

Chapter Five

Inference for one sample

In this chapter we will learn about two important statistical techniques: confidence intervals and hypothesis tests. Confidence intervals answer the question “based on my sample, what are the plausible values for the population parameter (eg population mean)?” Hypothesis tests answer a similar question—“is parameter value μ_0 a plausible value for my population?” The mechanics of probability are used to define exactly what we mean by ‘plausible,’ and construct intervals or tests that satisfy this definition. For that reason, the results of these techniques are called statistical inferences.

Objectives

- Be able to construct and interpret confidence intervals for a population proportions and population means.
- Be able to specify the sample size needed to obtain a confidence interval of a specified width.
- Be able to read a description of a scientific question and translate it into a pair of hypotheses about population parameters.
- Be able to perform hypothesis tests for population proportions and means.
- Be able to perform the sign test, and understand when testing for a value of the population median is desirable.
- Be able to explain type I and type II errors, and compute the power of a test against a specified alternative.

5.1 Estimators and Standard Errors

(In Chapter 5 instead?) As we learned in Chapter 5, the sample mean \bar{x} and sample proportion \hat{p} can be thought of as random

variables, with a mean and standard deviation. They can also be thought of as formulae for finding a 'best guess' for the corresponding population parameters. When we look at a random variable in this light, we call it an *estimator*; \bar{x} is an estimator for the population mean μ , \hat{p} is an estimator for the population proportion p . We also have a special name for the standard deviation of an estimator: a 'standard error.'

For any sample, we can observe the value of our estimator; this observed value is called an *estimate*. However, the standard error of an estimator typically depends on population parameters, which are unobservable. For instance, the standard error of \hat{p} is $\sqrt{p(1-p)/n}$; this depends on the unknown population parameter p . We naturally substitute the observed value of \hat{p} into the expression for the standard error to obtain a guess at the value of the standard error. This guess is called the *estimated standard error*. The standard error of \bar{x} is σ/\sqrt{n} . What would you use as the estimated standard error?

Estimators and their estimated standard errors are important building blocks for constructing confidence intervals. Keep in mind we understand their properties only when the estimator is based on a random sample of the population.

5.2 Confidence Intervals

We have stated before confidence intervals contain values that are plausible for the population parameter based on the observed sample. Our estimate for the population parameter should of course be a plausible value for it. In fact, for all the confidence intervals we will study, the interval is centered at the estimate for the population parameter. A *margin of error* is then computed; values within this margin are considered plausible values. The margin of error depends on two things: the standard error of our estimator, and the level of confidence we want. The confidence level is used to compute a multiplier; the confidence interval then takes the form

$$\text{estimate} \pm \text{estimated standard error} \times \text{multiplier}.$$

We discuss how to compute the multiplier from the confidence level in detail shortly, but roughly speaking larger confidence levels require larger multipliers and result in wider intervals. As we have already learned, less variable populations or larger sample sizes will lead to smaller standard errors; this naturally results in smaller confidence intervals.

5.2.1 The multiplier and the confidence level

The percentage of randomly sampled datasets for which the confidence interval procedure results in an interval that covers the true parameter value is called the *confidence level*. The procedure for constructing a 95% confidence interval will cover the true population parameter for 95% of datasets to which it is applied. Based on this we can say that p is usually somewhere in the interval, so we consider all values within the interval plausible values for p .

The multiplier used for a certain confidence level is based on the distribution of our estimator. We have learned the sample proportion has roughly a $\text{Normal}(p, \sqrt{p(1-p)/n})$ distribution. Based on this fact, we can say that for 95% of samples, the distance between \hat{p} and p is less than $1.96 \times \sqrt{p(1-p)/n}$, using the following reasoning:

- $(\hat{p} - p)/\sqrt{p(1-p)/n}$ has a $\text{Normal}(0,1)$ distribution.
- Using a normal table or software, we can determine a value z_C such that $C\%$ of the probability of the $\text{Normal}(0,1)$ distribution lies between $\pm z$. (For $z_{0.95} = 1.96$).
- Therefore, $(\hat{p} - p)/\sqrt{p(1-p)/n}$ lies between ± 1.96 for 95% of samples, and $\hat{p} - p$ lies between $\pm 1.96 \times \sqrt{p(1-p)/n}$ for 95% of samples.
- Finally, p lies within $\hat{p} \pm 1.96 \times \sqrt{p(1-p)/n}$ for 95% of samples.

This reasoning is still approximately true when we substitute the estimated standard error $\sqrt{\hat{p}(1-\hat{p})/n}$ for $\sqrt{p(1-p)/n}$. We need not restrict ourselves to 95% confidence intervals; by changing the multiplier (1.96) we can obtain any confidence level we like. Other common confidence levels are 98% and 99%. The needed multipliers are denoted $z_{0.98}$, $z_{0.99}$ (z is used since these intervals are based on the normal distribution).

