Elementary Statistics

Brian E. Blank

May 1, 2018

FIRST PRESIDENT UNIVERSITY PRESS

Chapter 1

10. Hypothesis Testing

10.1 Two Approaches to Hypothesis Testing

In this chapter we consider statistical tests that help us decide between two competing hypotheses about unknown parameters of distributions. One hypothesis will be a status quo or "no action required" hypothesis. It is called the ***null hypothesis*** and denoted H_0. The competing hypothesis is called the ***alternative hypothesis*** (or just "the alternative"). It is denoted H_a.[1] As the result of a statistical test, we will either retain H_0, when the evidence for overturning it is not convincing, or reject H_0, when the evidence against it is convincing. Either way, our conviction will be based on probability.

Two types of error are possible when we perform a hypothesis test. We may reject a true null hypothesis: this is called a ***type I error***. Or, we may retain a false null hypothesis: we say that this is a ***type II error***. The probability of making a type I error is denoted by α. That is,

$$\alpha = \mathrm{P}(\text{Reject } \mathrm{H}_0 \,|\, \mathrm{H}_0).$$

The probability of making a type II error is denoted by β. The classical approach to hypothesis testing is to specify an acceptable size α for the type I error, usually 0.01 or 0.05, and find a test that results in the size β of the type II error being as small as possible, given the specified size α of the type I error. In this context α is called the ***significance level***.

A statistic is chosen so that its value makes the determination between the null and alternative hypotheses. It is called the ***test statistic***. The set of values of the test statistic for which the null hypothesis is rejected is called the ***critical region***. In these lectures (and the exams that are based on these lectures), the sample mean will always be the test statistic for a hypothesis test about a population mean. The sample proportion will will always be the test statistic for a hypothesis test about a population proportion. The sample variance will always be the test statistic for a hypothesis test about a population variance.

There is a natural asymmetry between the null hypothesis and the alternative. When a hypothesis test is performed, it is usually because we want to reject the null hypothesis. That is because we consider departing from the status quo when we think there is a better alternative. Why else change? Nevertheless, making a change entails a cost. So we require strong evidence to justify the rejection of a null hypotheis. In analogy to the presumption that a defendant is innocent until proven guilty, a null hypothesis is true until proven false, with the significance level determining what is beyond a reasonable doubt.

The asymmetry extends to the two types of error. A type II error means sticking with the status quo when there is a better alternative. That would be like sticking with the DVD format instead of switching to blu-ray. There are worse things. A type I error means changing to an alternative when the status quo is

[1] Many older textbooks use H_1.

better. That would be like changing from DVD to VHS. Who does that? There is a reason the significance level is linked to the type I error.

The Classical Approach to Hypothesis Testing

A manufacturing process results in 4% of the units produced having defects. A new process is proposed and trialed. In 500 trials, only 16 manufactured units are found to have defects. Does the new process result in a true proportion of defects p that is smaller than the proportion of defects that occur with the existing process?

The observed sample proportion of defects, 16 out of 500, or 0.032 is indeed less than the proportion, 0.04, of defects that result from the existing manufacturing process. But before the company will commit the resources for a full scale upgrade of its plant, it must decide whether the *true* proportion p of defects that occur with the new process is less than p_0 where $p_0 = 0.04$. (Working with a general fixed value p_0 rather than the specific value 0.04 will enable us to more easily generalize the calculations of this example. For that reason we will also use n as the sample size rather than the specific value 500.) The company wishes to decide between the null (status quo) hypothesis

$$\mathrm{H}_0 : p = p_0$$

versus the alternative

$$\mathrm{H}_1 : p < p_0$$

The test statistic will be the sample proportion, $\widehat{p}$. The form of the alternative tells us that the set of values for which we reject the null hypothesis should be of the form $\widehat{p} < p_*$ for some fixed number p_* that will be determined. That is, the critical region will be a set of the form $\widehat{p} < p_*$. Finding the specific value p_*, the endpoint of the critical region, is the crux of the problem.

The first step is common sense. It makes no sense to take $p*$ *greater* than p_0. If we did, then we could have an observed sample proportion $\widehat{p}$ in the range $p_0 < \widehat{p} < p_*$ leading us to conclude that $p < p_0$. It would be crazy for an observed proportion greater than p_0 to lead us to conclude that the true proportion is less than p_0.

Thus, our task is to find a suitable value p_* less than p_0. The question is, How much less? If we observe $\widehat{p} < p_0$ and intend to conclude that the sample proportion $\widehat{p}$ is less than p_0 because the true proportion p is less than p_0, we would like to rule out, so far as is possible, the possibility that $\widehat{p} < p_0$ has been observed, $p = p_0$ not withstanding, because of the variability of sampling. We control for variability by selecting a significance level α. Ultimately we will use 0.05 for the value of α, but we will substitute this specific size only at the conclusion of our calculations.

As the development of the hypothesis test unfolds, the reader is advised to notice that, in contrast to the problem of estimation, *no* use of a standard error is made. Instead, we make use of the standard deviation of $\widehat{p}$ under the assumption that the null hypothesis is true. Thus, whereas the confidence interval for p involves $\sqrt{\widehat{p}_0(1-\widehat{p}_0)/n}$, the hypothesis test involves $\sqrt{p_0(1-p_0)/n}$ instead.

The determination of the critical region begins with the assignment of the size of the type I error:

$$\begin{aligned}
\alpha &= \text{P(Reject true null hypothesis)} \\
&= \text{P}(\widehat{p} < p_* \,|\, \text{H}_0) \\
&= \text{P}(\widehat{p} < p_* \,|\, p = p_0) \\
&= \text{P}\left(\frac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} < \frac{p_* - p_0}{\sqrt{p_0(1-p_0)/n}} \;\middle|\; p = p_0 \right) \\
&= \text{P}\left(\text{Z} < \frac{p_* - p_0}{\sqrt{p_0(1-p_0)/n}} \right) \qquad (p = p_0 \text{ was used to make the normal approximation}) \\
&= \text{P}\left(\text{Z} < -\frac{p_0 - p_*}{\sqrt{p_0(1-p_0)/n}} \right) \qquad (\text{remember that } p_0 - p_* > 0) \\
&= \text{P}\left(\text{Z} > \frac{p_0 - p_*}{\sqrt{p_0(1-p_0)/n}} \right) \qquad (\text{by symmetry}).
\end{aligned}$$

This chain of equalities tells us that

$$\frac{p_0 - p_*}{\sqrt{p_0(1-p_0)/n}} = z_\alpha,$$

or

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

In other words, we reject the null hypothesis if and only if

$$\widehat{p} < p_0 - z_\alpha \sqrt{\frac{p_0(1-p_0)}{n}}.$$

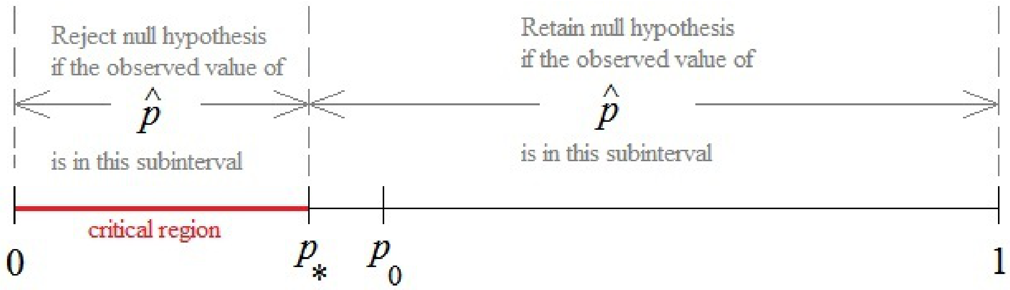


Figure 10.1.1. Critical Region.

All the work has been done. All that is left to do is notice the side of p_* to which $\widehat{p}$ lies. In this example, we substitute $p_0 = 0.04, \alpha = 0.05, z_\alpha = 1.6449, n = 500$ and $\widehat{p} = 0.032$. With these values we have $p_* = 0.04 - 1.6449 \times \sqrt{(0.04)(0.96)/500}$, or $p_* = 0.02558$. The observed value $\widehat{p} = 0.032$ lies outside the critical region $\widehat{p} < 0.02558$ and so we retain the null hypothesis. Notice that even though the sample proportion is less than p_0, we do not accept the alternative that the true proportion is less than p_0: the evidence is not strong enough to rule out sample variance as the cause of the sample proportion being less than p_0. Four procedural observations should also be made:

- The calculation begins with the specification of a significance level α, which was chosen before the experiment.

- The crux of the calculation is a comparative proportion p_* called a critical value.
- The observed value of the test statistic is *not* used in the calculation: it is used only in a last minute comparison.
- There is no waffling: a target is set, the target is reached or not, and a decisive conclusion is made accordingly.

The significance levels $\alpha = 0.05$ and $\alpha = 0.01$ are so standard that established jargon is associated with them. If a null hypothesis is rejected at significance level 0.05, then the result is said to be ***significant***. If a null hypothesis is rejected at significance level 0.01, then the result is said to be ***highly significant***. These are technical terms that convey nothing more than the meanings just ascribed to them. They are not value judgments pronounced in anticipation of the world's citizenry exclaiming, "Bravo! Bravo!"

The Contemporary Approach to Hypothesis Testing

In this subsection, we revisit the decision problem just discussed. For this go-round, we follow a more fashonable path. Who does not want to be in vogue?

Our starting point will be the observed value 0.32 of $\widehat{p}$. We will consider the likelihood of that observation, or one even more extreme, assuming the truth of the null hypothesis. If that probability, which is called a ***p-value*** turns out to be very small, then we will question the validity of our assumption that the null hypothesis is true. Thus, we calculate

$$\begin{aligned}
\mathrm{P}\left(\widehat{p} \leq 0.032 \,|p = p_0\right) &= \mathrm{P}\left(\widehat{p} - p_0 \leq 0.032 - 0.040 \;|\; p = p_0\right) \\
&= \mathrm{P}\left(\frac{\widehat{p} - p_0}{\sqrt{p_0\left(1 - p_0\right)/n}} \leq -\frac{0.008}{\sqrt{p_0\left(1 - p_0\right)/n}} \;\middle|\; p = p_0\right) \\
&= \mathrm{P}\left(Z \leq -\frac{0.008}{\sqrt{(0.04)(0.96)/500}}\right) \\
&= \mathrm{P}\left(Z \leq -0.9129\right) \\
&= \mathrm{P}\left(Z \geq 0.9129\right) \\
&= 1 - \Phi(0.9129) \\
&= 0.1806.
\end{aligned}$$

This p-value is not particularly small. Football fans will understand this. Take any rotten, pathetic football team. Your pick, but students at First President University do not have to look far. In the course of a 16 game season, that team, no matter the level of its ineptitude, will very likely win 3 games. That means that if you choose a game at random to watch this woeful team, there is a 3/16, or 0.1875, chance that you will witness a victory. Now, 0.1875 is scarcely more than the p-value of our new manufacturing process. That p-value is just not that small. There is no strong evidence to reject the null hypothesis. Four procedural observations should be made:

- In a contemporary hypothesis test, no significance level α is chosen.
- The observed value of the test statistic is used from the get-go and is the basis of the entire calculation.
- The crux of the computation is a probability: no comparative proportion value p_* is calculated.
- The test concludes with the report of a p-value.

Some of the terminology of the traditional hypothesis test carries over to the contemporary test. In particular, if a p-value is less than 0.05, then the null hypothesis rejection is said to be *significant*. If a p-value is less than 0.01, then the null hypothesis rejection is said to be *highly significant*. The p-value is sometimes said to be the probability that the null hypothesis is true. Strictly speaking, such a statement is goofy. A null hypothesis is either true or false. It is not the sort of thing to which you attach a probability.

10.2 Testing a Hypothesis about a Proportion

In this section, the value of a population proportion p is presumed to be p_0 (or perhaps to satisfy $p \leq p_0$ or $p \geq p_0$). The presumption about the value of p will be the null hypothesis. The null hypothesis will be challenged by an alternative hypothesis that takes one of the forms $p < p_0$, $p > p_0$, or $p \neq p_0$. Throughout this section, it is assumed that the sample size n is sufficiently large that the normal approximation is accurate.

Hypothesis Test (One-Sided)
$H_0 : p = p_0$ (or $p \geq p_0$)
$H_a : p < p_0$

Test Statistic: $\widehat{p}$ (sample proportion)
Sample Size: n
Significance Level: α
Critical Region: $\widehat{p} < p_0 - z_\alpha \sqrt{\dfrac{p_0(1-p_0)}{n}}$

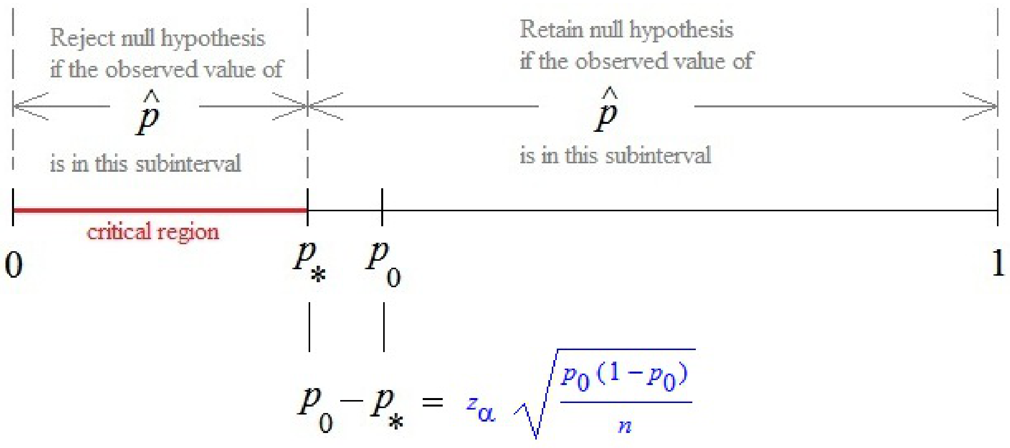


Figure 10.2.1. Critical Region.

Hypothesis Test (One-Sided)
$H_0 : p = p_0$ (or $p \leq p_0$)
$H_a : p > p_0$
Test Statistic: $\widehat{p}$ (sample proportion)
Sample Size: n
Significance Level: α
Critical Region: $\widehat{p} > p_0 + z_\alpha \sqrt{\dfrac{p_0(1-p_0)}{n}}$

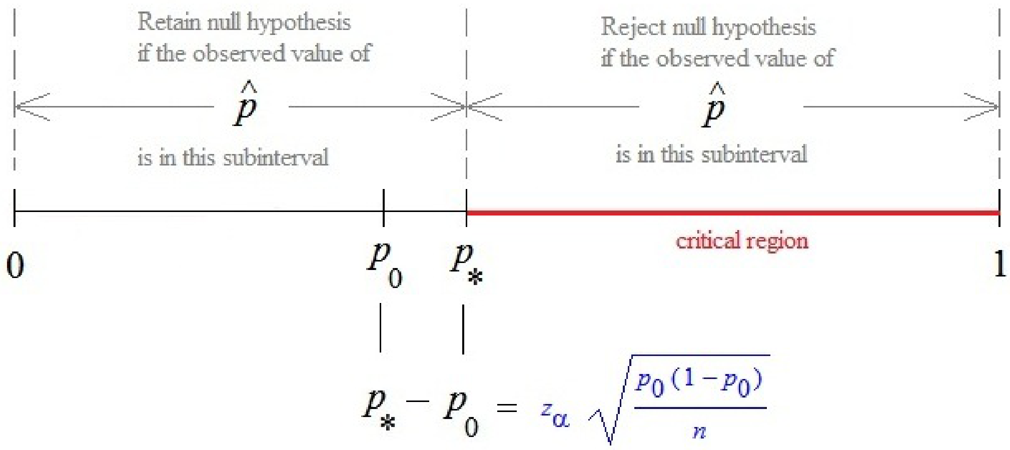


Figure 10.2.2. Critical Region.

Hypothesis Test (Two-Sided)
$H_0 \; : \; p = p_0$
$H_a \; : \; p \neq p_0$

Test Statistic: $\widehat{p}$ (sample proportion)
Sample Size: n
Significance Level: α
Critical Region: $|\widehat{p} - p_0| \geq z_{\alpha/2}\sqrt{\dfrac{p_0\,(1-p_0)}{n}}$
Equivalently $\widehat{p} < p_0 - z_{\alpha/2}\sqrt{\dfrac{p_0\,(1-p_0)}{n}}$ or $\widehat{p} > p_0 + z_{\alpha/2}\sqrt{\dfrac{p_0\,(1-p_0)}{n}}$

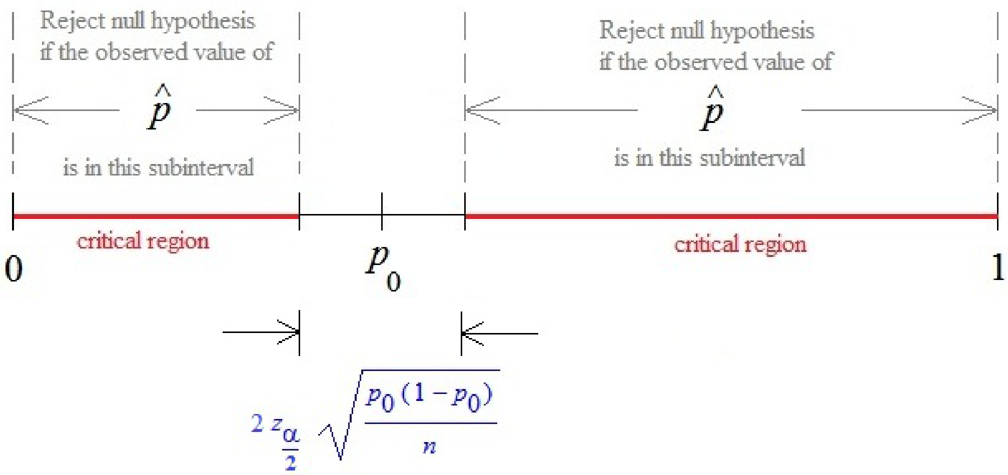


Figure 10.2.3. Critical Region.

