Probability and Statistics (part 4) : Hypothesis Testing (by Evan Dummit, 2020, v. 1.01)

Contents

4 Hypothesis Testing

In this chapter, we discuss statistical hypothesis testing, which (broadly speaking) is the process of using statistical analysis to deduce information about the plausibility or implausibility of a given hypothesis. We begin with a broad overview of the basics of hypothesis testing and terminology, and motivate our general framework of hypothesis testing using a number of examples.

Next, we discuss z tests, which are used to make inferences about normally-distributed variables whose standard deviation is known, and discuss the connections between hypothesis testing and our previous development of confidence intervals. We apply these ideas to extend our discussion to situations involving the binomial distribution and other distributions that are approximately normal, and then treat topics related to errors in hypothesis testing.

We then treat t tests, which are used to make inferences about normally-distributed variables whose standard deviation is unknown. We discuss the associated t distributions and then describe how to perform t tests of various flavors, which allow us to make inferences about sampling data that is approximately normally distributed using only information derived from the sampling data itself. Finally, we discuss the χ^2 distribution and the χ^2 tests for independence and goodness of fit, which allow us to assess the quality of statistical models.

4.1 Principles of Hypothesis Testing

- In the last chapter, we discussed methods for estimating parameters, and for constructing confidence intervals that quantify the precision of the estimate.
  - In many cases, parameter estimations can provide the basic framework to decide the plausibility of a particular hypothesis.
  - For example, to decide how plausible it is that a given coin truly is fair, we can flip the coin several times, examine the likelihood of obtaining that given sequence of outcomes, construct an estimate for the true probability of obtaining heads and associated confidence intervals, and then decide based on the position of the confidence interval whether it is reasonable to believe the coin is fair.
  - As another example, to decide how plausible it is that the average part size in a manufacturing lot truly is equal to the expected standard, we can measure the sizes of a sample from that lot, construct an estimate and confidence intervals for the average size of the lot from the sample data, and then decide whether it is reasonable to believe that the average part size is within the desired tolerance.
  - We can use a similar procedure to do things like decide whether one class's average on an exam was higher than another (by studying the difference in the class average), decide whether a ballot measure has a specified level of support (by conducting a random poll and constructing an appropriate confidence interval), or decide which of two medical interventions yields better outcomes (by comparing the average outcomes from two appropriate samples).
- However, in most of these situations, we are seeking a binary decision about a hypothesis: namely, whether or not it is justified by the available evidence.
  - The procedure of deciding whether or not a given hypothesis is supported by statistical evidence is known as <u>statistical hypothesis testing</u>.
  - Our goal is to describe how to use our analysis of random variables and their underlying distributions to perform hypothesis testing.

4.1.1 Null and Alternative Hypotheses, p-Values, and Decision Rules

- If we are making a binary decision, our first step is to explicitly identify the two possible results.
  - <u>Example</u>: "The coin is fair" versus "The coin is not fair".
  - <u>Example</u>: "The coin has probability 2/3 of landing heads" versus "The coin does not have probability 2/3 of landing heads".
  - <u>Example</u>: "Class 1 has the same average exam score as Class 2" versus "Class 1 does not have the same average exam score as Class 2".
  - <u>Example</u>: "Treatment A is more effective than a placebo" versus "Treatment A is not more effective than a placebo".
  - We must then test a hypothesis using a statistical model. In order to do this, we must formulate the hypothesis in a way that allows us to analyze the underlying statistical distribution.
- In the four examples above, only one of the two possible hypotheses provides grounds for a statistical model:
  - <u>Example</u>: "The coin is fair" provides us a model that we can analyze; namely, the distribution of the number of heads obtained by flipping a fair coin. The other hypothesis, "The coin is not fair" does not provide us with such a model, since the probability of heads could be one of many possible values, each of which would give a different distribution.
  - <u>Example</u>: "The coin has probability 2/3 of landing heads" likewise provides us a model we can analyze, unlike the hypothesis "The coin does not have probability 2/3 of landing heads".
  - <u>Example</u>: "Class 1 has the same average exam score as Class 2" provides us a model we can analyze, at least, under the presumption that the full set of exam scores have some underlying known distribution, such as a normal distribution, possibly with unknown parameters. Under the same presumptions, however, the other hypothesis "Class 1 does not have the same average exam score as Class 2" does not give us an underlying model, since there are many ways in which the average scores could be different.

  - Example: "Treatment A is not more effective than a placebo" provides us a model we can analyze (making the same sorts of presumptions as above, that the full set of treatment results has some known type of distribution but with unknown parameters). However, we do have to discard the possibility that Treatment A is actually less effective than a placebo in order to obtain a model. We would want to rephrase this hypothesis as "Treatment A is equally effective with a placebo" in order to test it using the model.

- Here is some more specific terminology regarding the hypotheses we wish to test.
  - The type of hypothesis we are testing in each case is a null hypothesis, which typically states that there is no difference or relationship between the groups being examined, and that any observed results are due purely to chance.
  - The other hypothesis is the alternative hypothesis, which typically asserts that there is some difference or relationship between the groups being examined.
  - The alternative hypothesis generally captures the notion that "something is occurring", while the null hypothesis generally captures the notion that "nothing is occurring". (Of course, there are occasional exceptions, such as the situation where we are postulating that the heads probability of the coin is a specific number, which will serve as our null hypothesis.)
- Because of the structure of our statistical approach, we are only able to test the null hypothesis directly.
- Our choices are either to reject the null hypothesis in favor of the alternative hypothesis (in the event that our analysis indicates that the observed data set was too unlikely to arise by random chance) or to fail to reject the null hypothesis (in the event that the data set could plausibly have arisen by chance).
  - Note that we do not actually "accept" any given hypothesis: we either reject the null hypothesis, or fail to reject the null hypothesis.
  - The reason for this (pedantic, but important) piece of terminology is that when we perform a statistical test that does not give strong evidence in favor of the alternative hypothesis, that does not constitute actual proof that the null hypothesis is true (merely some evidence, however strong it may be).
  - The principle is that, although we may have gathered some evidence that suggests the null hypothesis may be true, we have not actually proven that there is no relationship between the given variables. It is always possible that there is indeed some relationship between the variables we have not uncovered, no matter how much sampling data we may collect.
  - Likewise, rejecting the null hypothesis does not mean that we accept the alternative hypothesis: it merely means that there is strong evidence that the null hypothesis is false. It is always possible that the data set was unusual (merely because of random variation) and that there actually is no relationship between the given variables.
- With the hypothesis tests we will study, the null hypothesis H_0 will be of the form "The parameter equals a specific value".
  - We can recast all of our examples into this format.
  - Example: "The probability of obtaining heads when flipping a coin is 1/2".
  - Example: "The probability of obtaining heads when flipping a coin is 2/3".
  - Example: "The difference in the average scores of Class 1 and Class 2 is zero".
  - Example: "The difference between the average outcome using Treatment A and the average outcome using a placebo is zero".
- The alternative hypothesis H_a may then take one of several possible forms.
  - Two-sided: "The parameter is not equal to the given value".
  - One-sided: "The parameter is less than the given value" or "The parameter is greater than the given value".

- The two-sided alternative hypothesis is so named because it includes both possibilities listed for the one-sided hypotheses.
- Example: "The probability of obtaining heads when flipping a coin is not 1/2" is two-sided.
- Example: "The probability of obtaining heads when flipping a coin is not 2/3" is also two-sided.
- Example: "The difference in the average scores of Class 1 and Class 2 is not zero" is two-sided, while "The difference in the average scores of Class 1 and Class 2 is positive" is one-sided.
- Example: "The average outcome of using Treatment A is better than the average outcome using a placebo" is one-sided.
- The specific nature of the alternative hypothesis will depend on the situation. As in the third example, there may be several reasonable options to consider, depending on what result we want to study.

- Example: We wish to test whether a particular coin is fair, which we do by flipping the coin 100 times and recording the proportion p of heads obtained. Give the null and alternative hypotheses for this test.
  - The null hypothesis is H_0: $p = 0.5$, since this represents the result that the coin is fair.
  - The alternative hypothesis is H_a: $p \neq 0.5$, since this represents the result that the coin is not fair. Here, the alternative hypothesis is two-sided.

- Example: We wish to test whether the exams given to two classes were equivalent, which we do by comparing the average scores μ_A and μ_B in the two classes. Give the null and alternative hypotheses for this test.
  - The null hypothesis is H_0: $\mu_A = \mu_B$, since this represents the result that the averages were equal.
  - The alternative hypothesis is H_a: $\mu_A \neq \mu_B$, since this represents the result that the averages were not equal. Here, the alternative hypothesis is two-sided.

- Example: We wish to test whether the exam given to class A was easier than the exam given to class B, which we do by comparing the average scores μ_A and μ_B in the two classes. Give the null and alternative hypotheses for this test.
  - The null hypothesis is H_0: $\mu_A = \mu_B$, since this represents the result that the averages were equal.
  - The alternative hypothesis is H_a: $\mu_A > \mu_B$, since this represents the result that the average in class A is higher than the average in class B (which would correspond to an easier exam). Here, the alternative hypothesis is one-sided.

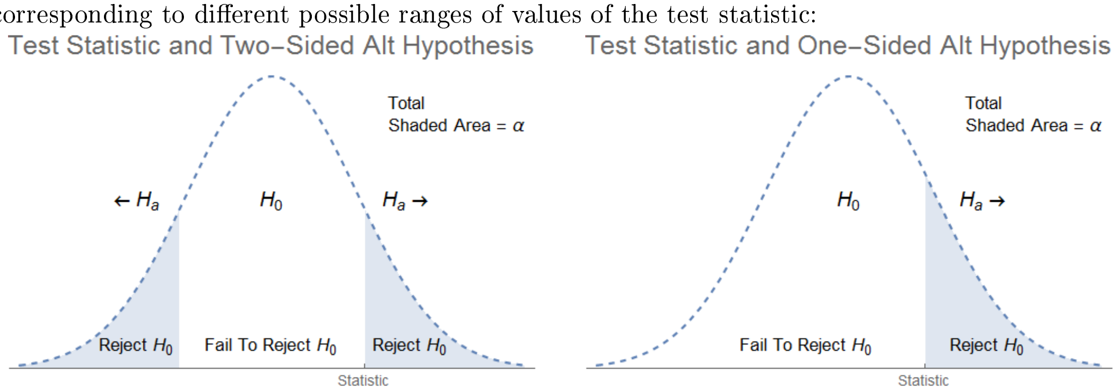
- Example: We wish to test whether a particular baseball player performs better in the playoffs than during the regular season, which we do by comparing the player's hitting percentage h_r during regular-season games to their hitting percentage h_p during playoff games. Give the null and alternative hypotheses for this test.
  - The null hypothesis is H_0: $h_r = h_p$, since this represents the result that the hitting percentages do not differ.
  - The alternative hypothesis is H_a: $h_r < h_p$, since this represents the result that the playoff percentage is better than the regular-season percentage. Here, the alternative hypothesis is one-sided.

