16 Statistical Inference: Hypothesis Tests

> *Judge*: Mr. Hutz we've been in here for four hours.
> Do you have any evidence at all?
> *Hutz*: Well, Your Honor. We've plenty of hearsay and conjecture.
> Those are kinds of evidence.
>
> From: *The Simpsons*

16.1 Introduction

In Chapter 15 we considered inference procedures that relied on estimation. In some situations, however, we want our statistical methods to provide a more direct guide for decision making. In such cases, confidence interval estimation may not be the most suitable form in which to present the statistical information. Hypothesis testing provides a useful alternative.

Hypothesis testing rests on a simple logical foundation, though with an important modification. Imagine we have assembled a group of people from different areas of the world. We want to make some determination of their origin by inquiring about their local weather. Consider for example the following true assertion: "In Barbados it never snows." From this we can deduce that if someone claims to have seen snow fall in his homeland, he cannot be from Barbados.

We can formulate this reasoning abstractly. We are given that a statement **S** of the form "If *H* occurs then *C* is impossible" is true. In our example, *H* would be the statement that the individual comes from Barbados. The condition *C* would be that you see snow fall. From the truth of **S** we can infer that the occurrence of *C* implies that *H* must be false, since otherwise we would never have seen event *C*.. We can use this reasoning to create a decision procedure. Whenever outcome *C* is observed, we will reject hypothesis *H,* thus asserting it to be a false hypothesis. From our assumptions, whenever the hypothesis *H* is actually true, this decision process can never falsely reject it, since outcome *C* will never be observed when *H* is actually true. In other words, when *H* is true there is zero chance of erroneously rejecting it.

On the other hand, when *H* is false (i.e., in our example the unknown location is not Barbados) we would like our decision procedure to reject *H* with high probability. To actually assess that probability we need more information regarding possible alternatives. For example, if we knew that many of the people in our group were from N.Y. then they would have likely observed snow and the procedure would have a high probability of correctly rejecting Barbados as the unknown location. We say that the test has a high *power* to distinguish between the original hypothesis and this particular alternative. On the other hand, if a plausible alternative original location were Jamaica, then there is little chance that the decision process will lead you to reject the Barbados hypothesis, since it never snows on Jamaica either. In this case, the test has little power to distinguish the two alternatives and if that distinction is important, this is a poor test to use.

Actual statistical hypothesis testing differs in one important respect from the simplified scenario described above. In most cases, when the hypothesis *H* is true, the outcome *C* whose occurrence is the basis for the rejection decision actually has a small chance of occurring. In other words, there will be a small chance, called the *significance level*, of erroneously rejecting a true hypothesis. Hypothesis testing is concerned with designing tests that have small significance level and also a high power to discriminate between the assumed hypothesis and plausible alternatives. We will now look more closely at how this is done for a number of common statistical scenarios.

16.2 General Notions

We will designate the hypothesis we wish to examine by the letter H_0. H_0 is called the *null hypothesis*. The name arises because often H_0 simply asserts that there is no difference between two competing claims. We attempt to test the validity of the hypothesis H_0 by performing an experiment whose outcome *C* is unlikely to occur when H_0 is true. We use this outcome as a decision criterion.

Definition 16.1: We say that outcome *C* is a *decision criterion* for hypothesis H_0 if the occurrence of *C* leads us to reject hypothesis H_0. When outcome *C* is not observed we do not reject hypothesis H_0.■

Since outcome *C* has a small chance of occurring when H_0 is true, the rejection of the hypothesis cannot be interpreted as an unequivocal assertion that H_0 is false. Rather, we are expressing a judgment based on a certain weight of evidence. Two numbers measure the quality of this evidence and the value of *C* as a decision criterion. We want a small probability of rejecting a true hypothesis and a high probability of rejecting a false one. We have already introduced these numbers informally and we now state more precise definitions.

Definition 16.2: The *significance level* of the decision criterion *C* for hypothesis H_0 is the probability that *C* will occur when H_0 is true. We denote the significance level by α (Greek letter “alpha”). ■

On the other hand, we are also concerned that when H_0 is false and some alternative H_a is true, the outcome *C* should occur with high probability, so that the false hypothesis will be rejected.

Definition 16.3: The *power* of the decision criterion *C* for hypothesis H_0 versus an alternative hypothesis H_a is the probability that *C* will occur when H_a is true. We denote the power of the test (for a specific alternative) by γ (Greek letter “gamma”). ■

Let’s compute these quantities in a simple example.

Example 16.1: Suppose we have a coin and we want to know whether the coin is fair, i.e. whether there is a 50% chance that it will land heads. Our decision criterion will be based on the number of heads obtained when the coin is tossed 25 times. Since a fair coin is unlikely to produce a large or small number of heads, we take as our decision criterion the event *C* that the number of heads is ≥ 17 or ≤ 8. Thus, if *C* occurs we will reject the null-hypothesis H_0 that the probability of heads is 0.5. Find

a) The significance level α of this decision criterion and

b) the power of this decision criterion to detect a coin with $p = 0.6$ and one with $p = .8$.

Solution:

a) The significance level is the probability that the decision criterion *C* occurs when the coin is fair. We can easily find this from the tables for the binomial distribution. Indeed from section B.2 we find that with $n = 25$ trials the probability of *C* is

$$P(\#\text{ heads} \le 8) + P(\#\text{ heads} \ge 17) = .054 + .054 = .108\,.$$

Hence, $\alpha = .108$. This means that the outcome *C* has about a 10% chance of occurring, even for a fair coin. Since the occurrence of *C* leads us to reject the null hypothesis, about 10.8% of fair coins would be incorrectly rejected by this procedure.

b) The power of this test to correctly reject the null hypothesis when it is false depends on the alternative. If the true probability of heads for the coin were $p = 0.6$ then from section B.2 we find that the likelihood of the decision criterion *C* occurring is

$$P(\#\text{ heads} \le 8\,|_{p=.6}) + P(\#\text{ heads} \ge 17\,|_{p=.6}) = .004 + .274 = .278\,.$$

If the true *p* is 0.8 then a similar calculation yields a power of $\gamma = 0.953$. Thus the test is extremely likely to detect very biased coins, but much less likely to rule out moderately biased ones. ■

