Tests of Hypotheses Using Statistics

Adam Massey*and Steven J. Miller†

Mathematics Department
Brown University
Providence, RI 02912

Abstract

We present the various methods of hypothesis testing that one typically encounters in a mathematical statistics course. The focus will be on conditions for using each test, the hypothesis tested by each test, and the appropriate (and inappropriate) ways of using each test. We conclude by summarizing the different tests (what conditions must be met to use them, what the test statistic is, and what the critical region is).

Contents

*E-mail: amassey3102@ucla.edu
†E-mail: sjmiller@math.brow.edu

1 Types of Hypotheses and Test Statistics

1.1 Introduction

The method of hypothesis testing uses tests of significance to determine the likelihood that a statement (often related to the mean or variance of a given distribution) is true, and at what likelihood we would, as statisticians, accept the statement as true. While understanding the mathematical concepts that go into the formulation of these tests is important, knowledge of how to appropriately use each test (and when to use which test) is equally important. The purpose here is on the latter skill. To this end, we will examine each statistical test commonly taught in an introductory mathematical statistics course, stressing the conditions under which one could use each test, the types of hypotheses that can be tested by each test, and the appropriate way to use each test. In order to do so, we must first understand how to conduct a statistical significance test (following the steps indicated in [MM]), and we will then show how to adapt each test to this general framework.

We begin by formulating the hypothesis that we want to test, called the alternative hypothesis. Usually this hypothesis is derived from an attempt to prove an underlying theory (for example, attempting to show that women score, on average, higher on the SAT verbal section than men). We do this by testing against the null hypothesis, the negation of the alternative hypothesis (using our same example, our null hypothesis would be that women do not, on average, score higher than men on the SAT verbal section). Finally, we set a probability level α; this value will be our significance level and corresponds to the probability that we reject the null hypothesis when it's in fact true. The logic is to assume the null hypothesis is true, and then perform a study on the parameter in question. If the study yields results that would be unlikely if the null hypothesis were true (like results that would only occur with probability .01), then we can confidently say the null hypothesis is not true and accept the alternative hypothesis. Now that we have determined the hypotheses and the significance level, the data is collected (or in this case provided for you in the exercises).

Once the data is collected, tests of hypotheses follow the following steps:

1. Using the sampling distribution of an appropriate test statistic, determine a critical region of size α.

2. Determine the value of the test statistic from the sample data.

3. Check whether the value of the test statistic falls within the critical region; if yes, we reject the null in favor of the alternative hypothesis, and if no, we fail to reject the null hypothesis.

These three steps are what we will focus on for every test; namely, what the appropriate sampling distribution for each test is and what test statistic we use (the third step is done by simply comparing values).

1.2 Types of Hypotheses

There are two main types of hypotheses we can test: one-tailed hypotheses and two-tailed hypotheses. Our critical region will be constructed differently in each case.

Example 1.1. *Suppose we wanted to test whether or not girls, on average, score higher than 600 on the SAT verbal section. Our underlying theory is that girls do score higher than 600, which would give us the following null (denoted H_0) and alternative (denoted H_1) hypotheses:*

$$\begin{array}{l} H_0: \ \mu \leq 600 \\ H_1: \ \mu > 600, \end{array} \tag{1.1}$$

where μ is the average score for girls on the SAT verbal section. This is an example of what is called a one-tailed hypothesis. The name comes from the fact that evidence against the null hypothesis comes from only one tail of the distribution (namely, scores above 600). When constructing the critical region of size α, one finds a critical value in the sampling distribution so that the area under the distribution in the interval $(critical\ value, \infty)$ is α. We will explain how to find a critical value in later sections.

Example 1.2. *Suppose instead that we wanted to see if girls scored significantly different than the national average score on the verbal section of the SAT, and suppose that national average was 500. Our underlying theory is that girls do score significantly different than the national average, which would give us the following null and alternative hypotheses:*

$$\begin{array}{l} H_0: \ \mu \ = \ 500 \\ H_1: \mu \ \neq \ 500 \end{array}, \tag{1.2}$$

where again μ is the average score for girls on the SAT verbal section. This is an example of a two-tailed hypothesis. The name comes from the fact that evidence against the null hypothesis can come from either tail of the sampling distribution (namely, scores significantly above AND significantly below 500 can offer evidence against the null hypothesis). When constructing the critical region of size α, one finds two critical values (when assuming the null is true, we take one above the mean and one below the mean) so that the region under the sampling distribution over the interval $(-\infty, critical\ value\ 1) \cup (critical\ value\ 2, \infty)$ is α. Often we choose symmetric regions so that the area in the left tail is $\alpha/2$ and the area in the right tail is $\alpha/2$; however, this is not required. There are advantages in choosing critical regions where each tail has equal probability.

There will be several types of hypotheses we will encounter throughout our work, but almost all of them may be reduced to one of these two cases, so understanding each of these types will prove to be critical to understanding hypothesis testing.

1.3 Types of Statistics

There are many different statistics that we can investigate. We describe a common situation. Let $X_1, \ldots, X_N$ be independent identically distributed random variables drawn from a population with density p. This means that for each $i \in \{1, \ldots, N\}$ we have that the probability of observing a value of X_i lying in the interval $[a, b]$ is just

$$\text{Prob}(X_i \in [a, b]) \ = \ \int_a^b p(x)dx. \tag{1.3}$$

We often use X to denote a random variable drawn from this population and x a value of the random variable X. We denote the mean of the population by μ and its variance by σ^2:

$$\begin{aligned} \mu &= \int_{-\infty}^{\infty} x p(x) dx \ = \ \mathbb{E}[X] \\ \sigma^2 &= \int_{\infty}^{\infty} (x-\mu)^2 p(x) dx \ = \ \mathbb{E}[X^2] - \mathbb{E}[X]^2. \end{aligned} \tag{1.4}$$

If X is in meters then the variance is in meters squared; the square root of the variance, called the standard deviation, is in meters. Thus it makes sense that the correct scale to study fluctuations is not the variance, but the square root of the variance. If there are many random variables with different underlying distributions, we often add a subscript to emphasize which mean or standard deviation we are studying.

If Y is some quantity we are interested in studying, we shall often study the related quantity

$$\frac{Y - \text{Mean}(Y)}{\text{StDev}(Y)} \ = \ \frac{Y - \mu_Y}{\sigma_Y}. \tag{1.5}$$

For example, if $Y = (X_1 + \cdots + X_N)/N$, then Y is an approximation to the mean. If we observe values $x_1, \ldots, x_N$ for $X_1, \ldots, X_N$, then the observed value of the sample mean is $y = (x_1 + \cdots + x_N)/N$. We have (assuming the random variables are independently and identically distributed from a population with mean μ_X and standard deviation σ_X), that

$$\begin{aligned} \mu_Y &= \mathbb{E}[Y] \\ &= \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) \\ &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[X_i] \\ &= \frac{1}{N} \cdot N\mu_X \ = \ \mu_X, \end{aligned} \tag{1.6}$$

and

$$\begin{aligned} \sigma_Y^2 &= \text{Var}(Y) \\ &= \text{Var}\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) \\ &= \frac{1}{N^2}\sum_{i=1}^{N} \text{Var}(X_i) \\ &= \frac{1}{N^2} \cdot N\text{Var}(X) \ = \ \frac{\sigma_X^2}{N}; \end{aligned} \tag{1.7}$$

thus

$$\sigma_Y \ = \ \text{StDev}(Y) \ = \ \sigma_X/\sqrt{N}. \tag{1.8}$$

Thus, as $N \to \infty$, we see that Y becomes more and more concentrated about μ_X; this is because the mean of Y is μ_X and its standard deviation is $\sigma_X/\sqrt{N}$, which tends to zero with N. If we believe $\mu_X = 5$, say, then for N large the observed value of Y should be close to 5. If it is, this

provides evidence supporting our hypothesis that the population has mean 5; if it does not, then we obtain evidence against this hypothesis.

Thus it is imperative that we know what the the distribution of Y is. While the exact distribution of Y is a function of the underlying distribution of the X_i's, in many cases the Central Limit Theorem asserts that Y is approximately normally distributed with mean 0 and variance 1. This is trivially true if the X_i are drawn from a normal distribution; for more general distributions this approximation is often fairly good for $N \geq 30$.

This example is typical of the statistics we shall study below. We have some random variable Y which depends on random variables $X_1, \dots, X_N$. If we observe values of $x_1, \dots, x_N$ for the $X_1, \dots, X_N$, we say these are the sample values. Given these observations we calculate the value of Y; in our case above where $Y = (X_1 + \cdots + X_N)/N$ we would observe $y = (x_1 + \cdots + x_N)/N$. We then normalize Y and look at

$$Z \;=\; \frac{Y - \text{Mean}(Y)}{\text{StDev}(Y)} \;=\; \frac{Y - \mu_Y}{\sigma_Y}. \tag{1.9}$$

The advantage is that Z has mean 0 and variance 1. This facilitates using a table to analyze the resulting value.

For example, consider a normal distribution with mean 0 and standard deviation σ. Are we surprised if someone says they randomly chose a number according to this distribution and observed it to be 100? We are if $\sigma = 1$, as this is over 100 standard deviations away from the mean; however, if $\sigma = 1000$ then we are not surprised at all. If we do not have any information about the scale of the fluctuations, it is impossible to tell if something is large or small – we have no basis for comparison. This is one reason why it is useful to study statistics such as $Z = (Y - \mu_Y)/\sigma_Y$, namely we *must* divide by the standard deviation.

Another reason why it is useful to study quantities such as $Z = (Y - \mu_Y)/\sigma_Y$ is that Z has mean 0 and variance 1. This allows us to create just *one* lookup table. If we just studied $Y - \mu_Y$, we would need a lookup table for each possible standard deviation. This is similar to logarithm tables. It is enough to have logarithm tables in one base because of the change of base formula:

$$\log_b x \;=\; \frac{\log_c x}{\log_c b}. \tag{1.10}$$

In particular, if we can calculate logarithms base e we can calculate logarithms in *any* base. The importance of this formula cannot be overstated. It reduced the problem of tabulating all logarithms (with any base!) to just finding logarithms in one base.

Exercise 1.3. *Approximate the probability of observing a value of 100 or larger if it is drawn from a normal distribution with mean 0 and variance 1. One may approximate the integrals directly, or use Chebyshev's Theorem.*

2 z-Tests and t-Tests

2.1 Testing Means I: Large Sample Size or Known Variance

The first type of test we explore is the most basic: testing the mean of a distribution in which we already know the population variance σ^2. Later we discuss how to modify these tests to handle the situation where we do not know the population variance.