The Standardized Test Statistic

Every one of the three classical hypothesis tests for a proportion p involves a critical region defined by $\widehat{p} < p_* = p_0 - z\sqrt{p_0\,q_0/n}$ and/or $\widehat{p} > p_* = p_0 + z\sqrt{p_0\,q_0/n}$, where z is equal to z_α for a one-sded test at significance level α or $z_{\alpha/2}$ for a two-sded test at significance level α. These inequalities are equivalent to $\mathrm{Z} < -z$ and/or $\mathrm{Z} > z$ where

$$\mathrm{Z} = \frac{\widehat{p} - p_0}{\sqrt{p_0\,q_0/n}}. \tag{10.2.1}$$

Many recent statistics textbooks choose Z, the standardization of $\widehat{p}$ under the null hypothesis, for the test statistic. Such a choice does have the advantage of simpler critical values: they are just plain z-scores. On the other hand, the hypothesis is about comparing a true but unknown proportion p with a specific value p_0: therefore, the statistic $\widehat{p}$ and critical value p_* are more closely related to the comparison than are z-scores. *Standardized test statistics will not be used in these notes.*

Classical and Contemporary Compared

No doubt the best way to appreciate the differences between the classical and contemporary approaches to hypothesis testing is to see the procedures performed back-to-back on the same problem. Recall that there are four hallmarks for both methods. The classical test begins with a significance level α and does not use the observed sample mean $\widehat{p}$ until the last step; the contemporary begins with the observed sample mean $\widehat{p}$ and dispenses with the significance level altogether. The key calculation of a classical test is a comparative proportion p_*, and no probability is computed; the key calculation of a contemporary test is a probability, and no comparative proportion p_* is computed. A classical test concludes decisively and objectively with a *Reject* or *Retain* decision, no ifs, ands, or buts; a contemporary test concludes with contemplation of a probability. Pay attention to these distinctions when you read the following two examples.

Example 1. *Based on past surveys, a company believes that 36% of the population recognizes its logo. However, in its most recent consumer survey, the company learned that only 127 of 400 surveyees recognized the logo. Test the hypothesis* $H_0 : p = 0.36$ *against the alternative* $H_a : p \neq 0.36$ *at significance level 0.05.*

Solution. *We calculate*

$$p_0 \pm z_{\alpha/2}\sqrt{\frac{p_0\,(1-p_0)}{n}} = 0.36 \pm 1.96\sqrt{\frac{(0.36)(0.64)}{400}} = 0.36 \pm 0.04704.$$

We reject the null hypothesis if either $\widehat{p} < 0.36 - 0.04704 = 0.31296$ *or* $\widehat{p} > 0.36 + 0.04704 = 0.40704$. *Now the observed value of* $\widehat{p}$ *is 127/400, or 0.3175, and this value does not satisfy either of the inequalities necessary for the rejection of the null hypothesis. We retain the null hypothesis.*

Example 2. *Reconsider the last example, but this time report a p-value.*

Solution. *The observed value of* $\widehat{p}$, *namely 0.3175, is a distance* $|0.3175 - 0.36|$, *or 0.0425 from the null* *hypothesis value 0.36 of p. The p-value is*

$$\begin{aligned}
P\left(|\widehat{p}-0.36| \geq 0.0425 \,|\, p=0.36\right) &= 2\,P\left(\widehat{p}-0.36 \geq 0.0425 \,|\, p=0.36\right) \\
&= 2\,P\left(\frac{\widehat{p}-0.36}{\sqrt{(0.36)(0.64)/400}} \geq \frac{0.0425}{\sqrt{(0.36)(0.64)/400}} \,\middle|\, p=0.36\right) \\
&= 2\,P\left(Z \geq \frac{0.0425}{\sqrt{(0.36)(0.64)/400}}\right) \\
&= 2\,P\left(Z \geq \frac{0.0425}{\sqrt{(0.36)(0.64)/400}}\right) \\
&= 2\,P(Z \geq 1.7708) \\
&= 2\,(1-\Phi(1.7708)) \\
&= 0.07659.
\end{aligned}$$

This p-value is not particularly small. We are inclined to retain the null hypothesis.

When should a test be two-sided? There is often no reason to suspect p will be on a particular side of a value p_0. Remember, the test should be part of the experimental design, not something put together after the experiment has been performed. Even when it is anticipated that p will lie on a particular side of a value p_0, the choice of a two-sided test in the experimental design acknowledges the possibility that the experiment may not go the way it was predicted.

Example 3. *Autopsies of people who die of Alzheimer's disease (AD) invariably reveal amyloid-β peptide plaques in the vicinity of the hippocampus. The role of these plaques is not well-understood at present. However, it is a working hypothesis that reducing these plaques, or at least slowing down their spread, can only be a good thing. As a first step in developing an immunotherapy to combat amyloid-β plaques, transgenic mice were infected with human amyloid-β. The mice were then inoculated with a synthetic substance called AN1792. As a result of this immunotherapy, the plaque burden of the mice was reduced and, as hoped, the mice suffered no further cognitive decline. In the next step, a human trial of mild to moderate AD sufferers, immunization was halted after 18 of the 300 patients who received AN1792 inoculations developed meningoencephalitis, a rare and often fatal condition.*[2] *Permanent brain injury resulted in some of the patients. Since then, some good has come out of the experiment. A follow-up that was conducted 4.6 years later, revealed that several of the patients, despite their incomplete immunotherapy, retained detectable levels of antibodies.*[3] *Twenty-five of these antibody responders participated in the follow-up, as did thirty subjects who, being in the control group, had received a placebo instead of AN1792. Of the 30 control group subjects, 14 were no longer able to stay at home because of the progression of AD. Let p be the population proportion of untreated mild to moderate AD sufferers who will need outside the home care within 4.6 years. Test the null hypothesis* $p = 0.24$ *against the alternative* $p \neq 0.24$. *What is the p-value?*

Solution. *The source of the hypothesized value* $p_0 = 0.24$ *will be revealed at the conclusion of the solution. For now, we have no reason to suspect* $p < 0.24$ *or* $p > 0.24$. *Accordingly, a two-sided test is appropriate. The sample proportion* $\widehat{p}$ *has observed value 14/30, or 0.4667, which is strikingly larger than* $p_0 = 0.24$. *We will start with a traditional test using 0.01 as the significance level. The critical values are* $p_0 \pm z_{0.005}\sqrt{p_0\,(1-p_0)\,/n}$, *or* $0.24 \pm 2.575829\sqrt{(0.24)\;(0.76)\;/30}$, *or 0.0391514 (with the minus) and 0.4408486 (with the plus). We reject the null hypothesis if either* $\widehat{(p)} < 0.0391514$ *or* $\widehat{(p)} > 0.4408486$. *The sample proportion, namely 0.4667, does*

[2] *Clinical effects of Abeta immunization (AN1792) in patients with AD in an interrupted trial*, S. Gilman et al., Neurology. 2005 May 10, vol 64 (9), 1553-62.

[3] *Long-Term Follow-Up of Patients Immunized with AN1792: Reduced Functional Decline in Antibody Responders*, Bruno Vellas et al., Current Alzheimer Research, 2009, 6, 144-151.

in fact exceed the larger of the two critical values, so the null hypothesis is rejected, and we may exclaim, "Highly significant!"

Now we will determine the requested p-value. Using symmetry at a key point, we calculate

$$\begin{aligned} p\text{-value} &= P(|\widehat{p} - 0.24| \geq |0.4667 - 0.24| \mid p = 0.24) \\ &= P\left(\frac{|\widehat{p} - p|}{\sqrt{p(1-p)/30}} \geq \frac{|0.4667 - 0.24|}{\sqrt{(0.24)(0.76)/30}} \,\middle|\, p = 0.24 \right) \\ &= P(|Z| \geq 2.907366) \\ &= 2\,P(Z \geq 2.907366) \\ &= 2\,(1 - \phi(2.907366)) \\ &= 0.003644865. \end{aligned}$$

Notice that the p-value is less than 0.01. Again we may proclaim, Highly significant!

It is time to answer the question, Where did the hypothesized value $p_0 = 0.24$ *come from? The answer tp this question is the key to appreciating the study that is the basis of this example. Of the 25 antibody responders who participated in the follow-up, only 6 required out-of-home care after 4.6 years. Observe that 6/25 comes to 0.24. Thus, the immunotherapy resulted in a highly significant reduction in the need for out-of-home care. Several other positive measures of the immunotherapy are discussed in the cited report. The take-away is that treatment of AD by immunotherapy is promising even if AN1792 is not the appropriate agent.*[4]

10.3 Testing a Hypothesis about a Population Mean (Standard Deviation Known)

In this section, we use a sample $X_1, X_2, \ldots, X_n$ of any size n drawn from a normal distribution, or a sufficiently large sample ($n \geq 30$) for the normal approximation to apply, with the goal of using the sample mean $\widehat{\mu} = \overline{X}$ to test a hypothesis about the value of the population mean μ. The null hypothesis is $\mu = \mu_0$ and the alternative hypothesis is either $\mu < \mu_0$ (in which case the null may be taken as $\mu \geq \mu_0$ with no change), or the alternative is $\mu > \mu_0$ (in which case the null may be taken as $\mu \leq \mu_0$ with no change), or the alternative is $\mu \neq \mu_0$. Notice that in all cases, it is the null hypothesis that includes equality. The alternative hypothesis involves an inequality in the case of a two-sided test or a strict inequality in the case of a one-sided test. The tests are so similar to the tests for proportions that deriving them would be repetitious. We set out the traditional tests first. As with the tests for a proportion, traditional tests of a population mean begin with the specification of a significance level α.

Hypothesis Test
  $H_0 : \mu = \mu_0$ (or $\mu \geq \mu_0$)
  $H_a : \mu < \mu_0$

Test Statistic: $\widehat{\mu}$ (sample mean)
Sample Size: n
Significance Level: α
Critical Region: $\widehat{\mu} < \mu_0 - z_\alpha \dfrac{\sigma}{\sqrt{n}}$

Hypothesis Test
  $H_0 : \mu = \mu_0$ (or $\mu \leq \mu_0$)
  $H_a : \mu > \mu_0$

[4]See *Can Alzheimer disease be prevented by amyloid-β immunotherapy?*, Cynthia A. Lemere and Eliezer Masliah, Nat Rev Neurol., 2010 February; **6**(2), 108119. `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2864089/pdf/nihms-194354.pdf`

Test Statistic: $\widehat{\mu}$ (sample mean)
Sample Size: n
Significance Level: α
Critical Region: $\widehat{\mu} > \mu_0 + z_\alpha \dfrac{\sigma}{\sqrt{n}}$

Hypothesis Test
$H_0 : \mu = \mu_0$
$H_a : \mu \neq \mu_0$

Test Statistic: $\widehat{\mu}$ (sample mean)
Sample Size: n
Significance Level: α
Critical Region: $|\widehat{\mu} - \mu_0| \geq z_{\alpha/2} \dfrac{\sigma}{\sqrt{n}}$ Equivalently $\widehat{\mu} < \mu_0 - z_{\alpha/2} \dfrac{\sigma}{\sqrt{n}}$ or $\widehat{\mu} > \mu_0 + z_{\alpha/2} \dfrac{\sigma}{\sqrt{n}}$

If the sample size n is large and if the population standard deviation σ is unknown, then we may replace the unknown parameter σ with the sample standard deviation S when we perform the hypothesis tests in this section. This replacement can be made when n is large whether or not we sample from a distribution that is normal. However, if the sampling is from a normal distribution, then we can avoid making the approximation $\sigma \approx S$ by using t-tests described in the next section.

Example 1. *The average ocean depth μ in a certain area has been listed as 71.6 fathoms. A random sample of 36 readings is taken with more modern equipment. The sample mean is 72.4 and the population standard deviation is 2.05.*
a) Test the null hypothesis $\mu = 71.6$ against the alternative $\mu > 71.6$ at 5 % significance level.
b) Test the null hypothesis $\mu = 71.6$ against the alternative $\mu \neq 71.6$ at 5 % significance level.

Solution. *a) Because $n = 36 > 30$, we may use the normal approximation and we may use the approximation $\sigma \approx S$. For the one-sided test, we have $H_0 : \mu = 71.6$ and $H_a : \mu > 71.6$. At 5% significance level for a one-sided test, we use $z_{0.05} = 1.6449$. The critical region is $\widehat{\mu} > 71.6 + 1.6449 \frac{2.05}{\sqrt{36}}$, or $\widehat{\mu} > 72.162$. The sample mean, namely 72.4, satisfies this inequality, so we reject the null hypothesis.*
b) For a two-sided test, we have $H_0 : \mu = 71.6$ and $H_a : \mu \neq 71.6$. At 5% significance level for a two-sided test, we use $z_{0.025} = 1.96$. The critical region is $\widehat{\mu} < 71.6 - 1.96 \frac{2.05}{\sqrt{36}}$ or $\widehat{\mu} > 71.6 + 1.96 \frac{2.05}{\sqrt{36}}$. The first inequality is $\widehat{\mu} < 70.93$. The second inequality is $\widehat{\mu} > 72.27$. The sample mean, namely 72.4, satisfies this second inequality, so we reject the null hypothesis (although this rejection was a closer call).

Example 2. *Repeat parts (a) and (b) of the preceding example, but use a contemporary test and state the p-values.*

Solution. *a) We calculate*

$$\begin{aligned}
\textit{p-value} &= P(\widehat{\mu} \geq 72.4 | \mu = 71.6) \\
&= P\left(\frac{\widehat{\mu} - 71.6}{2.05/\sqrt{36}} \geq \frac{72.4 - 71.6}{2.05/\sqrt{36}} \middle| \mu = 71.6 \right) \\
&= P(Z \geq 2.341463) \\
&= 1 - \Phi(2.341463) \\
&= 1 - 0.9903958 \\
&= 0.0096.
\end{aligned}$$

b) We calculate

$$\begin{aligned}
p\text{-value} &= P\left(|\widehat{\mu} - 71.6| \geq 72.4 - 71.6 \,|\, \mu = 71.6\right) \\
&= P\left(\frac{|\widehat{\mu} - 71.6|}{2.05/\sqrt{36}} \geq \frac{72.4 - 71.6}{2.05/\sqrt{36}} \,\middle|\, \mu = 71.6 \right) \\
&= P(|Z| \geq 2.341463) \\
&= 2\,P(Z \geq 2.341463) \\
&= 2\left(1 - \Phi(2.341463)\right) \\
&= 2\,(1 - 0.9903958) \\
&= 0.0192.
\end{aligned}$$

10.4 Testing a Hypothesis about a Population Mean (Standard Deviation Unknown)

In this section, we test hypotheses concerning the value of the population mean μ of a normal distribution (or a distribution that is well-approximated by a normal distribution). Because the underlying distribution is normal, or very nearly normal, we do not need to rely on the normal approximation that results when a sample size is large. Therefore, the tests in this section do not depend on whether n is large or small. However, we repeat that the tests of this section can be applied only to samples from distributions that are normal or very nearly normal.

In this section, the hypotheses about μ, both the null and the alternatives, are exactly the same as those of the preceding section. There are two differences:

- In the preceding section it was assumed that either σ is known or that n is sufficiently large to make the approximation $\sigma \approx S$. In this section we do not assume that σ is known and we do not assume that n is large. It may be large, but it does not have to be.
- In the preceding section, if the underlying distribution is not normal, then we can proceed nevertheless provided that the sample size is large. In this section, the underlying distribution must be normal.

Thus, in this section, we use a sample $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n$ of any size n drawn from a normal distribution with the goal of using the sample mean $\widehat{\mu} = \overline{\mathrm{X}}$ to test a hypothesis about the value of the population mean μ. The null hypothesis is $\mu = \mu_0$ and the alternative hypothesis is either $\mu < \mu_0$ (in which case the null may be taken as $\mu \geq \mu_0$ with no change), or the alternative is $\mu > \mu_0$ (in which case the null may be taken as $\mu \leq \mu_0$ with no change), or the alternative is $\mu \neq \mu_0$. Notice that in all cases, it is the null hypothesis that includes equality. The alternative hypothesis involves an inequality in the case of a two-sided test or a strict inequality in the case of a one-sided test. The tests are so similar to the tests for proportions that deriving them would be repetitious. We set out the traditional tests first. As with the tests for a proportion, traditional tests of a population mean begin with the specification of a significance level α.

The key idea is that when we draw a sample of size n from a normal distribution with mean μ_0, if S is the sample standard deviation, then

$$\frac{\widehat{\mu} - \mu_0}{S/\sqrt{n}} \quad \sim \quad \mathrm{t}_{n-1}\,. \tag{10.4.1}$$

Two features of line (10.4.1) should be noted: (1) The parameter σ is not involved in either the sampling statistic on the left or the distribution on the right, and (2) The distribution on the right is the exact distribution of the sampling statistic on the left: there is no distribution. We deduce the following hypothesis tests from those of Section 10.3 by replacing σ with the sample standard deviation S, z_α with $\mathrm{t}_{\alpha,n-1}$, and $z_{\alpha/2}$ by $\mathrm{t}_{\alpha/2,n-1}$.

