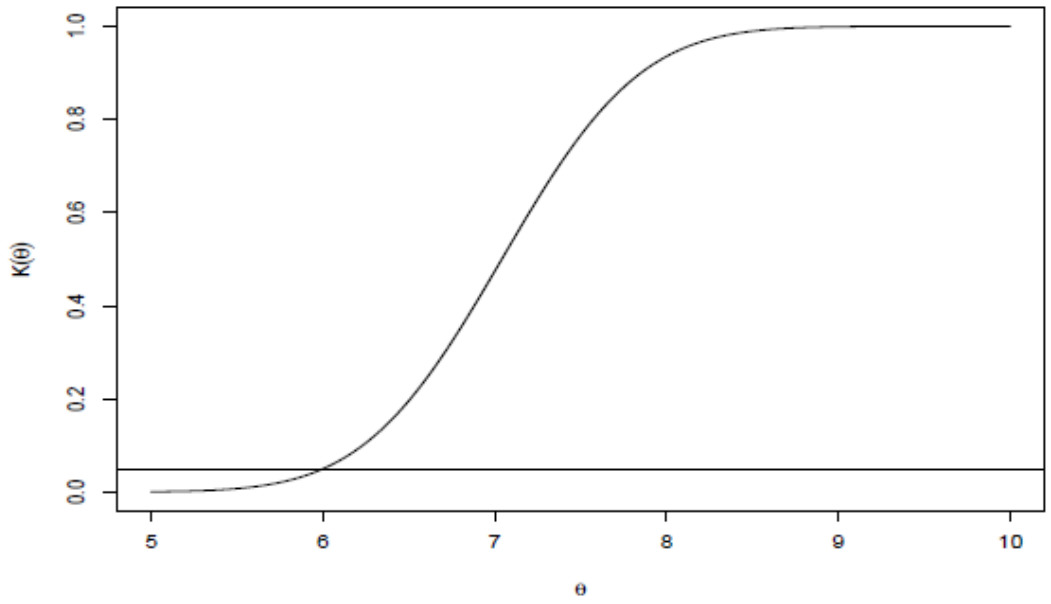

Figure 10.2: Power function $K(\theta)$ in Example 10.10 with $\alpha = 0.05$, $\theta_0 = 6$, $\sigma_0^2 = 4$, and $n = 10$. A horizontal line at $\alpha = 0.05$ is drawn.

ILLUSTRATION: Figure 10.2 displays the graph of $K(\theta)$ when $\alpha = 0.05$, $\theta_0 = 6$, $\sigma_0^2 = 4$, and $n = 10$. That is, we are testing

$$H_0 : \theta = 6$$

versus

$$H_a : \theta > 6.$$

We make the following observations.

- Note that $K(6) = 0.05$, that is, the power of the test when $H_0 : \theta = 6$ is true is equal to $\alpha = 0.05$.
- Note that $K(\theta)$ is an increasing function of θ. Therefore, the probability of rejecting H_0 increases as θ increases. For example, $K(6.5) \approx 0.1965$, $K(7) \approx 0.4746$, $K(8) \approx 0.9354$, $K(9) \approx 0.9990$, etc. $\square$

10.9.2 The Neyman-Pearson Lemma

TERMINOLOGY: In this course, we will usually take the null hypothesis to be sharp, or **simple**; that is, there is just one value of θ possible under H_0. The alternative may be simple or **composite**. Here is an example of a simple-versus-simple test:

$$H_0 : \theta = 5$$

versus

$$H_a : \theta = 6.$$

Here is an example of a simple-versus-composite test:

$$H_0 : \theta = 5$$

versus

$$H_a : \theta > 5.$$

Note that there are an infinite number of values of θ specified in a composite alternative hypothesis. In this example, H_a consists of any value of θ larger than 5.

GOAL: For a level α simple-versus-simple test, we seek the **most powerful rejection region**; i.e., the rejection region that maximizes the probability of rejecting H_0 when H_a is true. The Neyman-Pearson Lemma tells us how to find this "most powerful test."