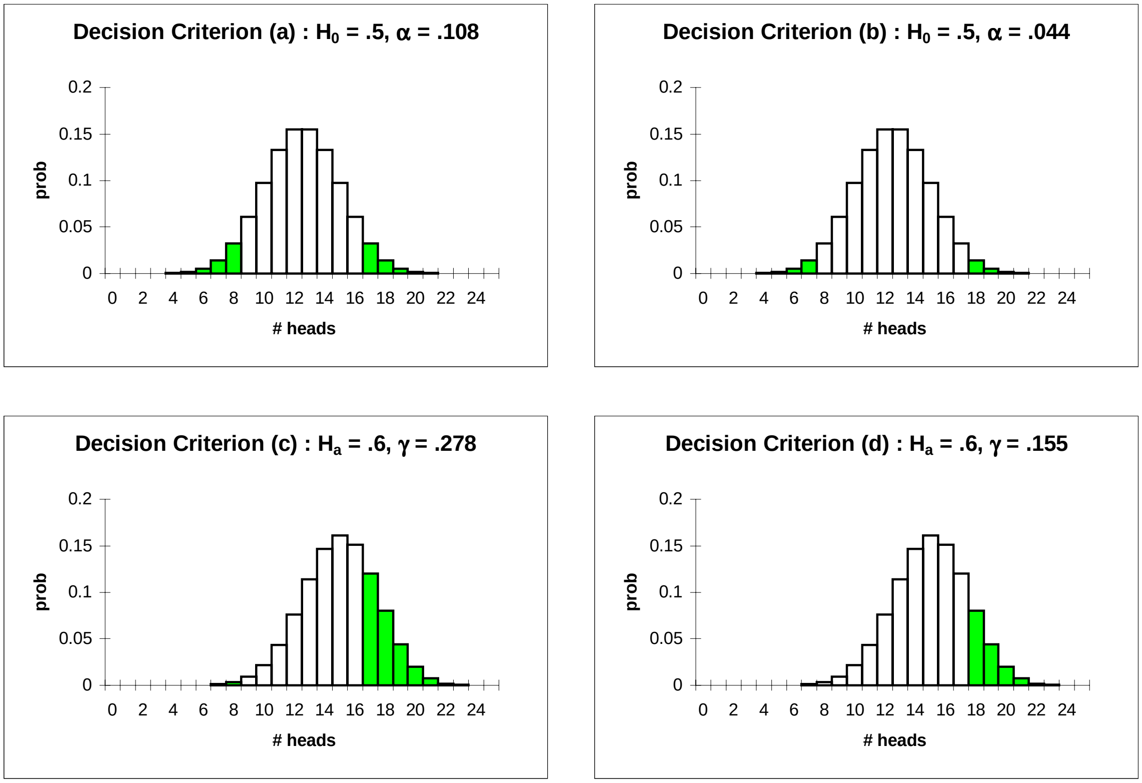
Some comments are in order regarding the decision criterion and the two probabilities associated with it.

- *Nature of the decision criterion*:

Since we have no prior knowledge indicating whether a biased coin will produce an excess or a deficit of heads, we must adopt a decision criterion that is sensitive to either large positive or large negative deviations from the hypothesized 50/50 balance. Such a statistical test is called *two-sided*. In some situations a decision criterion will only be triggered by exceptional outcomes in one direction from the assumed hypothetical state. Such tests are therefore called *one-sided*. The principles for constructing these are very similar to those required in two-sided tests, though it is not always a clear-cut decision as to which is appropriate.

- *Selecting the decision criterion:*

Since we want the decision criterion *C* to have a low probability of occurrence when the null hypothesis H_0 is true, the outcomes we select for *C* must come from the tail(s) of the distribution for the number of heads. The farther out in the tails we select the outcomes in *C*, the smaller α, which of course is desirable. Unfortunately, the same process will reduce the power of the test, for any given alternative. This is illustrated in the figure below. The top panels (a) and (b) show the probability distribution of heads for 25 tosses of a fair coin. In each case, the shaded region specifies a decision criterion *C*. In the bottom panel we show the distribution of heads for 25 tosses of a moderately biased coin ($p = .6$). The shaded regions in (c) and (d) use the same decision criteria as (a) and (b), respectively. As we move from (a) to (b) in the top panels α decreases. At the same time, the area of the shaded regions in the lower panels also decreases, implying a decrease in the power to detect this specific alternative.



Decision Criterion (a) : H_0 = .5, α = .108

Decision Criterion (b) : H_0 = .5, α = .044

Decision Criterion (c) : H_a = .6, γ = .278

Decision Criterion (d) : H_a = .6, γ = .155

Figure 16.1

- *Meaning of* α*:*

In Example 16.1 we found that the significance level of the decision criterion was 0.108**.** ***The significance level is often misinterpreted. Here are two common misconceptions***:

Misconception i) *Hypothesis tests having, say,* $\alpha = 0.1$ *reject about 10% of all null hypotheses.* This is only true when examining known true hypotheses. If the test of Example 16.1 were applied to a collection of fair coins we would indeed reject about 10% of these. However, if all of the tested coins had a probability of heads of $p = 0.8$ then we have seen that 95% of them will be rejected. In general, we will not know the mix of true and false null hypotheses that we are examining so the exact percentage that will be rejected cannot be determined.

Misconception ii) *The significance level measures the chance that your decision is incorrect.* Again, this is only true provided we add the words, "when examining true null hypotheses." For example, if we are examining 1000 coins all of which have a 0.6 probability of tossing a head then, according to Example 16.1, we will correctly reject about 278 of these. The remaining 722 will not be rejected, though they should have been. If we regard non-rejection as a decision, then the rate of incorrect decisions would be 72%. In general, if we are testing coins with a specific alternative bias denoted by H_a, then the fraction of incorrect decisions will be $1-\gamma$, where γ is the power of the decision criterion when the alternative hypothesis H_a is valid.

Although the significance level is a well-defined probability, it measures a probability that has no practical interpretation. After all, if one knows that a certain null hypothesis is true, one would never bother to test it! From the perspective of hypothesis testing the significance level provides a quantitative means of weighing the statistical importance of the experimental results. ***Indeed, the smaller the significance level then the greater is the weight to be accorded a result that leads to the rejection of*** H_0**.** For example, when we use a test with significance level $\alpha = .05$ the decision criterion *C* has a 5% chance of occurring when H_0 is true. Since we may have other grounds for believing that a particular H_0 is true, in practice, we may sometimes regard a statistical decision to reject H_0 with $\alpha = 0.05$ as insufficient evidence against the proposed model. On the other hand, when our experiment results in an outcome *C* with a significance level of 0.01, then in continuing to accept H_0 you would need to argue against an event that rarely occurs, but which has now been seen. If there is compelling evidence to the contrary one may occasionally dismiss such a statistical result as a coincidence, but to do so consistently is irrational.

If there is a need for greater moral certainty ("guilt beyond a reasonable doubt") then we can adopt a more stringent decision criterion with a smaller α. The price for this will be a decrease in the power of the test to detect differences from the hypothesized state. Unfortunately, such differences may be of great practical interest. It is possible to have a significance test with a small value of α and high power to detect important alternatives. As might be expected, though, such tests require more information about the population that we are examining. In section 16.7 we will see how to construct tests meeting such dual objectives.

16.3 Tests for Proportions

Example 16.1 illustrated the computation of the significance level and power for a simple test for a proportion. In that example we tested a coin by tossing it 25 times. When using so few repetitions we observed that the typical two-sided test with a significance level about 0.1 has very low power for alternatives, such as $p = .6$ in the example, that one might want to detect. For that reason, hypothesis tests for an unknown proportion that use a small number of trials are usually done with higher significance levels. More ideally one would like to perform such tests with larger sample sizes. We will therefore describe the construction of these tests for an unknown proportion *p* when the samples are large enough for us to apply the normal approximation to the binomial distribution. In this section we will concentrate on the construction of tests with a specified significance level, postponing the computation of the power to section 16.7

Example 16.2: Suppose we wish to evaluate the effectiveness of a new drug compared to an old one that is known to be 30% effective in treating a certain ailment. The new drug is administered to 75 randomly selected patients and it is found to be effective for 32 of them. Would this be sufficient evidence to reject the hypothesis that the drug is only as effective as the old medication? Give your response using both 0.05 and 0.01 significance levels.