Hypothesis Test
$H_0 : \mu = \mu_0$ (or $\mu \geq \mu_0$)
$H_a : \mu < \mu_0$
Test Statistic: $\widehat{\mu}$ (sample mean)
Sample Size: n
Significance Level: α
Critical Region: $\widehat{\mu} < \mu_0 - t_{\alpha,n-1} \dfrac{S}{\sqrt{n}}$

Hypothesis Test
$H_0 : \mu = \mu_0$ (or $\mu \leq \mu_0$)
$H_a : \mu > \mu_0$

Test Statistic: $\widehat{\mu}$ (sample mean)
Sample Size: n
Significance Level: α
Critical Region: $\widehat{\mu} > \mu_0 + t_{\alpha,n-1} \dfrac{S}{\sqrt{n}}$

Hypothesis Test
$H_0 : \mu = \mu_0$
$H_a : \mu \neq \mu_0$

Test Statistic: $\widehat{\mu}$ (sample mean)
Sample Size: n
Significance Level: α
Critical Region: $|\widehat{\mu} - \mu_0| \geq t_{\alpha/2,n-1} \dfrac{S}{\sqrt{n}}$ Equivalently $\widehat{\mu} < \mu_0 - t_{\alpha/2,n-1} \dfrac{S}{\sqrt{n}}$ or $\widehat{\mu} > \mu_0 + t_{\alpha/2,n-1} \dfrac{S}{\sqrt{n}}$

Some remarks may be helpful:

- These tests do not depend on the normal approximation resulting from the Central Limit Theorem: n need not be large.
- If n is large, then $\sigma \approx S$, $t_{\alpha,n-1} \approx z_\alpha$, and $t_{\alpha/2,n-1} \approx z_{\alpha/2}$.
- If n is large and if the underlying distribution is normal, then the tests of this section and the tests of the preceding section both apply. The three approximations mentioned in the preceding remark tell us that the two tests will produce similar results. If it is certain that the underlying distribution is normal and not very nearly normal, then the tests of this section have the advantage of not requiring any approximation.

Example 1. *In a test of 900 antique incandescant lightbulbs of the sort that cannot now be legally sold, the mean lifetime was 794 hours and the sample standard deviation was 70 hours. The null hypothesis printed on the packaging asserts a mean lifetime of $\mu = 800$ hours. Use a one-sided test of this hypothesis at 5% significance level. Assume that the lifetimes are normally distributed.*

Solution. *We are testing $H_0 : \mu = 800$ against the alternative $H_a : \mu < 800$. We use $n = 900$, $S = 70$, and $t_{0.05,899} \approx z_{0.05} = 1.6449$. The critical region is*

$$\widehat{\mu} < 800 - 1.6449 \cdot \frac{70}{\sqrt{900}} = 796.16.$$

The observed value of $\widehat{\mu}$, namely 794, satisfies this inequality, so we reject the null hypothesis.

Example 2. *Report a p-value for the hypothesis test of the preceding example.*

Solution. *The requested p-value is*

$$\begin{aligned} P(\widehat{\mu} \leq 794 \,|\, \mu_0 = 800) &= P(\widehat{\mu} - \mu_0 \leq 794 - 800 \,|\, \mu_0 = 800) \\ &= P\left(\frac{\widehat{\mu} - \mu_0}{S/\sqrt{n}} \leq \frac{794 - 800}{70/\sqrt{900}} \,\middle|\, \mu_0 = 800 \right) \\ &= P(t_{899} \leq -2.5714) \\ &\approx P(Z \leq -2.5714) \\ &= P(Z \geq 2.5714) \\ &= 1 - P(Z \leq 2.5714) \\ &= 1 - \Phi(2.5714) \\ &= 0.00506. \end{aligned}$$

This p-value is very small. It would lead us to reject the null hypothesis.

Example 3. *A treatment of 29 anorexic girls resulted in a 3 pound mean weight gain with sample standard deviation 7.32. Find a two-sided p-value for a test of the null hypothesis* $\mu = 0$ *where* μ *is the true mean weight change resulting from the treatment. Assume that the change of weight that results from the treatment is normally distributed.*

Solution. *What is of interest is the possibility that* $\mu > 0$*. However, a two-sided test includes the possibility that the test might actually have unintended negative results. The standard deviation shows that some of the girls did lose weight under the treatment.*

The null hypothesis is $\mu = \mu_0$ *where* $\mu_0 = 0$*. The observed value of* $\widehat{\mu}$ *is 3, which is 3 away from* μ_0*. The p-value is*

$$\begin{aligned} P(|\widehat{\mu} - \mu_0| \geq 3 \,|\, \mu = \mu_0) &= P\left(\left| \frac{\widehat{\mu} - \mu_0}{S/\sqrt{n}} \right| \geq \frac{3}{7.32/\sqrt{29}} \,\middle|\, \mu = \mu_0 \right) \\ &= P(\,|t_{28}| \geq 2.207) \\ &= 2\,P(t_{28} \geq 2.207)\,. \end{aligned}$$

From the tables, we find $P(t_{28} > 2.0484) = 0.025$ *and* $P(t_{28} > 2.4671) = 0.010$*. The equation of the straight line through the two points (2.0484,0.025) and (2.4671,0.010) has slope* $(0.010 - 0.025)/(2.4671 - 2.0484)$*, or -0.0358. The equation is therefore* $p = -0.0358\,(t - 2.0484) + 0.025$*. Setting* $t = 2.207$*, we obtain* $p = 0.0193$*. Thus,* $P(t_{28} \geq 2.207) \approx 0.0193$ *and the requested p-value is approximately* 2×0.0193*, or 0.038.*[5] *This small value suggests that the null hypothesis should be rejected.*

10.5 Two-Sample Tests

It is often necessary to compare the means of two populations. The effect of a new drug might be compared to the effect of an exisitng drug, for example. Suppose that $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ are *independent* random samples from distributions with means and standard deviations given by μ_X, σ_X and μ_Y, σ_Y respectively. As the notation suggests, the sizes need not be the same. We wish to test the null hypothesis

[5] Software provides a more accurate value for $P(t_{28} \geq 2.207)$: 0.0178403938.

that $\mu_X = \mu_Y$. As usual, there are three possible alternative hypotheses leading to two one-sided tests and one two-sided test: $\mathrm{H}_a : \mu_x < \mu_Y$ or $\mathrm{H}_a : \mu_x > \mu_Y$ or $\mathrm{H}_a : \mu_x \neq \mu_Y$.

Let $\delta = \mathrm{X}-\mathrm{Y}$ be the difference of the random variables. Then $\mu_\delta = \mu_\mathrm{X}-\mu_\mathrm{Y}$ and $\widehat{\delta} = \widehat{\mu_\mathrm{X}}-\widehat{\mu_\mathrm{Y}} = \overline{\mathrm{X}}-\overline{\mathrm{Y}}$ is the test statistic. If the alternative hypothesis is that $\mu_X < \mu_Y$, then, in a classical test at significance level α, we seek a critical value $d_* < 0$ for which the critical region is $\overline{\mathrm{X}} - \overline{\mathrm{Y}} < d_*$, and $\mathrm{P}\left(\widehat{\mu_X} - \widehat{\mu_Y} < d_* \,|\, \mu_\mathrm{X} = \mu_\mathrm{Y}\right) = \alpha$. It follows that

$$\mathrm{P}\left(\frac{\widehat{\mu_\mathrm{X}} - \widehat{\mu_\mathrm{Y}}}{\mathrm{Sd}\left(\widehat{\mu_\mathrm{X}} - \widehat{\mu_\mathrm{Y}}\right)} < \frac{d_*}{\mathrm{Sd}\left(\widehat{\mu_\mathrm{X}} - \widehat{\mu_\mathrm{Y}}\right)} \,\middle|\, \mu_\mathrm{X} = \mu_\mathrm{Y} \right) = \alpha. \tag{10.5.1}$$

Determining d_* requires an identification of the sample statistic

$$\frac{\widehat{\mu_\mathrm{X}} - \widehat{\mu_\mathrm{Y}}}{\mathrm{Sd}\left(\widehat{\mu_\mathrm{X}} - \widehat{\mu_\mathrm{Y}}\right)}. \tag{10.5.2}$$

If σ_X and σ_Y are known, then we know

$$\mathrm{Sd}\left(\widehat{\mu_\mathrm{X}} - \widehat{\mu_\mathrm{Y}}\right) = \sqrt{\mathrm{Var}\left(\widehat{\mu_\mathrm{X}} - \widehat{\mu_\mathrm{Y}}\right)} = \sqrt{\mathrm{Var}\left(\widehat{\mu_\mathrm{X}}\right) + \mathrm{Var}\left(\widehat{\mu_\mathrm{Y}}\right)} = \sqrt{\frac{\sigma_\mathrm{X}^2}{n} + \frac{\sigma_\mathrm{Y}^2}{m}}. \tag{10.5.3}$$

Expression (10.5.2) becomes

$$\frac{\widehat{\mu_\mathrm{X}} - \widehat{\mu_\mathrm{Y}}}{\sqrt{\frac{\sigma_\mathrm{X}^2}{n} + \frac{\sigma_\mathrm{Y}^2}{m}}}. \tag{10.5.4}$$

Testing Two-Population Means Using a z-Score Test

In this subsection, we assume either that X and Y are normally distributed or that the sample sizes drawn from the two distributions are large enough that we may use the normal approximation. We also assume that the population standard deviations σ_X and σ_Y are known.

Recall that to test the null hypothesis $\mu_\mathrm{X} = \mu_\mathrm{Y}$ against the alternative $\mu_\mathrm{X} < \mu_\mathrm{Y}$ at significance level α, we seek a critical value $d_* < 0$ such that the null hypothesis is rejected if $\overline{X} - \overline{Y} < d_*$. If n and m are sufficiently large to use the normal approximation, then, using formulas (10.5.1) and (10.5.4), we obtain

$$\begin{aligned}
\alpha &= \mathrm{P}\left(\frac{\widehat{\mu_\mathrm{X}} - \widehat{\mu_\mathrm{Y}}}{\mathrm{Sd}\left(\widehat{\mu_\mathrm{X}} - \widehat{\mu_\mathrm{Y}}\right)} < \frac{d_*}{\mathrm{Sd}\left(\widehat{\mu_\mathrm{X}} - \widehat{\mu_\mathrm{Y}}\right)} \,\middle|\, \mu_\mathrm{X} = \mu_\mathrm{Y} \right) \\
&= \mathrm{P}\left(\frac{\widehat{\mu_\mathrm{X}} - \widehat{\mu_\mathrm{Y}}}{\sqrt{\frac{\sigma_\mathrm{X}^2}{n} + \frac{\sigma_\mathrm{Y}^2}{m}}} < \frac{d_*}{\sqrt{\frac{\sigma_\mathrm{X}^2}{n} + \frac{\sigma_\mathrm{Y}^2}{m}}} \,\middle|\, \mu_\mathrm{X} = \mu_\mathrm{Y} \right) \\
&= \mathrm{P}\left(\mathrm{Z} < \frac{d_*}{\sqrt{\frac{\sigma_\mathrm{X}^2}{n} + \frac{\sigma_\mathrm{Y}^2}{m}}} \right) \\
&= \mathrm{P}\left(\mathrm{Z} > \frac{-d_*}{\sqrt{\frac{\sigma_\mathrm{X}^2}{n} + \frac{\sigma_\mathrm{Y}^2}{m}}} \right),
\end{aligned}$$

or

$$\frac{-d_*}{\sqrt{\frac{\sigma_\mathrm{X}^2}{n} + \frac{\sigma_\mathrm{Y}^2}{m}}} = z_\alpha,$$

or

$$d_* = -z_\alpha \sqrt{\frac{\sigma_\mathrm{X}^2}{n} + \frac{\sigma_\mathrm{Y}^2}{m}}.$$

The critical region is

$$\overline{X} - \overline{Y} < -z_\alpha \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}.$$

The other two tests can be derived in the same way. We state all three formally.

Hypothesis Test (n and m both large enough for the normal approximation to be used)
$H_0 : \mu_X = \mu_Y$ (or $\mu_X \geq \mu_Y$)
$H_a : \mu_X < \mu_Y$
Test Statistic: $\overline{X} - \overline{Y}$ (difference of sample means)
Sample Size: n for X and m for Y
Significance Level: α
Critical Region: $\overline{X} - \overline{Y} < -z_\alpha \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}$

Hypothesis Test (n and m both large enough for the normal approximation to be used)
$H_0 : \mu_X = \mu_Y$ (or $\mu_X \leq \mu_Y$)
$H_a : \mu_X > \mu_Y$
Test Statistic: $\overline{X} - \overline{Y}$ (difference of sample means)
Sample Size: n for X and m for Y
Significance Level: α
Critical Region: $\overline{X} - \overline{Y} > z_\alpha \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}$

Hypothesis Test (n and m both large enough for the normal approximation to be used)
$H_0 : \mu_X = \mu_Y$
$H_a : \mu_X \neq \mu_Y$
Test Statistic: $\delta = \overline{X} - \overline{Y}$ (difference of sample means)
Sample Size: n for X and m for Y
Significance Level: α
Critical Region: $|\delta| \geq z_{\alpha/2} \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}$ Equivalently $\delta < -z_{\alpha/2} \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}$ or $\delta > z_{\alpha/2} \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}$

If the sample sizes n and m are both large and if the population standard deviations σ_X and σ_Y are unknown, then we may replace the unknown parameters σ_X and σ_Y with the sample standard deviations S_X and S_Y when we perform the hypothesis tests in this subsection. These replacements can be made when n and m are large whether or not we sample from distributions that are normal. However, if X and Y are normally distributed, then we can avoid making the approximations $\sigma_X \approx S_X$ and $\sigma_Y \approx S_Y$ by using t-tests described in the next subsection.

Example 1. *In a comparison of the tar content (in mg) of two cigarettes produced by different processes, the following data was observed:*

	Sample Size	*Sample Mean*	*Sample Variance*
X	*60*	*15.2*	*2.68*
Y	*44*	*15.5*	*3.07*

Assume that the sample variances are the true population variances and test the null hypothesis that $\mu_x = \mu_y$ against the alternative that $\mu_x \neq \mu_y$ at significance level 0.05.

Solution. *We calculate*

$$Sd(\widehat{\mu_X} - \widehat{\mu_Y}) = \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} = \sqrt{\frac{2.68}{60} + \frac{3.07}{44}} = 0.3383$$

and

$$z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}} = (1.96)(0.3383) = 0.6630.$$

Because $\left|\overline{X}-\overline{Y}\right|$*, or* $|15.2-15.5|$*, or* 0.3 *is not greater than* 0.6630*, we do not reject the null hypothesis: there is insufficient evidence to conclude that* $\mu_X \neq \mu_Y$*.*

Remark: *A contemporary p-value test can be easily performed instead. The p-value is*

$$\begin{aligned}
P\left(|\widehat{\mu_X}-\widehat{\mu_Y}| \geq |15.2-15.5| \;\mid \mu_X=\mu_Y\right) &= P\left(|\widehat{\mu_X}-\widehat{\mu_Y}| \geq 0.3 \,|\, \mu_X=\mu_Y\right) \\
&= P\left(\left|\frac{\widehat{\mu_X}-\widehat{\mu_Y}}{Sd\left(\widehat{\mu_X}-\widehat{\mu_Y}\right)}\right| \geq \frac{0.3}{0.3383} \,|\, \mu_X=\mu_Y\right) \\
&= P(|Z| \geq 0.8868) \\
&= 2\,P(Z \geq 0.8868) \\
&= 2\,(1-\Phi(0.8868)) \\
&= 2\,(1-0.8124) \\
&= 0.3752,
\end{aligned}$$

which is considerably greater than conventional cut-offs for rejecting the null hypothesis.

Testing Two-Population Means Using the Student-t Distribution

The hypothesis tests of the preceding subsection have a serious limitation: If either sample size n or m is small, then the population standard deviations σ_{X} and σ_{Y} must be known. If at least one of the population standard deviations σ_{X} and σ_{Y} is unknown, and if at least one of the sample sizes is not large, then we must resort to a different test. In the replacement test, we use the sample variances $\mathrm{S}_{\mathrm{X}}^2$ and $\mathrm{S}_{\mathrm{Y}}^2$ instead of σ_{X}^2 and σ_{Y}^2. For this to work when at least one of the sample sizes is small, the distributions of X and Y must both be normal. We continue to use the difference δ of the sample means $\widehat{\mu_{\mathrm{X}}}-\widehat{\mu_{\mathrm{Y}}}$ as the test statistic. However, we replace the population variances in formula (10.5.4) with the sample variances:

$$\frac{\widehat{\mu_{\mathrm{X}}}-\widehat{\mu_{\mathrm{Y}}}}{\sqrt{\frac{\mathrm{S}_{\mathrm{X}}^2}{n}+\frac{\mathrm{S}_{\mathrm{Y}}^2}{m}}}. \tag{10.5.5}$$

This test statistic is approximately a Student-t distribution where we conservatively estimate the number of degress of freedom by $\mathrm{df}=\min(m,n)-1$.

Hypothesis Test (X and Y are normal, or at least approximately normal)
$\mathrm{H}_0 : \mu_{\mathrm{X}}=\mu_{\mathrm{Y}}$ (or $\mu_{\mathrm{X}} \geq \mu_{\mathrm{Y}}$)
$\mathrm{H}_{\mathrm{a}} : \mu_{\mathrm{X}}<\mu_{\mathrm{Y}}$
Test Statistic: $\overline{\mathrm{X}}-\overline{\mathrm{Y}}$ (difference of sample means)
Degrees of Freedom: $\mathrm{df}=\min(m,n)-1$
Sample Size: n for X and m for Y
Significance Level: α
Critical Region: $\overline{\mathrm{X}}-\overline{\mathrm{Y}}<-\,\mathrm{t}_{\alpha,\mathrm{df}}\sqrt{\frac{\mathrm{S}_{\mathrm{X}}^2}{n}+\frac{\mathrm{S}_{\mathrm{Y}}^2}{m}}$

Hypothesis Test (X and Y are normal, or at least approximately normal)

$\mathrm{H_0} : \mu_X = \mu_Y \qquad (\text{or } \mu_X \leq \mu_Y)$

$\mathrm{H_a} : \mu_X > \mu_Y$

Test Statistic: $\overline{X} - \overline{Y}$ (difference of sample means)

Degrees of Freedom: $\mathrm{df} = \min(m, n) - 1$

Sample Size: n for X and m for Y

Significance Level: α

Critical Region: $\overline{X} - \overline{Y} > t_{\alpha,\mathrm{df}} \sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}$

Hypothesis Test (X and Y are normal, or at least approximately normal)

$\mathrm{H_0} : \mu_X = \mu_Y$

$\mathrm{H_a} : \mu_X \neq \mu_Y$

Test Statistic $\delta = \overline{X} - \overline{Y}$ (difference of sample means)

Degrees of Freedom: $\mathrm{df} = \min(m, n) - 1$

Sample Size: n for X and m for Y

Significance Level: α

Critical Region: $|\delta| \geq t_{\alpha/2,\mathrm{df}} \sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}$

Equivalently, $\delta < -t_{\alpha/2,\mathrm{df}} \sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}$ or $\delta > t_{\alpha/2,\mathrm{df}} \sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}$

Example 2. *As in Example 1, the following data was observed in a comparison of the tar content (in mg) of two cigarettes produced by different processes:*

	Sample Size	*Sample Mean*	*Sample Variance*
X	*60*	*15.2*	*2.68*
Y	*44*	*15.5*	*3.07*

As in Example 1, test the null hypothesis that $\mu_x = \mu_y$ against the alternative that $\mu_x \neq \mu_y$ at significance level 0.05. This time, however, do not assume that the sample variances are the true population variances. Do assume that X and Y are normally distributed.