Example Dean et al (2006) reported that 25/81 mouse litters collected from high density populations showed evidence of multiple paternity. If these litters can be regarded as a random sample of all mouse litters in high density populations, what is a 95% confidence interval for the true frequency of multiple paternity?

Estimate $\hat{p} = 25/81 = 0.31$

Estimated Standard Error

$$\begin{aligned} & \sqrt{\hat{p}(1-\hat{p})/n} \\ &= \sqrt{(25/81 \times 56/81)/81} \\ &= 0.05 \end{aligned}$$

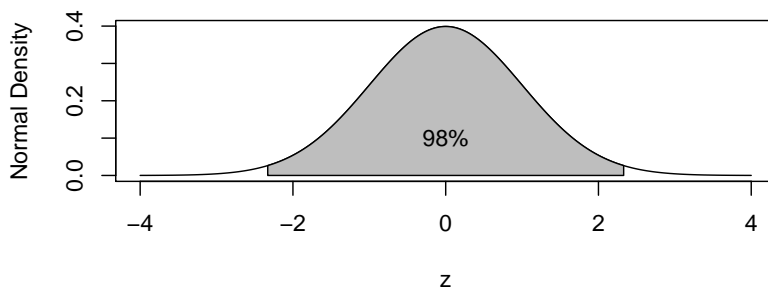
Multiplier $z_{0.95} = 1.96$

Margin of error $0.05 \times 1.96 = 0.10$

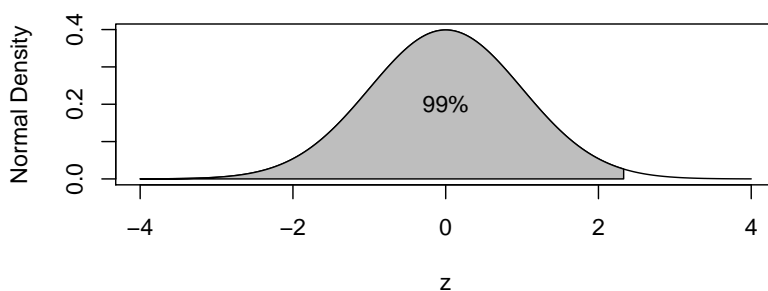
Confidence Interval (0.21, 0.41)

Example Lloyd and Frommer (2004) report that of 3000 people identified as “suspect” by a bowel cancer screening procedure at St. Vincents Hospital, Sydney, over several years, only 196 actually had bowel cancer. What is a 98% confidence interval for the proportion of people who test positive on the screening procedure who actually have bowel cancer?

First we consider how to compute the multiplier. We would like to find z such that the area under the standard normal curve between $\pm z$ is 0.98. (See Figure 5.1a). Most tables and software do not give this central probability, but rather the lower tail probability (Figure 5.2b). However, we know that the probability that we are



(a) Central 98% of the Normal density



(b) Lower 99% of the Normal density

Figure 5.1: Our confidence interval requires $z_{0.98}$, the number such that the probability a standard normal is within $\pm z_{0.98}$ is 98%; this is obtained from a table by looking up the number that gives a lower tail probability of 99%.

interested in a lower tail probability of 0.99. In other words, z is the 0.99 quantile of the standard normal distribution. Using software for computing quantiles, or a table A, we find $z = 2.33$. (To use Table A to find this value, we would search through the body of

the table to find 0.99. Then we would look at the row heading to find the first two digits, and the column heading to find the last digit.) Now we have all the building blocks necessary for our confidence interval:

Estimate $\hat{p} = 196/3000 = 0.065$

Estimated Standard Error

$$\begin{aligned} & \sqrt{\hat{p}(1 - \hat{p})/n} \\ = & \sqrt{(196/3000 \times 2804/3000)/3000} \\ = & 0.0045 \end{aligned}$$

Multiplier $z_{0.98} = 2.33$

Margin of Error $2.33 \times 0.0045 = 0.011$

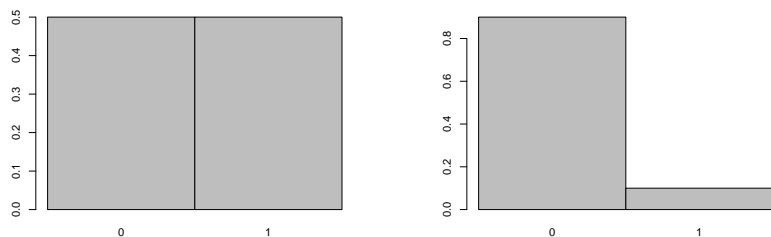
Confidence Interval (0.054, 0.076)

The distribution of \hat{p} we have used is an approximation that improves with the size of the sample. When is the sample size large enough for this confidence interval to be reasonable? A rule of thumb is $n\hat{p}$ and $n(1 - \hat{p})$ are greater than 10. Another way of saying this is that our sample for estimating p includes at least 10 successes and at least 10 failures. This incorporates the idea that the more normal looking the distribution is to start with, the fewer samples are needed to make the sample mean (in this case \hat{p}) roughly normal. If $p = 0.5$, each sample has the distribution pictured in Figure 5.2a). This does not look very normal, but at least it is symmetric around the mean, making it 'more normal' than the distribution when $p = 0.1$, pictured in Figure 5.2b. Our rule of thumb says we need about 20 samples for the approximation to be reasonable in the former case, about 100 in the latter case.