Solution:

Notice that rather than specifying a decision criterion, we have stated significance levels with which we wish to operate. The levels selected are conventional. If the corresponding decision criteria are met, we speak of the resulting rejections as *significant* ($\alpha = 0.05$), respectively, *highly significant* ($\alpha = 0.01$).

Our null hypothesis H_0 states that the new drug has the same effectiveness as the established treatment. Denoting the effectiveness of the new drug by *p* this can be written as $H_0 : p = .30$, versus the alternative $H_a : p > .3$ or $p < .3$. If our experiment rejects the null hypothesis we ordinarily report whether the result was indicative of an improvement ($p > .3$) or a worsening ($p < .3$) compared to the standard treatment

A natural candidate to use for testing the null hypothesis is the fraction of tested patients for whom the new drug was found effective. We introduced this quantity in Chapter 15, as the random variable $\hat{p}$. Recall that by Theorem 15.4 when *n*, the sample size, exceeds approximately 30 and the trials are conducted independently, $\hat{p}$ has a normal distribution with mean *p* and standard deviation $\sqrt{\frac{pq}{n}}$. Our decision criterion should reject the null hypothesis when the sample frequency $\hat{p}$ is significantly larger or smaller than the hypothesized value of 0.3. It is customary, (though not logically necessary) to select cutoff values for $\hat{p}$ that are symmetric around the hypothesized value $p = .3$. In other words, the decision region *C* should encompass the shaded portions in the probability histogram of $\hat{p}$ shown below.

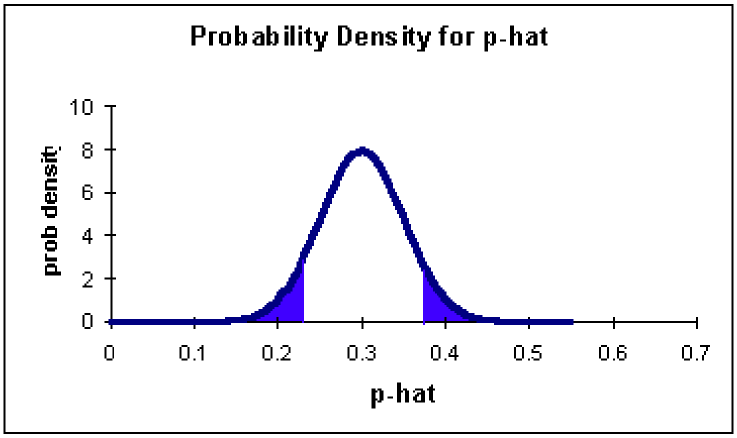


Figure 16.2

If we want $\alpha = .05$ then the area of the shaded regions should total .05. Since the regions are symmetric, the total area below the right cutoff value should be 0.975. From Table B.3 for the standard normal distribution, the right cutoff point must be 1.96 standard deviations above the hypothesized mean $p = .3$ and the left cutoff 1.96 standard deviations below the mean. Thus, describing our test using *z* values we have:

$H_0: \ p = 0.3$ versus. $H_a : p \neq 0.3$

Test: With $\alpha = 0.05$ reject H_0 if $z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{\hat{p} - .3}{\sqrt{\frac{(.3)(.7)}{75}}} > 1.96$ or $z < -1.96$.

Returning to the data of the example, for the new drug there were 32 effective treatments out of 75 trials, giving a value of $\hat{p} = \frac{32}{75} \approx .43$. The associated *z* value is $z = \frac{.43 - p}{\sqrt{\frac{pq}{n}}} = \frac{.43 - .3}{\sqrt{\frac{(.3)(.7)}{75}}} = 2.39$. Since this value exceeds 1.96, the observed frequency is large enough for us to reject the null hypothesis at the 0.05 significance level and to conclude that the true *p* is greater than 30%.

To perform a two-sided test at the 0.01 significance level we have only to determine the *z* value that cuts off a total area of 0.01 in the two tails or equivalently, total area .995 below the right tail. Using Table B.3 again we obtain the following formulation of the test:

$H_0: \ p = 0.3$ versus. $H_a : p \neq 0.3$

Test: With $\alpha = 0.01$ reject H_0 if $z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{\hat{p} - .3}{\sqrt{\frac{(.3)(.7)}{75}}} > 2.57$ or $z < -2.57$.

We have already computed that $z = 2.39$, which is not large enough to reject the null hypothesis at the 0.01 significance level. By this standard, the observed result (43% successful treatment rate in the sample) is within the realm of acceptable variation from the hypothesized value, which therefore cannot be rejected. ■

Example 16.2 illustrates a common feature of hypothesis testing. Adopting a more stringent standard of proof by lowering the significance level will often change a finding from significant to non-significant. Raising the significance level of course has the opposite effect. Setting the level too high will certify outcomes as significant that are more likely to be the result of chance variation. Too small a significance level will prevent us from seeing real effects that may be meaningful. Let us summarize the procedure used in Example 16.2 for the two-sided hypothesis for proportions.

Rule 16.1 (Two-sided hypothesis test for an unknown proportion): To test the null hypothesis $H_0 : p = p_0$ vs. an alternative $H_a : p \neq p_0$ at a significance level α using a random sample of size $n > 30$ ::

a) Use the table of the *Z* distribution to find the value $z_{\alpha/2}$ with the property that the area in the right tail, i.e. beyond $z_{\alpha/2}$, is $\alpha/2$. (For example with $\alpha = .05$, we have $z_{.025} = 1.96$, since $P(Z > 1.96) = .025$.)

b) Compute the *z* value for the observed proportion $\hat{p}$, assuming the true *p* is p_0, in other words compute $z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$. (Remember, *z* expresses the number of standard deviations of the observed value $\hat{p}$ from its hypothesized expected value of p_0.)

c) **Reject** H_0 if the value of *z* found in b) exceeds $z_{\alpha/2}$ or is smaller than $-z_{\alpha/2}$. ■

16.4 Tests for Means

The previous section laid out the principles of hypothesis testing for proportions when using large samples. Although the sample frequency $\hat{p}$ provides the ultimate justification for the decision we reach, the decision criterion in practice relies on the computation and value of *z*, as specified in Rule 16.1. In this context we call *z* a *test statistic*.

Fortunately, the ideas elaborated in section 16.3 apply with only slight modifications to many other hypotheses testing situations. In this section we consider hypothesis tests as they apply to means, examining first the case of large sample tests, followed by a more general discussion relating to arbitrary sample sizes.

Example 16.3: Prior to the introduction of a new production process, the lightbulbs produced in a certain factory had an average lifetime of 750 hours. The quality control department tests 100 randomly selected bulbs produced using the new process. The tested bulbs lasted on average 740 hours with a standard deviation of 35 hours. At the 0.05 significance level, is this evidence sufficient to conclude that there has been a change in bulb quality?