Solution. *We have*

$$\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}} = \sqrt{\frac{2.68}{60} + \frac{3.07}{44}} = 0.3383.$$

Under the current assumptions, we must obtain $t_{0.025,df}$ where $df = \min(60, 44) - 1 = 43$. Its value is 2.0167, which we obtained using software. Its value can be approximated from the tables by interpolation. To perform the interpolation,we look up $t_{0.025,40} = 2.0211$ and $t_{0.025,50} = 2.0086$ (because the required degrees of freedom, 43, is between 40 and 50, two values that are tabulated. The equation of the line that joins the two points (2.0211,40) and (2.0086,50) is

$$t = \frac{2.0086 - 2.0211}{50 - 40}(df - 40) + 2.0211,$$

or $t = 2.0711 - 0.00125\, df$. Substituting $df = 43$, we obtain $t_{0.025,43} = 2.01735$. The error of this approximation is only 0.00065, but, even so, we will proceed with the slightly more accurate value obtained by technology. We calculate

$$t_{0.025,43} \sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}} = 2.0167 \cdot 0.3383 = 0.6822.$$

Because $\left|\overline{X} - \overline{Y}\right|$, *or* $|15.2 - 15.5|$, *or* 0.3 *is not greater than 0.6822, we do not reject the null hypothesis: there is insufficient evidence to conclude that* $\mu_X \neq \mu_Y$.
Remark: *A contemporary p-value test can be performed instead. We indicate the calculation, which we bring to a close using technology. (In this case, the conclusion of the calculation using tables would be unpleasantly tedious because three interpolations would be required. Individually, they are no more sophisticated or difficult than the interpolations already done.) The p-value is*

$$\begin{aligned} P(|\widehat{\mu_X} - \widehat{\mu_Y}| \geq |15.2 - 15.5| \mid \mu_X = \mu_Y) &= P(|\widehat{\mu_X} - \widehat{\mu_Y}| \geq 0.3 \mid \mu_X = \mu_Y) \\ &= P\left(\left|\frac{\widehat{\mu_X} - \widehat{\mu_Y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}\right| \geq \frac{0.3}{0.3383} \mid \mu_X = \mu_Y\right) \\ &= P(|t_{43}| \geq 0.8868) \\ &= 2\,P(t_{43} \geq 0.8868) \\ &= 2\,(0.1901)) \\ &= 0.3802, \end{aligned}$$

which is considerably greater than conventional cut-offs for rejecting the null hypothesis.

Testing Two-Population Means of Distributions With Unknown but Equal Variances (Pooling)

In this subsection, we retain the assumptions of the preceding subsection. To them, we add one further assumption: the equality of S_X^2 and S_Y^2, even though they are unknown.[6] In this situation, we calculate the pooled variance:

$$\mathrm{S}_p^2 = \frac{(n-1)\,\mathrm{S}_\mathrm{X}^2 + (m-1)\,\mathrm{S}_\mathrm{Y}^2}{n+m-2}$$

and the resulting standard error

$$\mathrm{SE} = S_p\sqrt{\frac{1}{n} + \frac{1}{m}}.$$

Using these sample statistics, we obtain the following three tests:

Hypothesis Test (X and Y are normal, or at least approximately normal, with equal, unknown variances)
$\mathrm{H}_0 : \mu_\mathrm{X} = \mu_\mathrm{Y}$ (or $\mu_\mathrm{X} \geq \mu_\mathrm{Y}$)
$\mathrm{H}_\mathrm{a} : \mu_\mathrm{X} < \mu_\mathrm{Y}$
Test Statistic: $\overline{\mathrm{X}} - \overline{\mathrm{Y}}$ (difference of sample means)
Degrees of Freedom: $\mathrm{df} = n + m - 2$
Sample Size: n for X and m for Y
Significance Level: α
Critical Region: $\overline{\mathrm{X}} - \overline{\mathrm{Y}} < -\,\mathrm{t}_{\alpha,\mathrm{df}}\,\mathrm{SE}$

Hypothesis Test (X and Y are normal, or at least approximately normal)
$\mathrm{H}_0 : \mu_\mathrm{X} = \mu_\mathrm{Y}$ (or $\mu_\mathrm{X} \leq \mu_\mathrm{Y}$)
$\mathrm{H}_\mathrm{a} : \mu_\mathrm{X} > \mu_\mathrm{Y}$
Test Statistic: $\overline{\mathrm{X}} - \overline{\mathrm{Y}}$ (difference of sample means)
Degrees of Freedom: $\mathrm{df} = n + m - 2$

[6] It often happens that we can say two things are equal even if we do not know their common value. For example, actors Edward Norton and Christian Slater were born on the same day. Now you know that their ages are the same, even if you don't know their age. Same thing with actresses Ashley Olsen and Mary-Kate Olsen. They were born on the same day and have the same surname too. What a coincidence!

Sample Size: n for X and m for Y
Significance Level: α
Critical Region: $\overline{\mathrm{X}} - \overline{\mathrm{Y}} > \mathrm{t}_{\alpha,\mathrm{df}}$ SE

Hypothesis Test (X and Y are normal, or at least approximately normal)
 $\mathrm{H_0} : \mu_{\mathrm{X}} = \mu_{\mathrm{Y}}$
 $\mathrm{H_a} : \mu_{\mathrm{X}} \neq \mu_{\mathrm{Y}}$
Test Statistic: $\delta = \overline{\mathrm{X}} - \overline{\mathrm{Y}}$ (difference of sample means)
Degrees of Freedom: $\mathrm{df} = n + m - 2$
Sample Size : n for X and m for Y
Significance Level: α
Critical Region: $|\delta| \geq \mathrm{t}_{\alpha/2,\mathrm{df}}$ SE Equivalently $\delta < -\mathrm{t}_{\alpha/2,\mathrm{df}}$ SE or $\delta > \mathrm{t}_{\alpha/2,\mathrm{df}}$ SE

Example 3. *As in Examples 1 and 2, the following data was observed in a comparison of the tar content (in mg) of two cigarettes produced by different processes:*

	Sample Size	*Sample Mean*	*Sample Variance*
X	*60*	*15.2*	*2.68*
Y	*44*	*15.5*	*3.07*

As in Examples 1 and 2, test the null hypothesis that $\mu_X = \mu_Y$ *against the alternative that* $\mu_X \neq \mu_Y$ *at significance level 0.05. This time, however, assume that the sample variances are unknown but equal.*

Solution. *In this example in which we test for the equality of two means, we set the difference* d_0 *equal to 0. We have*

$$S_p^{\,2} = \frac{(n-1)\,S_X^2 + (m-1)\,S_Y^2}{n+m-2} = \frac{59 \times 2.68 + 43 \times 3.07}{60+44-2} = 2.8444$$

and

$$SE = S_p\sqrt{\frac{1}{n} + \frac{1}{m}} = \sqrt{2.8444} \cdot \sqrt{\frac{1}{60} + \frac{1}{44}} = 0.3347.$$

Under the current assumptions, we must obtain $t_{0.025,df}$ *where* $df = 60 + 44 - 2 = 102$. *Its value is 1.9835, which we obtained using software. Its value can be approximated from the tables by extrapolation (because the tables provided have maximum df equal to 100).*[7] *We calculate*

$$t_{0.025,102}\, SE = 1.9835 \times 0.3347 = 0.6639.$$

Because $\left|\overline{X} - \overline{Y} - d_0\right|$, *or* $|15.2 - 15.5 - 0|$, *or* 0.3 *is not greater than 0.6639, we do not reject the null hypothesis: there is insufficient evidence to conclude that* $\mu_x \neq \mu_y$.

Hypothesis Tests of Two-Sample Proportions

By now the transition between means and proportions should be familiar. The hypothesis tests for the null hypothesis that one proportion p_1 equals another proportion p_2 follow from line (9.4.5).
Hypothesis Test
 $\mathrm{H_0} : p_1 = p_2$ (or $p_1 \geq p_2$)
 $\mathrm{H_a} : p_1 < p_2$
Test Statistic: $\widehat{p}_1 - \widehat{p}_2$ (difference of sample proportions)
Sample Size: n and m

[7] Using df = 100 results in quite a good approximation: $\mathrm{t}_{0.025,\mathrm{df}} = 1.9840$. The normal approximation $\mathrm{z}_{0.025} = 1.96$ is not terrible.

Significance Level: α

Critical Region: $\widehat{p}_1 - \widehat{p}_2 < -z_\alpha \sqrt{\dfrac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \dfrac{\widehat{p}_2(1-\widehat{p}_2)}{m}}$

Hypothesis Test
$H_0 : p_1 = p_2$ (or $p_1 \leq p_2$)
$H_a : p_1 > p_2$
Test Statistic: $\widehat{p}_1 - \widehat{p}_2$ (difference of sample proportions)
Sample Size: n and m
Significance Level: α

Critical Region: $\widehat{p}_1 - \widehat{p}_2 > z_\alpha \sqrt{\dfrac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \dfrac{\widehat{p}_2(1-\widehat{p}_2)}{m}}$

Hypothesis Test
$H_0 : p_1 = p_2$
$H_a : p_1 \neq p_2$

Test Statistic: $\widehat{p}_1 - \widehat{p}_2$ (difference of sample proportions)
Sample Size: n and m
Significance Level: α

Critical Region: $|\widehat{p}_1 - \widehat{p}_2| \geq z_{\alpha/2} \sqrt{\dfrac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \dfrac{\widehat{p}_2(1-\widehat{p}_2)}{m}}$

Equivalently, $\widehat{p}_1 - \widehat{p}_2 < -z_{\alpha/2} \sqrt{\dfrac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \dfrac{\widehat{p}_2(1-\widehat{p}_2)}{m}}$ or $\widehat{p}_1 - \widehat{p}_2 > z_{\alpha/2} \sqrt{\dfrac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \dfrac{\widehat{p}_2(1-\widehat{p}_2)}{m}}$

Example 4. *Researchers in upper New York state investigated the relationship between watching television and aggressive behaviour. After sampling 707 teenagers, the researchers made follow-up observations for 17 years. The data from their investigation is tabulated.*

	Committed Aggressive Act	*Did Not Commit Aggressive Act*
Less than 1 hr/day TV	*5*	*83*
More than 1 hr/day TV	*154*	*465*

Let Group 1 be the population of teens who watch less than 1 hour of TV per day and let Group 2 be the population of teens who watch more than 1 hour of TV per day. Let p_1 be the proportion of teenagers in Group 1 who commit aggressive acts and let p_2 be the proportion of teenagers in Group 2 who commit aggressive acts. Test $p_1 = p_2$ against the alternative $p_1 \neq p_2$ at the 0.05 significance level.

Solution. *The sample proportions observed are* $p_1 = 5/(5+88) = 0.05376$ *and* $p_2 = 154/(154+465) = 0.2488$. *We calculate*

$$z_{\alpha/2} \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}} = 1.96 \sqrt{\frac{0.05376\,(1-0.05376)}{88} + \frac{0.2488\,(1-0.2488)}{619}} = 0.05814.$$

Because $|\widehat{p}_1 - \widehat{p}_2| = |0.05376 - 0.2488| = 0.19504$ *is greater than 0.05814, we reject the null hypothesis.*

Paired Samples

In this subsection we consider a hypothesis that may appear to be a two-sample hypotheis but which is treated by a one-sample test. The key distinction is that the two samples are not independent: each observation of one sample is paired with an observation from the second sample. Common examples include measurements before and after treatment.

Example 5. *The following table records the blood serum cholesterol levels of 7 patients, initially in the elevated to high range, before and after treatment.*

	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5	Patient 6	Patient 7
Before	219	249	221	209	217	237	213
After	205	243	223	200	221	235	198

Test the hypothesis that the mean population blood cholesterol level before treatment is equal to that after treatment against the altermative that the level after treatment is reduced. Calculate the p-value.

Solution. *Let X be the difference, before minus after, of cholesterol levels. The observations of X are 14, 6, -2, 9, -4, 2, 15. The sample statistics are* $\overline{X} = 5.714$ *and* $S_X = 7.455$. *Letting* μ *denote the mean difference of the population, the p-value is*

$$\begin{aligned} P((\widehat{\mu} \geq 5.714 \,|\, \mu = 0) &= P\left(\frac{\widehat{\mu} - \mu}{S/\sqrt{7}} \geq \frac{5.714}{7.455/\sqrt{7}} \,\middle|\, \mu = 0\right) \\ &= P((t_6 \geq 2.028) \\ &= 0.044. \end{aligned}$$

The p-value is generally accepted small enough to reject the null hypothesis.

10.6 Hypothesis Tests for Variance

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution with mean μ, variance σ^2, and standard deviation σ. Let S^2 be the sample variance. Then $(n-1)\,S^2/\sigma^2 \sim \chi^2_{n-1}$. It follows that, for a given positive number σ_0^2,

$$\begin{aligned} P\left(S^2 \leq \frac{\sigma_0^2}{n-1}\chi^2_{1-\alpha,n-1} \,|\, \sigma^2 = \sigma_0^2\right) &= P\left(\frac{(n-1)S^2}{\sigma_0^2} \leq \chi^2_{1-\alpha,n-1} \,|\, \sigma^2 = \sigma_0^2\right) \\ &= P\left(\frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{1-\alpha,n-1}\right) \\ &= P\left(\chi^2_{n-1} \leq \chi^2_{1-\alpha,n-1}\right) \\ &= 1 - P\left(\chi^2_{n-1} \geq \chi^2_{1-\alpha,n-1}\right) \\ &= 1 - (1-\alpha) \\ &= \alpha. \end{aligned}$$

This equation gives us our first hypothesis test for σ^2 and the two tests that follow can be derived using similar calculations.

Hypothesis Test
$H_0 : \sigma^2 = \sigma_0^2($ or $\sigma^2 \geq \sigma_0^2)$
$H_a : \sigma^2 < \sigma_0^2$
Test Statistic S^2 (sample variance)
Sample Size : n
Significance Level: α
Critical Region: $S^2 < \dfrac{\sigma_0^2}{n-1}\chi^2_{1-\alpha,n-1}$

Hypothesis Test
$H_0 : \sigma^2 = \sigma_0^2($ or $\sigma^2 \leq \sigma_0^2)$
$H_a : \sigma^2 > \sigma_0^2$

Test Statistic S^2 (sample variance)
Sample Size : n
Significance Level: α
Critical Region: $S^2 > \dfrac{\sigma_0^2}{n-1}\chi^2_{\alpha,n-1}$

Hypothesis Test
$H_0 : \sigma^2 = \sigma_0^2$
$H_a : \sigma^2 \neq \sigma_0^2$
Test Statistic S^2 (sample variance)
Sample Size : n
Significance Level: α
Critical Region: $S^2 < \dfrac{\sigma_0^2}{n-1}\chi^2_{1-\alpha/2,n-1}$ or $S^2 > \dfrac{\sigma_0^2}{n-1}\chi^2_{\alpha/2,n-1}$

Example 1. *A random sample of 16 scores were drawn from a large statistics class at First President University. The sample standard deviation was calculated to be 5.651. Let σ denote the population standard deviation. Test the null hypothesis that $\sigma = 4.199$ against the alternative $\sigma > 4.199$ at significance level 0.05. (The hypothesized value 4.199 was the standard deviation of the entire class.)*

Solution. *We calculate $S^2 = 31.93$ and $\sigma_0^2 \times \chi^2_{0.05,n-1}/(n-1) = (4.199)^2 \times 24.9958/15 = 29.38$. Because $S^2 \geq \sigma_0^2 \times \chi^2_{0.05,n-1}/(n-1)$, we reject the null hypothesis. (Remember, about 5% of the time that we use this procedure, we will reject a true null hypothesis.)*

10.7 The Three χ^2 Tests

Suppose that a sample (random or not) is taken from a population (human or not) that is composed of different groups. How can we decide whether the makeup of the sample is the same as that of the population? For example, consider drivers who are stopped by the police for discretionary, "investigative" reasons (such as burned-out tail lights, improper lane usage, and so on). Is the ethnic make-up of this population the same as the ethnic make-up of the entire population (apart from differences that may be ascribed to the variability of sampling)? A similar question may be posed for the citizens who are selected for jury duty. The same type of question can be posed even when the populations that are compared have no overlap whatsoever. Consider a treatment that has been trialed on men. In a subsequent trial of the treatment on women, we will want to know if the distribution of successes and failures matches the distribution for men. A university may want to know if the composition of graduating majors one year is the same as the composition of those who graduated another year. In this section we learn how to answer this type of question and two others that are somewhat similar.

The χ^2 Goodness of Fit Test

Suppose that X is a distribution composed of r integer counts: $x_1, x_2, \ldots, x_r$. In the applications that we will be considering, such an X typically arises as the frequencies of the classes of a categorical random variable. Another possibility is that X consists of the bin counts that arise when the observations of a numerical random variable have been binned as if to create a histogram with r bars. We are actually interested in the relative frquencies $p_1 = x_1/n_X, p_2 = x_2/n_X, \ldots, p_r = x_r/n_X$ where $n_X = x_1 + x_2 + \cdots + x_r$. Indeed, if we are given $p_1, p_2, \ldots, p_r$ instead of $x_1, x_2, \ldots, x_r$, then we have come out ahead: we do not need the counts and, having been given the relative frequencies, we do not need to calculate them ourselves. It is assumed that $p_1 + p_2 + \cdots + p_r = 1$, apart, possibly, from a tiny rounding inaccuracy.