Additional corrections may improve the approximation for sample sizes that are just above those required by our 'rule of thumb'; we do not address these here, but you may notice your software package implements them and gives slightly different answers than those you obtain by hand.

5.2.2 Confidence interval for a mean

The confidence interval for a mean can be created using the same principles: \bar{x} has, approximately, a $\text{Normal}(\mu, \sigma/\sqrt{n})$ distribution, so μ will be within $1.96 \times \sigma/\sqrt{n}$ of \bar{x} for 95% of data samples. However, substituting s , the estimate of σ , for σ itself can result in an interval that has much less than the desired probability of covering μ ; the fewer data points that are available to compute s , the worse the problem is. This can be fixed by using a larger



(a) Population Histogram with $p = 0.5$ (b) Population Histogram with $p = 0.1$

Figure 5.2: Contrast two different populations, one where the distribution of a sample proportion is well estimated by a normal distribution for moderate sample size; and one which requires a larger sample.

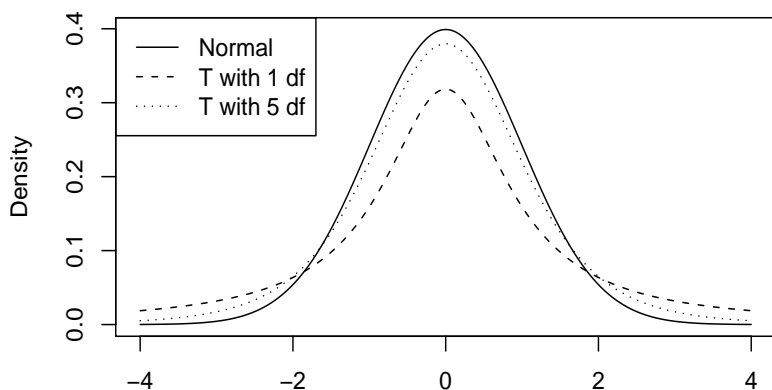


Figure 5.3: Contrast the normal distribution with two T-distributions. The T with one degree of freedom is sometimes called the Cauchy distribution.

multiplier. How large is determined by thinking about the distribution of $(\bar{x} - \mu)/(s/\sqrt{n})$. This is called the T-distribution, and its shape changes depending on n ; we say it has $n - 1$ degrees of freedom. . Using this distribution we can find bounds $\pm t$ such that $(\bar{x} - \mu)/(s/\sqrt{n})$ is within those bounds 95% (or other desired percentage) of the time; \bar{x} is then within $s/(\sqrt{n}) \times t_{0.95}$ 95% of the time. At first glance the T-distribution looks very similar to the normal, but carries more weight in the tails of the distribution. This is most pronounced in a T-distribution with 1 degree of freedom. As the degrees of freedom increase, the T-distribution looks more and more like a normal distribution (reflecting that the estimate s is closer and closer to the true parameter σ as the sample size n grows).

Example Mithas et al (2006) give summary statistics regarding the percentage of revenue invested in information systems for a sample of 487 large firms surveyed in 1999. The Average was 3.31%, with a standard deviation of 3.11%. If we regard these 487

as a random sample of large firms, what is a 95% confidence interval for the average percentage of revenue invested in information systems for this population?

Estimate 3.31%

Estimated Standard Error $3.11/\sqrt{487} = 0.14$

Multiplier 1.96, obtained from a T-distribution with 486 degrees of freedom. Note that the sample size is sufficiently large that the multiplier is essentially the same as that obtained from the normal distribution.

Margin of error $1.96 \times 0.14 = 0.28$

Confidence Interval (2.83%, 3.39%)

This confidence interval assumes that \bar{x} has (at least approximately) a normal distribution. This will be true either when the population has a normal distribution, or when the sample size is large enough so that \bar{x} has an approximately normal distribution. How large is 'large enough' for the distribution of \bar{x} to be approximately normal again depends on how non-normal the population is. If there is no extreme skewness or large outliers, a sample of size thirty is usually adequate.