Solution:

The null hypothesis H_0 states that there has been no change in mean bulb life, which remains 750 hours. We wish to assess whether the sample indicates evidence of a significant change from this value. Thus,

Null Hypothesis: $H_0 : \mu = 750$ vs. $H_a : \mu \neq 750$.

Intuitively, we would reject the null hypothesis when the sample mean $\overline{x}_{100}$ falls too far below or above the hypothesized value. Since the sample mean has an approximately normal distribution for large ($n > 30$) samples, we can use a *z* statistic to assess whether $\overline{x}_{100}$ is too far from the hypothetical value of 750. Thus

Test statistic: $z = \dfrac{\overline{x} - 750}{\frac{\sigma}{\sqrt{n}}}$, where $\overline{x}$ is the sample average for the tested bulbs, σ is the standard deviation for the entire population, and *n* is the sample size. When *n* is large and σ is not known we usually replace it by the sample standard deviation *s*.

What values of *z* will lead us to reject the null hypothesis? This is determined by the significance level. In order to attain a 0.05 the significance level, the total area in the rejection region must be 0.05. Selecting the rejection region to be symmetric, we use the value $z_{.025}$ described in Rule 16.1a). From table B.3 we obtain the value $z_{.025} = 1.96$. Thus

Decision Criterion: Reject H_0 if $z < -1.96$ or $z > 1.96$.

In this case we find using the values $\overline{x} = 740$, $s = 35$, and $n = 100$ that $z = \dfrac{740 - 750}{\frac{35}{\sqrt{100}}} = -\dfrac{10}{3.5} = -2.86$. Since this value is below the cutoff point of $z = -1.96$ we reject the null hypothesis and conclude that the new process produces bulbs with a shorter average life. ■

As in section 16.3 we can formulate the work in the example as a general procedure.

Rule 16.2: **(Two-sided large sample hypothesis test for a mean):** To test the null hypothesis $H_0 : \mu = \mu_0$ vs. the alternative $H_0 : \mu \neq \mu_0$ at a significance level α with a random sample of size $n > 30$:

a) Find the value $z_{\alpha/2}$ such that the $P(Z > z_{\alpha/2}) = \alpha/2$.

b) Compute the *z* value for the sample mean $\overline{x}$, assuming the population mean is μ_0. Use the expression $z = \dfrac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$, if σ is known or if σ is unknown, $z = \dfrac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$, where *s* is the standard deviation of the sample.

c) **Reject** the null hypothesis if the value *z* found in b) is less than $-z_{\alpha/2}$ or greater than $z_{\alpha/2}$.■

Rule 16.2 requires that we use a large sample to test the hypothesis. Under some mild restrictions we can use small samples, but the *z* statistic is only appropriate when the value of σ is known. When we substitute for the usually unknown σ the sample standard deviation *s* we can no longer use normal distributions, but rather the more general Student-*t* distributions discussed in section 15.4. The details are given in the Rule 16.3 and illustrated in Example 16.4.

Rule 16.3: (Small sample two-sided test hypothesis test for mean) Suppose a continuous random variable *X* has an approximately normal distribution. To test the null hypothesis $H_0 : \mu_X = \mu_0$ versus the alternative $H_a : \mu_X \neq \mu_0$ with significance level α using the average $\overline{x}$ from a random sample of size $n < 30$

a) Use Table B.5 to find the value $t_{f,1-\alpha}$, with $f = n-1$degrees of freedom, having the property that the central region under the Student density curve from $-t_{f,1-\alpha}$ to $t_{f,1-\alpha}$ has area $1-\alpha$. (The table limits the choice of α to 0.1, 0.05, 0.02, 0.01.)

b) Compute the *t* statistic for the observed sample mean $\overline{x}$. In other words, compute $t = \dfrac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}$, where *s* is the sample standard deviation.

c) **Reject** H_0 if the value of *t* found in b) is larger than $t_{f,1-\alpha}$ or smaller than $-t_{f,1-\alpha}$ ■.

Note the procedure is applicable only for testing the mean of a random quantity with an approximately normal or at least mound shaped distribution. Thus, we cannot use the Student distribution to perform small sample testing for the mean of a highly skewed distribution.

The value $t_{f,1-\alpha}$ found in a) determines the cutoff values for the test. To find this quantity we use a Student distribution with $f = n-1$ degrees of freedom. Roughly speaking, the number of degrees of freedom is related to the number of “independent” values that appear in the expression for the sample standard deviation *s*. For example, when we compute the standard deviation of three data values x_1, x_2 and x_3, we first find the sample variance s^2 using

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + (x_3 - \overline{x})^2}{2}.$$

Since $\overline{x} = x_1 + x_2 + x_3$, the three quantities in the numerator satisfy $(x_1 - \overline{x}) + (x_2 - \overline{x}) + (x_3 - \overline{x}) = 0$. Hence, one of these quantities may be computed from the other two. It is this fact that leads to the value $f = 2$ in this case and more generally $f = n - 1$.

Example 16.4: Suppose in Example 16.3 that only 15 randomly selected bulbs had been tested with the results as given ($\overline{x} = 740$ hours and $s = 35$ hours.) Can we conclude at the 0.05 significance level that the new production process produces bulbs with a shorter lifetime?

Solution:

We follow the steps given in Rule 16.3, using as our null hypothesis $H_0 : \mu = 750$ with a two-sided rejection region. Since our sample size is 15 we have $f = 14$ degrees of freedom. From table B.5 we obtain that $t_{14,1-\alpha} = t_{14,0.95} = 2.14$. Next we compute the value of *t* associated with the sample mean. We have

$$t = \frac{740 - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{740 - 750}{\frac{35}{\sqrt{15}}} \approx -1.11.$$

Since the value of *t* is in the interval $[-2.14, 2.14]$ there is insufficient evidence to reject the null hypothesis. The sampling has revealed no statistically significant change in the average life of the bulbs. ■

Why does the same sample average produce different conclusions in Example 16.3 and Example 16.4? Using a smaller sample, we expect larger variation in the sample average. Hence, a more deviant sample average is needed for us to believe that something is amiss. In the example, an average of 740 hours for a sample of size 15 is not extraordinary if the true mean is 750 hours. However, for a sample of size 100 such a mean (with the given value of *s*) would be highly unlikely. The precise testing procedures described above enable us to quantify these intuitive observations.

16.5 P-Values

In our discussion of hypothesis testing we have usually chosen to work with significance levels of 0.01 and 0.05. These are of course purely conventional and could just as well have been chosen as

0.008 or 0.06. The size of these conventional values has some basis in human psychology. For example, people begin to get suspicious about the "fairness" of a coin when it produces heads in five consecutive tosses (probability ≈ 0.03). However, the choice of conventional significance levels also derives from the fact that the computation of cutoff values relies in many cases on lengthy calculations (for instance for Student's *t*.) Computations by the statistical user were therefore limited by the information available in standard tables. Naturally, this led to various conventions on the choice of significance levels.

With the availability of software that can instantly compute probability values for even complicated distributions, there is no longer a need to impose artificial conventions as to what significance levels are appropriate. Rather, the researcher can report the smallest significance level for which the observed outcome would lead to the rejection of the null hypothesis. The precise concept is known as the *P*-value of the test. (The "*P*" stands for probability.)