Now suppose that Y is a distribution consisting of r counts $O_1, O_2, \ldots, O_r$. Typically these numbers are the observed counts of a categorical random variable with r classes. We will assume that each of these counts is at least 5. When this condition is not met, classes are merged to achieve the minimum count requirement. Let $n_Y = O_1 + O_2 + \cdots + O_r$ be the total number of observations for Y. The question we ask is, Do the distributions X and Y match? If they did match, then we could use the relative frequencies of the X distribution to deduce approximate expected values of Y. To be specific, if X and Y did match, then, for each j between 1 and r, we would expect the observed count O_j to be approximately $E_j = n_Y p_j$. Of course, due to variability, it is unreasonable to anticipate that the expected value E_j will be exactly equal to the observed value O_j. That said, if Y did match X, then we would not expect the deviations $O_j - E_j$ to be enormous. As usual in such situations, we square to prevent cancellation between deviations of opposite signs. And we scale by dividing by a measure, E_j in this case, of the size we are dealing with. Finally, we sum $(O_j - E_j)^2/E_j$ over all the r classes. If the distributions of X and Y match, which we will take as the null hypothesis in this setting, then the statistic

$$\sum_{j=1}^{r} \frac{(O_j - E_j)^2}{E_j} \tag{10.7.1}$$

is approximately χ^2_{r-1}:

$$\sum_{j=1}^{r} \frac{(O_j - E_j)^2}{E_j} \sim \chi^2_{r-1} \qquad \text{(approximately)}. \tag{10.7.2}$$

The summands $(O_j - E_j)^2/E_j$ in expression (10.7.1) are called ***components***. An immediate consequence of approximation (10.7.2) is that

$$\mathrm{P}\left(\chi^2_{r-1} \geq \sum_{j=1}^{r} \frac{(O_j - E_j)^2}{E_j}\right) \tag{10.7.3}$$

is the p-value of our hypothesis test. It is important to note a distinction between the observations and the expectations. The values of $O_1, O_2, \ldots, O_r$ are the counts from the sample. These values satisfy the equation $n = O_1 + O_2 + \cdots + O_r$. The numbers $E_1 = n\,p_1, E_2 = n\,p_2, \ldots, E_r = n\,p_r$ depend on the sample size n and the population proportions $p_1, p_2, \ldots, p_r$, which do not depend on the sample and which must be known *a priori*. Like the observations, the expectations sum to the sample size: $n = E_1 + E_2 + \cdots + E_r$.[8] Unlike the values $O_1, O_2, \ldots, O_r$, which are counts and must therefore be nonnegative integers, the values $E_1, E_2, \ldots, E_r$ are theoretical numbers that are, in general, not integers. In a classical hypothesis test, the critical region at significance level α is

$$\sum_{j=1}^{r} \frac{(O_j - E_j)^2}{E_j} \geq \chi^2_{\alpha, r-1}. \tag{10.7.4}$$

Example 1. *At the time of this writing, African Americans make up 23.7% of the population of St. Louis County, which does not include the city of St. Louis.*[9] *Suppose that, in a random sample of 100 traffic tickets, 32 tickets were issued to African Americans. Test the hypothesis that the distribution of tickets matches the population distribution. What if 33 of the violations were issued to African Americans?*

Solution. *In this example,* $r = 2$*: the two classes are African American and Other. The proportions are* $p_1 = 0.237$ *and* $p_2 = 0.763$*. The sample size is* $n = 100$*. The expectations are* $E_1 = n\,p_1 = 23.7$ *and* $E_2 = n\,p_2 = 76.3$*. The test statistic is*

$$\frac{(32 - 23.7)^2}{23.7} + \frac{(68 - 76.3)^2}{76.3}, \qquad \textit{or} \quad 3.8096.$$

The p-value is $P\left(\chi^2_{2-1} \geq 3.8096\right)$*, or 0.051. This value, greater than 0.05, is generally not taken to be small enough to reject the null hypothesis, but it is certainly a close call. In a classical hypothesis test, we*

[8] When the proportions are expressed as decimals, the expectations may not sum to exactly n due to rounding errors.

[9] In a lecture given on November 24, 2014, the author used the value 23.3%. That was based on the 2010 U.S. census. The value given here is a 2013 United States Census Bureau estimate (`http://quickfacts.census.gov/qfd/states/29/29189.html` Retrieved: 26 November 2014.).

would look up the values of $\chi^2_{0.05,1}$, which is 3.8415, and retain the null hyhpothesis because the value of the test statistic, namely 3.8096, is smaller. In the case of 33 citations, the value of the test statistic is $(33-23.7)^2/23.7+(67-76.3)^2/76.3$, or 4.7829. In either the classical or contemporary hypothesis test, the decision to reject the null hypothesis is not in doubt: in the former, 4.7829 is well above 3.8415, and in the latter, the p-value is 0.02874.

Example 2. *According to an American Medical Association 2010 census, practicing surgeons in the U.S. can be divided into the following specialties:*

Surgical Specialties	Count
General Surgery	22,061
Neurological Surgery	4,583
Orthopedic Surgery	18,281
Plastic Surgery	6,457
Thoracic Surgery	4,329
Vascular Surgery	2,594
Total	58,305

Table 10.7.1: U.S. Surgeons by Specialty, 2010

In a state with 496 general surgeons, 87 neurological surgeons, 381 orthopedic surgeons, 103 plastic surgeons, 89 thoracic surgeons, and 50 vascular surgeons, is the distribution os surgical specialties the same as the national distribution?

Solution. *In this example there are $r=6$ classes. We must compute the proportions from the given data: $p_1=22061/58305=0.3784$, $p_2=4583/58305=0.07860$, $p_3=18281/58305=0.3135$, $p_4=6457/58305=0.1107$, $p_5=4329/58305=0.07425$, $p_6=2594/58305=0.04449$. We must also compute the sample size: $n=516+90+382+99+89+30=1206$. The expected numbers of surgeons in the state are: $E_1=n\,p_1=1206\times0.3784=456.35$, $E_2=n\,p_2=1206\times0.07860=94.79$, $E_3=n\,p_3=1206\times0.3135=378.08$, $E_4=n\,p_4=1206\times0.1107=133.50$, $E_5=n\,p_5=1206\times0.07425=89.55$, $E_6=n\,p_6=1206\times0.04449=53.65$. All this data, as well as the components and their sum are recorded in Table 10.7.2:*

Surgical Specialties	National Count	National Proportion	State Count O	Expectation E	$(O-E)^2/E$
General Surgery	22,061	0.3784	496	456.4	3.436
Neurological Surgery	4,583	0.0786	87	94.79	0.6402
Orthopedic Surgery	18,281	0.3135	381	378.1	0.02224
Plastic Surgery	6,457	0.1107	103	133.5	6.968
Thoracic Surgery	4,329	0.0743	89	89.55	0.003378
Vascular Surgery	2,594	0.0445	50	53.65	0.2483
Total	58,305	1.0000	1206	1206	11.3182

Table 10.7.2: U.S. Surgeons by Specialty, 2010, And a Hypothetical State

The p-value is $P\left(\chi^2_5\geq 11.3182\right)$, or 0.0454, which is generally considered to be small enough to reject the null hypothesis. The national distribution of surgeons by specialty is not a good fit for the state distribution. Alternatively, i a classical test with significance level 0.05, we look up $\chi^2_{5,0.05}=11.07$. It follows that the critical region is $\sum_{j=1}^{r}\frac{(O_j-E_j)^2}{E_j}\geq 11.07$. Because our test statistic has value 11.3182, which is greater than 11.07, we reject the null hypothesis.

The χ^2 Test of Homogeneity

The chi-squared test for homogeneity tests the hypothesis that three or more populations are equal (or homogeneous) in several categories. The implementation of this test is similar to that of the chi-squared goodness of fit test, but the number of degrees of freedom depends not only on the number c of categories but also on the number r of populations. To be specific, $\text{df} = (r-1)(c-1)$. The next example will illustrate the details.

Example 3. *An admissions officer at Midwestern University tabulates the incoming class by race and school:*

Race \School	*A & S*	*Eng*	*Bus*	*Art-Arch*	*Total*
White	*500*	*160*	*122*	*58*	*840*
Black	*86*	*19*	*23*	*16*	*144*
Other	*114*	*46*	*30*	*26*	*216*
Total	*700*	*225*	*175*	*100*	*1200*

Table 10.7.3: Distribution of MU Incoming Class, by Race and School—Actual Counts

Are the distributions of Whites, Blacks, and the composite category of other races equal across the four schools?

Solution. *In this example there are $r = 3$ populations divided into $c = 4$ categories. We must compute the proportions from the given data:*
p_1 = *proportion of entering class enrolling in Arts and Sciences* $= 700/1200 = 0.58333$,
p_2 = *proportion of entering class enrolling in Engineering* $= 225/1200 = 0.18750$,
p_3 = *proportion of entering class enrolling in Business* $= 175/1200 = 0.14583$,
p_4 = *proportion of entering class enrolling in Business* $= 100/1200 = 0.08333$.
We must also The sample sizes of the three populations are $n_1 = 840$, $n_2 = 144$, and $n_3 = 216$. If the distributions were equal, the expected number of students would be:
$E_{1,1} = n_1\,p_1 = 840 \times 0.58333 = 490$ *White students in Arts & Sciences,*
$E_{2,1} = n_2\,p_1 = 144 \times 0.58333 = 84$ *Black students in Arts & Sciences,*
$E_{3,1} = n_3\,p_1 = 216 \times 0.58333 = 126$ *other students in Arts & Sciences,*
$E_{1,2} = n_1\,p_2 = 840 \times 0.18750 = 157.5$ *White students in Engineering,*
$E_{2,2} = n_2\,p_2 = 144 \times 0.18750 = 27$ *Black students in Engineering,*
$E_{3,2} = n_3\,p_2 = 216 \times 0.18750 = 40.5$ *other students in Engineering,*
$E_{1,3} = n_1\,p_3 = 840 \times 0.14583 = 122.5$ *White students in Business,*
$E_{2,3} = n_2\,p_3 = 144 \times 0.14583 = 21$ *Black students in Business,*
$E_{3,3} = n_3\,p_3 = 216 \times 0.14583 = 31.5$ *other students in Business,*
$E_{1,4} = n_1\,p_4 = 840 \times 0.08333 = 70$ *White students in Art & Architecture,*
$E_{2,4} = n_2\,p_4 = 144 \times 0.08333 = 12$ *Black students in Art & Architecture,*
$E_{3,4} = n_3\,p_4 = 216 \times 0.08333 = 18$ *other students in Art & Architecture.*
Let us tabulate these values

Race \School	*A & S*	*Eng*	*Bus*	*Art-Arch*	*Total*
White	*490*	*157.5*	*122.5*	*70*	*840*
Black	*84*	*27*	*21*	*12*	*144*
Other	*126*	*40.5*	*31.5*	*18*	*216*
Total	*700*	*225*	*175*	*100*	*1200*

Table 10.7.4: Expected Distribution by Race and School, if Homogeneous

The residuals are obtained by subtracting the values of the cells of Table 10.7.4 from the corresponding entries of Table 10.7.3. The test statistic is obtained by squaring the residuals, dividing each squared residual by the *expected value corresponding to the residual, and then summing the quotients:*

$$\frac{(500-490)^2}{490}+\frac{(160-157.5)^2}{157.5}+\frac{(122-122.5)^2}{122.5}+\frac{(58-70)^2}{70}+\frac{(86-84)^2}{84}+\frac{(23-21)^2}{21}$$

$$+\frac{(19-27)^2}{27}+\frac{(16-12)^2}{12}+\frac{(114-126)^2}{126}+\frac{(46-40.5)^2}{40.5}+\frac{(30-31.5)^2}{31.5}+\frac{(26-18)^2}{18},\ \textit{or}\ 11.762.$$

The p-value is $P\left(\chi^2_{(3-1)\times(4-1)} \geq 11.762\right)$*, or 0.0675, which is not particularly small. We retain the null hypothesis that the three distributions across schools are equal.*

The χ^2 Test of Independence

Consider the following contingency table of 500 randomly selected individuals.

		Blood	Type		
Eye Color	A	B	O	AB	Total
Blue	121	57	98	27	303
Brown	83	55	50	9	197
Total	204	112	148	36	500

Table 10.7.5: Blood Type–Eye Colour Contingency Table

Can we tell from this contingency table if eye colour depends on blood type? We will take the independence of these two attributes as the null hypothesis. The alternative hypothesis is that eye colour and blood type are dependent. Can these hypotheses be interchanged? As a practical matter, you will see that the choice of independence as the null hypothesis allows us to calculate either a p-value or a size of type I error, whichever is required. That is because with independence as the assumed null hypothesis, we can determine the number that is expected in each cell.

Of the 500 sampled individuals, 303 have blue eyes. Thus, the probability that a randomly selected individual from the sample has blue eyes is $p_{1,\cdot} = 0.606$. Similarly, the probability that a randomly selected individual from the sample has brown eyes is $p_{2,\cdot} = 197/500$, or $p_{2,\cdot} = 0.394$. Of the 500 members of the sample, 204 have blood type A. Thus, the probability that a randomly selected individual from the sample has blood type A is $p_{\cdot,1} = 0.408$. Similarly, the probability that a randomly selected individual from the sample has blood type B is $p_{\cdot,2} = 112/500$, or $p_{\cdot,2} = 0.224$, the probability that a randomly selected individual from the sample has blood type O is $p_{\cdot,3} = 148/500$, or $p_{\cdot,3} = 0.296$, and the probability that a randomly selected individual from the sample has blood type AB is $p_{\cdot,4} = 36/500$, or $p_{\cdot,4} = 0.0.72$.

Assuming the null hypothesis of independence, the probability that a randomly selected individual from the sample has blue eyes and blood type A is the product $p_{1,\cdot} \times p_{\cdot,1} = 0.606 \times 0.408 = 0.247248$. Therefore the expected number of individuals with blue eyes and blood type A is 0.247248×500, or 123.624. Similarly, the expected number of individuals with blue eyes and blood type B is $(0.606 \times 0.224) \times 500 = 67.872$. In general, the expected number in the i^{th} row j^{th} columns is $p_{i,\cdot} \times p_{\cdot,j} \times 500$. These expected numbers are tabulated below:

		Blood	Type	
Eye Color	A	B	O	AB
Blue	123.624	67.87	89.69	21.82
Brown	80.38	44.13	58.31	14.18

Table 10.7.6: Blood Type–Eye Colour: Expected Counts Assuming independence

We obtain a test statistic

$$\sum \frac{(O-E)^2}{E} \tag{10.7.5}$$

by subtracting expected counts E from corresponding actual counts O, squaring the differences, dividing each squared difference $(O-E)^2$ by the the corresponding expected count E, and then summing the quotients $(O-E)^2/E$ over all the cells. This statistic is approximately χ^2_{df} where

$$\text{df} = r \times c - (r-1) - (c-1) - 1 = (r-1)(c-1),$$

r being the number of rows of the contingency table (not including the marginal row for the totals) and c being the number of columns of the contingency table (not including the marginal column for the totals).

In this example, expression (10.7.5) comes to 9.6431 and df == $(2-1)(4-1) = 3$. The p-value is $\text{P}\left(\chi^2_3 \geq 9.6431\right)$, which is 0.02186. This small probability leads us to reject the null hypothesis. Alternatively, in a classical test at significance level 0.05, we find that $\chi^2_{0.05,3} = 7.8147$. The crtical region is $\sum (O-E)^2/E \geq$ 7.8147. Because the observed value, 9.6431, of the test statistic satisfies this inequality, i.e., lies in the critical region, we reject the null hypothesis: the evidence is that eye colour and blood type are dependent. (As usual, the existence of such a relationship does not imply causation.)

Remark: In order to calculate the entries of Table 10.7.6, probabilities $p_{i,\cdot}$ and $p_{\cdot,j}$ were employed. These probabilities were introduced for pedagogical purposes, but, as a practical matter, they are not necessary for the calculation. Let R_i be the sum of the values in the i^{th} row of Table 10.7.5—this is the value that appears in the marginal column to the right of the table. Let C_j be the sum of the values in the j^{th} column of Table 10.7.5—this is the value that appears in the marginal row below the table. Let n be the table total—this is the value that appears at the bottom right in the intersection of the two marginals. then the entry E_{ij} in the i^{th} row, j^{th} column of Table 10.7.6 is given by

$$\text{E}_{ij} = \frac{R_i \times C_j}{n}.$$

For example, the expected frequency of individuals with brown eyes and blood type O, namely 58.31, is $197 \times 148/500$.

Exercises

1. (Washington University Exam, Fall 2007)
   Census data for a certain county show that 19% of the adults residents are Hispanic. Suppose that 720 people are called for jury duty, and only 125 of them are Hispanic. Does the apparent underepresentation of Hispanics call into question the fairness of the jury selection system? Calculate the P-value of the two-sided test of the appropriate null hypothesis.

A) 0.2079	B) 0.2215	C) 0.2351	D) 0.2487	E) 0.2623
F) 0.2759	G) 0.3031	H) 0.3112	I) 0.3194	J) Other

2. (Washington University Exam, Fall 2007)
   It is widely believed that regular mammogram screening may detect breast cancer early, resulting in fewer deaths from that disease. One study that investigated this issue over a period of 18 years was published during the 1970s. Among 32,000 women who never had mammograms, 200 died of breast cancer, while only 150 0f 30,000 who had undergone screening died of breast cancer. Do these results suggest that regular mammogram screening may be an effective tool to reduce brest cancer deaths? Calculate the P-value for the appropriate one-sided hypothesis test.

3. (Washington University exam, Fall 2010)
In 2001, a national vital statistics report indicated that about 3% of all births produced twins. Is the rate of twin births the same among very young mothers? A randomly chosen sample found only 15 sets of twins born to 1000 teenage girls. If a 1000 such samples (of size 1000) were taken, what can we conclude?