- Once we have properly formulated the null and alternative hypotheses, we can set up a hypothesis test to decide on the reasonableness of rejecting the null hypothesis.
  - Ideally, we would like to assess how likely it is to obtain the data we observed if the null hypothesis were true.
  - We will compute a test statistic based on the data (this will usually be an estimator for a particular unknown parameter, such as the mean of the distribution), and then assess the likelihood of obtaining this test statistic by sampling the distribution in the situation where the null hypothesis is true.
  - In other words, we are using the projected distribution of the test statistic to calculate the likelihood that any apparent deviation from the null hypothesis could have occurred merely by chance.
  - In situations where the projected test statistic has a discrete distribution, we could, in principle, compute this exact probability. However, for continuous distributions, the likelihood of observing any particular data sample will always be zero.

- What we will do, as an approximate replacement, is instead compute the probability of obtaining a test statistic at least as extreme as the one we observed. This probability is called the p-value of the sample.
- Note that the definition of "extreme" will depend on the nature of the alternative hypothesis: if H_a is two-sided, then a deviation from the null hypothesis in either direction will be considered "extreme", whereas if H_a is one-sided, we only care about deviation from the null hypothesis in the corresponding direction of H_a.
- We then decide, based on the p-value, whether we believe this deviation in the test statistic plausibly occurred by chance.

• To decide whether to reject the null hypothesis, we adopt a decision rule of the following nature: we select a significance level α (often $\alpha = 0.1$, 0.05, or 0.01, but we could choose any value) and decide whether the p-value of the sample statistic satisfies $p < \alpha$ or $p \geq \alpha$.

- If $p < \alpha$, then we view the data as sufficiently unlikely to have occurred by chance: we reject the null hypothesis in favor of the alternative hypothesis and say that the evidence against the null hypothesis is statistically significant.
- If $p \geq \alpha$, then we view as plausible that the data could have occurred by chance: we fail to reject the null hypothesis and say that the evidence against the null hypothesis is not statistically significant.
- If we plot the projected distribution of values of the test statistic, then we can view these two situations as corresponding to different possible ranges of values of the test statistic:



- For a two-sided alternative hypothesis, there are two regions in which we would reject the null hypothesis: one where the test statistic is too high and the other where it is too low. Together, the total area of these regions is α.
- For a one-sided alternative hypothesis, there is a single region in which we would reject the null hypothesis, corresponding to a test statistic that is sufficiently far in the direction of the alternative hypothesis. The total area of this region is α.

• Historically, when it was difficult or time-consuming to compute exact p-values even for simple distributions like the normal distribution, the testing procedure above was phrased in terms of "critical values" or a "critical range", outside of which the null hypothesis would be rejected.

- Since we are now able to compute with arbitrary accuracy the exact distributions for the situations we will discuss, we will primarily work with explicit p-values and compare them to our significance level, rather than computing critical values for the test statistic.

• To summarize, we will adopt the following general procedure for our hypothesis tests:

1. Identify the null and alternative hypotheses for the given problem, and select a significance level α.
2. Identify the most appropriate test statistic and its distribution according to the null hypothesis (usually, this is an average or occasionally a sum of the given data values) including all relevant parameters.
3. Calculate the p-value: the probability that a value of the test statistic would have a value at least as extreme as the value observed.
4. Determine whether the p-value is less than the significance level α (reject the null hypothesis) or greater than or equal to the significance level α (fail to reject the null hypothesis).
   - Alternatively, in situations where the p-value may be difficult to calculate exactly, we may instead calculate a critical value, or critical range, beyond which the null hypothesis is rejected.

4.1.2 One-Sample and Two-Sample z Tests

- We now discuss concrete examples of hypothesis testing in one of the simplest possible situations: testing whether a normally-distributed quantity with a known standard deviation has a particular mean, which is known as a one-sample z test after the letter z traditionally used for normally-distributed quantities.
  - First, we must identify the appropriate null and alternative hypotheses and select a significance level α.
  - We will use the test statistic $\hat{\mu}$, the sample mean, since this is the minimum-variance unbiased estimator for the population mean. Under the assumption that H_0 is true, the test statistic is normally distributed with mean μ (the true mean postulated by the null hypothesis) and standard deviation σ (which we must be given).
    - In some cases we may prefer to work with a "normalized" test statistic given instead by $\dfrac{\hat{\mu}-\mu}{\sigma/\sqrt{n}}$, whose distribution follows the standard normal distribution of mean 0 and standard deviation 1. This corresponds to taking the test statistic to be the z-score.
  - If the hypotheses are $H_0 : \mu = c$ and $H_a : \mu > c$, then the p-value is $P(N_{\mu,\sigma} \geq z)$.
  - If the hypotheses are $H_0 : \mu = c$ and $H_a : \mu < c$, then the p-value is $P(N_{\mu,\sigma} \leq z)$.
  - If the hypotheses are $H_0 : \mu = c$ and $H_a : \mu \neq c$, then the p-value is $P(|N_{\mu,\sigma} - \mu| \geq |z-\mu|) = \begin{cases} 2P(N_{\mu,\sigma} \geq z) & \text{if } z \geq \mu \\ 2P(N_{\mu,\sigma} \leq z) & \text{if } z < \mu \end{cases}$.
  - In each case, we are simply calculating the probability that the normally-distributed random variable $N_{\mu,\sigma}$ will take a value further from the hypothesized mean μ (in the direction of the alternative hypothesis, as applicable) than the observed test statistic z.

- Example: The production of an assembly line is normally distributed, with mean 180 widgets and standard deviation 10 widgets for a 9-hour shift. The company wishes to test to see whether a new manufacturing technique is more productive. The new method is used for a 9-hour shift and produces a total of 197 widgets. Assuming that the standard deviation for the new method is also 10 widgets for a 9-hour shift, state the null and alternative hypotheses, identify the test statistic and its distribution, calculate the p-value, and test the claim at the 10%, 5%, and 1% levels of significance.
  - If μ represents the true mean of the new manufacturing process, then we want to decide whether $\mu > 180$ or not.
  - Thus, we have the null hypothesis $H_0 : \mu = 180$ and the alternative hypothesis $H_a : \mu > 180$.
  - Our test statistic is $z = 197$ widgets.
  - By assumption, the number of widgets on a shift is normally distributed with standard deviation 10 widgets.
  - Thus, because our alternative hypothesis is $H_a : \mu > 180$, the p-value is the probability $P(N_{180,10} \geq 197)$ that we would observe a result at least as extreme as the one we found, if the null hypothesis were actually true.
  - Using a normal cdf calculator, or a table of z-values, we can find $P(N_{180,10} \geq 197) = P(N_{0,1} \geq 1.7) = \boxed{0.04457}$. This is the p-value for our hypothesis test.
  - At the 10% level of significance ($\alpha = 0.10$), we have $p < \alpha$, and thus the result is statistically significant, so we would reject the null hypothesis in this case.
  - At the 5% level of significance ($\alpha = 0.05$), we have $p < \alpha$, and thus the result is statistically significant, so we would reject the null hypothesis in this case.
  - At the 1% level of significance ($\alpha = 0.01$), we have $p > \alpha$, and thus the result is not statistically significant, so we would fail to reject the null hypothesis in this case.

- Example: The Bad Timing Institute wants to raise awareness of the issue of improperly-set wristwatches. They believe that the average person's watch is set correctly, but with a standard deviation of 20 seconds. They poll 6 people, whose watches have errors of -39 seconds, $+14$ seconds, -21 seconds, -23 seconds, $+25$ seconds, and -31 seconds (positive values are watches that run fast while negative values are watches that run slow). Test at the 10% significance level the Bad Timing Institute's hypothesis that the true mean error μ is 0 seconds, if (i) the Institute is concerned about errors of any kind, and (ii) the Institute is only concerned about errors that make people late.

- For (i), our hypotheses are H_0: $\mu = 0$ and $H_a : \mu \neq 0$, since the Institute cares about errors in any direction.
- Our test statistic is the average error, which is $-75/6 = -12.5$ seconds.
- The distribution of the test statistic, under the null hypothesis, is normal with mean 0 seconds and standard deviation $30/\sqrt{6} = 8.1650$ seconds.
- Thus, because our alternative hypothesis is $H_a : \mu \neq 0$ (which is two-sided), the p-value is $2 \cdot P(N_{0,8.1650} \leq -12.5) = 2 \cdot P(N_{0,1} \leq -1.5309) = 0.1258$.
- Since the p-value is greater than $\alpha = 0.10$, it is not statistically significant at the 10% significance level, and we accordingly fail to reject the null hypothesis.
- For (ii), our hypotheses are H_0: $\mu = 0$ and $H_a : \mu_A < 0$, since the Institute cares about errors only in the direction that make people late (i.e., the negative direction).
- Our test statistic and distribution are the same as above.
- But now, because our alternative hypothesis is $H_a : \mu < 0$ (which is one-sided), the p-value is $P(N_{0,8.1650} \leq -12.5) = P(N_{0,1} \leq -1.5309) = 0.0629$.
- Since the p-value is less than $\alpha = 0.10$, it is statistically significant at the 10% significance level, and we accordingly reject the null hypothesis here.

• The example above illustrates that the decision about whether to reject the null hypothesis at a given significance level can depend on the choice of alternative hypothesis, even when the underlying data and test statistic are exactly the same.

- Ultimately, the decision about using a one-sided alternative hypothesis versus a two-sided alternative hypothesis depends on the context of the problem and the precise nature of the question being investigated.
- In situations where we are specifically trying to decide whether one category is better than another, we want to use a one-sided alternative hypothesis. In situations where we are trying to decide whether two categories are merely different, we want to use a two-sided alternative hypothesis.
- The statistical test itself cannot make this determination: it is entirely a matter of what question we are trying to answer using the observed data.
- This particular ambiguity also demonstrates one reason it is poor form simply to state the result of a test ("significant"/ "reject the null hypothesis" versus "not significant" / "fail to reject the null hypothesis") without clearly stating the hypotheses and giving the actual p-value.
- Here, even with the two-sided alternative hypothesis, we can see that $p = 0.0629$ is not that far below the (rather arbitrarily chosen) threshold value $\alpha = 0.10$, which is why there is a difference in the results of the one-sided test and the two-sided test. If the p-value had been much smaller than α, the factor of 2 would not have affected the statistical significance.