Thus, for now, we are assuming that our population is normal with known variance σ^2. Our test statistic is

$$z \;=\; \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}, \tag{2.11}$$

where n is the number of observations made when collecting the data for the study, and μ is the true mean when we assume the null hypothesis is true. So to test a hypothesis with given significance level α, we calculate the critical value of z (or critical values, if the test is two-tailed) and then check to see whether or not the value of the test statistic in (2.11) is in our critical region. This is called a z-test. We are most often concerned with tests involving either $\alpha = .05$ or $\alpha = .01$. When we construct our critical region, we need to decide whether or not our hypotheses in question are one-tailed or two-tailed. If one-tailed, we reject the null hypothesis if $z \geq z_\alpha$ (if the hypothesis is right-handed) or if $z \leq z_\alpha$ (if the hypothesis is left-handed). If two-tailed, we reject the null hypothesis if $|z| \geq z_{\alpha/2}$. So the most common z-values that we use are $z_{.05} = 1.645$, $z_{.01} = 2.33$, $z_{.025} = 1.96$ and $z_{.005} = 2.575$. These are good numbers to have memorized when performing hypothesis tests.

Example 2.1. *Suppose we want to test whether or not girls, on average, score higher than 600 on the SAT verbal section. This, as before, gives us the hypotheses:*

$$\begin{aligned} H_0 &: \;\mu \leq 600 \\ H_1 &: \;\mu > 600. \end{aligned} \tag{2.12}$$

Suppose we choose our α to be .05. Since this is a one-tailed test, we find our critical value in the upper tail of the sampling distribution, which is $z = 1.645$. Suppose we also happen to know that the standard deviation for girls SAT verbal section scores is 100. Since the true variance is known (knowing the true standard deviation is equivalent to knowing the variance), we may use the $z-$test. Now we collect the data using a random sample of 20 girls and their verbal section scores:

$$\begin{array}{lllll} 650 & 730 & 510 & 670 & 480 \\ 800 & 690 & 530 & 590 & 620 \\ 710 & 670 & 640 & 780 & 650 \\ 490 & 800 & 600 & 510 & 700. \end{array} \tag{2.13}$$

This gives us a sample mean of $\bar{x} = 641$. We use this to calculate our test statistic:

$$z \;=\; \frac{641-600}{100/\sqrt{20}} \;=\; 1.8336. \tag{2.14}$$

Since $1.8336 > 1.645$, we reject the null hypothesis in favor of the alternative explanation that girls score, on average, better than 600 on the verbal section of the SATs.

Exercise 2.2. *Test at the $\alpha = .05$ significance level whether the mean of a random sample of size $n = 16$ is statistically significantly less than 10 if the distribution from which the sample was taken is normal, $\bar{x} = 8.4$ and $\sigma^2 = 10.24$. What are the null and alternative hypotheses for this test?* This example is from [MM].

Exercise 2.3. *Suppose it is known from experience that the standard deviation of the weight of 10-ounce packages of cookies is 0.20 ounces. To check whether the true average is, on a given day, 10 ounces, employees select a random sample of 36 packages and find that their mean weight is $\bar{x} = 9.45$ ounces. Perform a two-tailed z-test on this data, checking at the $\alpha = .01$ significance level.* This example is from [MM].

The reason the z-test works is that the sum of normally distributed random variables is also normally distributed. We can perform z-tests in cases where the underlying population is not normal. If n is large and we know the population variance, then by the Central Limit Theorem the distribution of the random variable

$$Z \;=\; \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \;=\; \frac{\overline{X}-\mu}{\sigma_{\overline{X}}} \tag{2.15}$$

is approximately the standard normal, and we may therefore apply the z-test here.

A more difficult problem is if we do not know the population variance. If n is small we are in trouble; however, if n is large ($n \geq 30$ suffices for most distributions commonly encountered) the following approximation is quite good. We replace the unknown population variance with the sample variance s^2, where

$$s^2 \;=\; \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}. \tag{2.16}$$

Here x_i corresponds to each observation in the sample, and $\bar{x}$ the mean of the sample as always. Everything else in the analysis remains the same as before. Observe that the square root of the sample variance is the sample standard deviation s, and if given the sample standard deviation, one may use the analogous formula:

$$z \;=\; \frac{\bar{x}-\mu}{s/\sqrt{n}} \tag{2.17}$$

to calculate the test statistic. The point is that for n large, s^2 is a good approximation to the unknown population variance σ^2.

Remark 2.4. There are many reasons why we divide by $n-1$ and not n in calculating the sample variance s^2 in (2.16). First off, for many problems this gives us an unbiased estimator for the population variance. We give another explanation why it should not be n. Imagine the situation where we have just one observation, which for definiteness we'll say is 1701. We can use this observation to estimate the population's mean – our best guess is that the mean is 1701. What if we want to try to estimate the population's variance? With just one observation there *is* no variation – it is impossible for us to estimate the population variance from just one observation. Thus it is reasonable that (2.16) provides no information when $n = 1$ (we have the indefinite 0 over 0).

Example 2.5. *A brochure inviting subscriptions for a new diet program states that the participants are expected to lose over* 22 *pounds in five weeks. Suppose that, from the data of the five-week weight losses of* 56 *participants, the sample mean and sample standard deviation are found to be* 23.5 *and* 10.2*, respectively. Could the statement in the brochure be substantiated on the basis of these findings? Test at the* $\alpha = .05$ *level.*

To solve this problem, we first need to formulate the null and alternative hypotheses for this test:

$$\begin{aligned} H_0 &: \;\; \mu \leq 22 \\ H_1 &: \;\; \mu > 22. \end{aligned} \tag{2.18}$$

Since H_1 *is one-tailed to the right, our critical region must lie in the right tail of our sampling distribution. Using our statistics tables, we see that the interval to the right of* $z = 1.645$ *corresponds to a critical region of size* .05*. Now we simply calculate the test statistic:*

$$z \;=\; \frac{23.5-22}{10.2/\sqrt{56}} \;=\; 1.10. \tag{2.19}$$

Since 1.10 is outside of our critical region, we fail to reject the null hypothesis and cannot substantiate the brochure's claim based on these results. This example is from [JB].

Remark 2.6. It may seem surprising that, in the above example, our observations do not support the brochure's claim that you lose at least 22 pounds in 5 weeks. Our sample mean is 23.5, which *is* greater than 22, but not by a lot (only 1.5). What do we mean when we say that the sample mean isn't greater than the hypothesized value of 22 "by a lot"? The sample standard deviation equals $10.2/\sqrt{56} = 1.36$; this is of comparable size to the difference between the observed and conjectured means. The problem is that our sample size is very small; it would be a very different story if we observed a mean of 23.5 from a sample of size 1000. Returning to our numbers, consider what would happen if the true weight loss is 21.8 pounds in 5 weeks. Our observed value of 23.5 is quite close, within two sample standard deviations. This is why we cannot rule out the possibility that the true weight loss is less than 22 pounds in 5 weeks; however, we can easily rule out the possibility that it is 18 pounds or less in 5 weeks.

Exercise 2.7. *Suppose it is known that the average income for families living in New England last year was* \$50000*. We want to know whether or not the average yearly income of families in Providence is significantly less than the average yearly income in New England. Unfortunately, we don't know the variance of incomes in Providence or in New England. All we have is the following* 50 *incomes (from last year) taken randomly from families in Providence (all entries are in dollars):*

23500	37400	62600	19000	34700	81000	41500	32400	42000	18500
55000	47800	71200	29300	14900	101000	51700	32400	77000	21000
15000	34900	66700	91000	37200	70000	41900	42200	33000	28700
28500	67900	32700	14800	25800	51000	43400	44700	37000	18500
63500	42600	22600	49000	54600	47000	31500	32400	50000	38900.

(2.20)

Using this information, test the hypothesis that the average yearly income of families in Providence is below that of the average family in New England at the $\alpha = .05$ *level.*

We may also look at tests comparing the means of two different groups. For instance, we may want to explore whether or not girls score (on average) higher on the SAT verbal section than boys. Or to be even more specific, we may test whether or not girls score (on average) 50 points higher on the SAT verbal section than boys. To do so, we take random samples from the two populations we want to test (yielding an $\bar{x}_1$ and an $\bar{x}_2$) and then compare $\bar{x}_1 - \bar{x}_2$ to $\mu_1 - \mu_2$, where $\mu_1 - \mu_2$ is determined based on the hypothesis we're hoping to test. Now our only concern is the test statistic for this test, which turns out to be:

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \tag{2.21}$$

where σ_i^2 is the variance for the i^{th} distribution and n_i is the i^{th} sample size. Also, as before, if we don't know the population variances, we may substitute the sample variances s_1^2 and s_2^2 provided BOTH n_1 and n_2 are large enough (larger than 30).

Exercise 2.8. *Suppose we want to compare the average yearly income in Providence and Boston, two neighboring cities in New England. It is known from experience that the variance of yearly incomes in Providence is* \$40000 *and the variance for yearly incomes in Boston is* \$90000*. A random sample of* 20 *families was taken in Providence, yielding a mean yearly income of* \$47000*, while a random sample of* 30 *families was taken in Boston, yielding a mean yearly income of* \$52000*. At the* $\alpha = .01$ *significance level, test whether or not there is a significant different in average yearly income between the two cities.*

Exercise 2.9. *A study of the number of business lunches that executives in the insurance and banking industries claim as deductible expenses per month was based on random samples and yielded the following results:*

$$\begin{array}{cccccccc} n_1 & = & 40 & \bar{x}_1 & = & 9.1 & s_1 & = & 1.9 \\ n_2 & = & 50 & \bar{x}_2 & = & 8.0 & s_2 & = & 2.1. \end{array} \tag{2.22}$$

Test the null hypothesis $\mu_1 - \mu_2 = 0$ against the alternative hypothesis $\mu_1 - \mu_2 \neq 0$ at the $\alpha = .05$ significance level. This example is from [MM].

2.2 Testing Means II: Small Sample Size and Unknown Variance

Unfortunately, z-tests require one of two conditions: either the population is normally distributed with a known variance, or the sample size is large. For many applications the underlying distribution is approximately normal or the sample size is large, so often these conditions are met. There are times, however, when the sample sizes are not large. This is true when it is expensive to get sample points, which is often the case in the psychology studies (among many other situations).