A) The sample proportion of .015 would occur in only 1 in a 1000 such samples
B) The sample proportion of .015 or more extreme would occur in approximately 1 out of a 1000 such samples
C) The sample proportion of .015 or more extreme would occur in approximately 3 out of a 1000 such samples
D) The sample proportion of .015 or more extreme would occur in approximately 5 out of a 1000 such samples
E) The sample proportion of .015 or more extreme would occur in approximately 10 out of a 1000 such samples.
F) The sample proportion of .015 or more extreme would occur in approximately 15 out of a 1000 such samples.

4. (Washington University exam, Fall 2010)
We would like to know if St. Louis has a greater proportion of households that have pets than New York, so we do a hypothesis test using a sample of 100 houses in St. Louis, and 135 in New York. 53 of the 100 houses in St. Louis had pets, while 70 of the 135 houses in New York had pets. What is the model for the difference of the sample proportions?

A) $N(0, 0.066)$ B) $N(0, 0.142)$ C) $N(0, 0.254)$ D) $N(0, 0.416)$ E) $N(0, 1)$
F) $N(0.011, 0.323)$ G) $N(0.011, 0.262)$ H) $N(0.011, 0.521)$ I) $N(0.011, 0.416)$ J) $N(0.011, 1)$

5. (Washington University exam, Spring 2014) Based on data from two VERY LARGE independent samples, two students tested a hypothesis about equality of population means using $\alpha = 0.05$. One student used a one-tail test and rejected the null hypothesis, but the other used a two-tail test and failed to reject the null. Which of these might have been their calculated value of t-score?

A. 1:22 B. 1:55 C. 1:88 D. 2:22 E. 2:66

6. (Washington University exam, Spring 1014)
Do more than 50% of U.S. adults feel they get enough sleep? According to Gallup's December 2004 Lifestyle poll, 55.03% of U.S. adults said that that they get enough sleep. The poll was based on a random sample of 1003 U.S. adults. What was the p value? (Use a one-sided test)

A) 0.0004782 B) 0.0007212 C) 0.005782 D) 0.009156 E) 0.04337
F) 0.07005 G) 0.0960 H) 0.1615 I) 0.2010 J) 0.2350

7. (Washington University exam, Spring 2010)
To test whether cars get higher mileage per gallon with premium gas (as opposed to regular), we test 8 cars with a tankful of each of regular and premium gasoline. The resulting mileage (in miles per gallon) is below:

Regular	16	19	20	21	23	27	27	28
Premium	19	21	23	23	25	26	29	31

Do cars get better average mileage with premium gasoline? Find the P-value from the appropriate 1-sided hypothesis test.

A) 1.0000 B) 0.9982 C) 0.9830 D) 0.1876 E) 0.1780
F) 0.0169 G) 0.0017 H) 0.0008 I) 0.0000 J) Other

8. According to the United States Census Bureau, the composition of St. Louis County in 2013 was: White, not Hispanic or Latino 68.0%; Black or African American 23.7%; Asian 3.8%; White, Hispanic or Latino 2.3%; Other 2.2%. These are Census Bureau estimates. Assume that they are correct. According to the 2010 national census, the composition of St. Louis County that year was: White, not Hispanic or Latino 68.9%; Black or African American 23.3%, Asian 3.5%; White, Hispanic or Latino 1.4%; Other 2.9%. The 2010 census counted 998,954 individuals.

   Forty-eight weeks of the year, on Monday and on Wednesday, residents of the county are summoned to the jury assembly room in the Civil Court Building in Clayton. There they may be selected to serve as jurors. Suppose that of a random sample of 1000 jurors, 716 are White,not Hispanic or Latino, 214 are Black or African American, 41 are Asian, and 9 are Hispanic or Latino.

   a) If the sample had been taken in 2013, would the distribution of the sample of potential jurors matched the distribution of the entire county?
   b) If the sample had been taken in 2010, would the distribution of the sample of potential jurors matched the distribution of the entire county?

9. In a social study, 360 people of voting age were randomly selected and classified according to their age group and political leaning.

		Age	Group	
Leaning	18-35	36-50	Over 50	Total
Conservative	10	40	10	60
Moderate	80	85	45	210
Liberal	30	25	35	90
Total	120	150	90	360

   Test the null hypothesis that age and political leaning are not related. What is the endpoint of the critical region for test statistic (10.7.5) at significance level 0.05? What is the result of the hypothesis test?

10. A survey of 600 individuals (conducted in the 1980s) questioned surveyees about regular alcohol consumption and regular smoking (presumably tobacco, but, in any event, not limited to cigarettes):

	Smoker	Non-smoker
Drinker	188	166
Non-drinker	94	152

   Test the null hypothesis that drinking and smoking are not related. What is the endpoint of the critical region for test statistic (10.7.5) at significance level 0.01? What is the result of the hypothesis test?

11. The following data of histidine excretions (in mg) were collected from urine samples:

Men (X)	229	236	435	172	432					
Women (Y)	197	224	115	74	138	135	107	204	200	138

   Some sample statistics associated with this data are $\overline{X} = 300.8$, $\overline{Y} = 153.2$, $S_X^2 = 15291.7$, and $S_Y^2 = 2484.62$. Test the null hypothesis that $\mu_X = \mu_Y$ against the alternative that $\mu_X \neq \mu_Y$ at significance level 0.05.

12. A study at the University of Utah investigated the effect of cell phone usage on drivers' reaction times. In the control group, 32 randomly selected students listened to music, broadcast radio, or books-on-tape while engaged at a station that simulated driving. A different randomly selected group of 32 students used their cell phones. At irregular intervals, a simulated red light signal was emitted. The reaction times of the students in milliseconds (X for the control group, Y for the chatterboxes) were measured with the following results: $\overline{X} = 533.7$, $\overline{Y} = 585.2$, $S_X = 65.3$, and $S_Y = 89.6$. Test the null hypothesis that $\mu_X = \mu_Y$ against the alternative that $\mu_X < \mu_Y$ at significance level 0.05.

13. Redo the hypothesis test of the preceding exercise, but this time calculate a p-value.

14. The zinc content (in mg per kg of dry weight) of six samples of red leaf lettuce fertilized with sewage sludge was: 140, 190, 130, 150, 160, 140. Test $\sigma = 30$ against the alternative $\sigma < 30$ at 0.10 significance level.

15. A company that opened a turkey farm endeavoured to tempt local crop farmers with turkey dung as a substitute fertilizer for the cow manure that was in use. Random samples of size 25 each let to sample means $\widehat{\mu}_{CM} = 100$ and $\widehat{\mu}_{TD} = 110$ with sample variances $S_{CM}^2 = 400$ and $S_{TD}^2 = 625$. Test the null hypothesis $\mu_{CM} = \mu_{TD}$ against $\mu_{CM} < \mu_{TD}$ at significance level 0.05.

16. Find the p-value of the hypothesis test of Example 1 in Section 10.6.

17. Exercise 1 of this chapter states, "Census data for a certain county show that 19% of the adult residents are Hispanic. Suppose that 720 people are called for jury duty, and only 125 of them are Hispanic. Does the apparent underepresentation of Hispanics call into question the fairness of the jury selection system?" That exercise called for the calculation of the p-value of an appropriate two-sided hypothesis test. In this exercise, find the p-value of a goodness of fit test.

18. The R command `rpois(1:n,u)` returns n observations from a Poisson distribution with parameter u. (In these notes, λ has been the usual choice of letter, rather than u, to denote the Poisson parameter.) In an implementation of the command with $n = 50$ and $u = 1$, the values 0, 1, 2, 3, and 4 were observed with the following frequencies:

Poisson value	0	1	2	3	4
Number of Observations	17	19	8	5	1

Use these five Poisson values to create 5 classes. Use a chi-square goodness of fit test to determine if R's random generator has produced a good Poisson distribution. If a classical test were to be done, what would be the test statistic and the critical value at significance level 0.05? Now switch gears: what is the p-value? (The expected number of observations of 3 and of 4 are too small for an accurate test, but don't worry about it. This is a learning problem. Neither lives nor fortunes depend on the validity of the test.)

19. A uniform distribution U on an interval $[a, b]$ is a continuous random variable such that

$$P(u \le U \le v) = \frac{v-u}{b-a} \qquad \text{for } a \le u \le v \le b.$$

The R command `runif(50,0,8)` produced a random sample of size 50 from the uniform distribution on the interval [0,8]. The numbers of observations in the intervals $[0,1), [1,2), [2,3), [3,4), [4,5), [5,6), [6,7), [7,8]$ were 7,5,3,7,9,7,7,5 respectively. Use a chi-square goodness of fit test to determine if R's random generator has produced a good uniform distribution. If a classical test were to be done, what would be the test statistic and the critical value at significance level 0.5? (Not a typo!)

20. A psychology professor interested in art appreciation used an abstract painting in an experiment that will be described. Although the artist created the painting with an intended orientation, so that the painting did have an intended top, right side, bottom, and left side, the intended orientation was not so evident to the psychologist, who attached hinges to each side of the frame and asked 50 subjects to hang the painting in the orientation in which the painting looked best. The frequencies of their judgments can be found in the following table:

Top side up	Right side up	Bottom up	Left side up
18	7	17	8

The point of the experiment is not obvious to the author, but scorpions sting and psychologists experiment to find out what people think and do: it is in their nature. Can the chosen orientations be attributed entirely to chance, or were there genuine artistic sensibilities at work (even if those sensibilities were not uniform)?

21. There are two types of people: those who divide people into two types, and those who do not. A psychologist divided 200 subjects into two types: introverts and extroverts. (You may be interested to know that mathematicians can be of either type. The extroverts are the ones who look at *your* shoes when they talk to you.) Each of the subjects was asked to state a color preference for one of red, yellow, green, or blue. The cross-tabulation is given in the table below:

	Red	Yellow	Green	Blue
Introvert	16	8	12	24
Extrovert	84	12	24	20

Are personality type and color preference independent? Use a classical test with significance level 0.005.

22. The original AN1792 trial and its follow-up are discussed in the notes. One measure of the cognitive abilities of the participants was the Disability Assessment for Dementia exam (DAD). Of the patients who were treated with AN1792 and studied in the follow-up, 24 antibody responders retook the DAD 4.6 years after the trial, as did 27 patients who received the placebo. The scores for members of both groups declined. For each AN1792-treated patient, let X be the DAD score at the start of the trial minus the DAD score 4.6 years later. Let Y be the analogous difference for each placebo-treated patient. Assume that X and Y can be modelled by a normal distribution. The sample data for X and Y were $\overline{\mathrm{X}} = 42.28$, $\mathrm{S_X} = 30.88$, $\overline{\mathrm{Y}} = 56.34$, $\mathrm{S_Y} = 28.45$. (Notice that the positive values of X and Y indicate scores that *declined*.)

a) Perform a traditional two-sided hypothesis of $\mathrm{H_0} : \mu_\mathrm{X} < \mu_\mathrm{Y}$ at significance level 0.05.
b) Find the p-value for the test in part (a).
c) Repeat (a), but this time assume $\sigma_\mathrm{X} = \sigma_\mathrm{Y}$, as suggested by the values of $\mathrm{S_X}$ and $\mathrm{S_Y}$, and pool.
d) Find the p-value for the test in part (c).

Solutions to the Exercises

Chapter 10.

A preliminary word about interpolations. The author was asked the following question by a First President University student:

> How do you interpolate for a t or chi-square value x to find a p-value? I ask because in Chapter 10, Exercise 16, the solution interpolates essentially as:
>
> $$p = \frac{\text{(bigger significance level - smaller significance level)}}{\text{(value corresp to bigger alpha - value corresp to smaller alpha)}}\,(x - \text{value corresp to smaller alpha}) + \text{smaller alpha}$$
>
> However, in the final exam of Fall 2014, Question 16, the solution interpolates essentially
>
> $$p = \frac{\text{(smaller significance level - bigger significance level)}}{\text{(value corresp to smaller alpha - value corresp to bigger alpha)}}\,(x - \text{value corresp to bigger alpha}) + \text{bigger alpha}$$
>
> So essentially the order is switched between the two, and I was wondering which one we are supposed to use in which situation?

The answer is, both equations lead to the *same* value. Let's see by way of an example using these values:

bigger confidence level = 0.050
corresponding value to bigger alpha = 2.3
smaller confidence level = 0.025
corresponding value to smaller alpha = 2.7

with x = 2.6.

The first equation becomes

$$\frac{(0.050 - 0.025)}{(2.3 - 2.7)}\,(2.6 - 2.7) + 0.025, \qquad \text{or } 0.03125.$$

The second equation becomes

$$\frac{(0.025 - 0.050)}{(2.7 - 2.3)}\,(2.6 - 2.3) + 0.050, \qquad \text{or } 0.03125.$$

Now, on to the solutions.

1. Let p be the true proportion of Hispanics sentenced to jury duty. Let $p_0 = 0.19$. We will test
   $H_0 : p = p_0$
   $H_1 : p \neq p_0$
   The observed sample proportion is 125/720, or 0.1736. Let $\delta = |0.1736 - p_0| = |0.1736 - 0.1900| = 0.0164$. We must calculate of $|\widehat{p} - p_0| \geq \delta$ assuming that the null hypothesis is true.

$$
\begin{aligned}
&\mathrm{P}\left(|\widehat{p} - p_0| \geq \delta \,|\, p = p_0\right) \\
&\quad = \mathrm{P}\left(\left|\frac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)/720}}\right| \geq \frac{\delta}{\sqrt{p_0(1-p_0)/720}} \,\middle|\, p = p_0\right) \\
&\quad = \mathrm{P}\left(\left|\frac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)/720}}\right| \geq 1.1217 \,\middle|\, p = p_0\right) \\
&\quad = \mathrm{P}\left(|Z| \geq 1.1217\right) \\
&\quad = 2\,\mathrm{P}\left(Z \geq 1.1217\right) \\
&\quad = 2\left(1 - \mathrm{P}\left(Z \leq 1.1217\right)\right. \\
&\quad = 2\left(1 - \Phi(1.1217)\right. \\
&\quad = 0.262.
\end{aligned}
$$

2. Answered without pooling:
   Let p (and p_m) be the proportion of women who have not been screened (respectively, who have been screened) who die from breast cancer. We will test
   $H_0 : p_m = p$
   $H_1 : p_m < p$
   The observed values of $\widehat{p}$ and $\widehat{p_m}$ are 200/32000, or 0.00625, 150/30000, or 0.005. The observed difference in proportions is 0.00625 - 0.005, or 0.00125. We calculate

$$
\mathrm{SE}\left(\widehat{p} - \widehat{p_m}\right) = \sqrt{\frac{0.00625(1-0.00625)}{32000} + \frac{0.005(1-0.005)}{30000}} = 0.0006.
$$

We calculate the probability that the observed value of $\widehat{p} - \widehat{p_m}$ is greater than or equal to 0.00625 assuming that the expectation of $\widehat{p} - \widehat{p_m}$ is 0. We have

$$
\begin{aligned}
\mathrm{P}\left(\widehat{p} - \widehat{p_m} \geq 0.00125 \,|\, p = p_m\right) &= \mathrm{P}\left(\frac{\widehat{p} - \widehat{p_m}}{0.0006} \geq \frac{0.00125}{0.0006} \,\middle|\, p = p_m\right) \\
&= \mathrm{P}\left(Z \geq 2.0833\right) \\
&= 1 - \mathrm{P}\left(Z \leq 2.0833\right) \\
&= 0.01861.
\end{aligned}
$$

Answered with pooling (as intended on the exam):
The pooled sample proportion is (200+150)/(32000+30000), or 0.005645. Using this value,

$$
\mathrm{SE}\left(\widehat{p} - \widehat{p_m}\right) = \sqrt{\frac{0.005645(1-0.005645)}{32000} + \frac{0.005645(1-0.005645)}{30000}} = 0.0006021,
$$

and

$$
\begin{aligned}
\mathrm{P}\left(\widehat{p} - \widehat{p_m} \geq 0.00125 \,|\, p = p_m\right) &= \mathrm{P}\left(\frac{\widehat{p} - \widehat{p_m}}{0.0006021} \geq \frac{0.00125}{0.0006021} \,\middle|\, p = p_m\right) \\
&= \mathrm{P}\left(Z \geq 2.0761\right) \\
&= 1 - \mathrm{P}\left(Z \leq 2.0761\right) \\
&= 0.01894.
\end{aligned}
$$

3. Let $p_0 = 0.03$. Let p be the proportion of twin births among teenage mothers. Is $\mathrm{H}_0 : p = p_0$ or is $\mathrm{H}_a : p \neq p_0$? We calculate

$$\begin{aligned}
\mathrm{P}\left(|\widehat{p} - p_0| \geq 15/1000 - 0.03 \,|\, p = p_0\right) &= \mathrm{P}\left(\frac{|\widehat{p} - p_0|}{\sqrt{(p_0)\,(1-p_0)\,/1000}} \geq \frac{0.03 - 15/1000}{\sqrt{(0.03)(0.97)/1000}} \,|\, p = p_0\right) \\
&= \mathrm{P}\left(\frac{|\widehat{p} - p_0|}{\sqrt{(p_0)\,(1-p_0)\,/1000}} \geq 2.78064 \,|\, p = p_0\right) \\
&= 2\,\mathrm{P}\left(\frac{\widehat{p} - p_0}{\sqrt{(p_0)\,(1-p_0)\,/1000}} \geq 2.78064 \,|\, p = p_0\right) \\
&= 2\,\mathrm{P}\left(Z \geq 2.78064\right) \\
&= 2\left(1 - \Phi(2.78064)\right) \\
&= 0.005425 \\
&\approx \frac{5}{1000}.
\end{aligned}$$

4. We are testing $\mathrm{H}_0 : p_{\mathrm{NY}} = p_{\mathrm{StL}}$ versus $\mathrm{H}_a : p_{\mathrm{NY}} < p_{\mathrm{StL}}$. In the mechanics of the hypothesis test, either classical or modern, we assume the null hypothesis. The model for the difference therefore has mean 0. The standard deviation is

$$\sqrt{\frac{(53/100)(47/100)}{100} + \frac{(70/135)(65/135)}{135}}, \text{ or } 0.06588.$$

5. The emphasized words VERY LARGE tell us that the t-values that were used in the two tests were slightly greater than the corresponding z-values. The t-score calculated, the same for each student, would have been between the two t-values (in order for one to fail to reject and for the other to reject). Because $z_{0.05} = 1.645$ and $z_{0.025} = 1.96$, we see that 1.88 is the only t-score on offer that fits the bill.