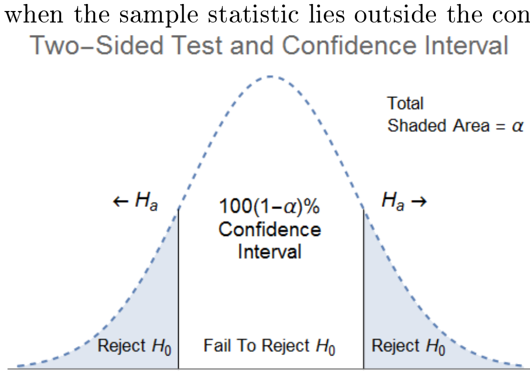
• In some situations, we want to compare two quantities to decide whether one of them is larger than the other.

- In situations where both quantities are normally distributed and independent, we can make this decision by analyzing the difference between the two quantities, which will also be normally distributed.
- We can then apply the same decision procedures described above to test the appropriate null hypothesis about the value of the difference of the quantities.
- Because there are now two samples involved and we are studying the properties of a normally distributed test statistic z, this method is referred to as a two-sample z-test.

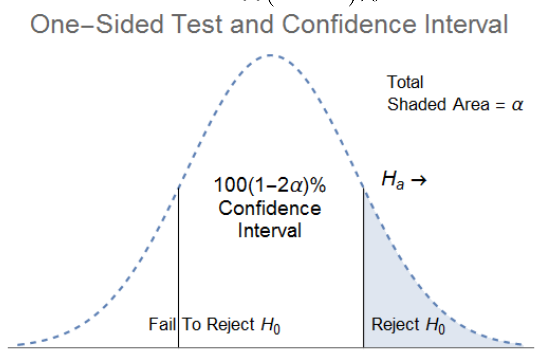
• Example: Exams are given to two different classes: a sample from Class A has 64 students and a sample from Class B has 100 students. The intention is that the exams are of equal difficulty, so that the average scores in the two classes are the same. In Class A's sample, the average score is 80.25 points, while in Class B's sample, the average is 81.16 points. The instructor believes the score for any individual student should be a normally distributed random variable with mean 80 points and standard deviation 5 points. Assuming the true standard deviation in each class is 5 points, test at the 10% and 3% significance levels whether (i) the average in Class A is equal to 80 points, (ii) the average in Class B is equal to 80 points, and (iii) the two class averages are equal.

- Let μ_A and μ_B be the respective class averages.
- For (i), our hypotheses are H_0: $\mu_A = 80$ and $H_a : \mu_A \neq 80$, since we do not care about a particular direction of error here.
- Our test statistic is $z = 80.25$ points, the average score of the 64 students in Class A.
- The distribution of the test statistic, under the null hypothesis, is normal with mean 80 points and standard deviation $5/\sqrt{64} = 0.625$ points.
- Thus, because our alternative hypothesis is $H_a : \mu_A \neq 80$ (which is two-sided), the p-value is $P(|N_{80,0.625} - 80| \geq 0.25) = 2 \cdot P(N_{80,0.625} \geq 80.25) = 2 \cdot P(N_{0,1} \geq 0.4) = 0.6892$.
- Since the p-value is quite large, it is $\boxed{\text{not significant}}$ at either the 10% or 3% significance level, and we accordingly fail to reject the null hypothesis in both cases.
- For (ii), our hypotheses are H_0: $\mu_B = 80$ and $H_a : \mu_B \neq 80$, as (like above) we do not care about a particular direction of error.
- Our test statistic is $z = 81.16$ points, the average score of the 100 students in Class B.
- The distribution of the test statistic, under the null hypothesis, is normal with mean 80 points and standard deviation $5/\sqrt{100} = 0.5$ points.
- Thus, because our alternative hypothesis is $H_a : \mu_A \neq 80$ (which is two-sided), the p-value is $P(|N_{80,0.5} - 80| \geq 1.16) = 2 \cdot P(N_{80,0.5} \geq 81.16) = 2 \cdot P(N_{0,1} \geq 2.32) = 0.0203$.
- Since the p-value is quite small, the result is statistically significant at both the 10% and 3% significance levels, and we accordingly reject the null hypothesis in both cases.
- For (iii), our hypotheses are H_0: $\mu_A = \mu_B$ and $H_a : \mu_A \neq \mu_B$, for the same reasons as above.
- Here, we want to use a two-sample test. Since our testing procedure requires testing the distribution of a specific quantity, we can rephrase our hypotheses as $H_0 : \mu_A - \mu_B = 0$ and $H_a : \mu_A - \mu_B \neq 0$.
- Our test statistic is $z = 80.25 - 81.16 = -0.91$ points, the difference in the two class averages.
- Under the null hypothesis, $\mu_A - \mu_B$ is normal with mean $80 - 80 = 0$ points and standard deviation $\sqrt{\sigma_A^2 + \sigma_B^2} = \sqrt{0.625^2 + 0.5^2} = 0.8004$ points.
- Thus, because our alternative hypothesis is $H_a : \mu_A - \mu_B \neq 0$ (which is two-sided), the p-value is $P(|N_{0,0.8004}| \geq 0.91) = 2 \cdot P(N_{0,0.8004} \leq -0.91) = 2 \cdot P(N_{0,1} \leq -1.1369) = 0.2556$.
- Since the p-value is relatively large, the result is not statistically significant at either the 10% or 3% significance level, and we accordingly fail to reject the null hypothesis in both cases.

- We will also mention that the results of a z test can also be interpreted in terms of confidence intervals.
  - For a two-sided alternative hypothesis, if we give a $100(1 - \alpha)\%$ confidence interval around the mean of a distribution under the conditions of the null hypothesis, then we will reject the null hypothesis with significance level α precisely when the sample statistic lies outside the confidence interval:



- Intuitively, this makes perfect sense: the $100(1-\alpha)\%$ confidence interval is precisely giving the range of values around the null hypothesis sample statistic that we would believe are likely to have occurred by chance, in the sense that if we repeated the experiment many times, then we would expect a proportion $1-\alpha$ of the results to land inside the confidence interval.
- If we interpret this probability as an area, what this means is that we would expect to see a test statistic "far away" from the null hypothesis value only with probability α: if we do obtain such an extreme value as our test statistic, we should take this as strong evidence (at the significance level α) that the true test statistic does not align with the prediction from the null hypothesis.
- Instead of quoting a confidence interval around the null-hypothesis prediction, we usually quote a confidence interval around the test statistic instead, and then check whether the null-hypothesis prediction lies within the confidence interval around the test statistic.
- We can do the same thing with a one-sided alternative hypothesis, but because of the lack of symmetry in the rejection region, we instead need to use a $100(1-2\alpha)\%$ confidence interval to get the correct area:



- In this case, the shaded region has area α, and there is a second region also of area α on the other side of the confidence interval, so the total area inside the confidence interval is $1-2\alpha$, meaning it is a $100(1-2\alpha)\%$ confidence interval.

- Example: Using the Class A (64 students, average 80.25) and Class B (100 students, average 81.16) data above, with individual score standard deviation 5 points, construct 90% and 97% confidence intervals for (i) the true average of Class A, (ii) the true average of Class B, and (iii) the difference between the averages of the two classes. Then use the results to test the hypotheses at the 10% and 3% significance levels that (iv) the average of Class A is 80 points, (v) the average of Class A is 79 points, (vi) the average of Class B is 80 points, (vii) the average of Class B is 82 points, (viii) the average scores in the classes are equal, and (ix) the average score in Class A is 1 points greater than the average in Class B.
  - For (i), as we calculated above, the estimator for the mean of Class A has $\hat{\mu}_A = 80.25$ and $\sigma_A = 5/\sqrt{64} = 0.625$.
  - Thus, the 90% confidence interval for the mean is $80.25 \pm 1.6449 \cdot 0.625 = \boxed{(79.22, 81.28)}$, and the 97% confidence interval is $80.25 \pm 2.1701 \cdot 0.625 = \boxed{(78.89, 81.61)}$.
  - For (ii), as we calculated above, the estimator for the mean of Class B has $\hat{\mu}_B = 81.16$ and $\sigma_B = 5/\sqrt{100} = 0.5$.
  - Thus, the 90% confidence interval for the mean is $81.16 \pm 1.6449 \cdot 0.5 = \boxed{(80.34, 81.98)}$, and the 97% confidence interval is $81.16 \pm 2.1701 \cdot 0.5 = \boxed{(80.07, 82.25)}$.
  - For (iii), as we calculated above, the estimator for the difference in the means has $\hat{\mu}_{A-B} = -0.91$ and $\sigma_{A-B} = \sqrt{0.625^2 + 0.5^2} = 0.8004$.
  - Thus, the 90% confidence interval for the difference in the means is $-0.91 \pm 1.6449 \cdot 0.8004 = \boxed{(-2.23, 0.41)}$, and the 97% confidence interval is $-0.91 \pm 2.1701 \cdot 0.8004 = \boxed{(-2.65, 0.83)}$.
  - For (iv), since 80 lies inside both confidence intervals, the result is not statistically significant at either the 10% or 3% significance levels: we fail to reject the null hypothesis that the true mean is 80 points.

- However, for (v), since 79 lies outside the first interval, the result is statistically significant at the 10% level (we reject the null hypothesis that the true mean is 79 points) but not statistically significant at the 3% level (we fail to reject the null hypothesis with this more stringent significance level).
- For (vi), since 80 lies outside both confidence intervals, the result is statistically significant at both the 10% and 3% levels: we reject the null hypothesis that the true mean is 80 points.
- For (vii), since 82 lies outside the first interval (barely!) but inside the second interval, the result is statistically significant at the 10% level (we reject the null hypothesis that the average is 82) but not statistically significant at the 3% level (we fail to reject the null hypothesis).
- For (viii), since 0 lies inside both intervals, the result is not statistically significant at either the 10% or 3% significance levels: we fail to reject the null hypothesis that the means are equal.
- For (ix), since +1 lies outside both intervals, the result is statistically significant at both the 10% and 3% levels: we reject the null hypothesis that the average in class A is 1 point higher than in Class B.