5.2.3 Sample size needed to obtain confidence intervals of a certain width

Before taking our sample, we may have in mind a level of precision we wish to attain. For instance, in the bowel screening example, perhaps we would like to expand our study so that our margin of error is 0.005 or less. This would allow us to tell patients who tested positive on the screening test their probability of having cancer to the nearest percentage point. In other words, we would like

$$2.33 \times \sqrt{\hat{p}(1 - \hat{p})/n} \leq 0.005 \quad (5.1)$$

This requires an estimate for p . In this case, where we are expanding an existing study, we have such an estimate: $196/3000=0.065$. To find the minimum size necessary, we need to solve the following equation for n :

$$2.33 \times \sqrt{\frac{196}{3000} \times \frac{2804}{3000}} \times \frac{1}{\sqrt{n}} = 0.005.$$

This tells us n must be at least 13260.6. Since our sample size must be an integer, the minimum sample size to attain a margin of error of 0.005 for our 98% confidence interval is 13261.

Note the exact number is sensitive to our estimate of p . Try using 0.065 or even 0.0653 rather than 196/3000. So what do we do if we want to be certain our target margin of error is attained (rather than approximately attained), or we don't have an estimate of p to start with? We note that the left hand side of equation 6.1 is maximized for $p = 0.5$. This will provide a generous estimate of the sample size needed (suggested sample size will always be large enough). For the example above we get:

$$2.33 \times \sqrt{0.5 \times 0.5} \times \frac{1}{\sqrt{n}} = 0.005.$$

Solving n gives us 54289. This illustrates both the strength and drawback of this approach. If the true population proportion is 196/3000, this is indeed "big enough;" but unless collecting samples is effortless, we may regret using such a generous estimate.

If we would like a 95% rather than 98% confidence interval, and round the necessary multiplier to 2.0, the generous estimate takes a particularly simple form. For specified margin of error me :

$$\begin{aligned} 2 \times \sqrt{0.5 \times 0.5} \times \frac{1}{\sqrt{n}} &= me \\ \frac{1}{\sqrt{n}} &= \frac{me}{2} \\ n &= \frac{4}{me^2} \end{aligned}$$

The approximate sample size necessary to attain a certain margin of error for a population mean can be computed in a similar way. (See exercise 5). However, for means an estimate of the population standard deviation is always necessary.

5.3 Hypothesis Tests

A confidence intervals consist of a range of plausible values for a parameter. This suggests it is not possible to conclude that a parameter is exactly equal to some value—even with a large sample size, other values very close to our best guess for a parameter will also be plausible. In some situations, however, it is possible to decisively conclude that a value is *not* plausible. This is of interest in many situations. For instance, a journalist covering a two candidate election might do a survey to see if one candidate has the election 'locked up.' This would involve concluding that 0.5 is not a plausible value for the proportion of the population supporting candidate A. In this situation we do what is called a *hypothesis test*.

5.3.1 Null and alternative hypotheses

A hypothesis test is always framed in terms of two competing hypotheses. The null hypothesis is that our population parameter, say p , equals some particular value p_0 (called the null value). In our example above, the parameter is the proportion of people favoring candidate A, and the null value is 0.5. The alternative hypothesis is what we are trying to conclusively demonstrate—for the journalist, the alternative hypothesis (representing the case that the election can be ‘called’) is $p \neq 0.5$. For the campaign manager of candidate A, a different alternative might be of interest—he would like to know if his candidate will win, ie $p > 0.5$. Candidate B’s campaign manager’s alternative hypothesis would be that $p < 0.5$. Which alternative (\neq , $<$, $>$) is appropriate must be selected based on the context. The \neq alternative is called a *two sided* hypothesis; the other two are *one sided* hypotheses. Usually when we use a one sided hypothesis, we rephrase the null hypothesis so that the pair of hypotheses still include all possible parameter values. For instance, if the alternative hypothesis is $p < 0.5$, we might write the Null hypothesis $p \geq 0.5$ —but the mechanics of the test are not changed by writing the null hypothesis this way rather than $p = 0.5$.

5.3.2 Test statistics and p-values

Next, we want to quantify how plausible the null hypothesis is. This is done with two quantities, called the *test statistic*, and the *p-value*. The test statistic summarizes the data, and does so in a way that the distribution of the test statistic under the null hypothesis is easy to compute. For a testing whether a population proportion is equal to some value p_0 , the test statistic is:

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}.$$

We use z to denote the test statistic, because when the null hypothesis is true, the test statistic has a roughly normal distribution.

The p-value is the probability, when the null hypothesis is true, of observing a test statistic ‘as or more extreme’ than the observed test statistic. High p-values indicate the Null hypothesis is plausible, low p-values indicate the alternative hypothesis is preferred. We usually specify a cutoff for what we will consider a small p-value, called the *significance level*. Common significance levels are 0.05, 0.01, and 0.001. When the p-value is less than the significance level, we say we ‘reject the null hypothesis.’”

What constitutes an ‘extreme’ test statistic depends on the alternative hypothesis. For the journalist, with his two sided hypothesis, either large or small values of the test statistic are ex-

treme. If Z is a standard normal random variable (ie has the same distribution as the test statistic if the null hypothesis is true), the journalist is interested in cases where the probability that $|Z| > |\text{observed } z|$ is small. This corresponds to cases when \hat{p} is far from 0.5. For candidate A's campaign manager, an extreme statistic is when the probability $Z > \text{observed } z$ is small, corresponding to cases with \hat{p} much larger than 0.5. What sort of test statistic is candidate B's campaign manager looking for?