6. We calculate

$$\begin{aligned}
\mathrm{P}\left(\widehat{p} \geq 0.5503 \,|\, p = 0.5000\right) &= \mathrm{P}\left(\frac{\widehat{p} - 0.5000}{\sqrt{(0.5000)(0.5000)/1003}} \geq \frac{0.5503 - 0.5000}{\sqrt{(0.5000)(0.5000)/1003}} \,|\, p = 0.5000\right) \\
&= \mathrm{P}\left(Z \geq 3.186019629\right) \\
&= 1 - \mathrm{P}\left(Z \leq 3.186019629\right) \\
&= 1 - \Phi\left(3.186019629\right) \\
&= 0.0007212.
\end{aligned}$$

7. Perhaps the wording with which the problem is stated might have been more clear, but only 8 cars in total are tested. Each of the eight cars is tested with regular and premium gasolene. The data sets on the two lines are *not* independent. The appropriate test is therefore a one-sided one-sample t-test of the differences. Thus, let X be the difference obtained by subtracting each entry on the first line from the entry below it: X : 3, 2, 3, 2, 2, -1, 2, 3. Let μ be the true mean of X. Our hypothesis test is

$\mathrm{H}_0 \,:\, \mu = 0$
$\mathrm{H}_0 \,:\, \mu > 0$

We calculate $\overline{X} = 2, \mathrm{S}^2 = 1.7143, \mathrm{S} = 1.3093$. Then

$$\begin{aligned} \mathrm{P}(\widehat{\mu} > 2 \,|\, \mu = 0) &= \mathrm{P}(\widehat{\mu} - \mu > 2 - 0 \,|\, \mu = 0) \\ &= \mathrm{P}\left(\frac{\widehat{\mu} - \mu}{\mathrm{S}/\sqrt{8}} > \frac{2}{1.3093/\sqrt{8}} \,\middle|\, \mu = 0 \right) \\ &= \mathrm{P}(\mathrm{t}_7 > 4.3205) \\ &= 0.0017. \end{aligned}$$

8. First, we calculate the number of members of the sample that come from the "Other" class: $1000 - (716 + 214 + 41 + 9)$, or 20.

   a) 2013. The expected numbers are: White, not Hispanic or Latino: 0.680×1000, or 680; Black or African American: 0.237×1000, or 237; Asian: 0.380×1000, or 38; White, Hispanic or Latino: 0.023×1000, or 23; Other: 0.022×1000, or 22. The test statistic is $(716 - 680)^2/680 + (214 - 237)^2/237 + (41 - 38)^2/38 + (9 - 23)^2/23 + (20 - 22)^2/22$, or 13.08. The p-value is $\mathrm{P}\left(\chi^2_4 \geq 13.08\right)$, or 0.011. We reject the null hypothesis: the composition of the sample does *not* match that of the county population.

   b) 2010. The expected numbers are: White, not Hispanic or Latino: 0.689×1000, or 689; Black or African American: 0.233×1000, or 233; Asian: 0.350×1000, or 35; White, Hispanic or Latino: 0.014×1000, or 14; Other: 0.029×1000, or 29. The test statistic is $(716 - 689)^2/689 + (214 - 233)^2/233 + (41 - 35)^2/35 + (9 - 14)^2/14 + (20 - 29)^2/29$, or 8.215. The p-value is $\mathrm{P}\left(\chi^2_4 \geq 8.215\right)$, or 0.084. The evidence for rejecting the null hypothesis is insufficient: the composition of the sample *does* match that of the county population.

9. The expected values are:

		Age	Group
Leaning	18-35	36-50	Over 50
Conservative	20	25	15
Moderate	70	87.5	52.5
Liberal	30	37.5	22.5

   The test statistic comes to 29.35. Because $\chi^2_{0.05,(3-1)\times(3-1)} = 9.4877$, we reject the null hypothesis (and it isn't even close): political leaning and age are related.

10. The expected values are:

	Smoker	Non-smoker
Drinker	166.38	187.62
Non-drinker	115.62	130.38

   The test statistic comes to 12.93. Because $\chi^2_{0.01,(2-1)\times(2-1)} = 6.6349$, we reject the null hypothesis (and it isn't even close): smoking and drinking were related in the 1980s.

11. We set $d_0 = 0$. By looking at the sample variances, it seems obvious that the assumption of equal variances (and hence pooling) is unwise. We calculate

$$\sqrt{\frac{\mathrm{S}^2_X}{n} + \frac{\mathrm{S}^2_Y}{m}} = \sqrt{\frac{15291.7}{5} + \frac{2484.62}{10}} = 57.505$$

and

$$\mathrm{t}_{\alpha/2,\mathrm{df}}\sqrt{\frac{\mathrm{S}_X^2}{n}+\frac{\mathrm{S}_Y^2}{m}} = \mathrm{t}_{\alpha/2,\min(5,10)-1}\times 57.505 = \mathrm{t}_{0.025,4}\times 57.505 = 2.7764\times 57.505 = 159.66.$$

Because $\left|\overline{X}-\overline{Y}-d_0\right| = |300.8-153.2| = 147.6$ is not greater than 159.66, we retain the null hypothesis that the population means are equal, a conclusion that might not have been evident from the values of the sample means.

12. We calculate

$$\sqrt{\frac{\mathrm{S}_X^2}{n}+\frac{\mathrm{S}_Y^2}{m}} = \sqrt{\frac{(65.3)^2}{32}+\frac{(89.6)^2}{32}} = 19.6$$

and

$$d_0 - \mathrm{t}_{\alpha,\mathrm{df}}\sqrt{\frac{\mathrm{S}_X^2}{n}+\frac{\mathrm{S}_Y^2}{m}} = 0-\mathrm{t}_{0.05,31}\times 19.6 = -1.6955\times 19.6 = -33.2318.$$

Because $\overline{X}-\overline{Y}-d_0 = 533.7-585.2-0 = -51.5$, which is less than -33.2318 (i.e., farther away from 0), we reject the null hypothesis that the population means are equal. There is sufficient evidence to conclude that cell phone usage while driving increases reaction time. So does old age.

13. The standard deviation $\mathrm{S_D}$ of $\widehat{\mu_\mathrm{X}}-\widehat{\mu_\mathrm{Y}}$ is given by

$$\mathrm{S_D} = \sqrt{\frac{\mathrm{S}_\mathrm{X}^2}{32}+\frac{\mathrm{S}_\mathrm{Y}^2}{32}} = \sqrt{\frac{(65.3)^2}{32}+\frac{(89.6)^2}{32}} = 19.6.$$

Let $\mu_\mathrm{D} = \mu_\mathrm{X}-\mu_\mathrm{Y}$ be the mean of $\widehat{\mu_\mathrm{X}}-\widehat{\mu_\mathrm{Y}}$. If $\mu_\mathrm{X}=\mu_\mathrm{Y}$, then $\mu_\mathrm{D}=0$ and

$$\frac{\widehat{\mu_\mathrm{X}}-\widehat{\mu_\mathrm{Y}}}{\mathrm{S_D}} = \frac{\widehat{\mu_\mathrm{X}}-\widehat{\mu_\mathrm{Y}}-\mu_\mathrm{D}}{\mathrm{S_D}} \sim \mathrm{t}_{31}.$$

The observed value of $\widehat{\mu_\mathrm{X}}-\widehat{\mu_\mathrm{Y}}$ is $533.7-585.2$, or -51.5. It follows that the p-value of the stated one-sided hypothesis test is

$$\begin{aligned}
\mathrm{P}\left(\widehat{\mu_\mathrm{X}}-\widehat{\mu_\mathrm{Y}} \le -51.5 \,|\, \mu_\mathrm{X}-\mu_\mathrm{Y}=0\right) &= \mathrm{P}\left(\frac{\widehat{\mu_\mathrm{X}}-\widehat{\mu_\mathrm{Y}}}{\mathrm{S_D}} \le \frac{-51.5}{19.6} \,\middle|\, \mu_\mathrm{X}-\mu_\mathrm{Y}=0\right) \\
&= \mathrm{P}\left(\mathrm{t}_{31} \le -2.6276\right) \\
&= \texttt{pt}(-2.6276, 31) \qquad \text{(This is R code. See below for tables.)} \\
&= 0.006626.
\end{aligned}$$

As indicated, the value 0.00626 was obtained using software. An approximation by interpolation is somewhat involved because there is no line in the table for df = 31. First, convert the left tail expressed mathematically in the third to last line to a right tail:

$$\mathrm{P}\left(\mathrm{t}_{31} \le -2.6276\right) = \mathrm{P}\left(\mathrm{t}_{31} \ge 2.6276\right).$$

Next, we will calculate $\mathrm{P}\left(\mathrm{t}_{30} \ge 2.6276\right)$ and $\mathrm{P}\left(\mathrm{t}_{40} \ge 2.6276\right)$. For each of df = 30 and df = 40, we observe that 2.6276 is between the values tabulated for $\alpha = 0.010$ and $\alpha = 0.005$. The two lines we seek can be written as

$$\alpha = \frac{0.010-0.005}{\mathrm{t}_{0.010,\mathrm{df}}-\mathrm{t}_{0.005,\mathrm{df}}}\left(t-\mathrm{t}_{0.005,\mathrm{df}}\right)+0.005.$$

For df = 30 and $t = 2.6276$ we obtain

$$\alpha = \frac{0.010 - 0.005}{2.4573 - 2.75}\,(2.6276 - 2.75) + 0.005 = 0.00709087803.$$

For df = 40 and $t = 2.6276$ we obtain

$$\alpha = \frac{0.010 - 0.005}{2.4233 - 2.7045}\,(2.6276 - 2.7045) + 0.005 = 0.00636735420.$$

Two down, one interpolation to go! For the final interpolation, we find the line segment in the line with df measured along the horizontal axis and α along the vertical axis. We interpolate between the points $(30, 0.00709087803)$ and $(40, 0.00636735420)$. The line is given by

$$\alpha = \frac{0.00636735420 - 0.00709087803}{40 - 30}\,(\mathrm{df} - 30) + 0.00709087803.$$

For df = 31 we obtain $\alpha = 0.007018525647$. Compare this value with the value 0.006626 obtained using software: the error is less than 0.0004. But it *was* quite a bit of work. On a multiple choice exam, make sure the answer choices require you to do the work before you do it. (In any event, the author does not make up exam questions that require interpolating both α and df. In this case, a little white lie—saying that the study involved 31 subjects instead of 32—would have eliminated two interpolations.)

The likelihood 0.006626 of the reported sample means being observed is less than a snowball's chance in hell. We reject the null hypothesis and conclude that cell phone usage while driving increases reaction time. So does old age.

14. The sample standard deviation is $\mathrm{S} = 21.37$ and the sample variance is $\mathrm{S}^2 = 456.68$. Also, $\chi^2_{1-\alpha,5} = \chi^2_{0.90,5} = 1.6103$, $\sigma_0 = 30$, and $\sigma_0^2 = 900$. We calculate $\sigma_0^2 \times \chi^2_{0.90,5}/5 = 289.854$. Because 456.68 is not less than 289.854, we retain the null hypothesis.

15. We calculate $\mathrm{t}_{0.05,24}\sqrt{\frac{400}{25} + \frac{625}{25}} = 1.7109 \times 6.403 = 10.955$. Because the observed value of the test statistic, namely $\widehat{\mu}_{CM} - \widehat{\mu}_{TD} = -10$, is not less than the critical value, namely -10.955, our test statistic falls outside the ciritical region and we retain the null hypothesis.

16. The p-value is

$$\begin{aligned}
\mathrm{P}\left(\mathrm{S}^2 \geq (5.651)^2 \,\middle|\, \sigma = 4.199\right) &= \mathrm{P}\left(15\,\frac{\mathrm{S}^2}{(4.199)^2} \geq 15\frac{(5.651)^2}{(4.199)^2} \,\middle|\, \sigma = 4.199\right) \\
&= \mathrm{P}\left(15\,\frac{\mathrm{S}^2}{\sigma^2} \geq 27.1675 \,\middle|\, \sigma = 4.199\right) \\
&= \mathrm{P}\left(\chi^2_{15} \geq 27.1675\right) \\
&= 0.0274.
\end{aligned}$$

The value in the last line was obtained using software. An interpolation between the two tabulated values $\chi^2_{0.025,15} = 27.4884$ and $\chi^2_{0.050,15} = 24.9958$ leads to the line segment

$$p = \frac{(0.050 - 0.025)}{(24.9958 - 27.4884)}\,(x - 27.4884) + 0.025, \qquad \text{or} \qquad p = -0.01003\,x + 0.3007.$$

Substituting $x = 27.1675$ results in 0.02822 as the approximate p-value.

17. Assuming the null hypothesis, 0.19 is the fraction of Hispanics and 0.81 is the fraction of Others. The expected numbers of Hispanics and Others are, respectively, 0.19×720, or 136.8, and 0.81×720, or 583.2. We calculate

$$\frac{(125-136.8)^2}{136.8} + \frac{\big((720-125)-583.2\big)^2}{583.2} = 1.2566.$$

The p-value is

$$\mathrm{P}\left(\chi_1^2 \geq 1.2566\right) = 0.2623,$$

which is very close to the value obtained in Exercise 1 using a different test.

18. We calculate the Poisson probability density function $f(k) = \exp(-u)\dfrac{u^k}{k!}$ for $u = 1$ and $k = 0, 1, 2, 3$. For $k = 4$, we will use $1-\big(f(0)+f(1)+f(2)+f(3)\big)$ so that the probabilities sum to 1. In R, the command `c( dpois(0:3,1), 1 - sum(dpois(0:3,1)) )` is all that is needed. The result is 0.36787944, 0.36787944, 0.18393972, 0.06131324, 0.01898816. These represent the probability of one observation from the Poisson distribution being 0, 1, 2, 3, or 4, respectively. If we multiply these probabilities by 50, then we obtain the expected number of observations, 18.3939721, 18.3939721, 9.1969860, 3.0656620, 0.9494078, of 0, 1, 2, 3, 4 (or more), respectively. The test statistic is 1.504599 and the critical value is $\chi^2_{0.05,4}$, or 9.487729. The test statistic is less than the critical value so we retain the null hypothesis: R's simulation is a good fit.

The p-value is $\mathrm{P}\left(\chi_4^2 > 1.504599\right)$, or 0.8259. This is a whopping probability. If the null hypothesis were true (i.e., R's simulation is a good one), then it would not at all be unlikely to observe a test statistic greater than or equal to 1.504599. We retain the null hypothesis.

19. The expected numbers are 50/8, 50/8, 50/8, 50/8, 50/8, 50/8, 50/8, 50/8. The observed value of the test statistic is 3.76. The critical value is $\chi^2_{0.5,7}$, or 6.345811. We retain the null hypothesis even though we have accepted a critical region so large that there is a 50-50 chance of rejecting a true null hypothesis.

20. The null hypothesis expected frquencies would be 50/4, or 12.5, for each orientation. The observed test statistic is therefore

$$\frac{(18-12.5)^2}{12.5} + \frac{(7-12.5)^2}{12.5} + \frac{(17-12.5)^2}{12.5} + \frac{(8-12.5)^2}{12.5}, \qquad \text{or } 8.08.$$

The p-value is $\mathrm{P}\left(\chi_3^2 \geq 8.08\right)$, or 0.044387. This value can be approximated by interpolating using $\chi^2_{0.025,3} = 9.3484$ and $\chi^2_{0.050,3} = 7.8147$:

$$\text{p-value} = \frac{0.050-0.025}{7.8147-9.3484}(8.08-7.8147) + 0.050 = 0.04567549064.$$

The error that results from the interpolation is fairly small: about 0.001288. In any event, the p-value is small enough so that we may reject the hypothesis that the paintings were oriented entirely at random.

21. The introvert total is 60, the extrovert total is 140, the color preference totals are: red 100, yellow 20, green 36, and blue 44. The products of the row totals times the column totals divided by the table total results in the following expectations for the eight cells:

	Red	Yellow	Green	Blue
Introvert	30	6	10.8	13.2
Extrovert	70	14	25.2	30.8

The test statistic is $(16-30)^2/30+(8-6)^2/6+(12-10.8)^2/10.8+\cdots+(20-30.8)^2/30.8$, or 23.09956710. The critical value is $\chi^2_{(2-1)\times(4-1),0.100}$, or 12.8382. Because the test statistic exceeds the critical value, we reject the null hypothesis: Personality type and color preference are *dependent*.