In general hypothesis testing becomes very difficult, but there is an important special case which has been extensively tabulated, namely the case when the underlying population is normal (but of unknown variance). We define the Student t-distribution with ν degrees of freedom by

$$f(t) \;=\; \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)} \;\cdot\; \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad -\infty \leq t \leq \infty. \tag{2.23}$$

If we let $T = Z/\sqrt{Y/\nu}$ where Z has the standard normal distribution and Y has a chi-square distribution with ν degrees of freedom, then T has the Student t-distribution with ν degrees of freedom. Note that as $\nu \to \infty$ that the density tends to the density of the standard normal (the proof involves identities of the Gamma function, as well as the observation that $e^z = \lim_{n\to\infty}(1 + \frac{z}{n})^n$).

If we have a sample of size n from a normal distribution with mean μ and unknown variance σ^2, we study

$$t \;=\; \frac{\bar{x} - \mu}{s/\sqrt{n}} \tag{2.24}$$

and compare this to the Student t-distribution with $n-1$ degrees of freedom. If $n \geq 30$ then by the Central Limit Theorem we may instead compare it to the standard normal.

For example, if $n = 30$ then the critical values for the Student t-distribution are 2.045 (for $\alpha = .025$, which gives us a critical region of size .05 for a two-tail test) and 2.756 (for $\alpha = .005$, which gives us a critical region of size .01 for a two-tail test); these are close to the corresponding values from the standard normal, which are 1.960 for $\alpha = .025$ and 2.576 for $\alpha = .005$.

Example 2.10. *A city health department wishes to determine if the mean bacteria count per unit volume of water at a lake beach is within the safety level of* 200*. A researcher collected* 10 *water samples of unit volume and found the bacteria count to be:*

$$\begin{array}{cccccc} 175 & 190 & 215 & 198 & 184 \\ 207 & 210 & 193 & 196 & 180. \end{array} \tag{2.25}$$

Does the data indicate that the bacteria count is within the safety level? Test at the $\alpha = .01$ level. You may assume that the measurements constitute a sample from a normal population.

Before we begin we need to formulate our hypotheses. For the bacteria count to be within the safety level, we must have that $\mu \leq 200$*, and so if* $\mu > 200$ *we are out of the safety level. This gives us the hypotheses:*

$$\begin{array}{l} H_0 : \mu > 200 \\ H_1 : \mu \leq 200. \end{array} \tag{2.26}$$

Since our sample size is 10*, our desired* t*-distribution will have* 9 *degrees of freedom. Since* H_1 *is one-tailed to the left, we need to construct a left-handed critical region. Using a statistics table, one finds that, with* 9 *degrees of freedom,* $T \leq -2.821$ *gives us a critical region of size* $\alpha = .01$*. Our data yields a sample mean of* 194.8 *and a sample variance of* 172.66*. Now we calculate our test statistic:*

$$t \;=\; \frac{194.8 - 200}{\sqrt{172.66/10}} \;=\; -1.25. \tag{2.27}$$

Since $t = -1.25$ *is outside of our critical region, we fail to reject the null hypothesis and cannot conclude that the true mean is within the safety level.* This example is from [JB].

Exercise 2.11. *A random sample of size* 20*, from a normal population, has* $\bar{x} = 182$ *and* $s = 2.3$*. Test the hypotheses*

$$\begin{array}{l} H_0 : \mu \leq 181 \\ H_1 : \mu > 181 \end{array} \tag{2.28}$$

at the $\alpha \;=\; .05$ *significance level.* This example is from [JB].

Exercise 2.12. *The following measurements of the diameters (in feet) of Indian mounds in southern Wisconsin were gathered by examining reports in* Wisconsin Archeologist*:*

$$\begin{array}{ccccccc} 22 & 24 & 24 & 30 & 22 & 20 & 28 \\ 30 & 24 & 34 & 36 & 15 & 37. & \end{array} \tag{2.29}$$

Does the data substantiate the conjecture that the population mean diameter is larger than 21 *feet? Test at the* $\alpha \;=\; .01$ *significance level. You may assume the sample was taken from a normal distribution.* This example is from [JB].

Exercise 2.13. *A car manufacturer asserts that with the new collapsible bumper system, the mean body repair cost for the damage sustained in a collision impact of* 15 *miles per hour does not exceed* \$400*. To test the validity of this claim,* 6 *cars are crashed into a barrier at* 15 *miles per hour and their repair costs recorded. The mean and standard deviation are found to be* \$458 *and* \$48*, respectively. At the* $\alpha = .05$ *significance level, does the test data contradict the manufacturer's claim that the repair cost does not exceed* \$400*? You may assume the sample was taken from a normal population.* This example is from [MLM].

We may also test to compare two means when we don't know the population variances and when at least one of the sample sizes is less than 30. However, in order to do so, we must again make a few extra assumptions. If we assume that our samples are taken from two normal populations *with the same variance* σ^2, then we may use the following t-test:

$$t \;=\; \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{2.30}$$

where

$$s_p^2 \;=\; \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{2.31}$$

is the pooled estimate of the population variance. This test statistic will then have a t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

Remark 2.14. Whenever we encounter a formula such as (2.31), it is good to examine extreme cases to try and get a feel for what it says. For example, let's consider the case when n_1 is much greater than n_2. In this case we find that s_p^2 is approximately s_1^2. This makes sense, as we may interpret this as saying that we have so many observations from the first population that this is a very good approximation of the true variance. For example, imagine that $n_1 = 10^8$ and $n_2 = 5$. We don't expect the sample variance from a sample of size 5 to be a good estimator of the true variance: one bad observation would greatly influence the results. However, the sample variance from a sample of size 10^8 should be an excellent estimator of the true variance: a few bad observations will have negligible effect.

Example 2.15. *Suppose we wanted to test whether or not the average SAT score at Brown University was significantly different from the average SAT score at Wassamatta University. Taking a random sample, we manage to get the following* 7 *scores from Brown students:*

$$\text{Brown University}: \quad 1340 \quad 1500 \quad 1430 \quad 1440 \quad 1380 \quad 1470 \quad 1290 \tag{2.32}$$

and the following 9 *scores from Wassamatta students:*

$$\begin{array}{lccccc} \text{Wassamatta University}: & 1540 & 1480 & 1390 & 1450 & 1440 \\ & 1350 & 1520 & 1400 & 1600. & \end{array} \tag{2.33}$$

Assuming that the variance in SAT scores for students at Brown and students at Wassamatta are the same, and assuming that both samples are taken from normal populations, test (at the $\alpha = .05$ significance level) whether or not there is a significant difference in the average SAT score between these two schools.

To do so, we need to first formulate our null and alternative hypothesis:

$$\begin{aligned} H_0: &\quad \mu_B - \mu_H = 0 \\ H_1: &\quad \mu_B - \mu_H \neq 0. \end{aligned} \tag{2.34}$$

Next, we calculate the mean and variance for each sample. Doing so gives us

$$\begin{array}{llll} \bar{x}_B = 1407.14 & & s_B^2 = 5523.81 \\ \bar{x}_H = 1463.33 & & s_H^2 = 6375.00. \end{array} \tag{2.35}$$

Using this information, we calculate the pooled estimate of the population variance:

$$s_p^2 = \frac{(6)(5523.81) + (8)(6375.00)}{14} = 6010.20. \tag{2.36}$$

Now we have enough to calculate our test statistic, but before doing so, we need to find our .05 *critical values (our alternative hypothesis is two-tailed). Using a statistics table, we find the* .05 *critical values for a T-distribution with* 14 *degrees of freedom to be* -2.145 *and* 2.145*. So to reject the null, we need out test statistic to satisfy $|t| \geq 2.145$.*

Now we calculate our test statistic to be:

$$t = \frac{(1407.14 - 1463.33)}{(77.5255)\sqrt{\frac{1}{7} + \frac{1}{9}}} = \frac{-56.19}{39.07} = -1.4382. \tag{2.37}$$

Since our test statistic does not lie within our critical region, we fail to reject the null and do not have any evidence that there is a significant difference in the mean SAT scores for the two schools.

Exercise 2.16. *To compare two kinds of bumper guards, six of each kind were mounted on a certain make of compact car. Then each car was run into a concrete wall at* 5 *miles per hour, and the following are the costs of the repairs (in dollars):*

$$\begin{array}{lllllll} \text{Bumper guard 1}: & 127 & 168 & 143 & 165 & 122 & 139 \\ \text{Bumper guard 2}: & 154 & 135 & 132 & 171 & 153 & 149. \end{array} \tag{2.38}$$

Test, at the $\alpha = .01$ *significance level, whether or not there is a significant difference between the means of these two samples.* This example is from [MM].

3 Testing the Variance

Up to this point, we've only discussed methods for testing the mean. Now we discuss how to test the other important population parameter: the variance. As it turns out, the sample variance s^2 will be our estimator for σ^2 just like $\bar{x}$ was our estimator for μ. Our test statistic for the variance will be:

$$\chi^2 \;=\; \frac{(n-1)s^2}{\sigma^2}, \tag{3.39}$$

a χ^2 test with $n-1$ degrees of freedom. Recall the density function of a chi-square distribution with ν degrees of freedom is given by $f(x) = 0$ for $x < 0$ and

$$f(x) \;=\; \frac{1}{2^{\nu/2}\Gamma(\nu/2)}\; x^{\frac{\nu-2}{2}}\; e^{-x/2}. \tag{3.40}$$

The reason the chi-square distribution arises so frequently is that if X has the standard normal distribution then X^2 has a chi-square distribution with 1 degree of freedom. Another nice properties of chi-square distributions is that if X_i has a chi-square distribution with ν_i degrees of freedom, then $X_1 + \cdots + X_N$ has a chi-square distribution with $\nu_1 + \cdots + \nu_N$ degrees of freedom. When we study the sample variance, we have from (2.16) that

$$s^2 \;=\; \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}. \tag{3.41}$$

Thus if X_i is normally distributed then we basically have a sum of squares of normally distributed random variables. This observation explains why the chi-square distribution arises in these tests.

The null hypothesis we always test is $\sigma^2 = \sigma_0^2$, and our alternative hypothesis is always either $\sigma^2 < \sigma_0^2$ (the one-tailed, left-handed case), $\sigma^2 > \sigma_0^2$ (the one-tailed, right-handed case), or $\sigma^2 \neq \sigma_0^2$ (the two-tailed case). We construct our critical region of size α in the one-tailed hypothesis case as either $\chi^2 \geq \chi^2_{\alpha,\, n-1}$ in the right-handed case and $\chi^2 \leq \chi^2_{1-\alpha,\, n-1}$ in the left-handed case. In the two-tailed case, our critical region of size α will be all values of χ^2 such that either $\chi^2 \leq \chi^2_{1-\alpha/2,\, n-1}$ or $\chi^2 \geq \chi^2_{\alpha/2,\, n-1}$. All such χ^2 values, for a given α and given degrees of freedom may be found online at

`http://www.fourmilab.ch/rpkp/experiments/analysis/chiCalc.html`

or in a statistics table of χ^2 values. We only require that our sample of size n be taken from a normal population for this test to work correctly.