RECALL: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from $f_Y(y;\theta)$. The **likelihood function** for θ is given by

$$L(\theta) = L(\theta|\boldsymbol{y}) = L(\theta|y_1, y_2, ..., y_n) = \prod_{i=1}^{n} f_Y(y_i;\theta).$$

NEYMAN-PEARSON LEMMA: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from $f_Y(y;\theta)$, and let $L(\theta)$ denote the likelihood function. Consider the following simple-versus-simple hypothesis test:

$$H_0 : \theta = \theta_0$$

versus

$$H_a : \theta = \theta_a.$$

The level α test that maximizes the power when $H_a : \theta = \theta_a$ is true uses the rejection region

$$\text{RR} = \left\{ \boldsymbol{y} : \frac{L(\theta_0)}{L(\theta_a)} < k \right\},$$

where k is chosen so that

$$P(\text{Reject } H_0 | H_0 \text{ is true}) = \alpha.$$

This is called the **most-powerful level α test** for H_0 versus H_a.

Example 10.11. Suppose that Y is a single observation (i.e., an iid sample of size $n = 1$) from an exponential distribution with mean θ. Using this single observation, we would like to test

$$H_0 : \theta = 2$$

versus

$$H_a : \theta = 3.$$

Use the Neyman-Pearson Lemma to find the most powerful level $\alpha = 0.10$ test.

REMARK: Note that even though we have found the most powerful level $\alpha = 0.10$ test of H_0 versus H_a, the test is not all that powerful; we have only about a 21.5 percent chance of correcting rejecting H_0 when H_a is true. Of course, this should not be surprising, given that we have just a single observation Y. We are trying to make a decision with very little information about θ. $\square$

Example 10.12. Suppose that $Y_1, Y_2, ..., Y_{10}$ is an iid sample of Poisson(θ) observations and that we want to test

$$H_0 : \theta = 1$$

versus

$$H_a : \theta = 2.$$

Find the most-powerful level α test.

What is the power of the level $\alpha = 0.0487$ test when H_a is true?

RESULT: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from $f_Y(y;\theta)$ and let U be a sufficient statistic. The rejection region for the most powerful level α test of $H_0 : \theta = \theta_0$ versus $H_a : \theta = \theta_a$ always depends on U.

Proof. From the Factorization Theorem, we can write

$$\frac{L(\theta_0)}{L(\theta_a)} = \frac{g(u;\theta_0)h(\boldsymbol{y})}{g(u;\theta_a)h(\boldsymbol{y})} = \frac{g(u;\theta_0)}{g(u;\theta_a)},$$

where g and h are nonnegative functions. By the Neyman-Pearson Lemma, the most-powerful level α rejection region is

$$\text{RR} = \left\{\boldsymbol{y} : \frac{L(\theta_0)}{L(\theta_a)} < k\right\} = \left\{\boldsymbol{y} : \frac{g(u;\theta_0)}{g(u;\theta_a)} < k\right\},$$

where k is chosen so that $P(\text{Reject } H_0 | H_0 \text{ is true}) = \alpha$. Clearly, this rejection region depends on the sufficient statistic U. $\square$

10.9.3 Uniformly most powerful (UMP) tests

REMARK: For a simple-versus-simple test, the Neyman-Pearson Lemma shows us explicitly how to derive the most-powerful level α rejection region. We now discuss simple-versus-composite tests; e.g., $H_0 : \theta = \theta_0$ versus $H_a : \theta > \theta_0$ and $H_0 : \theta = \theta_0$ versus $H_a : \theta < \theta_0$.

TERMINOLOGY: When a test maximizes the power for all θ in the alternative space; i.e., for all $\theta \in H_a$, it is called the **uniformly most powerful (UMP) level α test**. In other words, if $K_U(\theta)$ denotes the power function for the UMP level α test of H_0 versus H_a, and if $K_{U^*}(\theta)$ denotes the power function for some other level α test, then $K_U(\theta) \geq K_{U^*}(\theta)$ for all $\theta \in H_a$.