4.1.3 z-Tests for Unknown Proportion

- In situations where the binomial distribution is well approximated by the normal distribution, we can adapt our procedure for using a z test to handle hypothesis testing with a binomially-distributed test statistic.
  - Thus, suppose we have a binomially distributed test statistic $B_{n,p}$ counting the number of successes in n trials with success probability p.
  - If np (the number of successes) and $n(1-p)$ (the number of failures) are both larger than 5, we are in the situation where the normal approximation to the binomial is good: then $P(a \le B_{n,p} \le b)$ will be well approximated by $P(a - 0.5 < N_{np,\sqrt{np(1-p)}} < b + 0.5)$, where N is normally distributed with mean $\mu = np$ and standard deviation $\sigma = \sqrt{np(1-p)}$. (Note that we have incorporated the continuity correction in our estimate[1].)
  - We can then test the null hypothesis $H_0 : p = c$ by equivalently testing the equivalent hypothesis $H_0 : np = nc$ using the normal approximation via a one-sample z test, where our test statistic is the number of observed successes k.
  - Therefore, if the hypotheses are $H_0 : p = c$ and $H_a : p > c$, the associated p-value is $P(B_{n,p} \ge k) \approx P(N_{np,\sqrt{np(1-p)}} > k - 0.5)$.
  - If the hypotheses are $H_0 : p = c$ and $H_a : p < c$, the associated p-value is $P(B_{n,p} \le k) \approx P(N_{np,\sqrt{np(1-p)}} < k + 0.5)$.
  - Finally, if the hypotheses are $H_0 : p = c$ and $H_a : p \ne c$, the associated p-value is $P(|B_{n,p} - nc| \ge |k - nc|) \approx \begin{cases} 2P(N_{np,\sqrt{np(1-p)}} > k - 0.5) & \text{if } k > nc \\ 2P(N_{np,\sqrt{np(1-p)}} < k + 0.5) & \text{if } k < nc \end{cases}$. (For completeness[2], in the trivial case $k = c$ the p-value is 1.)
  - We then compare the p-value to the significance level α and decide whether or not to reject the null hypothesis.
  - All of this still leaves open the question of what we can do in situations where the binomial distribution is not well approximated by the normal distribution.
  - In such cases, we can work directly with the binomial distribution explicitly, or (in the event n is large but np or $n(1-p)$ is small) we could use a Poisson approximation.
  - Of course, in principle, we could always choose to work with the exact distribution, but when n is large computing the necessary probabilities becomes cumbersome, which is why we usually use the normal approximation instead.

[1]As a practical matter, the continuity correction usually does not affect the resulting p-values very much, but in the interest of consistency with our previous discussion of the binomial distribution, we have included it here.

[2]We note that the equality of the binomial range and the normal range given in this formula is slightly erroneous in the case where np is not an integer, since in that case the two tail probabilities will be rounded to integers slightly differently. We shall ignore this very minor detail, since the practical effect of the difference is extremely small when the normal approximation is valid, we are already approximating anyway, and conventions occasionally differ on the proper handling of two-tailed binomial rounding calculations like this one.

- Example: A coin with unknown heads probability p is flipped $n = 100$ times, yielding 64 heads. Test at the 11%, 4%, and 0.5% significance levels the hypotheses (i) that the coin is fair, (ii) that the coin is more likely to land heads than tails, (iii) that heads is twice as likely as tails, (iv) heads is more than twice as likely as tails, and (v) heads is more than thrice as likely as tails.
  - For (i), our hypotheses are $H_0 : p = 1/2$ and $H_a : p \neq 1/2$, since we only want to know whether or not the coin is fair.
  - Here, we have $np = n(1-p) = 50$ so we can use the normal approximation. Note that $np = 50$ and $\sqrt{np(1-p)} = 5$.
  - We compute the p-value as $P(|B_{100,0.5} - 50| \geq 14) \approx 2P(N_{50,5} > 63.5) = 0.00693$.
  - Thus, the result is statistically significant at the 11% and 4% levels and we reject the null hypothesis in these cases. But it is not statistically significant at the 0.5% level, so we fail to reject the null hypothesis there. (We interpret this as giving fairly strong evidence that the heads probability is not 1/2.)
  - For (ii), it is easy to see that we want a one-sided alternative hypothesis; the only question is the appropriate direction.
  - Here, although we actually want to decide whether or not $p > 1/2$, this is the appropriate form of the alternative hypothesis. Thus, we take the null hypothesis as $H_0 : p = 1/2$ and the alternative hypothesis as $H_a : p > 1/2$.
  - We have the same parameters as above, so $np = 50$ and $\sqrt{np(1-p)} = 5$, and then the p-value is $P(B_{100,0.5} \geq 64) \approx P(N_{50,5} > 63.5) = 0.00347$.
  - Thus, the result is statistically significant at the 11%, 4%, and 0.5% significance levels, and we reject the null hypothesis in each case. (We interpret this as giving strong evidence that the heads probability is greater than 1/2.)
  - For (iii), our hypotheses are $H_0 : p = 2/3$ and $H_a : p \neq 2/3$, since we want to know whether or not the heads probability is 2/3 .
  - Our parameter values are now $n = 100$, $p = 2/3$ so that $np = 66.667$ and $\sqrt{np(1-p)} = 4.714$.
  - Since $n(1-p) = 33.333$ the normal approximation is still appropriate, so we compute the p-value as $P(|B_{100,2/3} - 66.667| \geq 2.667) \approx 2P(N_{66.667,4.714} < 63.5) = 0.5017$.
  - Thus, the result is not statistically significant at the 11%, 4%, or 0.5% levels, and we accordingly fail to reject the null hypothesis in each case. (We interpret this as giving minimal evidence against the hypothesis that the heads probability is 2/3.)
  - For (iv), our hypotheses are $H_0 : p = 2/3$ and $H_a : p > 2/3$, since we want to know whether or not heads is more than twice as likely as tails, and this is appropriately set as the alternative hypothesis.
  - As above, the parameter values are $n = 100$, $p = 2/3$ so that $np = 66.667$ and $\sqrt{np(1-p)} = 4.714$.
  - We compute the p-value as $P(B_{100,2/3} \geq 64) \approx P(N_{66.667,4.714} > 63.5) = 0.7491$.
  - Thus, the result is not statistically significant at the 11%, 4%, or 0.5% levels, and we accordingly fail to reject the null hypothesis in each case. (We interpret this, again, as giving minimal evidence against the hypothesis that the heads probability is 2/3.)
  - For (v), we first try taking the hypotheses as $H_0 : p = 3/4$ and $H_a : p > 3/4$, since we want to know whether or not heads is more than thrice as likely as tails, and this is appropriately set as the alternative hypothesis.
  - The parameter values now are $n = 100$ and $p = 3/4$ so that $np = 75$ and $\sqrt{np(1-p)} = 4.330$.
  - Since $np = 75$ and $n(1-p) = 25$ the normal approximation is still appropriate, so we compute the p-value as $P(B_{100,3/4} \geq 64) \approx P(N_{75,4.330} > 63.5) = 0.9960$.
  - The result is (extremely) not statistically significant at the 11%, 4%, or 0.5% levels, and we accordingly fail to reject the null hypothesis in each case. (We interpret this as giving essentially zero evidence against the hypothesis that the true heads probability is at most 3/4.)
  - Although it seems quite obvious that the true heads probability should be less than 3/4 based on the results of the last hypothesis test we performed, that is not how we can interpret the result of the calculation.

- Instead, we should test $H_0 : p = 3/4$ with alternative hypothesis $H_a : p < 3/4$: this will have a p-value of $P(B_{100,3/4} \geq 64) \approx P(N_{75,4.330} < 64.5) = 0.0077$.
- This latter test is statistically significant at the 11% and 4% levels, but not statistically significant at the 0.5% level. Our interpretation now is that we have fairly strong evidence against the hypothesis that the true heads probability is greater than or equal to 3/4.

- This last example illustrates another nuance with hypothesis testing, namely, that if we are using a one-sided alternative hypothesis, we may actually want to try testing the other version of the alternative hypothesis depending on what the result of the test will be.
  - In general, we interpret "rejecting the null hypothesis" as a much stronger statement than "failing to reject the null hypothesis", since rejecting the null hypothesis takes more evidence (the p-value must be less than the significance level α, which is usually a stringent requirement).
  - Thus, the version of the alternative hypothesis in which we reject the null hypothesis (if there is one) is usually the one we will want to discuss.

- Example: A 6-sided die is rolled 18 times, yielding six 4s. Test at the 15%, 4%, and 1% significance levels the hypothesis that the true probability of rolling a 4 is equal to 1/6.
  - Our hypotheses are $H_0 : p = 1/6$ and $H_a : p \neq 1/6$.
  - Under the conditions of the null hypothesis, the total number of 4s rolled is binomially distributed with parameters $n = 18$ and $p = 16$. Here, $np = 3$ is too small for us to apply the normal approximation to the binomial distribution, so we will work directly with the binomial distribution itself.
  - The desired p-value is $P(\left|B_{18,1/6} - 3\right| \geq |6 - 3|) = P(B_{18,1/6} \geq 6) + P(B_{18,1/6} \leq 0) = 0.1028$.
  - The result is statistically significant at the 15% significance level, and we accordingly reject the null hypothesis. However, it is not statistically significant at the 4% or 1% significance levels, and so we fail to reject the null hypothesis in these cases.
  - We interpret this result as saying that there is moderate evidence against the hypothesis that the probability of rolling a 4 is equal to 1/6.

- If we have two independent, binomially-distributed quantities each of which is well approximated by a normal distribution, we can use the method for a two-sample z test to set up a hypothesis test for the difference of these quantities: we refer to this as a two-sample z-test for unknown proportion.
  - Suppose the two proportions are A and B. Then we would use the null hypothesis $H_0 : A - B = 0$ to test whether $A = B$, and our test statistic would be the difference between the proportions.
  - By hypothesis, A is normally distributed with mean p_A and standard deviation $\sigma_A = \sqrt{p_A(1 - p_A)/n_A}$ while B is normally distributed with mean p_B and standard deviation $\sigma_B = \sqrt{p_B(1 - p_B)/n_B}$.
  - Under the assumption that H_0 is true, the test statistic $A - B$ is normally distributed with mean 0 (the true mean postulated by the null hypothesis).
  - However, the null hypothesis does not actually tell us the standard deviations of A and B (that would only be the case if the null hypothesis were to state a specific value for A and for B).
  - What we must do instead is estimate the standard deviations using the sample data.
  - Here, under the null hypothesis assumption that the two proportions are actually equal, we can calculate a pooled estimate for the true proportion p by putting the two samples together: if sample A has k_A successes in n_A trials and sample B had k_B successes in n_B trials, then together there were $k_A + k_A$ successes in $n_A + n_B$ trials, so our pooled estimate for both p_A and p_B is $p_{\text{pool}} = \dfrac{k_A + k_B}{n_A + n_B}$.
  - Then the standard deviation of A is $\sigma_A = \sqrt{\dfrac{p_{\text{pool}}(1 - p_{\text{pool}})}{n_A}}$ and the standard deviation of B is $\sigma_B = \sqrt{\dfrac{p_{\text{pool}}(1 - p_{\text{pool}})}{n_B}}$, so the standard deviation of $A - B$ is $\sigma_{A-B} = \sqrt{\sigma_A^2 + \sigma_B^2} = \sqrt{\dfrac{p_{\text{pool}}(1 - p_{\text{pool}})}{n_A} + \dfrac{p_{\text{pool}}(1 - p_{\text{pool}})}{n_B}}$.
  - The desired p-value is then the probability that the normally-distributed random variable $N_{\mu_{A-B},\sigma_{A-B}}$ will take a value further from the hypothesized value 0 (in the direction of the alternative hypothesis, as applicable) than the observed test statistic $z = \hat{p}_A - \hat{p}_B$.