Definition 16.4: (*P*-value) Suppose a hypothesis test uses a value of some random variable *X* (the test statistic) as a rejection criterion for a null hypothesis H_0 and we observe the value *x* of this test statistic. The *P-value* of the outcome is the probability that, when the null hypothesis is true, the random variable *X* will produce a value as extreme as the one observed. ■

We clarify the meaning of Definition 16.4 by finding the *P*-values for the experiments reported in Example 16.3 and Example 16.4.

Example 16.5: Find the *P*-values for the outcomes observed in the light bulb tests described in Example 16.3 and Example 16.4**.**

Solution:

In Example 16.3, which uses a large sample test, we use as a test statistic the *z* score of the sample mean, assuming the null hypothesis is true.

We have seen in Chapter 15 (Theorem 15.3) that the sample mean for large samples has a normal distribution. The *z* score of the sample mean is our test statistic. In Example 16.3 we found $z = -2.86$. According to Definition 16.4 the *P*-value is the probability of obtaining a result as or more extreme than this. Since we are dealing with a two-sided test that rejects the null hypothesis when we obtain either large positive or negative values of *z*, the *P*-value is $P(|Z| \geq 2.86) \approx 0.004$. This number means that if we perform a hypothesis test with an $\alpha > 0.004$, the null hypothesis will be rejected by the given observation. In effect, when $\alpha > 0.004$ the cutoff value for *z* has a tail area $\frac{\alpha}{2} > 0.002$, which is the tail area corresponding to $z = 2.86$. In order for the tail area to be larger than 0.002, the cutoff value must become ***smaller*** than 2.86, so the given observation will then be significant at the level α. In particular, the observation will be significant for $\alpha = 0.05$, as we have shown directly in Example 16.3.

For Example 16.4 we must use a software package to compute the *P*-value. In that example we used a Student-*t* with $f = 14$ degrees of freedom as the test statistic. The value for the Student random variable was $t_{14} = -1.11$. As we are doing a two-sided test, we need account for possibly positive or negative deviations about the hypothesized value. Thus we must find the $P(|t_{14}|) > 1.11$. In *Excel*, using the command *tdist()* (see Section 15.4), we find that $P(|t_{14}|) > 1.11 \approx 0.285$. Therefore, the smallest significance level for which the null hypothesis would be rejected by the observed outcome would be $\alpha = 0.285$. In particular, we would not have grounds to reject the null hypothesis at the 0.05 significance level. ■

Reporting a *P*-value instead of using a conventional significance level allows the reader to set his or her own standard with which to weigh the evidence. Extremely small values of *P*, say $P < 0.001$, provide a clear signal that the null hypothesis should be rejected. Chance effects alone are unlikely to produce such an improbable result. A *P*-value in excess of 0.2 provides no compelling reason to doubt the null hypothesis. Between these extremes is a large gray area where individual interpretations vary. From a practical viewpoint, for example, a result with a *P*-value of 0.054 should probably be accorded the same weight as one that is significant at the conventional 0.05 level. The *P*-value methodology allows us to make such distinctions and is the preferred way of describing the outcome of a test of significance

16.6 Comparisons

In Sections 16.3 and 16.4 we discussed tests that are used to examine whether some unknown proportion or mean has a specified value. In many applications, for example comparing the effectiveness of two treatments, it is more usual to test both treatments and to compare the outcomes to each other. We will now discuss the framework for doing this.

Suppose for instance that we are interested in studying the response of patients to two different treatments. The subjects are divided into two groups, ideally through some random allocation. Participants in one group receive treatment A, the others treatment B. Our purpose in doing the experiment is to see if there is evidence to believe there is a difference in the underlying frequencies of successful treatment for the population as a whole, p_A and p_B. When the experiment terminates we determine the fraction of patients who responded successfully to each treatment. These frequencies will be denoted by $\hat{p}_A$ and $\hat{p}_B$ respectively. Rule 16.4 below describes the steps needed to carry out a large sample hypothesis test in this situation:

Rule 16.4: (Large Sample Comparison of Proportions): Suppose p_A and p_B are unknown success frequencies for two treatments. Suppose two independent (large) random samples of sizes n_A and n_B yield x_A and x_B successes respectively with corresponding frequencies of $\hat{p}_A$ and $\hat{p}_B$. To test the hypothesis $H_0 : p_A = p_B$ versus the alternative that $p_A \neq p_B$ at a significance level α carry out the following steps:

a) Find the value $z_{\alpha/2}$ such that the $P(Z > z_{\alpha/2}) = \alpha/2$. (see Rule 16.1a)

b) Compute the *z* score defined by $z = \dfrac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p}\hat{q}(\frac{1}{n_A} + \frac{1}{n_B})}}$ where $\hat{p} = \dfrac{x_A + x_B}{n_A + n_B}$ and $\hat{q} = 1 - \hat{p}$.

c) If $z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$ reject the null hypothesis and conclude that $p_A \neq p_B$. ■

Steps a) and c) in Rule 16.4 are familiar from our earlier discussion. Step b) is somewhat mysterious and forbidding. The following paragraphs provide an explanation, but the reader who wishes may skip these comments and proceed to Example 16.6 for a concrete illustration.

We know that when n_A and n_B are large the sample frequencies $\hat{p}_A$ and $\hat{p}_B$ are each normally distributed, with expected values p_A and p_B, and variances $\sigma^2_{\hat{p}_A} = \dfrac{p_A q_A}{n_A}$ and $\sigma^2_{\hat{p}_B} = \dfrac{p_B q_B}{n_B}$, respectively. Our null hypothesis asserts that $p_A = p_B$. If we are looking for evidence to refute this we should consider the difference between the sample relative frequencies $\hat{p}_A - \hat{p}_B$. When this difference deviates substantially from zero it is clearly evidence <u>against</u> H_0, but we need to know how large a deviation is significant.

Fortunately, not only do we have the specific facts listed above regarding the individual random variables $\hat{p}_A$ and $\hat{p}_B$, but similar results hold for the difference $\hat{p}_A - \hat{p}_B$. Specifically, when both n_A and n_B are large (and $n_A p_A$, $n_A(1 - p_A)$, $n_B p_B$, and $n_B(1 - p_B)$ are all bigger than 5) then

- $\hat{p}_A - \hat{p}_B$ has a normal distribution.
- The expected value of $\hat{p}_A - \hat{p}_B$ is $p_A - p_B$.
- The variance of $\hat{p}_A - \hat{p}_B$ is the <u>sum</u> of the variances of $\hat{p}_A$ and $\hat{p}_B$. In other words, the standard deviation for $\hat{p}_A - \hat{p}_B$ is $\sqrt{\dfrac{p_A q_A}{n_A} + \dfrac{p_B q_B}{n_B}}$.