FINDING UMP TESTS: Suppose that our goal is to find the UMP level α test of

$$H_0 : \theta = \theta_0$$

versus

$$H_a : \theta > \theta_0.$$

Instead of considering this simple-versus-composite test, we first "pretend" like we have the level α simple-versus-simple test

$$H_0 : \theta = \theta_0$$

versus

$$H_a : \theta = \theta_a,$$

where $\theta_a > \theta_0$ is arbitrary. If we can then show that neither the test statistic nor the rejection region for the most powerful level α simple-versus-simple test depends on θ_a, then the test with the same rejection region will be UMP level α for the simple-versus-composite test $H_0 : \theta = \theta_0$ versus $H_a : \theta > \theta_0$.

CURIOSITY: Why does this work? Essentially we are showing that for a given θ_a, the level α simple-versus-simple test is most powerful, by appealing to the Neyman-Pearson Lemma. However, since the value θ_a is arbitrary and since the most powerful RR is free of θ_a, this same test must be most powerful level α for every value of $\theta_a > \theta_0$; i.e., it must be the **uniformly** most powerful (UMP) level α test for all $\theta > \theta_0$.

Example 10.13. Suppose that $Y_1, Y_2, ..., Y_{15}$ is an iid sample from a Rayleigh distribution with pdf

$$f_Y(y) = \begin{cases} \frac{2y}{\theta} e^{-y^2/\theta}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Find the UMP level $\alpha = 0.05$ test of

$$H_0 : \theta = 1$$

versus

$$H_a : \theta > 1.$$

What is the power function $K(\theta)$ for this test?

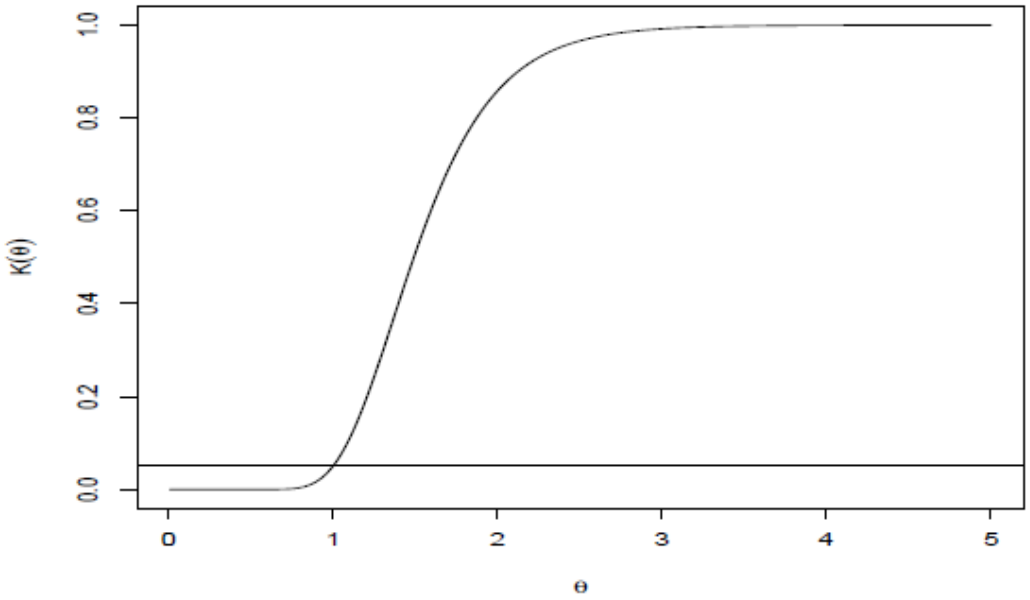

Figure 10.3: Power function $K(\theta)$ in Example 10.13 with $\alpha = 0.05$, $\theta_0 = 1$, and $n = 15$. A horizontal line at $\alpha = 0.05$ is drawn.