- Remark: We could also estimate the two standard deviations from their sample proportions separately as $\sigma_A = \sqrt{\hat{p}_A(1-\hat{p}_A)/n_A}$ and $\sigma_B = \sqrt{\hat{p}_B(1-\hat{p}_B)/n_B}$: these are called the unpooled standard deviations, and they give a slightly different estimate $\sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n_A} + \frac{\hat{p}_B(1-\hat{p}_B)}{n_B}}$ for the standard deviation of $A - B$. As a practical matter, if the sample proportions $\hat{p}_A$ and $\hat{p}_B$ are actually close to each other, these values will both also be close to p_{pool}, and thus the two estimates for σ_{A-B} will also be very close. We usually use the unpooled standard deviations if we want to perform more complicated tests on the observed proportions (e.g., if we wanted to test whether the proportion for A exceeded the proportion for B by 2% or more). There is not universal consensus on the usage of the pooled versus unpooled standard deviations.

- Example: In a sample from a statistics class taught with a traditional curriculum, 125 students out of 311 received an A (40.2%), whereas in a sample from a statistics class taught with a revised curriculum, 86 students out of 284 received an A (30.3%). If p_t is the proportion of students getting an A with the traditional curriculum and p_r is the proportion of students getting an A with the revised curriculum, test the hypothesis $p_t = p_r$ at the 10%, 5%, 1%, and 0.1% significance levels with alternative hypothesis (i) $p_t > p_r$, (ii) $p_t < p_r$, and (iii) $p_t \neq p_r$.
  - The proportion of students getting an A in each of the samples will be binomially distributed, and the parameters are in the range where the normal approximation is applicable in both samples.
  - For (i), the null hypothesis is $H_0 : p_t - p_r = 0$ with alternative hypothesis $p_t - p_r > 0$.
  - Here, we have $n_t = 311$, $n_r = 284$, $\hat{p}_t = 125/311 = 0.4019$, and $\hat{p}_r = 86/284 = 0.3028$, so that $\hat{p}_{t-r} = 0.0991$.
  - To find the pooled standard deviation, we have $p_{\text{pool}} = (125+86)/(311+284) = 0.3546$. Then $\sigma_{t-r,\text{pool}} = \sqrt{p_{\text{pool}}(1-p_{\text{pool}})\left[\frac{1}{n_A} + \frac{1}{n_B}\right]} = 0.03927$.
  - If we wanted to use the unpooled standard deviations, we would have $\sigma_t = \sqrt{\hat{p}_t(1-\hat{p}_t)/n_t} = 0.02780$, $\sigma_r = \sqrt{\hat{p}_r(1-\hat{p}_r)/n_r} = 0.02726$, and $\sigma_{t-r,\text{unpool}} = \sqrt{\hat{p}_t(1-\hat{p}_t)/n_t + \hat{p}_r(1-\hat{p}_r)/n_r} = 0.03894$.
  - Using the pooled standard deviation, the desired p-value is $P(N_{0,0.03927} \geq 0.0991) = P(N_{0,1} \geq 2.5242) \approx 0.00580$.
  - Thus, the result is statistically significant at the 10%, 5%, and 1% significance levels, and we accordingly reject the null hypothesis in these cases, but it is not statistically significant at the 0.1% significance level.
  - We interpret this result as saying that there is strong evidence for the hypothesis that the students with the traditional curriculum had a higher proportion of As than the students with the revised curriculum.
  - For (ii), the null hypothesis is $H_0 : p_t - p_r = 0$ with alternative hypothesis $p_t - p_r < 0$.
  - The parameters and values are the same as before: the only difference is that the p-value is now $P(N_{0,0.03927} \leq 0.0991) = P(N_{0,1} \leq 2.5242) \approx 0.99420$.
  - Thus, the result is (extremely!) not statistically significant at any of the indicated significance levels, and we fail to reject the null hypothesis in all cases.
  - We interpret this result as saying that there is essentially zero evidence for the hypothesis that the students with the revised curriculum had a higher proportion of As than the students with the traditional curriculum.
  - For (ii), the null hypothesis is $H_0 : p_t - p_r = 0$ with alternative hypothesis $p_t - p_r \neq 0$.
  - The parameters and values are still the same; the only difference is that the p-value is now $P(|N_{0,0.03927} - 0| \geq |0.0991 - 0|) = 2P(N_{0,0.03927} \geq 0.0991) = 2P(N_{0,1} \geq 2.5242) \approx 0.01160$.
  - Thus, the result is statistically significant at the 10% and 5% significance levels, so we accordingly reject the null hypothesis in those situations, but it is not statistically significant at the 1% or 0.1% significance levels, so we fail to reject the null hypothesis in those cases.
  - We interpret this result as saying that there is relatively strong evidence for the hypothesis that the students with the revised curriculum had a different proportion of As than the students with the traditional curriculum.

- Example: A pollster conducts a poll on the favorability of Propositions ♣ and ♡. They poll 2,571 people and find that 1,218 of them favor Proposition ♣ (47.4%). In a separate poll, also of 2,571 people, they find 1,344 of them favor Proposition ♡ (52.3%). Perform hypothesis tests at the 8% and 1% significance levels that (i) Proposition ♣ has at least 50% support, (ii) the support for Proposition ♣ is exactly 50%, (iii) Proposition ♡ has at least 50% support, (iv) the support for Proposition ♡ is exactly 55%, (v) Proposition ♡ has more support than Proposition ♣.
  - For (i), our hypotheses are $H_0 : p_{\clubsuit} = 0.50$ and $H_a : p_{\clubsuit} < 0.50$, since we want to test whether Proposition ♣ has at least 50% support.
  - We test this direction of the alternative hypothesis because the observed support level of Proposition ♣ is actually less than 50%, so we would like to reject the other possibility (i.e., that the support is not less than 50%).
  - Here, we have $np = n(1-p) = 1285.5$ so we can use the normal approximation. Note that $np = 1285.5$ and $\sqrt{np(1-p)} = 25.35$.
  - We compute the p-value as $P(B_{2571,0.5} \leq 1218) \approx P(N_{1285.5,25.35} < 1218.5) = P(N_{0,1} < -2.6427) = 0.00411$.
  - Thus, the result is statistically significant at both the 8% and 1% significance levels. (We interpret this as saying that there is strong evidence against the hypothesis that the support for Proposition ♣ is 50% or above.)
  - For (ii), our hypotheses are $H_0 : p_{\clubsuit} = 0.50$ and $H_a : p_{\clubsuit} \neq 0.50$, since we now only want to test whether Proposition ♣ has 50% support (not whether it is specifically higher or lower).
  - We have the same parameters as above, so the p-value is $\approx 2P(N_{1285.5,25.35} < 1218.5) = 2P(N_{0,1} < -2.6427) = 0.00822$.
  - Thus, the result is statistically significant at both the 8% and 1% significance levels. (We interpret this as saying that there is strong evidence against the hypothesis that the support for Proposition ♣ is equals 50%.)
  - For (iii), our hypotheses are $H_0 : p_{\heartsuit} = 0.50$ and $H_a : p_{\heartsuit} > 0.50$, since we want to test whether Proposition ♡ has at least 50% support.
  - Note that we test with a different alternative hypothesis than (i) because the observed support level of Proposition ♡ is actually greater than 50%, so we would like to reject the other possibility (i.e., that the support is not greater than 50%).
  - We still have the same parameters as above (only the actual test statistic value will differ), so we compute the p-value as $P(B_{2571,0.5} \geq 1344) \approx P(N_{1285.5,25.35} > 1343.5) = P(N_{0,1} > 2.2877) = 0.01107$.
  - Thus, the result is statistically significant at the 8% significance level, but not statistically significant at the 1% significance level. (We interpret this as saying that there is moderately strong evidence against the hypothesis that the support for Proposition ♡ is 50% or below.)
  - For (iv), our hypotheses are $H_0 : p_{\heartsuit} = 0.55$ and $H_a : p_{\heartsuit} \neq 0.55$, since we want to test whether or not Proposition ♡ has 55% support (and that any difference from 55% is not in any particular direction).
  - Here we have $n = 2571$ and $p = 0.55$ so that $np = 1414.05$ and $\sqrt{np(1-p)} = 25.225$.
  - Then the p-value is $\approx 2P(N_{1414.05,25.225} < 1344.5) = P(N_{0,1} > -2.7571) = 0.00291$.
  - Thus, the result is statistically significant at both the 8% and 1% significance levels, and we accordingly reject the null hypothesis. (We interpret this as saying that there is strong evidence against the hypothesis that the support for Proposition ♡ is 55%.)
  - For (v), this is a two-sample test, so we take our hypotheses as $H_0 : p_{\clubsuit} - p_{\heartsuit} = 0$ and $H_a : p_{\clubsuit} - p_{\heartsuit} < 0$, since we want to test whether or not Proposition ♡ has more support than Proposition ♣ (we want a one-sided alternative hypothesis) and because the sampling suggests that Proposition ♡ does actually have more support than Proposition ♣.
  - Here, we have $n_{\clubsuit} = n_{\heartsuit} = 2571$, $\hat{p}_{\clubsuit} = 1218/2571 = 0.4737$, and $\hat{p}_{\heartsuit} = 1344/2571 = 0.5228$, so that $\hat{p}_{\clubsuit-\heartsuit} = -0.04901$.
  - To find the pooled standard deviation, we have $p_{\text{pool}} = (1218 + 1344)/(2571 + 2571) = 0.4982$, so then $\sigma_{\clubsuit-\heartsuit,\text{pool}} = \sqrt{p_{\text{pool}}(1-p_{\text{pool}})\left[\frac{1}{n_{\clubsuit}} + \frac{1}{n_{\heartsuit}}\right]} = 0.01395$.

- Then the desired p-value is $P(N_{0,0.01395} < -0.04901) = P(N_{0,1} < -3.5143) = 0.000220$.
- Thus, the result is statistically significant at both the 8% and 1% significance levels, and we accordingly reject the null hypothesis. (We interpret this as saying that there is strong evidence that Proposition ♡ has more support than Proposition ♣.)