Example 3.1. *Suppose that the thickness of a part used in a semiconductor is its critical dimension and that measurements of the thickness of a random sample of* 18 *such parts have the variance* $s^2 = 0.68$*, where the measurements are in thousandths of an inch. The process is considered to be under control if the variation of the thickness is given by a variance not greater than* 0.36*. Assuming that the measurements constitute a random sample from a normal population, test the null hypothesis* $\sigma^2 = 0.36$ *against the alternative* $\sigma^2 > 0.36$ *at the* $\alpha = .05$ *significance level.*

We've already been given our hypotheses. Our alternative is one-tailed and right-handed, so we needed a right-handed critical region of size .05 *with* 17 *degrees of freedom. Looking at our statistics tables, we find our critical region to be* $\chi^2 \geq 27.587$*. So calculating our test statistic with* $s^2 = 0.68$*,* $n = 18$ *and* $\sigma^2 = 0.36$ *gives us*

$$\chi^2 \;=\; \frac{(17)(0.68)}{0.36} \;=\; 32.11. \tag{3.42}$$

Since our test statistic lies in our critical region, we reject the null hypothesis in favor of the alternative and conclude that $\sigma^2 > 0.36$ *based on this study.* This example is from [MM].

Exercise 3.2. *Nine determinations of the specific heat of iron had a standard deviation of* 0.0086*. Assuming that these determinations constitute a random sample from a normal population, test the null hypothesis* $\sigma = 0.0100$ *against the alternative hypothesis* $\sigma < 0.0100$ *at the* $\alpha = .05$ *significance level. (Hint: What is the relation between standard deviation and variance?)* This example is from [MM].

Exercise 3.3. *In a random sample, the weights of* 24 *Black Angus steers of a certain age have a standard deviation of* 238 *pounds. Assuming that the weights constitute a random sample from a normal population, test the null hypothesis* $\sigma = 250$ *pounds against the two-tailed alternative* $\sigma \neq 250$ *pounds at the* $\alpha = .01$ *significance level.* This example is from [MM].

4 Testing Proportions

4.1 Testing Proportions I: One Proportion

Oftentimes we come across data that lends itself well to discrete distributions, as opposed to the continuous ones we've been looking at. One such example is the binomial distribution, which is good for instances in which events we're studying has only two possible outcomes (such as "heads" or "tails" with a coin). In this instance we look for the number of successes (often called x) in n trials (leaving $n - x$ failures) and look at either the number of successes or look at x/n, the proportion of trials that was a success. Many times the two are interchangeable.

For this type of test, it is often more convenient to use the P-value to determine whether or not to reject the null hypothesis. Simply put, the P-value is the probability that, assuming the null is true, the random variable takes on values that far or farther away from the mean. It's equivalent to looking at values within the critical region; if the probability that such an observation occurs is less than α, we reject the null hypothesis. One may use the P-value in all the tests we've performed up to now, but this is the first time where there is any advantage to using it.

Remark 4.1. We strongly encourage anyone who will have to accept or reject a statistical hypothesis to read the article *On p-values* [Ia].

We may, of course, still construct an actual critical region if we'd like. To test a one-tailed alternative $\theta > \theta_0$ against the null hypothesis $\theta = \theta_0$, the critical region is $x \geq k_\alpha$, where k_α is the

smallest integer such that

$$\sum_{y=k_\alpha}^{n} b(y; n, \theta_0) \leq \alpha, \tag{4.43}$$

and x is the number of observed successes. Similarly, to test a one-tailed alternative $\theta < \theta_0$, the critical region is $x \leq k'_\alpha$, where k'_α is the largest integer such that

$$\sum_{y=0}^{k'_\alpha} b(y; n, \theta_0) \leq \alpha. \tag{4.44}$$

For the two-tailed alternative, we use $x \geq k_{\alpha/2}$ and $x \leq k'_{\alpha/2}$ as our critical region. Again, it is much easier to work with P-values, but you may construct these critical regions instead if you'd like.

Example 4.2. *If $x = 12$ of $n = 20$ patients suffered serious side effects from a new medication, test the null hypothesis $\theta = 0.50$ against the alternative hypothesis $\theta \neq 0.50$ at the $\alpha = .05$ significance level. Here θ is the true proportion of patients suffering serious side effects from the new medication.*

Our hypotheses are:

$$\begin{aligned} H_0 : \theta &= 0.50 \\ H_1 : \theta &\neq 0.50 \\ \alpha &= 0.05. \end{aligned} \tag{4.45}$$

Our test statistic in this case is simply going to be the number of successes, which in this case is $x = 12$. Since we're using the binomial distribution, we can find the probability that $X \geq 12$ when $\theta = 0.50$. Using a statistics table, we find this probability to be 0.2517. Since our alternative is two-tailed, we multiply this by two to get the P-value, which is 0.5034, much larger than 0.05. Therefore we fail to reject the null hypothesis that $\theta = 0.50$. This example is from [MM].

Exercise 4.3. *Suppose we're looking at a two-candidate mayoral race in Providence. Candidate A is declared the winner with 55 percent of the vote (i.e. a θ of .55). However, Candidate B is suspicious of these results. Having his group of close friends take a random sample of 17 voters from Providence, he finds that 7 voted for Candidate A while 10 voted for him. On the basis of this study performed at the $\alpha = .05$ level of significance, should Candidate B demand a recount?*

Exercise 4.4. *A doctor claims that less than 30 percent of all persons exposed to a certain amount of radiation will feel any ill effects. If, in a random sample, only 1 of 19 persons exposed to such radiation felt any ill effects, test the null hypothesis $\theta = 0.30$ against the alternative hypothesis $\theta < 0.30$ at the $\alpha = .05$ significance level.* This example is from [MM].

If our sample size n is large enough (typically $n \geq 30$ suffices), we may instead use a z-test based on the following test statistic:

$$z = \frac{x - n\theta}{\sqrt{n\theta(1-\theta)}}. \tag{4.46}$$

Here our sampling distribution will be the standard normal distribution, and we are able to construct critical regions as we did in §2.1.

Exercise 4.5. *In a random survey of 1000 households in the United States, it is found that 29 percent of the households have at least one member with a college degree. Does this finding refute the statement that the proportion of all such United States households is at least 35 percent? Test at the $\alpha = .05$ significance level.* This example is from [MM].

4.2 Testing Proportions II: K Proportions

We now expand upon the methods discussed in §4.1 to explore K proportions instead of just one. Such methods will allow us to see whether observed differences in proportions are due to chance or if they're actually significant differences.

Suppose we have $X_1, \ldots, X_K$ independent random variables with X_i having the binomial distribution with parameters n_i (number of trials) and θ_i (probability of success). Thus we have K populations, and with each population is associated exactly one proportion (the probability of success). Let $x_1, \ldots, x_K$ be the observed values from these distributions. If all n_i's are sufficiently large then we may use the central limit theorem to approximate each by the standard normal:

$$z_i \;=\; \frac{X_i - n_i\theta_i}{\sqrt{n_i\theta_i(1-\theta_i)}}. \tag{4.47}$$

Since each gives a normal distribution, we know that

$$\chi^2 \;=\; \sum_{i=1}^{K} \frac{(x_i - n_i\theta_i)^2}{n_i\theta_i(1-\theta_i)} \tag{4.48}$$

will be a χ^2 distribution with K degrees of freedom.

Therefore, to test the null hypothesis that $\theta_1 = \cdots = \theta_K = \theta_0$ against the alternative hypothesis that some $\theta_i \neq \theta_0$, we use $\chi^2 \geq \chi^2_{\alpha,\ K}$ as our critical region of size α, where

$$\chi^2 \;=\; \sum_{i=1}^{K} \frac{(x_i - n_i\theta_0)^2}{n_i\theta_0(1-\theta_0)}. \tag{4.49}$$

If θ_0 is not specified then we have to use the pooled estimate

$$\hat{\theta} \;=\; \frac{x_1 + \cdots + x_K}{n_1 + \cdots + n_K} \tag{4.50}$$

and the critical region becomes $\chi^2 \geq \chi^2_{\alpha,\ K-1}$, where

$$\chi^2 \;=\; \sum_{i=1}^{K} \frac{(x_i - n_i\hat{\theta})^2}{n_i\hat{\theta}(1-\hat{\theta})}. \tag{4.51}$$

Observe that when we have to use the pooled estimate, we lose a degree of freedom. Therefore, it is very important to keep track of when you use the pooled estimate and when you don't, as doing so affects your test statistic and critical region.

While you may use this method if you want, we introduce another method that is easier to generalize to the case when there are more than just two possible outcomes; these are the $r \times c$ contingency tables, which we study in §4.3. Suppose we set up a table of our observations in the following way:

	Successes	Failures
Sample 1	x_1	$n_1 - x_1$
Sample 2	x_2	$n_2 - x_2$
$\vdots$	$\vdots$	$\vdots$
Sample K	x_K	$n_K - x_K$.

(4.52)

These are our observations, which we label f_{ij}, where i refers to the sample, $j = 1$ if we're discussing successes, and $j = 2$ if we're discussing failures. Thinking of this information as a $K \times 2$ matrix, f_{ij} refers to the ij^{th} entry of the matrix. We use the letter f to denote frequency and e to denote expectation.

When we're considering the null hypothesis $\theta_1 = \cdots = \theta_K = \theta_0$, we have the following expected values:

$$e_{i1} \;=\; n_i\theta_0 \quad e_{i2} \;=\; n_i(1-\theta_0). \tag{4.53}$$

When we have to use the pooled estimate $\hat{\theta}$, the expected values are:

$$e_{i1} \;=\; n_i\hat{\theta} \quad e_{i2} \;=\; n_i(1-\hat{\theta}). \tag{4.54}$$

Simple algebra yields that our test statistic as:

$$\chi^2 \;=\; \sum_{i=1}^{K}\sum_{j=1}^{2}\frac{(f_{ij}-e_{ij})^2}{e_{ij}} \tag{4.55}$$

and we may use the exact same critical regions as before. This test statistic will come up again in §4.3.

Exercise 4.6. *Prove* (4.55).