22. a) We use the Student-t test for the differences of means, with $d_0 = 0$. The test statistic is $\delta = \overline{\mathrm{X}} - \overline{\mathrm{Y}}$ and the critical region is

$$|\,\delta\,| > \mathrm{t}_{0.025,23}\sqrt{\frac{(30.88)^2}{24}+\frac{(28.45)^2}{27}} = 2.068658 \times 8.34926 = 17.27176.$$

Because $\delta = \overline{\mathrm{X}} - \overline{\mathrm{Y}} = 42.28 - 56.34 = -14.06$, we retain the null hypothesis.

b) We calculate

$$\begin{aligned}
\text{p-value} &= \mathrm{P}\left(\left|\overline{\mathrm{X}} - \overline{\mathrm{Y}}\right| \geq |\,42.28 - 56.34\,|\,\middle|\, \mu_{\mathrm{X}} = \mu_{\mathrm{Y}}\right) \\
&= \mathrm{P}\left(\left|\overline{\mathrm{X}} - \overline{\mathrm{Y}}\right| \geq 14.06\,\middle|\, \mu_{\mathrm{X}} = \mu_{\mathrm{Y}}\right) \\
&= \mathrm{P}\left(\left|\overline{\mathrm{X}\text{ - }\mathrm{Y}} - \mu\mathrm{X}\text{ - }\mathrm{Y}\right| \geq 14.06\,\middle|\, \mu_{\mathrm{X}} = \mu_{\mathrm{Y}}\right) \\
&= \mathrm{P}\left(\left|\frac{\overline{\mathrm{X}\text{ - }\mathrm{Y}} - \mu\mathrm{X}\text{ - }\mathrm{Y}}{\sqrt{(30.88)^2/24 + (28.45)^2/27}}\right| \geq \frac{14.06}{\sqrt{(30.88)^2/24 + (28.45)^2/27}}\,\middle|\, \mu_{\mathrm{X}} = \mu_{\mathrm{Y}}\right) \\
&\approx \mathrm{P}\left(|\mathrm{t}_{23}| \geq 1.683982\right) \\
&= 2\,\mathrm{P}\left(\mathrm{t}_{23} \geq 1.683982\right) \\
&= 2\,(0.05285511) \\
&= 0.1057102.
\end{aligned}$$

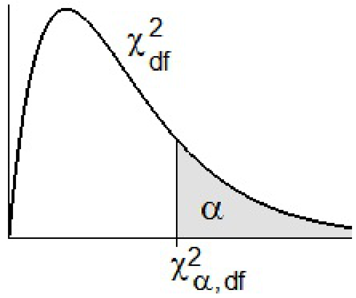
c) Here we must calculate:

$$\mathrm{S}_p = \sqrt{\frac{23\,(30.88)^2 + 26\,(28.45)^2}{24 + 27 - 2}} = 29.61545.$$

We use the Student-t test for the differences of means, with $d_0 = 0, \mathrm{df} = 24 + 27 - 2$, and pooling. The test statistic is $\delta = \overline{\mathrm{X}} - \overline{\mathrm{Y}}$ and the critical region is

$$|\,\delta\,| > \mathrm{t}_{0.025,49}\,\mathrm{S}_p\sqrt{\frac{1}{24}+\frac{1}{27}} = 2.009575 \times 29.61545 \times 0.2805418 = 16.6963.$$

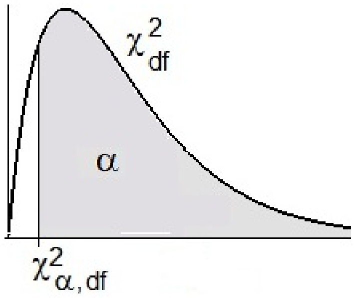
Because $\delta = \overline{\mathrm{X}} - \overline{\mathrm{Y}} = 42.28 - 56.34 = -14.06$, we retain the null hypothesis $\mu_{\mathrm{X}} = \mu_{\mathrm{Y}}$. In the literature, statistical tests for this insignificant improvement were not presented. The observations were simply described qualitatively as mild cognitive improvement over the control group.



Values of $\chi^2_{\alpha,df}$ $\quad P(\chi^2_{df} \geq \chi^2_{\alpha,df}) = \alpha$

df \ α	0.005	0.010	0.025	0.050	0.100	0.200	0.250	0.300	0.400	0.500
1	7.8794	6.6349	5.0239	3.8415	2.7055	1.6424	1.3233	1.0742	0.7083	0.4549
2	10.5966	9.2103	7.3778	5.9915	4.6052	3.2189	2.7726	2.4079	1.8326	1.3863
3	12.8382	11.3449	9.3484	7.8147	6.2514	4.6416	4.1083	3.6649	2.9462	2.3660
4	14.8603	13.2767	11.1433	9.4877	7.7794	5.9886	5.3853	4.8784	4.0446	3.3567
5	16.7496	15.0863	12.8325	11.0705	9.2364	7.2893	6.6257	6.0644	5.1319	4.3515
6	18.5476	16.8119	14.4494	12.5916	10.6446	8.5581	7.8408	7.2311	6.2108	5.3481
7	20.2777	18.4753	16.0128	14.0671	12.0170	9.8032	9.0371	8.3834	7.2832	6.3458
8	21.9550	20.0902	17.5345	15.5073	13.3616	11.0301	10.2189	9.5245	8.3505	7.3441
9	23.5894	21.6660	19.0228	16.9190	14.6837	12.2421	11.3888	10.6564	9.4136	8.3428
10	25.1882	23.2093	20.4832	18.3070	15.9872	13.4420	12.5489	11.7807	10.4732	9.3418
11	26.7568	24.7250	21.9200	19.6751	17.2750	14.6314	13.7007	12.8987	11.5298	10.3410
12	28.2995	26.2170	23.3367	21.0261	18.5493	15.8120	14.8454	14.0111	12.5838	11.3403
13	29.8195	27.6882	24.7356	22.3620	19.8119	16.9848	15.9839	15.1187	13.6356	12.3398
14	31.3193	29.1412	26.1189	23.6848	21.0641	18.1508	17.1169	16.2221	14.6853	13.3393
15	32.8013	30.5779	27.4884	24.9958	22.3071	19.3107	18.2451	17.3217	15.7332	14.3389
16	34.2672	31.9999	28.8454	26.2962	23.5418	20.4651	19.3689	18.4179	16.7795	15.3385
17	35.7185	33.4087	30.1910	27.5871	24.7690	21.6146	20.4887	19.5110	17.8244	16.3382
18	37.1565	34.8053	31.5264	28.8693	25.9894	22.7595	21.6049	20.6014	18.8679	17.3379
19	38.5823	36.1909	32.8523	30.1435	27.2036	23.9004	22.7178	21.6891	19.9102	18.3377
20	39.9968	37.5662	34.1696	31.4104	28.4120	25.0375	23.8277	22.7745	20.9514	19.3374
21	41.4011	38.9322	35.4789	32.6706	29.6151	26.1711	24.9348	23.8578	21.9915	20.3372
22	42.7957	40.2894	36.7807	33.9244	30.8133	27.3015	26.0393	24.9390	23.0307	21.3370
23	44.1813	41.6384	38.0756	35.1725	32.0069	28.4288	27.1413	26.0184	24.0689	22.3369
24	45.5585	42.9798	39.3641	36.4150	33.1962	29.5533	28.2412	27.0960	25.1063	23.3367
25	46.9279	44.3141	40.6465	37.6525	34.3816	30.6752	29.3389	28.1719	26.1430	24.3366
30	53.6720	50.8922	46.9792	43.7730	40.2560	36.2502	34.7997	33.5302	31.3159	29.3360
40	66.7660	63.6907	59.3417	55.7585	51.8051	47.2685	45.6160	44.1649	41.6222	39.3353
50	79.4900	76.1539	71.4202	67.5048	63.1671	58.1638	56.3336	54.7228	51.8916	49.3349
60	91.9517	88.3794	83.2977	79.0819	74.3970	68.9721	66.9815	65.2265	62.1348	59.3347
70	104.2149	100.4252	95.0232	90.5312	85.5270	79.7146	77.5767	75.6893	72.3583	69.3345
80	116.3211	112.3288	106.6286	101.8795	96.5782	90.4053	88.1303	86.1197	82.5663	79.3343
90	128.2989	124.1163	118.1359	113.1453	107.5650	101.0537	98.6499	96.5238	92.7614	89.3342
100	140.1695	135.8067	129.5612	124.3421	118.4980	111.6667	109.1412	106.9058	102.9459	99.3341

Chi-Squared Values—Right Tails.



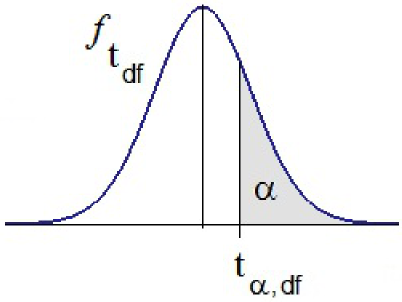
Values of $\chi^2_{\alpha,df}$ $\quad P(\chi^2_{df} \geq \chi^2_{\alpha,df}) = \alpha$

df \ α	0.600	0.700	0.750	0.800	0.900	0.950	0.975	0.990	0.995
1	0.2750	0.1485	0.1015	0.0642	0.0158	0.0039	0.0010	0.0002	0.0000
2	1.0217	0.7133	0.5754	0.4463	0.2107	0.1026	0.0506	0.0201	0.0100
3	1.8692	1.4237	1.2125	1.0052	0.5844	0.3518	0.2158	0.1148	0.0717
4	2.7528	2.1947	1.9226	1.6488	1.0636	0.7107	0.4844	0.2971	0.2070
5	3.6555	2.9999	2.6746	2.3425	1.6103	1.1455	0.8312	0.5543	0.4117
6	4.5702	3.8276	3.4546	3.0701	2.2041	1.6354	1.2373	0.8721	0.6757
7	5.4932	4.6713	4.2549	3.8223	2.8331	2.1673	1.6899	1.2390	0.9893
8	6.4226	5.5274	5.0706	4.5936	3.4895	2.7326	2.1797	1.6465	1.3444
9	7.3570	6.3933	5.8988	5.3801	4.1682	3.3251	2.7004	2.0879	1.7349
10	8.2955	7.2672	6.7372	6.1791	4.8652	3.9403	3.2470	2.5582	2.1559
11	9.2373	8.1479	7.5841	6.9887	5.5778	4.5748	3.8157	3.0535	2.6032
12	10.1820	9.0343	8.4384	7.8073	6.3038	5.2260	4.4038	3.5706	3.0738
13	11.1291	9.9257	9.2991	8.6339	7.0415	5.8919	5.0088	4.1069	3.5650
14	12.0785	10.8215	10.1653	9.4673	7.7895	6.5706	5.6287	4.6604	4.0747
15	13.0297	11.7212	11.0365	10.3070	8.5468	7.2609	6.2621	5.2293	4.6009
16	13.9827	12.6243	11.9122	11.1521	9.3122	7.9616	6.9077	5.8122	5.1422
17	14.9373	13.5307	12.7919	12.0023	10.0852	8.6718	7.5642	6.4078	5.6972
18	15.8932	14.4399	13.6753	12.8570	10.8649	9.3905	8.2307	7.0149	6.2648
19	16.8504	15.3517	14.5620	13.7158	11.6509	10.1170	8.9065	7.6327	6.8440
20	17.8088	16.2659	15.4518	14.5784	12.4426	10.8508	9.5908	8.2604	7.4338
21	18.7683	17.1823	16.3444	15.4446	13.2396	11.5913	10.2829	8.8972	8.0337
22	19.7288	18.1007	17.2396	16.3140	14.0415	12.3380	10.9823	9.5425	8.6427
23	20.6902	19.0211	18.1373	17.1865	14.8480	13.0905	11.6886	10.1957	9.2604
24	21.6525	19.9432	19.0373	18.0618	15.6587	13.8484	12.4012	10.8564	9.8862
25	22.6156	20.8670	19.9393	18.9398	16.4734	14.6114	13.1197	11.5240	10.5197
30	27.4416	25.5078	24.4776	23.3641	20.5992	18.4927	16.7908	14.9535	13.7867
40	37.1340	34.8719	33.6603	32.3450	29.0505	26.5093	24.4330	22.1643	20.7065
50	46.8638	44.3133	42.9421	41.4492	37.6886	34.7643	32.3574	29.7067	27.9907
60	56.6200	53.8091	52.2938	50.6406	46.4589	43.1880	40.4817	37.4849	35.5345
70	66.3961	63.3460	61.6983	59.8978	55.3289	51.7393	48.7576	45.4417	43.2752
80	76.1879	72.9153	71.1445	69.2069	64.2778	60.3915	57.1532	53.5401	51.1719
90	85.9925	82.5111	80.6247	78.5584	73.2911	69.1260	65.6466	61.7541	59.1963
100	95.8078	92.1289	90.1332	87.9453	82.3581	77.9295	74.2219	70.0649	67.3276

Chi-Squared Values—Central Hump + Right Tails.

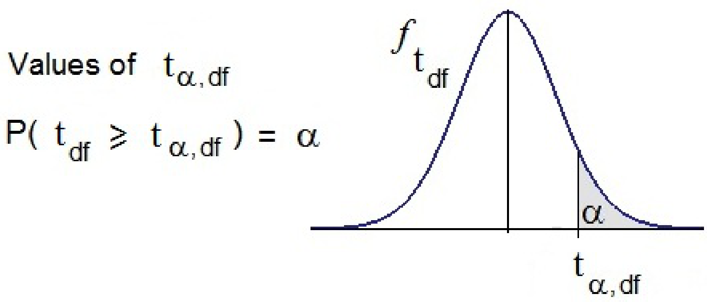
Values of $t_{\alpha,df}$

$P(t_{df} \geqslant t_{\alpha,df}) = \alpha$



df \ α	.450	.400	.350	.300	.250	.200
1	.1584	.3249	.5095	.7265	1.0000	1.3764
2	.1421	.2887	.4447	.6172	.8165	1.0607
3	.1366	.2767	.4242	.5844	.7649	.9785
4	.1338	.2707	.4142	.5686	.7407	.9410
5	.1322	.2672	.4082	.5594	.7267	.9195
6	.1311	.2648	.4043	.5534	.7176	.9057
7	.1303	.2632	.4015	.5491	.7111	.8960
8	.1297	.2619	.3995	.5459	.7064	.8889
9	.1293	.2610	.3979	.5435	.7027	.8834
10	.1289	.2602	.3966	.5415	.6998	.8791
11	.1286	.2596	.3956	.5399	.6974	.8755
12	.1283	.2590	.3947	.5386	.6955	.8726
13	.1281	.2586	.3940	.5375	.6938	.8702
14	.1280	.2582	.3933	.5366	.6924	.8681
15	.1278	.2579	.3928	.5357	.6912	.8662
16	.1277	.2576	.3923	.5350	.6901	.8647
17	.1276	.2573	.3919	.5344	.6892	.8633
18	.1274	.2571	.3915	.5338	.6884	.8620
19	.1274	.2569	.3912	.5333	.6876	.8610
20	.1273	.2567	.3909	.5329	.6870	.8600
21	.1272	.2566	.3906	.5325	.6864	.8591
22	.1271	.2564	.3904	.5321	.6858	.8583
23	.1271	.2563	.3902	.5317	.6853	.8575
24	.1270	.2562	.3900	.5314	.6848	.8569
25	.1269	.2561	.3898	.5312	.6844	.8562
26	.1269	.2560	.3896	.5309	.6840	.8557
27	.1268	.2559	.3894	.5306	.6837	.8551
28	.1268	.2558	.3893	.5304	.6834	.8546
29	.1268	.2557	.3892	.5302	.6830	.8542
30	.1267	.2556	.3890	.5300	.6828	.8538
40	.1265	.2550	.3881	.5286	.6807	.8507
50	.1263	.2547	.3875	.5278	.6794	.8489
60	.1262	.2545	.3872	.5272	.6786	.8477
70	.1261	.2543	.3869	.5268	.6780	.8468
80	.1261	.2542	.3867	.5265	.6776	.8461
90	.1260	.2541	.3866	.5263	.6772	.8456
100	.1260	.2540	.3864	.5261	.6770	.8452

Student-t Values—Right Tails $\alpha = 0.45,\ 0.40,\ 0.35,\ 0.30, 0.25, 0.20.$



df \ α	.150	.100	.050	.025	.010	.005
1	1.9626	3.0777	6.3138	12.7062	31.8205	63.6567
2	1.3862	1.8856	2.9200	4.3027	6.9646	9.9248
3	1.2498	1.6377	2.3534	3.1824	4.5407	5.8409
4	1.1896	1.5332	2.1318	2.7764	3.7469	4.6041
5	1.1558	1.4759	2.0150	2.5706	3.3649	4.0321
6	1.1342	1.4398	1.9432	2.4469	3.1427	3.7074
7	1.1192	1.4149	1.8946	2.3646	2.9980	3.4995
8	1.1081	1.3968	1.8595	2.3060	2.8965	3.3554
9	1.0997	1.3830	1.8331	2.2622	2.8214	3.2498
10	1.0931	1.3722	1.8125	2.2281	2.7638	3.1693
11	1.0877	1.3634	1.7959	2.2010	2.7181	3.1058
12	1.0832	1.3562	1.7823	2.1788	2.6810	3.0545
13	1.0795	1.3502	1.7709	2.1604	2.6503	3.0123
14	1.0763	1.3450	1.7613	2.1448	2.6245	2.9768
15	1.0735	1.3406	1.7531	2.1314	2.6025	2.9467
16	1.0711	1.3368	1.7459	2.1199	2.5835	2.9208
17	1.0690	1.3334	1.7396	2.1098	2.5669	2.8982
18	1.0672	1.3304	1.7341	2.1009	2.5524	2.8784
19	1.0655	1.3277	1.7291	2.0930	2.5395	2.8609
20	1.0640	1.3253	1.7247	2.0860	2.5280	2.8453
21	1.0627	1.3232	1.7207	2.0796	2.5176	2.8314
22	1.0614	1.3212	1.7171	2.0739	2.5083	2.8188
23	1.0603	1.3195	1.7139	2.0687	2.4999	2.8073
24	1.0593	1.3178	1.7109	2.0639	2.4922	2.7969
25	1.0584	1.3163	1.7081	2.0595	2.4851	2.7874
26	1.0575	1.3150	1.7056	2.0555	2.4786	2.7787
27	1.0567	1.3137	1.7033	2.0518	2.4727	2.7707
28	1.0560	1.3125	1.7011	2.0484	2.4671	2.7633
29	1.0553	1.3114	1.6991	2.0452	2.4620	2.7564
30	1.0547	1.3104	1.6973	2.0423	2.4573	2.7500
40	1.0500	1.3031	1.6839	2.0211	2.4233	2.7045
50	1.0473	1.2987	1.6759	2.0086	2.4033	2.6778
60	1.0455	1.2958	1.6706	2.0003	2.3901	2.6603
70	1.0442	1.2938	1.6669	1.9944	2.3808	2.6479
80	1.0432	1.2922	1.6641	1.9901	2.3739	2.6387
90	1.0424	1.2910	1.6620	1.9867	2.3685	2.6316
100	1.0418	1.2901	1.6602	1.9840	2.3642	2.6259

Student-t Values—Right Tails $\alpha = 0.15,\ 0.10,\ 0.05,\ 0.025, 0.010, 0.005.$