REMARK: UMP level α tests do not always exist. For example, a two-sided test $H_0 : \theta = \theta_0$ versus $H_a : \theta \neq \theta_0$ never has a UMP rejection region. This is because

- the power function of the UMP level α test of $H_0 : \theta = \theta_0$ versus $H_a : \theta < \theta_0$ will be larger than the power function of the UMP level α test of $H_0 : \theta = \theta_0$ versus $H_a : \theta > \theta_0$ when $\theta < \theta_0$.
- the power function of the UMP level α test of $H_0 : \theta = \theta_0$ versus $H_a : \theta > \theta_0$ will be larger than the power function of the UMP level α test of $H_0 : \theta = \theta_0$ versus $H_a : \theta < \theta_0$ when $\theta > \theta_0$.

For two-sided alternatives, the class of level α tests, say, $\mathcal{C}$, is too large, and finding one rejection region that uniformly beats all other level α rejection regions is impossible.

10.10 Likelihood ratio tests

TERMINOLOGY: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from $f_Y(y;\theta)$, where the parameter $\theta \in \Omega$. We call Ω the **parameter space**; that is, Ω represents the set of all values that θ (scalar or vector) can assume. For example, if

- $Y \sim b(1,\theta) \Longrightarrow \Omega = \{\theta : 0 < \theta < 1\}$
- $Y \sim \text{exponential}(\theta) \Longrightarrow \Omega = \{\theta : \theta > 0\}$
- $Y \sim \text{gamma}(\alpha, \beta) \Longrightarrow \Omega = \{\boldsymbol{\theta} = (\alpha, \beta)' : \alpha > 0,\ \beta > 0\}$
- $Y \sim \mathcal{N}(\mu, \sigma^2) \Longrightarrow \Omega = \{\boldsymbol{\theta} = (\mu, \sigma^2)' : -\infty < \mu < \infty,\ \sigma^2 > 0\}$.

TERMINOLOGY: Suppose that we partition Ω into two sets Ω_0 and Ω_a, that is, we write

$$\Omega = \Omega_0 \cup \Omega_a,$$

where Ω_0 and Ω_a are mutually exclusive. A hypothesis test can be stated very generally as $H_0 : \theta \in \Omega_0$ versus $H_a : \theta \in \Omega_a$. We call Ω_0 the **null space** and Ω_a the **alternative space**.

TERMINOLOGY: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from $f_Y(y;\theta)$, where $\theta \in \Omega$. A level α **likelihood ratio test** (LRT) for

$$H_0 : \theta \in \Omega_0$$

versus

$$H_a : \theta \in \Omega_a$$

employs the test statistic

$$\lambda = \frac{L(\widehat{\Omega}_0)}{L(\widehat{\Omega})} \equiv \frac{\sup_{\theta \in \Omega_0} L(\theta)}{\sup_{\theta \in \Omega} L(\theta)}$$

and uses the rejection region

$$\text{RR} = \{\lambda : \lambda \leq k\},$$

where k is chosen such that

$$P(\text{Reject } H_0 | H_0 \text{ is true}) = \alpha.$$

From the definition, we see that $0 \leq \lambda \leq 1$, because $L(\cdot)$ is positive and $\Omega_0 \subset \Omega$. Also,

- $L(\widehat{\Omega}_0)$ is the likelihood function evaluated at the maximum likelihood estimator (MLE) over Ω_0, the "restricted" parameter space.
- $L(\widehat{\Omega})$ is the likelihood function evaluated at the MLE over Ω, the "unrestricted" parameter space.

TECHNICAL NOTE: If H_0 is a **composite** hypothesis, we define

$$\alpha = \sup_{\theta \in \Omega_0} P(\text{Reject } H_0|\theta) = \sup_{\theta \in \Omega_0} K(\theta),$$

where $K(\theta)$ denotes the power function.