Example 4.7. *Determine, on the basis of the sample data shown below, whether the true proportion of shoppers favoring detergent A over detergent B is the same in all three cities:*

	# favoring Det. A	# favoring Det. B	
Los Angeles	232	168	400
San Diego	260	240	500
Fresno	197	203	400.

(4.56)

Use the $\alpha = .05$ significance level.

$$\begin{aligned} H_0:&\quad \theta_1 \;=\; \theta_2 \;=\; \theta_3 \\ H_1:&\quad \theta_1,\ \theta_2,\ \text{and } \theta_3 \text{ are not all equal.} \end{aligned} \tag{4.57}$$

So we construct the critical region, and since we have to use the pooled estimate $\hat{\theta}$, we have $3-1=2$ degrees of freedom. Using a statistic table, one sees that $\chi^2_{.05,\ 2} = 5.991$. So we reject the null if our test statistic satisfies $\chi^2 \geq 5.991$. We will use the test statistic described in (4.55).

So now we calculate the pooled estimate of θ to be

$$\hat{\theta} \;=\; \frac{232+260+197}{400+500+400} \;=\; \frac{689}{1300} \;=\; 0.53. \tag{4.58}$$

Using this, we calculate our expected cell frequencies to be:

$$\begin{aligned} e_{11} &= 400(.53) = 212 \quad e_{12} = 400(.47) = 188 \\ e_{21} &= 500(.53) = 265 \quad e_{22} = 500(.47) = 235 \\ e_{31} &= 400(.53) = 212 \quad e_{32} = 400(.47) = 188. \end{aligned} \tag{4.59}$$

So we use (4.55) to calculate χ^2, which turns out to be 6.48. Since $6.48 > 5.991$, we reject the null hypothesis in favor of the alternative that not all true proportions are the same. This example is from [MM].

Exercise 4.8. *If* 26 *of* 200 *tires of brand A failed to last* 30000 *miles, whereas the corresponding figures for* 200 *tires of brands B, C, and D were* 23, 15, *and* 32 *respectively, test the null hypothesis that there is no difference in the quality of the four kinds of tires at the* $\alpha = .05$ *significance level.* This example is from [MM].

Exercise 4.9. *In random samples of* 250 *persons with low incomes,* 200 *persons with average incomes, and* 150 *persons with high incomes, there were, respectively,* 155, 118 *and* 87 *who favor a certain piece of legislation. Use the* $\alpha = .05$ *significance level to test the null hypothesis* $\theta_1 = \theta_2 = \theta_3$ *against the alternative that not all three* θ*'s are equal.* This example is from [MM].

4.3 Testing $r \times c$ Contingency Tables

We now can expand further our view of the situation in §4.2 to allow for an arbitrary number of outcomes (whereas in §4.2 there were only two, namely success and failure). This method is our best method when we're testing an association between categorical variables. In this situation, we have distinct samples from r populations, and we have c distinct outcomes, and we categorize each member of each sample into their appropriate outcome. The results can be written up in a table, which is often called an $r \times c$ contingency table. Realizing that each population comes from a multinomial distribution $\text{mult}(n_i, \theta_{i,1}, ..., \theta_{i,c})$, we may formulate our hypotheses in the following way:

$$\begin{aligned} H_0: &\quad \text{For all indices } 1 \leq \text{j} \leq \text{c}, \text{we have } \theta_{\text{i,j}} = \theta_{\text{k,j}} \text{ for all} 1 \leq \text{i}, \text{k} \leq \text{r} \\ H_1: &\quad \text{There exist indices i, j, k such that } \theta_{\text{i,j}} \neq \theta_{\text{k,j}}. \end{aligned}$$

These will always be the hypotheses we're testing with this particular test. Our test statistic is going to be

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}, \tag{4.60}$$

where the f_{ij}'s come from the collected data (f_{ij} is the number of observations in population i of outcome j) and the e_{ij}'s are calculated in the following way:

$$e_{ij} = \frac{n_i f_j}{n}. \tag{4.61}$$

Here n_i is the size of the sample taken from population i, $f_j = \sum_i f_{ij}$ is the total number of times we observe outcome j, and $n = n_1 + \cdots + n_r$. This test statistic gives us a t-distribution with $(r-1)(c-1)$ degrees of freedom.

Example 4.10. *Use the following data to test at the* $\alpha = .01$ *significance level whether a person's ability in mathematics is independent of his or her interest in statistics:*

	Low Ability	Average Ability	High Ability
Low Interest	63	42	15
Average Interest	58	61	31
High Interest	14	47	29.

(4.62)

We begin, as always, by stating the hypotheses:

$$\begin{aligned} H_0: &\quad \text{For all indices } 1 \leq \text{j} \leq 3, \text{we have } \theta_{\text{i,j}} = \theta_{\text{k,j}} \text{ for all} 1 \leq \text{i}, \text{k} \leq 3 \\ H_1: &\quad \text{There exist indices i, j, k such that } \theta_{\text{i,j}} \neq \theta_{\text{k,j}}. \end{aligned} \tag{4.63}$$

In the context of this problem, H_0 states that ability in mathematics and interest in statistics are independent, whereas H_1 says they are not. We are testing at the $\alpha = .01$ level and we have $(3-1)(3-1) = (2)(2) = 4$ degrees of freedom, giving us a critical value of 13.277. So to reject H_0, we need $\chi^2 \geq 13.277$.

Now we need to calculate the e_{ij}'s. Calculating e_{11}, we need to know the total number with low interest in statistics, the total with low ability in math, and the total studied in this study. There are $63+42+15 = 120$ with low interest in statistics, $63+58+14 = 135$ with low ability in math, and the total number studied is $63+42+15+58+61+31+14+47+29 = 360$. So $e_{11} = \frac{(120)(135)}{360} = 45$. To find e_{12}, we take $e_{12} = \frac{(120)(150)}{360} = 50$. Continuing in this way we can find all the e_{ij}'s, which are given below:

$$\begin{array}{lll} e_{11} = 45.00 & e_{12} = 50.00 & e_{13} = 25.00 \\ e_{21} = 56.25 & e_{22} = 62.50 & e_{23} = 31.25 \\ e_{31} = 33.75 & e_{32} = 37.50 & e_{33} = 18.75. \end{array} \tag{4.64}$$

We now use these to calculate our test statistic, which gives us:

$$\chi^2 = \frac{(63-45.00)^2}{45.0} + \frac{(42-50.00)^2}{50.0} + \cdots + \frac{(29-18.75)^2}{18.75} = 32.14. \tag{4.65}$$

Since $32.14 > 13.277$, we reject the null hypothesis in favor of the alternative. This example is from [MM].

Exercise 4.11. *Tests of the fidelity and the selectivity of 190 radios produced the following results:*

	Low Fidelity	Average Fidelity	High Fidelity
Low Selectivity	7	12	31
Average Selectivity	35	59	18
High Selectivity	15	13	0.

(4.66)

At the $\alpha = .01$ significance level, test the null hypothesis that fidelity is independent from selectivity. This example is from [MM].

Exercise 4.12. *Four coins were tossed 160 times and 0, 1, 2, 3, or 4 heads showed up, respectively, 19, 54, 58, 23, and 6 times. At the $\alpha = .05$ significance level, test whether it is reasonable to suppose that the coins are balanced and randomly tossed.* This example is from [MM].

4.4 Incomplete $r \times c$ Contingency Tables Tables

Consider the following question: are runs scored and allowed in a baseball game independent? Assume that it is appropriate to analyze this with an $r \times c$ contingency table. The various populations will be the different runs scored (say 0, 1, 2, 3, ...), while the various outcomes will be the runs allowed (say 0, 1, 2, 3, ...). Except for one All-Star game, baseball games *cannot* end in ties. Thus we know that all the diagonal entries of the above table *must* be zero! These special entries are called structural zeros. They are very different from an entry that happens to be zero because nothing was observed: these entries are zero because it is *impossible* to observe something here.

For more details on handling structural zeros, see [BF, SD]; for the analysis of the baseball example above, see [Mil1].

We describe the modified χ^2 test for an incomplete two-dimensional contingency table with diagonal entries forced to be zero. We assume there are N populations and N outcomes for each population (i.e., $r = c = N$). Thus there are N structural zeros.

We describe the iterative fitting procedure given in the appendix to [BF] to obtain maximum likelihood estimators for the $E_{r,c}$, the expected frequency of cell (r,c), under the assumption that the diagonal cells must be empty. For $1 \le r,c \le N$, let $E_{r,c}^{(0)} = 1$ if $r \ne c$ and 0 if $r = c$. Set

$$X_{r,+} \;=\; \sum_{c=1}^{N} F_{r,c}, \qquad X_{+,c} \;=\; \sum_{r=1}^{N} F_{r,c}. \tag{4.67}$$

Then

$$E_{r,c}^{(\ell)} \;=\; \begin{cases} E_{r,c}^{(\ell-1)} X_{r,+} \;/\; \sum_{c=1}^{N} E_{r,c}^{(\ell-1)} & \text{if } \ell \text{ is odd} \\ \\ E_{r,c}^{(\ell-1)} X_{+,c} \;/\; \sum_{r=1}^{N} E_{r,c}^{(\ell-1)} & \text{if } \ell \text{ is even,} \end{cases} \tag{4.68}$$

and

$$E_{r,c} \;=\; \lim_{\ell\to\infty} E_{r,c}^{(\ell)}; \tag{4.69}$$

the iterations converge very quickly in practice. Then

$$\sum_{r=1}^{N} \sum_{\substack{c=1 \\ c\ne r}}^{N} \frac{(F_{r,c} - E_{r,c})^2}{E_{r,c}} \tag{4.70}$$

is approximately a χ^2 distribution with $(N-1)^2 - N$ degrees of freedom.

Exercise 4.13. *Show that if we had a complete two-dimensional contingency table, then the iteration reduces to the standard values, namely $E_{r,c} = \sum_{c'} F_{r,c'} \cdot \sum_{r'} F_{r',c} \;/\; F$, where $F = \sum_r \sum_c F_{r,c}$ is the total number of observations.*

5 Normal Regression Analysis

In this section, we discuss tests relating to one of the most important topics in statistics: establishing relationships between variables so that knowledge of one allows us to make inferences about the other. The most common form comes about in linear regression; that is, when the relationship between two variables is a line. One can use the method of least squares (see [Mil2]) to take given data and give it a best fit line $\hat{y} = \hat{\alpha} + \hat{\beta}x$. To find the coefficients $\hat{\alpha}$ and $\hat{\beta}$, we use the following formulas:

$$\begin{aligned} \hat{\beta} &= \frac{S_{xy}}{S_{xx}} \\ \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x}, \end{aligned} \tag{5.71}$$

where $\bar{x}$ and $\bar{y}$ are the sample means for the x_i's and y_i's respectively, and

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \\ S_{yy} &= \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 \\ S_{xy} &= \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right). \end{aligned} \tag{5.72}$$

All of these values will be very important to our work.