Now let's add to this mix the additional assumption of the null hypothesis that $p_A - p_B = 0$. Based on the usual procedures for a normal distribution, the *z* score for $\hat{p}_A - \hat{p}_B$ is

$$z = \frac{\hat{p}_A - \hat{p}_B - (p_A - p_B)}{\sqrt{\frac{p_A q_A}{n_A} + \frac{p_B q_B}{n_B}}} = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\frac{p_A q_A}{n_A} + \frac{p_B q_B}{n_B}}}, \qquad (16.1)$$

since $p_A - p_B = 0$ by hypothesis. Of course the denominator in the last expression contains the population parameters p_A and p_B which are unknown. (Note that unlike Rule 16.1 the null hypothesis here provides no assumption about the individual values of p_A and p_B.) We can simply replace the values of p_A and p_B by the sample frequencies $\hat{p}_A$ and $\hat{p}_B$, but a better estimate can be made using the null hypothesis. Namely, under the null hypothesis $p_A = p_B$ we can view both samples as having been drawn from a population with the same underlying frequency of success *p*. We can estimate this common frequency by pooling or combining the results from the two treatments. Altogether there were $x_1 + x_2$ successes out of a total sample of $n_A + n_B$. Thus a reasonable estimate for the hypothesized common value of p_A and p_B is

$$\hat{p} = \frac{x_A + x_B}{n_A + n_B}. \qquad (16.2)$$

If we replace p_A and p_B in the right side of (16.1) by the formula for $\hat{p}$ in (16.2) and set both q_A and q_B equal to $1 - \hat{p}$ we obtain the formula for *z* given in Rule 16.4.

Example 16.6: A drug company wishes to test a new antidepressant medication versus a placebo, a harmless treatment that has no active medical ingredient. Patients with similar diagnoses are randomly selected to either receive the treatment or the placebo. 150 patients received the active treatment and 130 (the so-called control group) received the placebo. After 4 weeks a team of psychiatrists conducted standard evaluations of the patients. In the group receiving the medication, 70 patients showed marked improvements according to the psychiatric evaluation while in the control group 47 patients received a similar evaluation

i) Does the trial provide sufficient evidence at the 0.05 significance level to conclude that the new medication is more effective than a placebo.

ii) What is the *P*-value for the outcome of the trial?

Solution:

i) Let p_A denote the frequency of successful treatments using the medication in a population of depressed patients. Similarly, let p_B denote the success rate in such a population of using a placebo treatment. Our null hypothesis H_0 is that $p_A = p_B$. We will use the data from the trial to test this hypothesis according to the steps in Rule 16.4.

a) Based on the significance level the cutoff value for the *z* score is 1.96.

b) We have $n_A = 150$, $n_B = 130$, $x_A = 70$, and $x_B = 47$. From this data we compute the other quantities needed to find the z score: $\hat{p}_A = \frac{x_A}{n_A} = \frac{70}{150} \approx .467$, $p_B = \frac{x_B}{n_B} = \frac{47}{130} \approx .361$ and $\hat{p} = \frac{x_A + x_B}{n_A + n_B} = \frac{117}{280} \approx .418$. Therefore we obtain

$$z = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p}\hat{q}(\frac{1}{n_A} + \frac{1}{n_B})}} \approx \frac{.467 - .361}{\sqrt{(.418)(.582)(\frac{1}{150} + \frac{1}{130})}} \approx 1.79$$

c) Since the observed value of z is neither larger than 1.96 nor smaller than -1.96, there is insufficient evidence to reject the null hypothesis. The medication may in fact be no better than a placebo. Although new drug performed better than a placebo in the trial, the10 % improvement in effectiveness is not sufficiently large at this significance level to reach a definitive judgement in favor of the new drug.

ii) The *P*-value is the probability that under the null hypothesis we will obtain an experimental outcome as extreme as the one obtained. Since we obtained a z value of 1.79 we need the probability $P(|Z| > 1.79)$. From the tables the latter probability is 0.073. This is the smallest significance level at which the observed result would allow us to reject the null hypothesis. ■

We can also formulate hypothesis tests comparing two means, two variances or many other statistics that can be computed for an underlying population and estimated from data. The details of these tests are all slightly different and some rely on probability distributions not covered in this text. However, the logic and mechanics of the testing are the same in all cases and the reader who needs to be acquainted with a particular method can consult any standard statistics textbook. Most standard statistical packages provide the user with templates that can be used to perform the actual computations.

16.7 Power Computations

In Definition 16.3 we introduced the power of a test. This measures the likelihood that an incorrect null hypothesis is rejected when an alternative is correct. If possible, our test should be designed so that when a plausible alternative is true there is a high probability of rejecting the null hypothesis. In this section we consider how to find the power for some of the tests considered in this chapter

Example 16.7: Suppose the new drug in Example 16.2 is actually about 40% effective in treating the illness for which it is prescribed. Using a significance level of $\alpha = .05$ and a patient pool of 75, what is the power of the test used in Example 16.2 to detect this alternative?

Solution:

The power γ is the probability that the experimental outcome will satisfy the decision criterion, if the alternative hypothesis is true. The specific alternative hypothesis we are interested in is $H_a : p = .40$. We must first cast the decision criterion in terms of $\hat{p}$ before we can compute the power of the test.

Recall from Example 16.2 that the null hypothesis will be rejected if the quotient $\frac{\hat{p}-.3}{\sqrt{\frac{(.3)(.7)}{75}}}$ is greater than 1.96 or less than -1.96. In other words we reject the null hypothesis if

$$\frac{\hat{p}-.3}{\sqrt{\frac{(.3)(.7)}{75}}} > 1.96 \text{ or } \frac{\hat{p}-.3}{\sqrt{\frac{(.3)(.7)}{75}}} < -1.96 . \qquad (16.3)$$

Solving these two inequalities for $\hat{p}$ enables us to express the decision criterion in terms of the observed frequency $\hat{p}$. We obtain

Reject H_0 if $\hat{p} > 0.404$ or $\hat{p} < 0.196$.

The power γ for the alternative hypothesis $p = 0.4$ is the probability of this rejection criterion occurring when the true *p* is 0.4. We can again use the normal distribution to compute this. We have

$$\begin{aligned}\gamma &= P(\hat{p} > .404\,|_{p=.4}) + P(\hat{p} < .196\,|_{p=.4}) \\ &= P(Z > \frac{.404-.4}{\sqrt{\frac{(.4)(.6)}{75}}}) + P(Z < \frac{.196-.4}{\sqrt{\frac{(.4)(.6)}{75}}}) \\ &= P(Z > .07) + P(Z < -3.6) \\ &= .53 + .0002 \approx .53\end{aligned}$$

Thus if there is actually an improvement in effectiveness to 40%, this has only about a 50% chance of being detected by this experiment, when testing at the commonly accepted .05 significance level. ■

When a change in treatment success, say from 30% to 40%, is not statistically significant, researchers often say that the test showed a statistically non-significant change compared with the old treatment, though there was a clinically significant improvement. ***When our experiment results in a value of*** $\hat{p}$ ***that is not large enough to reject the null hypothesis but is of possibly practical interest, we need to consider whether the test had a high enough power for the relevant alternative***.