Example Suppose the journalist's newspaper does a poll based on a simple random sample of 1000 people, and 550 are for candidate A. Can the next days headline read 'Election Locked Up?'

Null Hypothesis $p = 0.5$

Alternative Hypothesis Since the journalist has a headline if either candidate is winning, use a two-sided alternative: $p \neq 0.5$.

Test Statistic

$$\frac{\bar{x} - \mu_0}{\sqrt{p(1-p)/n}} = \frac{0.55 - 0.5}{\sqrt{0.55 \times 0.45/1000}} = 3.17$$

p-value 0.0015, incorporating the probability of observing a test statistic > 3.17 or < -3.17 when the Null hypothesis is true.

Conclusion The observed data is very unlikely under the null hypothesis, so H_0 is rejected. We reject the null hypothesis and conclude that one candidate (candidate A) has a clear majority.

How does a one sided test, like the test candidate A's manager would perform differ?

Null Hypothesis $p \leq 0.5$

Alternative Hypothesis The manager would like to demonstrate his candidate will win th election: $p > 0.5$.

Test Statistic Identical to the two-sided case.

$$\frac{\bar{x} - \mu_0}{\sqrt{p(1-p)/n}} = \frac{0.55 - 0.5}{\sqrt{0.55 \times 0.45/1000}} = 3.17$$

p-value Now we are only interested in the probability of observing a test statistic larger than 3.17 if the null hypothesis is true. This is half of the probability for the two-sided case, about 0.0008.

Conclusion The observed data is very unlikely under the null hypothesis so H_0 is rejected. We reject the null hypothesis and conclude that candidate A has a clear majority.

5.3.3 Hypothesis tests for means

When we are interested in whether the population mean is equal to a certain value μ_0 , our test statistic is different. It is:

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

When μ_0 is the true population mean, ie the null hypothesis is true, this has a T-distribution with $n-1$ degrees of freedom, which can be used to compute a p-value.

Example Tippett (1950) gives the following observations for the number of yarn breaks for each of 9 looms (where a loom corresponds to a fixed length of yarn) using the same type of wool and tension: 18, 21, 29, 17, 12, 18, 35, 30, 36. Suppose the yarn manufacturer specifies the average number of warp breaks per loom should be 20 under these conditions. Does this data give us reason to doubt the manufacturer's claim?

Null Hypothesis $\mu \leq 20$

Alternative Hypothesis We are only concerned if the number of breaks exceeds the manufacturer's claim (one sided hypothesis). $\mu > 20$

Test Statistic $\bar{x} = 24, s = 8.66$

$$\frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{24 - 20}{8.66/\sqrt{9}} = 1.39$$

p-value Compare with a t distribution with 8 degrees of freedom: p-value=0.10.

Conclusion If we use a significance level of 0.05 (or lower), we do not reject H_0 . Based on this data, 20 is a plausible value for the population mean.

One of the most common situations where we perform a test for a single population mean is when we are looking at a *paired* experiment. A paired experiment is when the sample units occur in pairs, where units within each pair are similar. Each pair is split so that one unit receives treatment one and the other unit treatment two. The difference in outcome within each pair is then recorded, and we test whether the mean of the population of differences is zero. The pairing of experimental units removes some of the variability due to differences between units, making it easier to detect differences between the treatments. Some examples of paired experiments follow.

The before and after experiment Imagine we are evaluating an athsma treatment. For each subject we might look at the difference in their number of attacks per week before and after the treatment. The change in number of attacks may be similar for each person even when the variability in the number of attacks between people is large. Looking at the change for each person rather than comparing 2 separate groups of people makes it easier to detect the treatment effect.

Experiments with twins or siblings Suppose we are comparing two exercise programs that promote agility. We may enroll twins (or siblings of similar age) in our study, assigning one from each pair to the two exercise programs. We base our conclusions on the difference between twins in agility recorded at the end of the program. Each pair of experimental units has a similar home environment and genetic makeup, so differences in the efficacy of the exercise programs are less likely to be obscured by the population variability in physical ability or attitude toward excercise.

The split plot experiment Agricultural experiments are often performed on split-plots: two treatments are used on several pairs of adjacent plots. Again, the difference in yield between the two plots is recorded. The soil, moisture, and other conditions ought to be similar for the adjacent plots, making it easier to detect treatment differences.

Once the differences between the pairs have been computed, the mechanics of the test are just like any test of a single population mean, as we see in the example below.