Example 10.14. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from an exponential(θ) distribution. Find the level α likelihood ratio test (LRT) for

$$H_0 : \theta = \theta_0$$

versus

$$H_a : \theta \neq \theta_0.$$

SOLUTION. Here, the restricted parameter space is $\Omega_0 = \{\theta_0\}$, that is, Ω_0 contains only one value of θ. The alternative parameter space is $\Omega_a = \{\theta : \theta > 0,\ \theta \neq \theta_0\}$, and the unrestricted parameter space is $\Omega = \{\theta : \theta > 0\}$. Note that $\Omega = \Omega_0 \cup \Omega_a$. The likelihood function for θ is given by

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \frac{1}{\theta} e^{-y_i/\theta} \\ &= \frac{1}{\theta^n} e^{-\sum_{i=1}^{n} y_i/\theta} \\ &= \theta^{-n} e^{-u/\theta}, \end{aligned}$$

where the sufficient statistic $u = \sum_{i=1}^{n} y_i$. Over the restricted (null) space, we have

$$L(\widehat{\Omega}_0) = \sup_{\theta \in \Omega_0} L(\theta) = L(\theta_0),$$

because Ω_0 contains only the singleton θ_0. Over the unrestricted space,

$$L(\widehat{\Omega}) = \sup_{\theta \in \Omega} L(\theta) = L(\widehat{\theta}),$$

where $\widehat{\theta}$ is the maximum likelihood estimator (MLE) of θ. Recall that for the exponential(θ) model, the MLE is

$$\widehat{\theta} = \overline{Y}.$$

Therefore, the likelihood ratio test statistic is

$$\lambda = \frac{L(\widehat{\Omega}_0)}{L(\widehat{\Omega})} = \frac{L(\theta_0)}{L(\overline{y})} = \frac{\theta_0^{-n} e^{-u/\theta_0}}{\overline{y}^{-n} e^{-u/\overline{y}}}.$$

Because $u = \sum_{i=1}^n y_i = n\overline{y}$, we can rewrite λ as

$$\lambda = \left(\frac{\overline{y}}{\theta_0}\right)^n \frac{e^{-u/\theta_0}}{e^{-n\overline{y}/\overline{y}}} = \left(\frac{e}{\theta_0}\right)^n \overline{y}^n e^{-n\overline{y}/\theta_0}.$$

Therefore, to find the level α LRT, we would choose k so that

$$P\left[\left(\frac{e}{\theta_0}\right)^n \overline{Y}^n e^{-n\overline{Y}/\theta_0} \leq k \middle| \theta = \theta_0\right] = \alpha.$$

This is an unfriendly request, so let's approach the problem of choosing k in another way.

EXCURSION: For $a > 0$, define the function

$$g(a) = \left(\frac{e}{\theta_0}\right)^n a^n e^{-na/\theta_0}$$

so that

$$\ln g(a) = \ln c_0 + n \ln a - \frac{na}{\theta_0},$$

where the constant $c_0 = (e/\theta_0)^n$. Note that

$$\frac{\partial \ln g(a)}{\partial a} = \frac{n}{a} - \frac{n}{\theta_0}.$$

If we set this derivative equal to 0 and solve for a, we get the first order critical point

$$a = \theta_0.$$

This value of a maximizes $\ln g(a)$ because

$$\frac{\partial^2 \ln g(a)}{\partial a^2} = -\frac{n}{a^2} < 0,$$

by the Second Derivative Test. Also, note that

$$\frac{\partial \ln g(a)}{\partial a} = \begin{cases} \frac{n}{a} - \frac{n}{\theta_0} > 0, & \text{if } a < \theta_0 \\ \frac{n}{a} - \frac{n}{\theta_0} < 0, & \text{if } a > \theta_0, \end{cases}$$

so $\ln g(a)$ is strictly increasing for $a < \theta_0$ and strictly decreasing for $a > \theta_0$. However, because the log function is 1:1, all of these findings apply to the function $g(a)$ as well:

- $g(a)$ is strictly increasing when $a < \theta_0$.
- $g(a)$ is strictly decreasing when $a > \theta_0$.
- $g(a)$ is maximized when $a = \theta_0$.

Therefore, there exist constants $c_1 < c_2$ such that

$$g(a) \leq k \Longleftrightarrow a \leq c_1 \text{ or } a \geq c_2.$$

This is easy to see from sketching a graph of $g(a)$, for $a > 0$.