4.1.4 Errors in Hypothesis Testing

- When we perform a hypothesis test, there are two possible outcomes (reject H_0 or fail to reject H_0).
  - The correctness of the result depends on the actual truth of H_0: if H_0 is false then it is correct to reject it, while if H_0 is true than it is correct not to reject it.
  - The other two situations, namely "rejecting a correct null hypothesis" and "failing to reject an incorrect null hypothesis" are refered to as hypothesis testing errors.
  - Since these two errors are very different, we give them different names:
- Definition: If we are testing a null hypothesis H_0, we commit a type I error if we reject H_0 when H_0 was actually true. We commit a type II error if we fail to reject H_0 when H_0 was actually false.
  - We usually summarize these errors with a small table:

H_0 \ Result	Fail to Reject H_0	Reject H_0
H_0 is true	Correct Decision	Type I Error
H_0 is false	Type II Error	Correct Decision

  - The names for these two errors are very unintuitive[3], and it must simply be memorized which one is which.
  - If we view a positive result as one in which we reject the null hypothesis, which in most cases is the practical interpretation, then a type I error corresponds to a false positive (a positive test on an actual negative sample) while a type II error corresponds to a false negative (a negative test on an actual positive sample).
  - For example, if the purpose of the hypothesis test is to determine whether or not to mark an email as spam (with H_0 being that the email is not spam), a type I error would be marking a normal email as spam, while a type II error would be marking a spam email as normal.
- We would like, in general, to minimize the probabilities of making a type I or type II error.
  - The probability of committing a type I error is the significance level α of the test, since by definition this is the probability of rejecting the null hypothesis when it is actually true.
  - The probability of committing a type II error is denoted by β. This value is more difficult to calculate, since it will depend on the actual nature in which H_0 is false.
  - If we postulate the actual value of the test statistic, we can calculate the probability of committing a type II error.
- Example: A new mathematics curriculum is being tested in schools to see if students score more highly on standardized tests. The scores for students using the old curriculum are normally distributed with mean 200 and standard deviation 20. It is assumed that scores using the new curriculum are also normally distributed with mean μ and standard deviation 20. The hypothesis $H_0 : \mu = 200$ is tested against the alternative $H_a : \mu > 200$ using a sample of 400 students using the new curriculum. The null hypothesis will be rejected if the sample mean $\hat{\mu} > 202$. Find (i) the probability of making a type I error, and also find the probability of making a type II error if the true mean is actually (ii) 201, (iii) 202, (iv) 203, (v) 204, and (vi) 205.
  - For (i), we want to calculate the probability of rejecting the null hypothesis when it is true. If the null hypothesis is true, then the sample mean $\hat{\mu}$ will be normally distributed with mean 200 and standard deviation $20/\sqrt{400} = 1$.

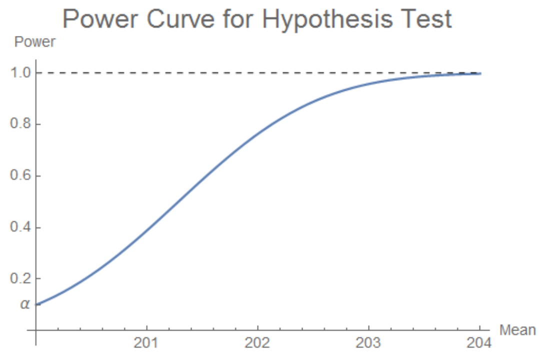
[3] The terminology for Type I and Type II errors is directly from the original 1930 paper of the originators of this method of hypothesis testing, Jerzy Neyman and Egon Pearson.

- Then, the probability of rejecting the null hypothesis is $P(N_{200,1} > 202) = P(N_{0,1} > 2) = \boxed{0.02275}$. (Note that this value is the significance level α for this hypothesis test.)
- For (ii), we want to calculate the probability of failing to reject the null hypothesis when it is false. Under the assumption given, the sample mean $\hat{\mu}$ will be normally distributed with mean 201 and standard deviation $20/\sqrt{400} = 1$.
- Then, the probability of failing to reject the null hypothesis is $P(N_{201,1} \le 202) = P(N_{0,1} \le 1) = \boxed{0.8413}$: quite large.
- For (iii), the assumption now is that $\hat{\mu}$ is normally distributed with mean 202 and standard deviation 1, so the probability of failing to reject the null hypothesis is $P(N_{202,1} \le 202) = P(N_{0,1} \le 0) = \boxed{0.5}$: still quite large.
- For (iv), $\hat{\mu}$ is normally distributed with mean 203 and standard deviation 1 so the probability of failing to reject the null hypothesis is $P(N_{203,1} \le 202) = P(N_{0,1} \le -1) = \boxed{0.1587}$: smaller than before, but still fairly high.
- For (v), $\hat{\mu}$ is normally distributed with mean 204 and standard deviation 1 so the probability of failing to reject the null hypothesis is $P(N_{204,1} \le 202) = P(N_{0,1} \le -2) = \boxed{0.02275}$: reasonably small.
- For (vi), $\hat{\mu}$ is normally distributed with mean 205 and standard deviation 1 so the probability of failing to reject the null hypothesis is $P(N_{205,1} \le 202) = P(N_{0,1} \le -3) = \boxed{0.00135}$: very small.

- We can see that as the true mean gets further away from the mean predicted by the null hypothesis, the probability of making a type II error drops.
  - The idea here is quite intuitive: the bigger the distance between the true mean and the predicted mean, the better our hypothesis test will be better at picking up the difference between them.
- If we use the same rejection rule, but instead vary the sample size, the probability of making either type of error will change:
- <u>Example</u>: The school wants to gather more data on the effectiveness of the new curriculum. Assume as before the scores with the old curriculum are normally distributed with mean 200 and standard deviation 20, and the new curriculum scores also have standard deviation 20. We again test $H_0 : \mu = 200$ against $H_a : \mu > 200$ and reject the null hypothesis if $\hat{\mu} > 202$. Find the probabilities of a type I error, and the probability of a type II error if the true mean is actually $\mu = 203$, using a sample size (i) $n = 100$, (ii) $n = 400$, and (iii) $n = 1600$.
  - For (i), to find the probability of a type I error we assume $\mu = 200$. Then the sample mean $\hat{\mu}$ is normally distributed with mean 200 and standard deviation $\sigma = 20/\sqrt{100} = 2$, so the probability of a type I error is $P(N_{200,2} > 202) = P(N_{0,1} > 1) = \boxed{0.1587}$.
  - For a type II error, we assume $\mu = 203$. Then the sample mean $\hat{\mu}$ is normally distributed with mean 203 and standard deviation $\sigma = 20/\sqrt{100} = 2$, so the probability of a type II error is $P(N_{203,2} \le 202) = P(N_{0,1} \le -0.5) = \boxed{0.3085}$.
  - For (ii), for a type I error, the sample mean $\hat{\mu}$ is normally distributed with mean 200 and standard deviation $\sigma = 20/\sqrt{400} = 1$, so the probability of a type I error is $P(N_{200,1} > 202) = P(N_{0,1} > 2) = \boxed{0.02275}$.
  - For a type II error, the sample mean $\hat{\mu}$ is normally distributed with mean 203 and standard deviation $\sigma = 20/\sqrt{400} = 1$, so the probability of a type II error is $P(N_{203,1} \le 202) = P(N_{0,1} \le -1) = \boxed{0.1587}$.
  - For (iii), for a type I error, the sample mean $\hat{\mu}$ is normally distributed with mean 200 and standard deviation $\sigma = 20/\sqrt{1600} = 0.5$, so the probability of a type I error is $P(N_{200,0.5} > 202) = P(N_{0,1} > 4) = \boxed{0.0000316}$.
  - For a type II error, the sample mean $\hat{\mu}$ is normally distributed with mean 203 and standard deviation $\sigma = 20/\sqrt{1600} = 0.5$, so the probability of a type II error is $P(N_{203,0.5} \le 202) = P(N_{0,1} \le -2) = \boxed{0.02275}$.
- If we fix the significance level α but vary the sample size, the probability of a type II error will change:

- Example: The school wants to determine how large a sample size would have been necessary to determine the effectiveness of the new curriculum. Assume the scores with the old curriculum are normally distributed with mean 200 and standard deviation 20, and the new curriculum scores are normally distributed with true mean 203 and standard deviation 20. We test $H_0 : \mu = 200$ against $H_a : \mu > 200$ at the 1% significance level. Find the probabilities of a type II error using a sample size (i) $n = 100$, (ii) $n = 400$, (iii) $n = 900$, and (iv) $n = 1600$.
  - Under the assumptions of the hypothesis test, $\hat{\mu}$ is normally distributed with mean 200 and standard deviation $\sigma = 20/\sqrt{n}$.
  - Since the test is one-tailed, the critical value of $\hat{\mu}$ is the value c such that $P(N_{200,20/\sqrt{n}} > c) = 0.01$. Equivalently, this says $P(N_{0,1} > \frac{c-200}{20/\sqrt{n}}) = 0.01$, which occurs when $\frac{c-200}{20/\sqrt{n}} = 2.3263$ and thus $c = 200 + 2.3263 \cdot 20/\sqrt{n}$.
  - In reality, the sample mean $\hat{\mu}$ is normally distributed with mean 203 and standard deviation $\sigma = 20/\sqrt{n}$.
  - This means that the probability of a type II error is $P(N_{203,20/\sqrt{n}} \le c) = P(N_{0,1} \le 2.3263 - \frac{3\sqrt{n}}{20})$.
  - For (i), evaluating this probability for $n = 100$ yields the type-II error probability as $P(N_{0,1} \le 0.8263) = \boxed{0.7957}$.
  - For (ii), evaluating this probability for $n = 400$ yields the type-II error probability as $P(N_{0,1} \le -0.6737) = \boxed{0.2503}$.
  - For (iii), evaluating this probability for $n = 900$ yields the type-II error probability as $P(N_{0,1} \le -2.1737) = \boxed{0.01486}$.
  - For (iv), evaluating this probability for $n = 1600$ yields the type-II error probability as $P(N_{0,1} \le -3.6737) = \boxed{0.0001195}$.
- We can glean a few general insights from from the examples above.
  - First, by adjusting the significance level α, we can affect the balance between the probabilities of a type I error and a type II error.
  - A smaller α gives a smaller probability of a type I error but a greater probability of a type II error: we are more stringent about rejecting the null hypothesis (so we make fewer type I errors) but at the same time that means we also incorrectly fail to reject the null hypothesis more (so we make more type II errors).
  - Second, by increasing the sample size, we decrease the probabilities of both error types together (though they do not necessarily drop similar amounts). This is also quite reasonable: the larger the sample, the closer the sample mean should be to the true mean and the less variation around the true mean it will have.
  - What this means is that with a larger sample size, the test will have a better ability to distinguish smaller deviations away from the null hypothesis. This property has a name:
- Definition: If we are testing a null hypothesis H_0, the probability $1-\beta$ of correctly rejecting the null hypothesis when it is false is called the power of the test.
  - The power of the test will depend on the significance level α, the true value of the test parameter, and the size n of the sample.
  - For a fixed α and n, we can plot the dependence of the power on the true value of the test parameter to produce what are called power curves.
  - To plot a power curve, we need only perform a calculation like the one we did above (first calculating the critical value, and then calculating the probability of correctly rejecting the null hypothesis based on the value of the parameter).
  - For the test we analyzed above, of testing $H_0 : \mu = 200$ against $H_a : \mu > 200$ with significance level $\alpha = 0.10$ and a sample size $n = 400$, we want to reject the null hypothesis if $\hat{\mu} > 201.282$, and so the power of the test if the true mean is x is $P(N_{x,1} \ge 201.282 = P(N_{0,1} \ge 201.282 - x)$, whose graph is

plotted below:



- As is suggested by the plot given above, the limit of the power as the true mean approaches the null hypothesis mean is equal to α. (This follows by noting the moderately confusing fact that the type II error coincides with the complement of the type I error in the limit.)
- Furthermore, the power increases monotonically as the true parameter value moves away from the null hypothesis mean, and approaches 1 as the true parameter value becomes large.