If we wish to increase the power for a given alternative, while using the same significance level, then we must increase the size of our sample. Let's illustrate this point. Fixing the significance level, the hypothesized value of *p* and the sample size, we can compute the power of our (two-sided) test for any alternative hypothesis, exactly as we did in Example 16.7. We can plot the results as a graph, known as the *operating characteristic curve* for the test. Varying the sample size produces different curves, from which we can easily determine an appropriate sample size for alternatives that are of interest. The table below shows four such curves in which $\alpha = 0.05$, the null hypothesis is $p_0 = 0.4$, and *n* takes on values 50, 100, 150 and 200.

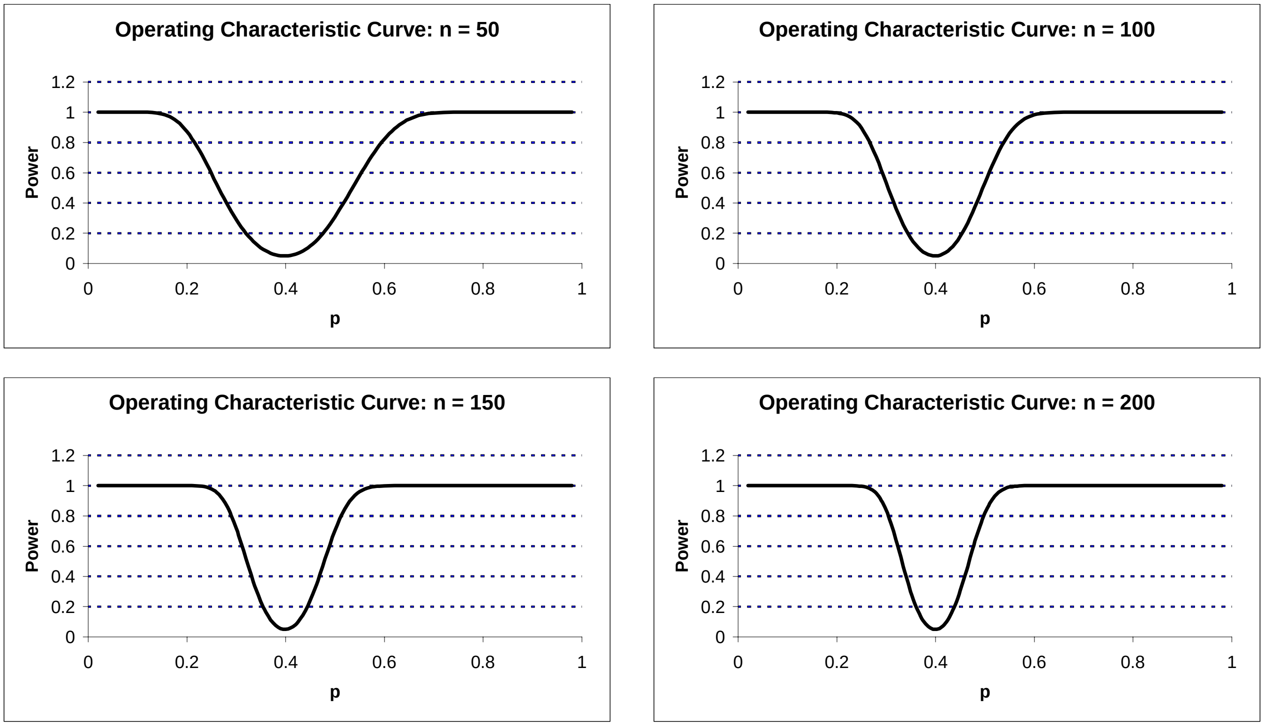


Figure 16.3

Example 16.8: Based on the Operating Characteristic Curves in Figure 16.3, approximately what power does this test have for the alternative hypothesis $p = 0.5$, as *n* varies from 50 to 200?

Solution:

Using the graphs we can estimate the power to detect the alternative hypothesis $p_a = 0.5$ for the specified values of *n*. We obtain the results tabulated below.

n	50	100	150	200
power	0.3	0.4	0.6	0.8

Thus**,** if it is important to reject the null hypothesis when the true value of $p = .5$, we should use a sample of size at least 150 or 200. Smaller sample sizes run the risk of producing non-significant results, even when the true *p* is 0.5. ■

16.8 Summary

Hypothesis testing is a methodology that uses information from a sample to make a decision regarding some aspect of a population. Typical features of the population that lend themselves to this type of analysis include assessing an unknown probability, an unknown mean or standard deviation, or comparing two such quantities. Although not discussed in this text, more general procedures known as "goodness of fit" tests can be used to decide on the suitability of a proposed probability model (distribution) for the population.

In all cases we begin by selecting a ***null hypothesis*** H_0 regarding the particular quantity of interest. We then need to find some "***test statistic***", a random quantity computed from a sample, whose values will be used to determine the outcome of the decision. In the applications discussed, the test statistic is usually a *z* score or a Student *t*. Typically, when the null hypothesis is true, the tail values of the test statistic have a small probability of occurring. When such an improbable value occurs in our sample we take that as evidence against the null hypothesis.

The weight to be accorded a given observation is measured by its ***P-value***. This is the probability of obtaining a result at least as extreme as the one actually observed. A small *P*-value indicates that, if the null hypothesis were true, we have observed an event not likely to occur by chance. We therefore have strong grounds for rejecting H_0. Such an observation is called ***significant*** or ***highly significant*** depending on the size of *P* (the smaller *P*, the more highly significant!). A large *P*-value implies that the observation has a reasonable chance of occurring, given the null hypothesis, and therefore provides no grounds for rejecting the latter. The conventional distinction between low and high *P*-values is $P = 0.05$.

When the *P*-value is high we must be careful to phrase our conclusion as a provisional "Do not reject H_0," rather than "Accept H_0." The true quantity for the population may well differ from the hypothesized value and even by a practically important amount, yet because of the sample size, the selected test may have little chance of correctly rejecting H_0. Once we have settled on a choice of *P*-value to separate rejection from non-rejection (the ***significance level*** of the test), we can compute the probability that the test will reject H_0 when it is false and some alternative of practical importance is true. The latter probability is the ***power*** of the test for the particular alternative.

16.9 Exercises

1. Suppose that the transit authority claims that 80% of its commuter trains arrive on time (meaning within 5 minutes of their scheduled arrival time). You examine this claim by randomly selecting 30 trains and determining how many arrive on time. If *p* is the probability that a train arrives on time, you wish to test the null hypothesis that $p = .80$ versus the alternative that $p \neq .8$.

   a) Formulate a rejection criterion and find the significance level for your choice.

b) What is the power of your criterion to reject the null hypothesis when the true *p* is actually 70%?

2. During a presidential campaign, published polls show the leading candidate with 60% of the popular vote in your state. You decide to test whether this percent of support is correct among people at your school. You pick 20 people at random and obtain their voting preference. Using a null hypothesis that $p = 0.60$, where *p* is the probability that a voter will support the candidate, you will perform a two-sided hypothesis test. Suppose your rejection criterion is that at most 7 (≤ 7) of the sampled voters or more than 16 (≥ 17) indicated their support for the candidate.

   a) What is the significance level of this test?

   b) What is the power of this test against the alternative that $p = 0.5$?