Example The following experiment was conducted in Sweden to determine the influence of a speed limit on road safety (data from Svensson, 1981). Pairs of similar days in 1961 and 1962 were used to control for seasonal differences. Data was collected on 92 days in each year. In one year, there was no speed limit on a section of the road network, and in the other a speed limit was set. (Which day got the speed limit was randomized for each pair of days to eliminate any systematic differences between years.) The number of accidents was recorded for each day. The average difference between the no speed limit day and the speed limit day was 3.62, with standard deviation 8.04. Does a speed limit reduce the number of accidents?

Null Hypothesis $\mu_d = 0$ The average difference between the no speed limit day and speed limit day is zero.

Alternative hypothesis $\mu_d > 0$ The average difference (no speed limit day - speed limit day) is positive; the speed limit reduces the number of accidents (one-sided hypothesis).

Test Statistic

$$\frac{\bar{d} - \mu_{d0}}{s_d/\sqrt{n}} = \frac{3.62 - 0}{(8.04/\sqrt{92})} = 4.31$$

p-value Compare to a T-distribution with 91 degrees of freedom. The p-value is 2.06×10^{-5}

Conclusion Reject the null hypothesis. There were fewer accidents on days with a speed limit.

5.3.4 Sign test

The testing methods we have introduced all have distributional assumptions—for proportions, the test statistic must be approximately normal, for means, it must have a T-distribution. These assumptions generally hold if the sample size is moderate or large. But sometimes we would like to do a test that does not require these assumptions to be made. Such a test is called a *nonparametric* test. One such test is the sign test, which tests whether the median of a continuous distribution is equal to a certain value m_0 . The only requirement for performing this test is that a random sample of the population is taken.

Even when the sample size is fairly large so that the assumptions for use of the T-distribution are satisfied, if the distribution is very skewed, a test for the median may be of more interest than a test for the mean. As discussed in Chapter 1, for skewed distributions the median corresponds more closely to most people's idea of the center of the distribution than the mean does.

The test statistic for the sign test is the number of observations greater than m_0 . (In many applications m_0 is zero, so the test statistic is the number with positive sign, thus the name 'sign test.'). If the null hypothesis is true, the number of observations greater than m_0 is binomial(0.5, $n - n_0$, where n_0 is the number of observations that are exactly 0). This distribution (or the normal approximation to it, for large $n - n_0$) can be used to compute a p-value.

Example Revisiting the Swedish speed limit data, we note that for 59 of the pairs of days, the number of accidents with no speed limit was greater than the number of accidents with a speed limit (the difference was positive). For 3 of the 92 pairs of days, there were identical numbers of accidents with or without a speed limit.

Null Hypothesis $m_d = 0$ The median difference between the no speed limit day and speed limit day is zero.

Alternative hypothesis $m_d > 0$ The median difference (no speed limit day - speed limit day) is positive; the speed limit reduces the number of accidents (one-sided hypothesis).

Test Statistic The difference is positive for 59 pairs, out of 89 with a difference.

p-value Compare to a binomial distribution with $p = 0.5$ and $n = 92 - 3 = 89$. Probability of seeing 59 or more successes is 0.0014.

Conclusion Reject the null hypothesis. There were fewer accidents on days with a speed limit.

5.3.5 Type I and Type II error

A hypothesis test does not always reach the right conclusion. Sometimes the null hypothesis will be rejected when the null value is correct; this is called *type I error*. The significance level that we use for our test controls the probability of type I error. If we reject the null hypothesis when data as or more extreme would occur less than 0.05 of the time under the null hypothesis, when the null hypothesis is true we end up rejecting it 0.05 of the time. We also may fail to reject the null hypothesis when it is false. This is called *type II error*. The probability of type II error is affected by the distance between the null value and the true value; and the sample size. The closer the null and true parameter values are, the higher the probability of type II error. In addition, the smaller the sample size is, the higher the probability of type II error is. (See exercise 10.) Sometimes instead of saying we want low type II error, we say we want our experiment to have high *power*. Power is $1 - P(\text{type II error})$.

5.3.6 Connection between hypothesis tests and confidence intervals

You will have noticed that confidence intervals and hypothesis tests have similar, though not identical, aims. A confidence interval includes plausible values for a parameter; a hypothesis test returns the conclusion “reject the null hypothesis” if the null value of the parameter is not plausible. Do hypothesis tests and confidence intervals always give the same answer about whether a value is plausible? In some situations, yes. A two-sided hypothesis test for a mean using significance level α will reject the null hypothesis in exactly the cases where a $(1 - \alpha) \times 100\%$ confidence interval does

not include the null value. There is not an exact correspondence for proportions, because in a hypothesis test the null value is used to compute the standard error (because we are interested in the distribution of our statistic when the null hypothesis is true); in a confidence interval, the observed proportion \hat{p} is used in the estimated standard error. (See Exercise 3.) Confidence intervals have a 'two-sided' character, as they typically seek to limit the values considered plausible on both the high end and low end. A level α one sided hypothesis test corresponds to a $(1 - \alpha) \times 100\%$ confidence interval that has a limit on one side, but extends to infinity on the other!