When we analyze a set of paired data $\{(x_i, y_i)|i = 1, 2, ..., n\}$ by regression analysis, we look at the x_i's as constant and the y_i's as values of corresponding independent random variables Y_i. This allows us to look at the conditional density of the random variable Y_i (for a fixed x_i) is the normal density

$$\frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{y_i-(\alpha+\beta x_i)}{\sigma}\right)^2}, \quad -\infty < y_i < \infty \tag{5.73}$$

where α, β, and σ are our parameters (σ is the standard deviation for this distribution, and will be used in our analysis).

Just as we calculated $\hat{\alpha}$ and $\hat{\beta}$ above to estimate α and β, we construct $\hat{\sigma}$ to estimate σ. As it turns out, one can show that

$$\hat{\sigma} \;=\; \sqrt{\frac{1}{n}(S_{yy} - \hat{\beta} \cdot S_{xy})}. \tag{5.74}$$

Then one can show (see [MM] for details) that under the assumptions of normal regression analysis, the test statistic

$$t \;=\; \frac{\hat{\beta} - \beta}{\hat{\sigma}}\sqrt{\frac{(n-2)S_{xx}}{n}} \tag{5.75}$$

is a random variable having the t distribution with $n - 2$ degrees of freedom. Therefore, we can use our earlier understanding of the t distribution to now test hypotheses about the regression coefficient β.

Example 5.1. *Suppose a collection of 10 paired observations has yielded $\hat{\beta} = 3.471$, $\hat{\sigma} = 4.720$, and $S_{xx} = 376$. Test the null hypothesis $\beta = 3$ against the alternative $\beta > 3$ at the $\alpha = .01$ significance level.*

Look at a table of values for the t distribution with 8 degrees of freedom, we see our one-tailed critical region is $t \geq 2.896$. We calculate the test statistic to be

$$t \;=\; \frac{3.471 - 3}{4.720}\sqrt{\frac{8 \cdot 376}{10}} \;=\; 1.73. \tag{5.76}$$

Since 1.73 is not in our critical region, we fail to reject the null hypothesis that $\beta = 3$. This example is from [MM].

Exercise 5.2. *The following data shows the assessed values and the selling prices (in thousands of dollars) of eight houses, constituting a random sample of all houses sold recently in the metropolitan area:*

Assessed value	*Selling price*
70.3	114.4
102.0	169.3
62.5	106.2
74.8	125.0
57.9	99.8
81.6	132.1
110.4	174.2
88.0	143.5.

(5.77)

Use this data to calculate a best fit line (that is, find $\hat{\beta}$ and $\hat{\alpha}$). Then test the null hypothesis $\beta = 1.30$ against the alternative hypothesis $\beta > 1.30$ at the $\alpha = .05$ significance level. This example is from [MM].

6 Non-parametric Tests

6.1 Tests of Signs

The sign test is one of our first examples of a nonparametric test; that is, a test where we make no assumptions about the functional form of the underlying population. In the case of the signs test, we only assume that the population is continuous and symmetrical about the mean (we discuss this assumption in greater detail in Remark 6.3). It is often used as an alternative to the one-sample t-test where we test the null hypothesis $\mu = \mu_0$ against some alternative.

In the sign test, we replace each sample value exceeding μ_0 with a plus sign and replace each sample value less than μ_0 with a minus sign (since the population is continuous, then the probability of getting a value equal to μ_0 is zero; in cases where we do get a value of μ_0, we simply discard it) and then test the equivalent null hypothesis that the number of plus signs is a random variable having a binomial distribution with parameters n (number of plus signs and minus signs) and $\theta = 1/2$. Our two-tailed alternative $\mu \neq \mu_0$ becomes $\theta \neq 1/2$, while one-tailed alternatives such as $\mu > \mu_0$ and $\mu < \mu_0$ become $\theta > 1/2$ and $\theta < 1/2$, respectively. When n is small, we refer to a table of binomial probabilities directly, and when n is large, one can use the normal approximation introduced in §4.1 with $\theta = 1/2$.

Example 6.1. *The following are measurements of the breaking strength of a certain kind of* 2*-inch cotton ribbon in pounds:*

$$\begin{array}{cccccccccc} 163 & 165 & 160 & 189 & 161 & 171 & 158 & 151 & 169 & 162 \\ 163 & 139 & 172 & 165 & 148 & 166 & 172 & 163 & 187 & 173. \end{array} \tag{6.78}$$

Use the sign test to test the null hypothesis $\mu = 160$ *against the alternative* $\mu > 160$ *at the* $\alpha = .05$ *significance level.*

Since we already have our hypotheses, we replace each value exceeding 160 *with a plus sign and each value less than* 160 *with a minus sign:*

$$\begin{array}{cccccccccc} + & + & + & + & + & - & - & + & + & + \\ - & + & + & - & + & + & + & + & + & \end{array} \tag{6.79}$$

(after having removed the value equal to 160*). So* $n = 19$ *and* $x = 15$*. Using a table of binomial probabilities, we find that* $P(x \geq 15;\ n = 19) = 0.0095$*. Since* $0.0095 < 0.05$*, we reject the null hypothesis in favor of the alternative that* $\mu > 160$*.* This example is from [MM].

Exercise 6.2. *The following are the amounts of time, in minutes, that it took a random sample of* 20 *technicians to perform a certain task:* 18.1*,* 20.3*,* 18.3*,* 15.6*,* 22.5*,* 16.8*,* 17.6*,* 16.9*,* 18.2*,* 17.0*,* 19.3*,* 16.5*,* 19.5*,* 18.6*,* 20.0*,* 18.8*,* 19.1*,* 17.5*,* 18.5*, and* 18.0*. Assuming that this sample came from a symmetrical continuous population, use the sign test at the* $\alpha = .05$ *significance level to test the null hypothesis that* $\mu = 19.4$ *against the alternative hypothesis* $\mu \neq 19.4$*. First perform the test using a table for binomial distributions, and then perform the test using the normal approximation to the binomial distribution.* This example is from [MM].

Remark 6.3. While many populations are symmetric about their mean, not all are. More generally, for a continuous population we look at fluctuations above and below the median. Recall the median $\tilde{\mu}$ is such that the probability of observing a value less than $\tilde{\mu}$ is 50%, as is the probability of observing a value greater than $\tilde{\mu}$; we must be careful with discrete populations, as there singletons can have positive probability.

6.2 Tests of Ranked Signs

While the test for signs from §6.1 is effective and easy to perform, it ignores a lot of information; namely it only concerns itself with whether or not a value is above or below μ_0, but it doesn't consider HOW far above or below that value is. That is what we explore in this section, using a new test called the Signed-Rank Test. We again assume that the population is symmetric and continuous. In this test, we first calculate each of the differences between the supposed mean μ_0 and every data entry collected (entries equal to μ_0 are again discarded). Then we rank the absolute values of the differences in the following way: the smallest difference in absolute value is ranked 1, the next smallest is ranked 2, and so forth until all are ranked. If two or more entries share the same absolute value, we assign them the mean of the ranks that they would have occupied (for instance, if three entries have the same absolute value and would have taken ranks 2, 3, and 4, then we assign each the rank $\frac{2+3+4}{3} = 3$). This allows us to find the following three test statistics:

$$\begin{aligned} T^+ &= \sum(\text{ranks assigned to positive differences}) \\ T^- &= \sum(\text{ranks assigned to negative differences}) \\ T &= \min\{\mathrm{T}^+, \mathrm{T}^-\}. \end{aligned} \tag{6.80}$$

We then use each of these test statistics depending on our alternative hypothesis, as seen below:

$$\begin{array}{cc} \text{Alternative Hypothesis} & \text{Reject Null If :} \\ \mu \neq \mu_0 & T \leq T_\alpha \\ \mu > \mu_0 & T^- \leq T_{2\alpha} \\ \mu < \mu_0 & T^+ \leq T_{2\alpha}. \end{array} \tag{6.81}$$

Here α is the significance level, and T_α or $T_{2\alpha}$ can be found in a table of critical values for the Signed-Rank Test (for a given n, our sample size).

Example 6.4. *The following are* 15 *measurements of the octane rating of a certain kind of gasoline:* 97.5*,* 95.2*,* 97.3*,* 96.0*,* 96.8*,* 100.3*,* 97.4*,* 95.3*,* 93.2*,* 99.1*,* 96.1*,* 97.6*,* 98.2*,* 98.5*, and* 94.9*. Use the Signed-Rank test with* $\alpha = .05$ *to test whether the mean octane rating of the given kind of gasoline is* 98.5*.*

So we begin by stating our hypotheses:

$$\begin{aligned} H_0 &: \mu = 98.5 \\ H_1 &: \mu \neq 98.5. \end{aligned} \tag{6.82}$$

This means we'll be looking at the test statistic T*, and our critical region will be* $T \leq T_{.05}$*. There is only one value equal to the mean so we remove it and are left with* $n = 14$*. Using our statistics table, we see that* $T_{.05} = 21$ *when* $n = 14$*, so our critical region will be* $T \leq 21$*. So subtracting* 98.5

from each entry and ranking the absolute value of the difference, we get the following:

$$\begin{array}{ccc} Measurement & Difference & Rank \\ 97.5 & -1.0 & 4 \\ 95.2 & -3.3 & 12 \\ 97.3 & -1.2 & 6 \\ 96.0 & -2.5 & 10 \\ 96.8 & -1.7 & 7 \\ 100.3 & 1.8 & 8 \\ 97.4 & -1.1 & 5 \\ 95.3 & -3.2 & 11 \\ 93.2 & -5.3 & 14 \\ 99.1 & 0.6 & 2 \\ 96.1 & -2.4 & 9 \\ 97.6 & -0.9 & 3 \\ 98.2 & -0.3 & 1 \\ 94.9 & -3.6 & 13. \end{array} \tag{6.83}$$

We can use this to calculate T^- and T^+.

$$T^- \;=\; 4+12+6+10+7+5+11+14+9+3+1+13 \;=\; 95 \tag{6.84}$$

and

$$T^+ \;=\; 8+2 \;=\; 10. \tag{6.85}$$

So $T = 10$. Since $10 < 21$, we reject the null hypothesis and accept the alternative that the mean octane rating for this kind of gasoline is not 98.5*.*

Exercise 6.5. *Redo Exercise 6.2 using the Signed-Rank Test instead of the signs test.*

6.3 Tests Based on Runs

The final type of test we explore is based on the theory of runs, where a run is a succession of identical letters (or other kinds of symbols). It could refer to consecutive heads in flips of a fair coin, for instance (when written on paper, this run would be written $\ldots, H, H, H, H, H, \ldots$; this is a run of 5 consecutive heads).