3. A new drug for insomniacs claims to increase nightly sleep on average by 1.25 hours. A sleep lab tests the drug on 50 volunteers and finds that the average increase in sleep was 1.6 hours with a standard deviation of 1 hour.

   a) State a null hypothesis and a test statistic.

   b) Carry out the hypothesis test using a significance level of $\alpha = 0.05$. When you perform this test what, if any, implicit assumptions are you using about the effect of the medication?

4. a) To test whether a die is fair we toss it 100 times and count the number of times a one appears. Using a significance level of 0.05 describe the appropriate null hypothesis, test statistic and rejection criterion for the test.

   b) If we obtained a one on 22 out of 100 tosses, would that be sufficient evidence to reject the null hypothesis for the test described in a)?

   c) What is the *P*-value for the outcome in which 22 ones are observed in 100 tosses?

   d) In a) we considered testing whether a die was fair by examining the frequency of one of the six possible outcomes. It might be thought that a better procedure would be to apply the same criterion developed in a) (with 0.05 significance level) to all six values. We would therefore reject the die if at least one of the frequencies falls in the appropriate rejection region. Does this test have a significance level of 0.05? Explain? Test your answer by simulating 20 replications of 100 tosses of a fair die and counting how often the null hypothesis is incorrectly rejected by the procedure we just enunciated. The file *dice.xls* or *Excel's* random number generator can be used for this purpose.

5. A tire manufacturer claims that a certain brand of tire will last on average 40,000 miles before it needs to be replaced. A consumer group wishes to check this claim by doing a test in which a mechanical device simulates the driving conditions. Ten tires are tested in this way until the tread wear indicators are visible and therefore the tires are in need of replacement. The table below gives the mileage data for the experiment:

Tire #	1	2	3	4	5
Mileage (000s)	35	37.4	36	41	33

Tire #	6	7	8	9	10
Mileage (000s)	37	41	39	38	42.8

a) Use hypothesis testing ($\alpha = 0.05$) with a suitable a null hypothesis to evaluate the manufacturer’s claim. How would you report your findings?

b) What is the *P*-value of the results obtained in the test? Explain what this number means.

6. In 1999 a psychologist reported that young women with low cholesterol levels (serum cholesterol levels below 160 mg/dl) were more likely to score high on measures of depression and anxiety than a group of comparable women with normal or elevated cholesterol. Specifically, the study found that in the low cholesterol group of 55 women a total of 21 scored high or very high on personality traits indicating they were prone to depression, while the corresponding figure for the high cholesterol group was 12 out of 65.

a) Formulate a suitable null hypothesis and determine whether the observed difference is significant at the $\alpha = 0.05$ significance level?

b) What is the exact *P*-value for the observed outcome?

c) Does the result of the study imply that low cholesterol levels are a contributing factor in depression in young women

7. Sepsis is a life-threatening reaction of the body to major infection. It often leads to death (approximately 225,000 deaths in the U.S. each year). A genetically engineered drug for the treatment of sepsis, Zovant, was tested on 1690 patients with severe sepsis at 150 medical centers. Half were given Zovant and half a placebo, in addition to the standard medical treatment for sepsis. In the group receiving Zovant, 24.7% of the patients died, while the death rate in the placebo group was 30.8 %.

a) Formulate a suitable null hypothesis and determine whether the observed difference is significant at the $\alpha = 0.05$ significance level?

b) What is the *P*-value for the observed outcome?

c) Comment on of each of the following two possible ways of reporting the result. You should address the question of accuracy as well as clarity of meaning in each statement.

i) “The trials produced an absolute reduction in risk of death of 6%.”

ii) “The trials produced a 20% reduction in the death rate.”

8. A long-term study was carried out in Japan to assess the benefits of drinking green tea on reducing the incidence of gastric (stomach) cancer. The study compared the incidence of gastric cancer per person per year for heavy consumers of green tea compared with light users. There were a total of 206 cases of gastric cancer in persons reporting a consumption of more than 500-ml of green tea per day. Altogether, this group constituted 85,299 person-years.

There were 66 cases of gastric cancer out of a total of 36,572 person-years in the group who consumed less than 100 ml of green tea per day.

a) Formulate a suitable null hypothesis and determine whether the observed difference in cancer rates is significant at the 0.05 significance level.

b) What conclusion do you draw from the study regarding the usefulness of green tea in preventing stomach cancer?

9. Fosamax is a drug that acts to prevent bone resorption and thereby inhibit or reverse osteoporosis. To test the efficacy of the drug in lowering the incidence of vertebral fractures a three-year study was conducted with 2027 patients having similar indicators of osteoporosis. 1022 patients received Fosamax and 1005 received a placebo. In the three years of the study 15.0% of the placebo group were diagnosed with a least one new vertebral fracture, while the corresponding incidence for the Fosamax patients was 7.9%.

   Formulate an appropriate null hypothesis and find the *P*-value for the observed outcome. Do the results provide support for the claim that the drug reduces the fracture rate for at risk patients?

10. Until 1991 the state of California allowed persons who had served a prison term for a violent misdemeanor to legally purchase a handgun. This was prohibited beginning in 1991. A group of researchers obtained the names of 787 violent misdemeanants who made legal handgun purchases in the period 1989-90 (the approved group). A similar group of 986 misdemeanants tried to make a purchase in 1991 but were denied by the new law (the denied group). The researchers studied the number of violent crimes committed by each group over the following three years. The results are given below:

	Number of person-years	Number of violent crimes
Approved Group $(n = 727)$	1757	174
Denied Group $(n = 927)$	2325	186

   Does the study provide evidence supporting the contention that restriction of gun sales to violent misdemeanants reduces subsequent violent crimes in that group? Justify your answer using a suitable hypothesis test.

11. a) If two hypothesis tests have significance levels α and α', where $\alpha > \alpha'$, which will have the larger power for a given alternative hypothesis? Explain (Hint: Think about the rejection regions for the two tests.)

    b) Explain why we can make the power of a test arbitrarily close to one for any alternative hypothesis by taking a large enough sample size. Why then might a very large sample size be counterproductive?

12. When there is a low incidence of a condition in a population, a large sample size is needed to attain high power to detect the effect of an intervention. For example, suppose an illness

occurs at a rate of 10 per 10,000 persons per year. A certain intervention is thought likely to reduce the incidence of the illness. A 20% reduction to a rate of 8 per 10,000 persons per year would be considered desirable.

a) Suppose 1000 persons participating in the intervention are followed for 10 years. What is the power of a test with significance level 0.05 for the alternative hypothesis that the incidence of the illness is reduced to 8 cases per 10,000 persons?

b) Repeat the calculation of power in a) but assuming 10,000 individuals in the study for 10 years.

13. Suppose we wish to test the null hypothesis: $H_0 : p = p_0$ versus the alternative $H_a : p \neq p_0$ using a significance level of 0.05. Your colleague proposes the following large sample test: With $\hat{p}$ denoting the sample frequency, construct the 95% confidence interval based on $\hat{p}$. Reject H_0 if this confidence interval does not contain the hypothesized value of *p*.

a) Explain why this is a suitable test, i.e. why it has a significance level of 0.05.

b) Show that in fact this test is equivalent to the usual test using *z* scores.