LRT: Now, returning to the problem at hand, we need to choose k so that

$$P\left[\left(\frac{e}{\theta_0}\right)^n \overline{Y}^n e^{-n\overline{Y}/\theta_0} \leq k \middle| \theta = \theta_0\right] = \alpha.$$

The recent excursive argument should convince you that this is equivalent to choosing c_1 and c_2 so that

$$P(\{\overline{Y} \leq c_1\} \cup \{\overline{Y} \geq c_2\} | \theta = \theta_0) = \alpha.$$

However, because $c_1 < c_2$, the sets $\{\overline{Y} \leq c_1\}$ and $\{\overline{Y} \geq c_2\}$ must be mutually exclusive. By Kolmolgorov's third axiom of probability, we have

$$\begin{aligned} \alpha &= P(\{\overline{Y} \leq c_1\} \cup \{\overline{Y} \geq c_2\} | \theta = \theta_0) \\ &= P(\overline{Y} \leq c_1 | \theta = \theta_0) + P(\overline{Y} \geq c_2 | \theta = \theta_0). \end{aligned}$$

We have changed the problem to now specifying the constants c_1 and c_2 that satisfy this most recent expression. This is a much friendlier request because the distribution of $\overline{Y}$ is tractable; in fact, a simple mgf argument shows that, in general,

$$\overline{Y} \sim \text{gamma}\left(n, \frac{\theta}{n}\right),$$

Therefore, when $H_0 : \theta = \theta_0$ is true, we can take c_1 and c_2 to satisfy

$$\begin{aligned}
\int_0^{c_1} \frac{1}{\Gamma(n)\left(\frac{\theta_0}{n}\right)^n} a^{n-1} e^{-a/\left(\frac{\theta_0}{n}\right)} da &= \alpha/2 \\
\int_{c_2}^{\infty} \frac{1}{\Gamma(n)\left(\frac{\theta_0}{n}\right)^n} a^{n-1} e^{-a/\left(\frac{\theta_0}{n}\right)} da &= \alpha/2,
\end{aligned}$$

that is, c_1 is the lower $\alpha/2$ quantile of the gamma$(n, \theta_0/n)$ distribution and c_2 is the corresponding upper $\alpha/2$ quantile. R makes getting these quantiles simple. It is possible to get closed-form expressions for c_1 and c_2. In fact, it can be shown that

$$\begin{aligned}
c_1 &= \left(\frac{\theta_0}{2n}\right) \chi^2_{2n,1-\alpha/2} \\
c_2 &= \left(\frac{\theta_0}{2n}\right) \chi^2_{2n,\alpha/2},
\end{aligned}$$

where $\chi^2_{2n,1-\alpha/2}$ and $\chi^2_{2n,\alpha/2}$ are the lower and upper $\alpha/2$ quantiles of the $\chi^2(2n)$ distribution. Therefore, the level α likelihood ratio test (LRT) uses the rejection region

$$\text{RR} = \{\overline{y} : \overline{y} \leq c_1 \text{ or } \overline{y} \geq c_2\}.$$

ILLUSTRATION: Suppose that $\alpha = 0.05$, $\theta_0 = 10$, and $n = 20$, so that

$$\begin{aligned}
c_1 &= \left(\frac{10}{40}\right) \chi^2_{40,0.975} = 6.1083 \\
c_2 &= \left(\frac{10}{40}\right) \chi^2_{40,0.025} = 14.8354.
\end{aligned}$$

Therefore, the level $\alpha = 0.05$ LRT employs the rejection region

$$\text{RR} = \{\overline{y} : \overline{y} \leq 6.1083 \text{ or } \overline{y} \geq 14.8354\}.$$

For this rejection region, the power function is given by

$$\begin{aligned}
K(\theta) &= P(\text{Reject } H_0|\theta) \\
&= P(\overline{Y} \leq c_1|\theta) + P(\overline{Y} \geq c_2|\theta) \\
&= \int_0^{c_1} \frac{1}{\Gamma(20)\left(\frac{\theta}{20}\right)^{20}} a^{20-1} e^{-a/\left(\frac{\theta}{20}\right)} da + \int_{c_2}^{\infty} \frac{1}{\Gamma(20)\left(\frac{\theta}{20}\right)^{20}} a^{20-1} e^{-a/\left(\frac{\theta}{20}\right)} da
\end{aligned}$$