So we consider a string of observations A and B, and we want to test whether or not this string is random by exploring the number of runs the string has. This is always the hypothesis we test when using a runs test; whether or not the given string is random. Let n_1 be the number of A's in the strong, n_2 the number of B's, and u the number of runs. For example, if $n_1 = 3$ and $n_2 = 2$ then there are $5!/3!2! = 10$ distinct strings (remember all A's are identical and all B's are identical).

We list the strings[1], and the number of runs in each:

$$\begin{array}{ll} AAABB & 2 \\ AABBA & 3 \\ ABBAA & 3 \\ BBAAA & 2 \\ AABAB & 4 \\ ABABA & 5 \\ BABAA & 4 \\ ABAAB & 4 \\ BAABA & 4 \\ BAAAB & 3. \end{array} \tag{6.86}$$

Thus 40% of the time there are 4 runs, 30% of the time there are 3 runs, 20% of the time there are two runs and only 10% of the time are there 5 runs. Thus, if we observe 5 runs, we might be moderately surprised. We would only be moderately surprised because the sample size is so small. If instead we had $n_1 = 20$ and $n_2 = 20$, then we would be astonished if there were 39 or 40 runs; the runs test quantifies our astonishment, telling us how likely it is to observe a given number of runs *under the assumption that, once we have specified how many A's and B's we have, the order of the A's and B's is random.*

We now describe how to perform the runs test. When n_1 and n_2 are small, the test is performed by constructing the critical region $u \le u'_{\alpha/2}$ or $u \ge u_{\alpha/2}$, where $u_{\alpha/2}$ and $u'_{\alpha/2}$ are found in a statistics table of critical values for tests based on runs (these are determined by the pair (n_1, n_2) as well as the value α).

Example 6.6. *Checking on elm trees that were planted many years ago along a county road, a county official obtained the following arrangement of healthy, H, and diseased, D, trees:*

$$H\ H\ H\ H\ D\ D\ D\ H\ H\ H\ H\ H\ H\ H\ D\ D\ H\ H\ D\ D\ D\ D. \tag{6.87}$$

Test at the $\alpha = .05$ significance level whether this arrangement may be regarded as random.

We begin by stating our hypotheses:

$$\begin{array}{ll} H_0: & \text{Arrangement is random} \\ H_1: & \text{Arrangement is not random.} \end{array} \tag{6.88}$$

Since $\alpha = .05$, $n_1 = 13$, and $n_2 = 9$, we use our table to find $u'_{.025}$ and $u_{.025}$, which happen to be 6 and 17, respectively. So we reject the null hypothesis if $u \le 6$ or $u \ge 17$. By inspection of the data, $u = 6$, which lies within our critical region. Therefore we reject the null hypothesis and conclude that this arrangement is not random. This example is from [MM].

[1]There are 5!/3!2! distinct strings (as all A's are the same and all B's are the same). This is similar to counting how many different words can be written with the letters in MISSISSIPPI. There are 11 letters: 4 S's, 4 I's, 2 P's and 1 M. The number of distinct words is $11!/4! \cdot 4! \cdot 2! \cdot 1!$. In listing the 10 possible strings, it is important to choose a method of enumeration so that nothing is missed. Here we first looked at all the strings with the two B's adjacent, then with one A between the two B's, then two A's between the two B's, and finally three A's between the two B's. There are, of course, other ways of enumerating the 10 possiblities.

Exercise 6.7. *Suppose we flip a coin 15 times and come up with the following arrangement:*

$$H\ T\ T\ T\ H\ H\ T\ T\ T\ T\ H\ H\ T\ H\ H. \tag{6.89}$$

Test at the $\alpha = .01$ significance level whether this arrangement may be regarded as random.

When both n_1 and n_2 are larger than 10, we may approximate the distribution corresponding to u with a normal curve with mean

$$\mu \ = \ \frac{2n_1n_2}{n_1+n_2}+1 \tag{6.90}$$

and variance

$$\sigma^2 \ = \ \frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}. \tag{6.91}$$

Our test statistic then becomes

$$z \ = \ \frac{u-\mu}{\sigma} \tag{6.92}$$

and we construct our critical region as if we were looking at a two-tailed hypothesis.

Example 6.8. *The following is an arrangement of men, M, and women, W, lined up to purchase tickets for a rock concert:*

$$\begin{array}{cccccccccccc} M & W & M & W & M & M & M & W & M & W & M & M \\ M & W & W & M & M & M & M & W & W & M & W & M \\ M & M & W & M & M & M & W & W & W & M & W & M \\ M & M & W & M & W & M & M & M & M & W & W & M. \end{array} \tag{6.93}$$

Test for randomness at the $\alpha = .05$ significance level.

As always, we first state our hypotheses:

$$\begin{array}{l} H_0 : \text{Arrangement is random.} \\ H_1 : \text{Arrangement is not random.} \end{array} \tag{6.94}$$

We construct our critical region using the normal distribution, which tells us to reject the null hypothesis if $z \leq -1.96$ or $z \geq 1.96$. Since $n_1 = 30$, $n_2 = 18$, and $u = 27$, we get:

$$\mu \ = \ \frac{2 \cdot 30 \cdot 18}{30+18}+1 \ = \ 23.5 \tag{6.95}$$

and

$$\sigma \ = \ \sqrt{\frac{2 \cdot 30 \cdot 18(2 \cdot 30 \cdot 18-30-18)}{(30+18)^2(30+18-1)}} \ = \ 3.21. \tag{6.96}$$

Therefore we have

$$z \ = \ \frac{27-23.5}{3.21} \ = \ 1.09. \tag{6.97}$$

Since 1.09 is outside our critical region, we fail to reject the null hypothesis, and conclude there is no real evidence to indicate the arrangement is not random. This example is from [MM].

Exercise 6.9. *Suppose you are given the following results of* 100 *tosses of a coin:*

$$\begin{array}{cccccccccc} H & T & H & T & H & H & H & T & H & T \\ T & H & T & T & H & T & H & H & T & T \\ T & H & T & H & H & T & T & T & H & T \\ H & T & T & H & H & H & T & T & T & T \\ H & T & H & H & H & T & H & T & H & H \\ H & T & T & H & T & H & H & T & T & T \\ H & T & T & H & H & T & H & H & H & T \\ T & H & T & H & T & H & H & H & T & H \\ T & H & H & T & T & T & H & T & H & H \\ H & T & T & H & T & H & T & T & H & H; \end{array} \tag{6.98}$$

for ease of display we have written the data in a square form, though it should be regarded as one long sequence. Thus the last observation on the first line T, is followed by the first observation of the second line, T. Test the randomness of this string at both the $\alpha = .05$ level and the $\alpha = .01$ level.

7 Summary

In this section, we summarize the conditions under which each test may be used, the test statistic used in each test, and the critical regions (of each type) that we use in each test. This section will serve as a good, quick reference for those who already know HOW to apply the above tests, and just want to know when it's appropriate to apply a given test.

7.1 z-tests

Conditions for Use: We may only use the z-test if our sample is taken from a normal distribution with known variance or if our sample size is large enough to invoke the Central Limit Theorem (usually $n \geq 30$ is a good rule of thumb). In the latter case, we use s^2, the sample variance, instead of the population variance σ^2.

The Test Statistic: When our sample is taken from a normal distribution with known variance, then our test statistic is:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}, \tag{7.99}$$

where $\bar{x}$ is the sample mean, μ is the population mean, σ is the population standard deviation, and n is the sample size. If our population standard deviation is unknown but $n \geq 30$, we can invoke the Central Limit Theorem and use the test statistic:

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}}, \tag{7.100}$$

where s is the sample standard deviation.

The Critical Region: To construct a critical region of size α, we first examine our alternative hypothesis. If our hypothesis is one-tailed, our critical region is either $z \geq z_\alpha$ (if the hypothesis is right-handed) or $z \leq -z_\alpha$ (if the hypothesis is left-handed). If our hypothesis is two-tailed, then our critical region is $|z| \geq z_{\alpha/2}$.

7.2 t-tests

Conditions for Use: When we have a small sample size ($n < 30$) taken from a normal distribution of unknown variance, we may use the t-test with $n - 1$ degrees of freedom.

The Test Statistic: Our test statistic for the t-test is:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}, \tag{7.101}$$

where where $\bar{x}$ is the sample mean, μ is the population mean, s is the sample standard deviation, and n is the sample size.

The Critical Region: To construct a critical region of size α, we first examine our alternative hypothesis. If our hypothesis is one-tailed, our critical region is either $t \geq t_{\alpha,\ n-1}$ (if the hypothesis is right-handed) or $t \leq -t_{\alpha,\ n-1}$ (if the hypothesis is left-handed). If our hypothesis is two-tailed, then our critical region is $|t| \geq t_{\alpha/2,\ n-1}$.

7.3 Tests comparing means

Conditions for Use: If both distributions are normal with known variances, then we can use the z-test comparing means. If not, then if both samples sizes are large enough to invoke the central limit theorem, we may still use the z-test comparing means, with s_1^2 for σ_1^2 and s_2^2 for σ_2^2. If (at least) one sample is smaller than 30, then we can invoke the t-test comparing means provided both are taken from a normal population with equal variance. This t-test has $n_1 + n_2 - 2$ degrees of freedom.

The Test Statistic: If both distributions are normal with known variances, then our test statistic is:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}. \tag{7.102}$$

If we don't know the variance of both distributions, then we may use:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \tag{7.103}$$

provided n_1 and n_2 are large enough to invoke the Central Limit Theorem. If they aren't both large enough to invoke the Central Limit Theorem, then (provided our conditions are satisfied) our test statistic is:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{7.104}$$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}. \tag{7.105}$$

The Critical Region: When we use the z-test comparing means, our critical region is constructed in exactly the same way as in §7.1. When using the t-test comparing means, we again refer to our alternative hypothesis. If our hypothesis is one-tailed, our critical region is either $t \geq t_{\alpha,\ n_1+n_2-2}$ (if the hypothesis is right-handed) or $t \leq -t_{\alpha,\ n_1+n_2-2}$ (if the hypothesis is left-handed). If our hypothesis is two-tailed, then our critical region is $|t| \geq t_{\alpha/2,\ n_1+n_2-2}$.

7.4 Variance Test

Conditions for Use: The only condition required for using this test is that our sample is drawn for a normal distribution.