- Although it may seem that we would always want the power of the test to be as large as possible, there are certain non-obvious drawbacks to this desire.
  - Specifically, if the power is very large even for small deviations away from the null hypothesis parameter, then the test will frequently yield statistically significant results even when the sample parameter is not very far away from the null hypothesis parameter.
  - In some situations this is good, but in others it is not: for example, suppose we want to test whether the new curriculum actually improves scores above the original mean $\mu = 200$.
  - If the power is sufficiently high, the hypothesis test will indicate a statistically significant result whenever the the sample mean $\hat{\mu} > 200.001$. Now, it certainly is useful to know that the true mean is statistically significantly different from 200, but in most situations we would not view this difference as substantial.
  - This issue is usually framed as "statistical significance" versus "practical significance": with large samples, we may obtain a statistically significant difference from the hypothesized mean (perhaps even with an exceedingly small p-value), yet the actual difference is negligibly small and not actually important in practice.
  - This highlights one issue with relying solely on p-values on a measure of evidence quality: it is possible to set up tests (e.g., by using a very large sample) that yield extremely small p values even if the actual result is practically meaningless.
  - Another viewpoint here is that the null hypothesis is rarely (if ever) exactly true: thus, if we take a sufficiently large sample size, we can identify as statistically significant whatever tiny deviation actually exists, even if this deviation is not practically relevant.

- With these observations in mind, we can see that that the precise choice of the significance level α is entirely arbitrary (which has been illustrated by the somewhat eclectic selection of values in the examples we have given so far).
  - The only particular considerations we have are whether the choice of α yields acceptably low probabilities of making a type I or type II error.
  - In some situations, we would want to be extremely sure, when we reject the null hypothesis, that it was truly outlandishly unlikely to have observed the given data by chance: this corresponds to requiring α to be very small.
  - For example, if the result of the hypothesis test is regarding whether the numbers in a company's accounting ledgers are real or manufactured to cover up embezzling, we would want to be very sure that any seeming discrepancies were not merely random chance.

- However, in other situations (e.g., in the sciences) where the statistical test is merely one component of broader analysis of a topic, we should view the result of a hypothesis test as more of a suggestion for what to investigate next. If the p-value is very small, then it suggests that the alternative hypothesis may be correct, and further study is warranted. If the p-value is large, then it suggests that the null hypothesis is correct, and that additional study is not likely to yield different results.

- For various historical reasons, the significance level $\alpha = 0.05$ is very commonly used, since it strikes a balance between requiring strong evidence (only a 5% chance that the result could have arisen by chance) but not so strong as to tend to ignore good evidence suggesting the null hypothesis is false (which becomes likely with smaller values of α).
  - Indeed, many authors, both in the past and the present, often call a result with $p < 0.05$ "statistically significant" (with no qualifier) and a result with $p < 0.01$ "very statistically significant" (and if $p < 0.001$, one also sometimes sees "extremely statistically significant").
  - Such statements entirely ignore the actual nuances of what p-values measure, and should be assiduously avoided: a hypothesis test with $p = 0.051$ provides almost the same level of evidence against the null hypothesis as a hypothesis test with $p = 0.049$, and there is simply no practical distinction that should be made between the two.
  - Nonetheless, the prevalence of the view that results are not worth reporting unless they have $p < 0.05$ has led to various undesirable, and very real, consequences. One such problem is the lack of reporting of experiments that had negative or "statistically insignificant" results (which is also partly a cultural issue in research, more generally), which leads to a bias in the resulting literature.

- There are various other related factors that can also contribute to an overall bias in reported results of hypothesis tests.
  - When analyzing collected data, it is important to examine outliers (points far away from the norm), since they may be the results of errors in data collection or otherwise unrepresentative of the desired sample. The presence of outliers often has a large effect on the results of a hypothesis test, especially one that relies on an estimate of a standard deviation or variance, and in some situations it is entirely reasonable to discard outliers.
  - However, this process can rise to the level of scientific misconduct if it is done after the fact: the phenomenon called <u>p-hacking</u> involves massaging the underlying data used for a statistical test (e.g., by removing additional outliers, or putting back outliers that were previously removed) so that it yields a p-value less than 0.05 rather than greater than 0.05.

- Another related issue is that of performing multiple comparisons on the same set of data.
  - This procedure is sometimes (more uncharitably) referred to as <u>data dredging</u>: sifting through data to find signals in the underlying noise.
  - The difficulty with performing multiple comparisons is that there is a probability α that any given hypothesis test will yield a statistically significant result even though the null hypothesis is true, and these probabilities add up if we perform more tests.
  - For example, if we perform 40 hypothesis tests where the null hypothesis is actually true at the $\alpha = 0.05$ significance level, we will have a probability $1-0.95^{40} \approx 87\%$ of getting at least one statistically significant result (i.e., making a type I error), even though there is no actual result to find.
  - If we test a large number of hypotheses, then (depending on the actual likelihood of the hypotheses we test and the significance level α) it is entirely possible that most of the statistically significant results we obtain will be spurious. It is the same situation of testing for a rare disease: positive tests are much more likely to be a false positive than a correct positive.
  - When actually performing a large number of hypothesis tests, one should correct for the fact that multiple tests were performed on the same data. Various methods exist for this, such as the <u>Bonferroni correction</u>, which states that the desired confidence level α should be divided by the total number of tests performed: the idea is simply that we want a total probability of approximately α (among all the tests) of obtaining a type I error.

- Thus, if we perform 5 different tests on the same data using the typical $\alpha = 0.05$, we should actually test at the level $\alpha = 0.01$ in order to have an overall total probability of approximately 0.05 of obtaining at least one type I error.
- Multiple hypothesis testing on the same data is not necessarily a problem if we report the results of all of the tests and give the actual p-values for each test, since then it is straightforward to apply such correction methods to identify which results are likely to be real.
- However, a much more serious issue occurs when we only report the statistically significant results without noting (or correcting for) the fact that other hypothesis tests were also performed and not reported: as noted above, it is then entirely possible that most of the reported results are false.
- The extent to which false research findings are an actual problem in scientific research is disputed[4], and varies substantially by field, but is obviously a fundamental concern!
- When spurious results are reported as significant, followup studies will (at least in theory) eventually show that the original results were erroneous; this phenomenon of having subsequent studies widely being unable to replicate the results of the originals has led to a replication crisis in various fields, since it suggests that most of these original results were actually false.
- Although one can reasonably adopt the viewpoint that eventually incorrect results will be identified and extirpated, having many false results believed to true creates a substantial waste of resources (in having to perform unnecessary replication studies and, more broadly, building additional research on a faulty foundation).

- We have mentioned these issues because it is very important to be sanguine about the limitations of hypothesis testing, and how easy it is to misuse or misinterpret the results of hypothesis tests.
  - Ultimately, there can be no "magic fix" for these issues: statistical testing is fundamentally an approximation, and there is always a positive probability of getting an incorrect result.
  - When designing an experiment and a hypothesis test, the best we can do is to identify the right significance level α (which balances the possibility of making a type I error against the possibility of making a type II error) and the right sample size n (which balances the possibility of making any type of error with the difficulty and expense of obtaining the necessary data, and with the likelihood that there probably is some practically irrelevant deviation from the null hypothesis), and conduct followup analyses and replication studies to make sure any observed results are truly real and practically significant.
- In 2016, the American Statistical Association[5] released guidelines for interpretation and usage of p-values:
  1. p-values can indicate how incompatible the data are with a specified statistical model.
  2. p-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
  3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
  4. Proper inference requires full reporting and transparency.
  5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
  6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.
- We also quote their conclusion, summarizing proper use of hypothesis tests, p-values, and statistics generally:

> Good statistical practice, as an essential component of good scientific practice, emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean. No single index should substitute for scientific reasoning.

[4] See Ioannadis, "Why Most Published Research Findings are False", PLoS Medicine (2005) for an argument that this is a serious problem.

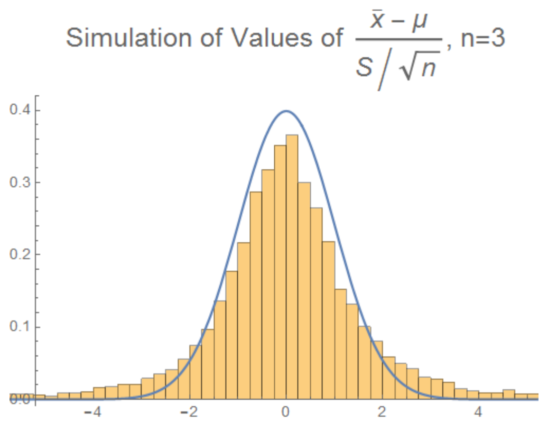
[5] Wasserstein and Lazar, "The ASA Statement on p-Values: Context, Process, and Purpose", https://amstat.tandfonline.com/doi/full/10.1080/00031305.2016.1154108 .