This power function is shown in Figure 10.4. □

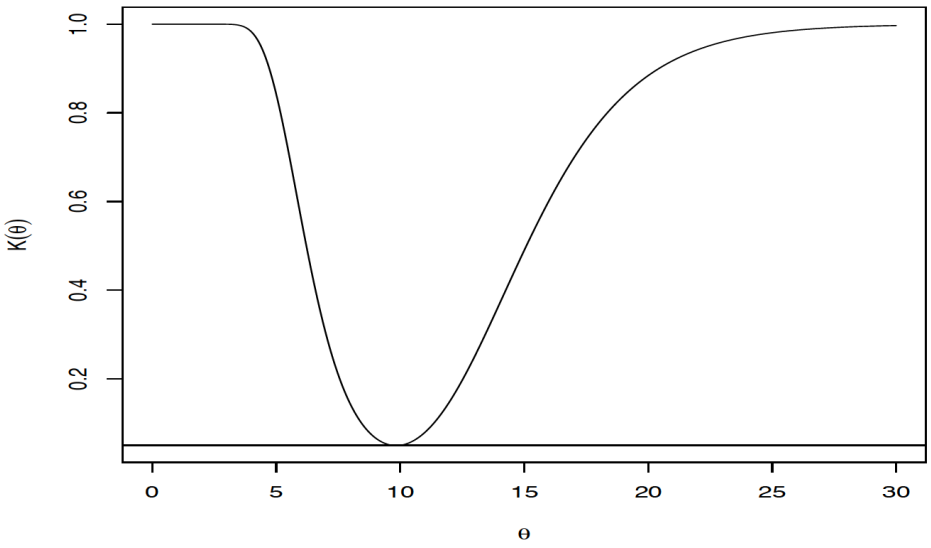

Figure 10.4: Power function $K(\theta)$ in Example 10.14 with $\alpha = 0.05$, $\theta_0 = 10$, and $n = 20$. A horizontal line at $\alpha = 0.05$ is drawn.

REMARK: In Example 10.14, we were fortunate to know the sampling distribution of $\overline{Y}$ when H_0 was true. In other situations, the distribution of the test statistic may be intractable. When this occurs, the following large-sample result can prove to be useful.

ASYMPTOTIC RESULT: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from $f_Y(y; \theta)$, where $\theta \in \Omega$, and that we are to test

$$H_0 : \theta \in \Omega_0$$

versus

Under certain "regularity conditions" (which we will omit), it follows that, under H_0,

$$-2 \ln \lambda \xrightarrow{d} \chi^2(\nu),$$

as $n \to \infty$, where ν is the difference between the number of free parameters specified by $\theta \in \Omega_0$ and the number of free parameters specified in by $\theta \in \Omega$.

Example 10.15

SITUATION: Suppose that we have two **independent** samples; i.e.,

Sample 1: $Y_{11}, Y_{12}, ..., Y_{1n_1}$ are iid with mean μ_1 and variance σ_1^2

Sample 2: $Y_{21}, Y_{22}, ..., Y_{2n_2}$ are iid with mean μ_2 and variance σ_2^2,

and that interest lies in testing

$$H_0 : \mu_1 - \mu_2 = d_0$$

versus

$$H_a : \mu_1 - \mu_2 \neq d_0$$

(or any other H_a), where d_0 is a known constant. When the population variances are equal; that is, when $\sigma_1^2 = \sigma_2^2$, we know that

$$\frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2),$$

where

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

is the pooled sample variance. Therefore, to perform a level α (two-sided) test, we use the test statistic

$$t = \frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - d_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

and the rejection region

$$\text{RR} = \{t : |t| > t_{n_1+n_2-2,\alpha/2}\}.$$

Probability values are also computed from the $t(n_1 + n_2 - 2)$ distribution. One-sided tests use a suitably-adjusted rejection region.