The Test Statistic: Under these conditions, our test statistic becomes a χ^2 statistic with $n-1$ degrees of freedom. Our test statistic is given by:

$$\chi^2 \ = \ \frac{(n-1)s^2}{\sigma^2}. \tag{7.106}$$

The Critical Region: The null hypothesis we always test is $\sigma^2 = \sigma_0^2$, and our alternative hypothesis is always either $\sigma^2 < \sigma_0^2$ (the one-tailed, left-handed case), $\sigma^2 > \sigma_0^2$ (the one-tailed, right-handed case), or $\sigma^2 \neq \sigma_0^2$ (the two-tailed case). We construct our critical region of size α in the one-tailed hypothesis case as either $\chi^2 \geq \chi^2_{\alpha,\ n-1}$ in the right-handed case and $\chi^2 \leq \chi^2_{1-\alpha,\ n-1}$ in the left-handed case. In the two-tailed case, our critical region of size α will be all values of χ^2 such that either $\chi^2 \leq \chi^2_{1-\alpha/2,\ n-1}$ or $\chi^2 \geq \chi^2_{\alpha/2,\ n-1}$. All such χ^2 values, for a given α and given degrees of freedom may be found online at

`http://www.fourmilab.ch/rpkp/experiments/analysis/chiCalc.html`

or in a statistics table of χ^2 values.

7.5 Proportions

Conditions for Use: For both the one proportion case and the K proportion case, we assume that our sample is drawn from a binomial distribution. In the one proportion case, for large values of n we may invoke the Central Limit Theorem and use the normal approximation to the binomial distribution. For K proportions, we construct a contingency table to test the data. Note for the K proportion case we have K populations; each population is a binomial (thus there are two possible outcomes, success of failure), and the proportion is the probability of success.

The Test Statistic: For the one proportion case, if n is small, our test statistic is simply X, the observed number of successes. If n is large, we can invoke the Central Limit Theorem, giving us the test statistic:

$$z \ = \ \frac{x - n\theta}{\sqrt{n\theta(1-\theta)}}. \tag{7.107}$$

For the K proportion case, our test statistic is:

$$\chi^2 \;=\; \sum_{i=1}^{K}\sum_{j=1}^{2}\frac{(f_{ij}-e_{ij})^2}{e_{ij}}. \tag{7.108}$$

The f_{ij}'s are our observed successes and failures; f_{i1} is the observed number of successes in population i, and $f_{i2} = n_i - f_{i1}$ is the observed number of failures in population i. The e_{ij}'s are calculated differently depending on the null hypothesis. If our null hypothesis is that $\theta_1 = \cdots = \theta_K = \theta_0$, then our distribution has K degrees of freedom and we have $e_{i1} = n\theta_0$ and $e_{i2} = n(1-\theta_0)$. If our null hypothesis is that $\theta_1 = \cdots = \theta_K$, then our distribution has $K-1$ degrees of freedom and we have $e_{i1} = n\hat{\theta}$ and $e_{i2} = n(1-\hat{\theta})$, where

$$\hat{\theta} \;=\; \frac{x_1 + \cdots + x_K}{n_1 + \cdots + n_K}. \tag{7.109}$$

The Critical Region: For the one proportion case, we calculate the P-value, and we reject the null if $P \le \alpha$. If our alternative hypothesis is one-tailed, then (given θ) $P = \text{Prob}(X \le x)$ (where x is our observed number of successes) in the left-handed case and $P = \text{Prob}(X \ge x)$ in the right-handed case. If our hypothesis is two-tailed, then $P = 2 \cdot \text{Prob}(X \ge x)$ if $x > \frac{n}{2}$ and $P = 2 \cdot \text{Prob}(X \le x)$ if $x < \frac{n}{2}$. When we use the normal approximation to the binomial distribution, our critical regions are constructed in the exact same way as in §7.1.

For the K proportion case, our critical region is $\chi^2 \ge \chi^2_{\alpha,\ K}$ if our null hypothesis is $\theta_1 = \cdots = \theta_K = \theta_0$ and our critical region is $\chi^2 \ge \chi^2_{\alpha,\ K-1}$ if our null hypothesis is $\theta_1 = \cdots = \theta_K$.

7.6 Contingency Tables

Conditions for Use: We assume we have r samples, each drawn from a multinomial distribution with c distinct outcomes. This is our most effective method of performing statistical analysis of categorical variables.

The Test Statistic: Our test statistic is:

$$\chi^2 \;=\; \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij}-e_{ij})^2}{e_{ij}}. \tag{7.110}$$

Here the f_{ij}'s are our observed values (f_{ij} is the number of times in the sample from population i that we observe outcome j) and the e_{ij}'s are calculated by

$$e_{ij} \;=\; \frac{n_i f_j}{n}, \tag{7.111}$$

with n_i the size of the sample from population i, $n = n_1 + \cdots + n_r$ is the total number of observations (the sum of the sample sizes of all the populations), and $f_j = \sum_i f_{ij}$ is the total number of times we observe outcome j.

The Critical Region: Our critical region for this test is $\chi^2 \ge \chi^2_{\alpha,\ (r-1)(c-1)}$.

7.7 Regression Analysis

Conditions for Use: If we have paired data (x_i, y_i) so that, for each fixed x_i, the conditional density of the corresponding random variable Y_i is the normal density:

$$\omega(y_i|x_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y_i-(\alpha+\beta x_i))^2/2\sigma^2} \quad -\infty < y_i < \infty, \tag{7.112}$$

then we may perform a regression analysis test.

The Test Statistic: Our test statistic for this test is:

$$t = \frac{\hat{\beta} - \beta}{\hat{\sigma}} \sqrt{\frac{(n-2)S_{xx}}{n}}, \tag{7.113}$$

where

$$\begin{aligned} \hat{\beta} &= \frac{S_{xy}}{S_{xx}} \\ \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x}, \end{aligned} \tag{7.114}$$

$\bar{x}$ and $\bar{y}$ are the sample means for the x_i's and y_i's respectively, and

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \\ S_{yy} &= \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 \\ S_{xy} &= \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right). \end{aligned} \tag{7.115}$$

The test statistic t has a Student's t-distribution with $n-2$ degrees of freedom.

The Critical Region: To construct a critical region of size α, we first examine our alternative hypothesis. If our hypothesis is one-tailed, our critical region is either $t \geq t_{\alpha,\ n-2}$ (if the hypothesis is right-handed) or $t \leq -t_{\alpha,\ n-2}$ (if the hypothesis is left-handed). If our hypothesis is two-tailed, then our critical region is $|t| \geq t_{\alpha/2,\ n-2}$.

7.8 Signs and Ranked Signs

Conditions for Use: For both the signs test and signed-rank test, we only need that the underlying population is continuous and symmetric.

The Test Statistic: For the signs test, our test statistic will simply be the number of observed plus signs when n is small. If n is large enough to invoke the Central Limit Theorem, we may use the normal approximation to the binomial distribution:

$$z = \frac{x - n\theta}{\sqrt{n\theta(1-\theta)}} \tag{7.116}$$

with $\theta = 1/2$.

For the signed-rank test, we have three test statistics:

$$\begin{aligned} T^+ &= \sum(\text{ranks assigned to positive differences}) \\ T^- &= \sum(\text{ranks assigned to negative differences}) \\ T &= \min\{\mathrm{T}^+, \mathrm{T}^-\}. \end{aligned} \tag{7.117}$$

Each of these correspond to the different possible alternative hypotheses.

The Critical Region: For the signs test, when n is small we construct our critical region in exactly the same way as we did in §7.5, with the additional constraint that $\theta = 1/2$. When n is large and we use the normal approximation to the binomial theorem, we construct our critical region in exactly the same way as in §7.1.

For the signed-rank test, our critical regions are constructed based on the alternative hypothesis in the following way:

Alternative Hypothesis	Reject Null If :
$\mu \neq \mu_0$	$T \leq T_\alpha$
$\mu > \mu_0$	$T^- \leq T_{2\alpha}$
$\mu < \mu_0$	$T^+ \leq T_{2\alpha}$.

(7.118)

7.9 Tests on Runs

Conditions for Use: The runs test is the quintessential non-parametric test. We need only assume that there are two possible outcomes and the data is ordered.

The Test Statistic: We have two different test statistics for a runs test. Suppose we have our string of outcomes, let n_1 be the total frequency of outcome 1, let n_2 be the total frequency of outcome 2, and u the number of runs in the string; if our string is HHTTTHTHHH then $n_1 = 6$, $n_2 = 2$ and $u = 5$. When n_1 and n_2 are small, our test statistic is simply u. When $n_1, n_2 \geq 10$, our test statistic becomes the normal approximation:

$$z = \frac{u - \mu}{\sigma} \tag{7.119}$$

where

$$\mu = \frac{2n_1n_2}{n_1 + n_2} + 1, \quad \sigma^2 = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}. \tag{7.120}$$

The Critical Region: When n_1 and n_2 are small, we construct the critical region $u \leq u'_{\alpha/2}$ or $u \geq u_{\alpha/2}$, where $u_{\alpha/2}$ and $u'_{\alpha/2}$ are found in a statistics table of critical values for tests based on runs (these are determined by the pair (n_1, n_2) as well as the value α).

When $n_1, n_2 \geq 10$, we use the normal approximation given above, then we construct our critical regions in exactly the same way as in §7.1.

References

[BF] Y. M. M. Bishop and S. E. Fienberg, *Incomplete Two-Dimensional Contingency Tables*, Biometrics **25** (1969), no. 1, 119–128.

[Ia] D. Iacobucci, *On p-values*, Journal of Consumer Research **32** (June 2005), no. 1, 6–12. http://www.journals.uchicago.edu/cgi-bin/resolve?JCR320199PDF

[JB] R. Johnson and G. Bhattacharyya, *Statistics: Principles and Methods*; 3rd edition; John Wiley and Sons, Inc.; 1996.

[MLM] R. Mason, D. Lind, and W. Marchal, *Statistics, An Introduction*; 5th edition; Duxbury Press; Brooks/Cole Publishing Company; 1998.

[MM] I. Miller and M. Miller, *John E. Freund's Mathematical Statistics with Applications*; 7th edition; Pearson Prentice-Hall; Upper Saddle River, NJ; 2004.

[Mil1] S. J. Miller, *A Derivation of the Pythagorean Won-Loss formula in baseball*. http://www.math.brown.edu/∼sjmiller/math/papers/PythagWonLoss_Paper.pdf

[Mil2] S. J. Miller, *The Method of Least Squares*, course notes.

[SD] I. R. Savage and K. W. Deutsch, *A Statistical Model of the Gross Analysis of Transaction Flows* Econometrica **28** (1960), no. 3 551–572.