5.4 Exercises

In doing the exercises, please follow the format of the examples given in the chapter. For instance, when computing a confidence interval, give the estimate, estimated standard error, multiplier and margin of error as well as the final interval; for hypothesis tests, give the null and alternative hypotheses, test statistic, and pvalue, and give your conclusion in the original context as well as stating whether or not the null hypothesis has been rejected.

1. Masden (1976) describes a survey of 1681 Copenhagen residents on housing conditions. 668 of these reported high satisfaction with their housing conditions. Construct a confidence interval for the proportion of Copenhagen residents who are highly satisfied with their housing conditions.
2. Hoyt et al (1959) give a summary of a census of people who graduated from high school in Minnesota in 1938.
 - a) 3945 of the 14068 individuals recorded entered college immediately after high school. Give a confidence interval for the proportion.
 - b) The data is from a census, not a sample. However, what sort of population might the confidence interval in (a) pertain to?
3. A local veterinary clinic is provided with several trial boxes of a diet dog food. The manufacturer claims 73% of dogs switched to this food for 6 weeks loose weight. The clinic takes a sample of 30 of their overweight patients and provides 6 weeks worth of food to their owners. At the end of the trial period, 17 have lost weight.
 - a) Construct a confidence interval for the percentage of dogs who loose weight on the food.

- b) Do a one-sided hypothesis test of whether the manufacturer's claim is reasonable. Does the conclusion agree with the answer obtained in (a)? Does the conclusion of a two-sided test agree with (a)?
 - c) What is the standard error used in the denominator of the test statistic computed for (b)? What is the estimated standard error used for the margin of error in (a)? Why are they different?
4. A sample of tires (Davies, 1947) has a hardness measurement recorded for each tire. The measurements are given in Column 3 of Table 5.1. What is a 99% confidence interval for the mean hardness in the population of tires from which these have been selected?
5. You would like to obtain a confidence interval for the average number of bids placed in online auctions for car seats, restricting your self to auctions that get at least one bid. You would like your margin of error for a 95% confidence interval to be one bid or less. Looking up the auctions over the last three weeks, you see a standard deviation of 18.42. Using this as an estimate of the population standard deviation, determine how large a sample of auctions you will need to obtain the desired width. You may assume that the sample size will be large enough that the multiplier should be taken from the normal distribution.
6. Aitkin (1978) provides a dataset collected by S. Quine, giving the number of days absent from school for a single school year, for 146 children. 69 of the children missed 10 or fewer days of school. Suppose a target is for 50% of children to miss 10 or fewer days. Regarding these 146 children as a sample of those that will attend the school over several years, do a test to see if these data provide evidence that the target is not being met.
7. Knight and Skagen (1988) observed one eagle attempting to pirate each fish caught by another eagle. Of 27 fish, the pirate eagle successfully obtained 17.
- a) Find a 95% confidence interval for this eagle's probability of successful pirating.
 - b) Could we use the same confidence interval procedure if we had observed, say, 15 successful attempts out of 16? Explain.
8. Gonzalez-Martinez et al (2006) defines saplings as trees between 4 and 25 years old. Could the average age of the sapling population actually lie at the center of this range (14.5)? They have sampled 201 saplings, and observed a

- sample mean of 15 with a sample standard deviation of 4. Perform the relevant hypothesis test.
9. The following are measurements of the extra hours of sleep obtained for each of 10 patients after using a soporific drug (Student, 1908): 0.7, -1.6, -0.2, -1.2, -0.1, 3.4, 3.7 0.8, 0.0, 2.0. Do a test to determine if the drug increases the average hours of sleep.
 10. Suppose we are testing whether the mean of a certain population is equal to 5, using a two-sided test with significance level 5%. A pilot study has estimated the population standard deviation at 1.4.
 - a) Suppose a sample of size 1000 will be taken. How much will the sample mean have to deviate from 5 for us to reject the null hypothesis? (Use the sample standard deviation from the pilot study in your calculation)
 - b) If the true mean is 0.52, and true standard deviation 1.4, what is the distribution of the sample mean with sample size 1000?
 - c) What is the probability that a random variable with the distribution in (b) exceeds the cutoff specified in (a)? This is the power of the test for sample size 1000 and the population parameters specified in (b).
 - d) Repeat parts (a)-(c) for a sample of size 500.
 11. Charles Darwin conducted a paired experiment to determine whether cross fertilized plants grew taller than self fertilized ones. 15 pairs (one cross fertilized, one self-fertilized) were grown, with members of the same pair germinated at the same time and grown in the same pot. The observed height differences (cross-fertilized - self-fertilized, in units of 1/8 of an inch) are listed below. 49, -67, 8, 16, 6, 23, 28, 41, 14, 29, 56, 24, 75, 60, -48
 - a) Perform the t-test on these data.
 - b) Perform the sign test on these data.
 - c) Compare your results. Which do you think is more appropriate?

Tire	Abrasion Loss in g/hr	Hardness in Shore units	Tensile Strength in kg/sq m
1	372	45	162
2	206	55	233
3	175	61	232
4	154	66	231
5	136	71	231
6	112	71	237
7	55	81	224
8	45	86	219
9	221	53	203
10	166	60	189
11	164	64	210
12	113	68	210
13	82	79	196
14	32	81	180
15	228	56	200
16	196	68	173
17	128	75	188
18	97	83	161
19	64	88	119
20	249	59	161
21	219	71	151
22	186	80	165
23	155	82	151
24	114	89	128
25	341	51	161
26	340	59	146
27	283	65	148
28	267	74	144
29	215	81	134
30	148	86	127

Table 5.1: Measurements on a population of Tires, from Davies, 1947.

Glossary

Confidence interval An interval containing the plausible values for a parameter.

Estimate Our best guess at a population parameter for a particular data set.

Estimator The formula applied to a data set to obtain an estimate.

Hypothesis A conjecture about the value of a population parameter.

Alternative Hypothesis Typically what we are trying to demonstrate: that the population parameter is \neq , $>$, or $<$ the null value.

Null Hypothesis The hypothesis that includes the possibility that the parameter equals the null value ($=$, \geq , le).

One sided alternative hypothesis The alternative where only extreme values of one sort (large or small) are of interest. Uses either $>$ or $<$.

Two sided alternative hypothesis The alternative where either extremely large or extremely small values are of interest. Uses \neq .

Margin of error A symmetric confidence interval contains all values within the margin of error of the estimate.

Multiplier Used in computing a confidence interval. The multiplier determines what the confidence level of the interval will be.

Nonparametric A procedure that does not require one to assume the data (or summary statistics computed from the data) has a specific type of distribution.

Paired data Data where the experimental units occur in pairs of similar units. Typically a unit from each pair is assigned to each of two treatments.

Power The probability of rejecting the null hypothesis when it is false. Depends on the sample size, standard error of the estimate, and distance between the true and null values.

P-value The probability under the null hypothesis of observing data “as or more extreme.”

Significance level A threshold for rejecting the null hypothesis specified before a test is carried out; we reject when the p-value is less than the significance level.

Sign test A test for whether the population median is equal to a specific value. Uses the fact that the number of observations greater than the population median has a binomial distribution.

Standard error The standard deviation of an estimator.

T-distribution A symmetric distribution with heavier tails than the normal distribution. The parameter *degrees of freedom* governs the shape of the distribution; the larger the degrees of freedom, the closer the distribution is to the normal distribution.

Type I error The probability the null hypothesis is rejected when it is actually true. Equal to the significance level.

Type II error The probability we fail to reject the null hypothesis when it is actually false. Power is $1 - (\text{type II error})$.

References

- Aitkin, M. (1978). The analysis of unbalanced cross classifications (with discussion). *Journal of the Royal Statistical Society, Series A 141*, 195–223.
- Darwin, C. (1876). *The effects of cross- and self-fertilisation in the vegetable kingdom*. John Murray.
- Davies, O. L. (1947). *Statistical Methods in Research and Production*, pp. 119. Oliver and Boyd. Table 6.1.
- Dean, M., K. Ardlie, and M. W. Nachman (2006). The frequency of multiple paternity suggests that sperm competition is common in house mice (*mus domesticus*). *Molecular Ecology 15*, 4141–4151.
- González-Martínez, S. C., J. Burczyk, R. Nathan, N. Nanos, L. Gil, and Alia (2006). Effective gene dispersal and female reproductive success in mediterranean maritime pine (*pinus pinaster aiton*). *Molecular Ecology 15*, 4577–4588.
- Hoyt, C., P. R. Krishnaiah, and E. P. Torrance (1959). Analysis of complex contingency tables. *Journal of Experimental Education 27*, 187–194.
- Knight, R. L. and S. K. Skagen (1988). Agonistic asymmetries and the foraging ecology of bald eagles. *Ecology 69*, 1188–1194.
- Lloyd, C. J. and D. Frommer (2004). Estimating the false negative fraction for a multiple screening test for bowel cancer when negatives are not verified. *Australia and New Zealand Journal of Statistics 46*, 531–542.
- Madsen, M. (1976). Statistical analysis of multiple contingency tables. *Scandinavian Journal of Statistics 3*, 97–106.
- Mithas, S., D. Almirall, and M. S. Krishnan (2006). Do crm systems cause one-to-one marketing effectiveness? *Statistical Science 21*, 223–233.
- Student (1908). The probable error of the mean. *Biometrika 6*, 20.
- Svensson, A. (1981). On the goodness-of-fit test for the multiplicative poisson model. *Annals of Statistics 9*, 697–704.
- Tippet, L. (1950). *Technological Applications of Statistics*, pp. 106. Wiley.