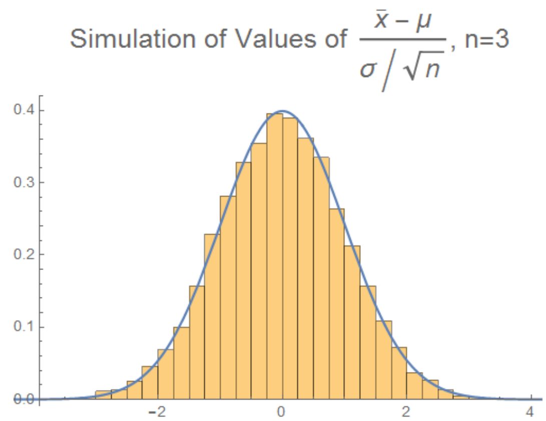
4.2 The t Distribution and t Tests

- Our goal in this section is to discuss the t distribution and t tests, which allow us to expand our hypothesis tests (and related discussion of confidence intervals) to approximately normally distributed quantities whose standard deviation is unknown.

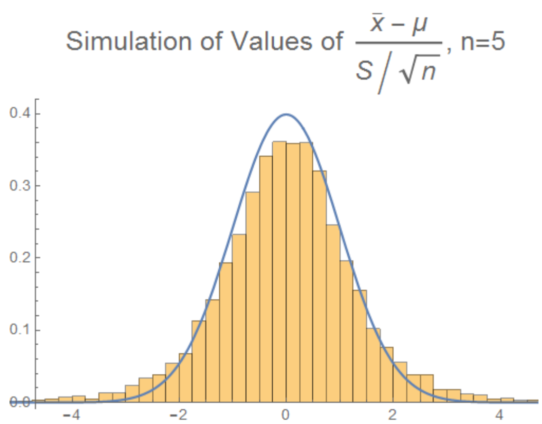
4.2.1 The t Distributions

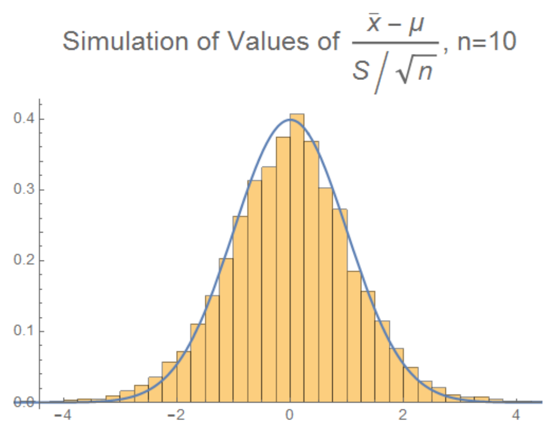
- In our discussion of hypothesis testing so far, we have relied on z tests, which require an approximately normally distributed test statistic whose standard deviation is known.
  - However, in most situations, it is unlikely that we would actually know the population standard deviation.
  - Instead, we must estimate the population standard deviation from the sample standard deviation.
  - As we have already discussed, for values $x_1, \dots, x_n$ drawn from a normal distribution with unknown mean μ and unknown standard deviation σ, the maximum likelihood estimate $\hat{\sigma} = \sqrt{\frac{1}{n}\left[(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2\right]}$ for the standard deviation is biased. (Note that $\overline{x} = \frac{1}{n}(x_1 + \cdots + x_n)$ is the sample mean.)
  - Thus, instead of the estimator $\hat{\sigma}$, we use the sample standard deviation $S = \sqrt{\frac{1}{n-1}\left[(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2\right]}$, whose square S^2 is an unbiased estimator of the population variance σ^2.
- It may seem reasonable to say that if we use the estimated standard deviation S in place of the unknown population σ, then we should be able to use a z test with the resulting approximation, much as we did with the normal approximation to the binomial distribution.
  - However, this turns out not to be the case! We can make clearer why not by converting the discussion to a distribution with a single unknown parameter by working with the normalized ratio $\frac{\overline{x} - \mu}{S/\sqrt{n}}$, which has mean 0 and standard deviation 1 and is analogous to the z-score $\frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$, whose distribution (under the assumptions of the null hypothesis) is the standard normal distribution of mean 0 and standard deviation 1.
  - If we take $\frac{\overline{x} - \mu}{S/\sqrt{n}}$ as our test statistic, then (as we will show) this test statistic is not normally distributed!
  - The distribution is similar in shape to the normal distribution, but it is in fact different, and is called the <u>t distribution</u>.
- We can illustrate visually the lack of normality of the normalized test statistic $\frac{\overline{x} - \mu}{S/\sqrt{n}}$ by simulating a sampling procedure.
  - Explicitly, suppose that X is normally distributed with mean $\mu = 0$ and standard deviation $\sigma = 1$, and we want to test the hypothesis that the mean actually is equal to 0 using the normalized test statistic $\frac{\overline{x} - \mu}{S/\sqrt{n}}$ with $n = 3$.
  - To understand the behavior of $\frac{\overline{x} - \mu}{S/\sqrt{n}}$, we sample the standard normal distribution to obtain 3 data points x_1, x_2, x_3 and then compute $\frac{\overline{x} - \mu}{S/\sqrt{n}}$ using the sample mean $\overline{x}$ and estimated standard deviation S.
  - The histogram below shows the results of performing this sampling 10000 times, along with the actual graph of the probability density function of the predicted normal distribution:

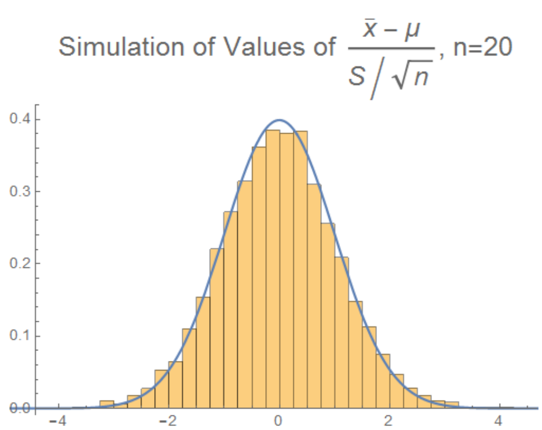


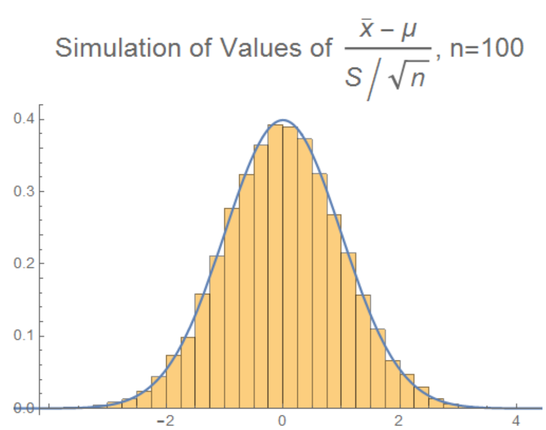


- Note how the actual histogram differs from the normal distribution: specifically, there are values occurring in the tails of the distribution far more often than they do for the normal distribution, while the values near the center occur slightly less often than predicted.
- In contrast, the same simulation using the test statistic $\frac{\overline{x}-\mu}{\sigma/\sqrt{n}}$ matches the normal distribution very closely, as can be seen in the second histogram above.
- The difference between the distribution of $\frac{\overline{x}-\mu}{S/\sqrt{n}}$ and the standard normal distribution is more pronounced when n is small. For larger n, the distribution looks much more approximately normal (this is related to the central limit theorem and the fact that the approximation of σ by S is unbiased in the limit as $n \to \infty$):









- What these plots indicate is that the actual distribution of the test statistic $\frac{\overline{x}-\mu}{S/\sqrt{n}}$ will depend on n, and that for larger n, it will be approximately normal.

- It is not a trivial matter to find the probability density function for the t distribution modeling the test statistic $\dfrac{\overline{x}-\mu}{S/\sqrt{n}}$.
  - The constants involved in the pdf involve a special function known as the gamma function, which generalizes the definition of the factorial function:
- Definition: If z is a positive real number[6], the gamma function $\Gamma(z)$ is defined to be the value of the improper integral $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\,dx$.
  - The gamma function arises naturally in complex analysis, number theory, and combinatorics, in addition to our use here in statistics.
  - By integrating by parts, one may see that $\Gamma(z+1) = z\Gamma(z)$ for all z. Combined with the easy observation that $\Gamma(1) = 1$, we can see that $\Gamma(n) = (n-1)!$ for all positive integers n.
  - The values of the gamma function at half-integers can also be computed explicitly: to compute $\Gamma(1/2)$ we may substitute $u = \sqrt{x}$ to see $\Gamma(1/2) = 2\int_0^\infty e^{-u^2}\,du = \sqrt{\pi}$ as we calculated before when analyzing the normal distribution.
  - Then, by using the identity $\Gamma(z+1) = z\Gamma(z)$, we can calculate $\Gamma(n+\frac{1}{2}) = (n-\frac{1}{2})(n-\frac{3}{2})\cdot\cdots\cdot\frac{1}{2}\sqrt{\pi} = \dfrac{(2n)!}{2^{2n}n!}\sqrt{\pi}$.
- Definition: The t distribution with k degrees of freedom is the continuous random variable T_k whose probability density function $p_{T_k}(x) = \dfrac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\,\Gamma(\frac{k}{2})}\cdot(1+x^2/k)^{-(k+1)/2}$ for all real numbers x.
  - As we will outline in a moment, the t distribution with $n-1$ degrees of freedom is the proper model for the test statistic $\dfrac{\overline{x}-\mu}{S/\sqrt{n}}$.
  - This distribution was first derived in 1876 by Helmert and Lüroth, and then appeared in several other papers.
  - It is often referred to as Student's t distribution, because an analysis was published under the pseudonym "Student" by William Gosset, who because of his work at Guinness did not publish the results under his own name[7].
  - Example: The t distribution with 1 degree of freedom has probability density function $p_{T_1}(x) = \dfrac{1}{\pi(1+x^2)}$, which is the Cauchy distribution.
  - Example: The t distribution with 2 degrees of freedom has probability density function $p_{T_2}(x) = \dfrac{1}{(2+x^2)^{3/2}}$.
  - Example: The t distribution with 3 degrees of freedom has probability density function $p_{T_3}(x) = \dfrac{6\sqrt{3}}{\pi(3+x^2)^2}$.
- We collect a few basic properties of the t distribution:
  - The probability density function of the t distribution is symmetric about 0, since $p_{T_k}(-x) = p_{T_k}(x)$.
  - Per the symmetry about 0, we would typically expect that the expected value of the distribution would be 0. This is true when $k \geq 2$, but in fact the expected value is undefined when $k = 1$ (as we noted previously, the expected value integral for the Cauchy distribution is a non-convergent improper integral).

[6] In fact, this definition also makes perfectly good sense if z is a complex number whose real part is positive (which is why we used the letter z here).

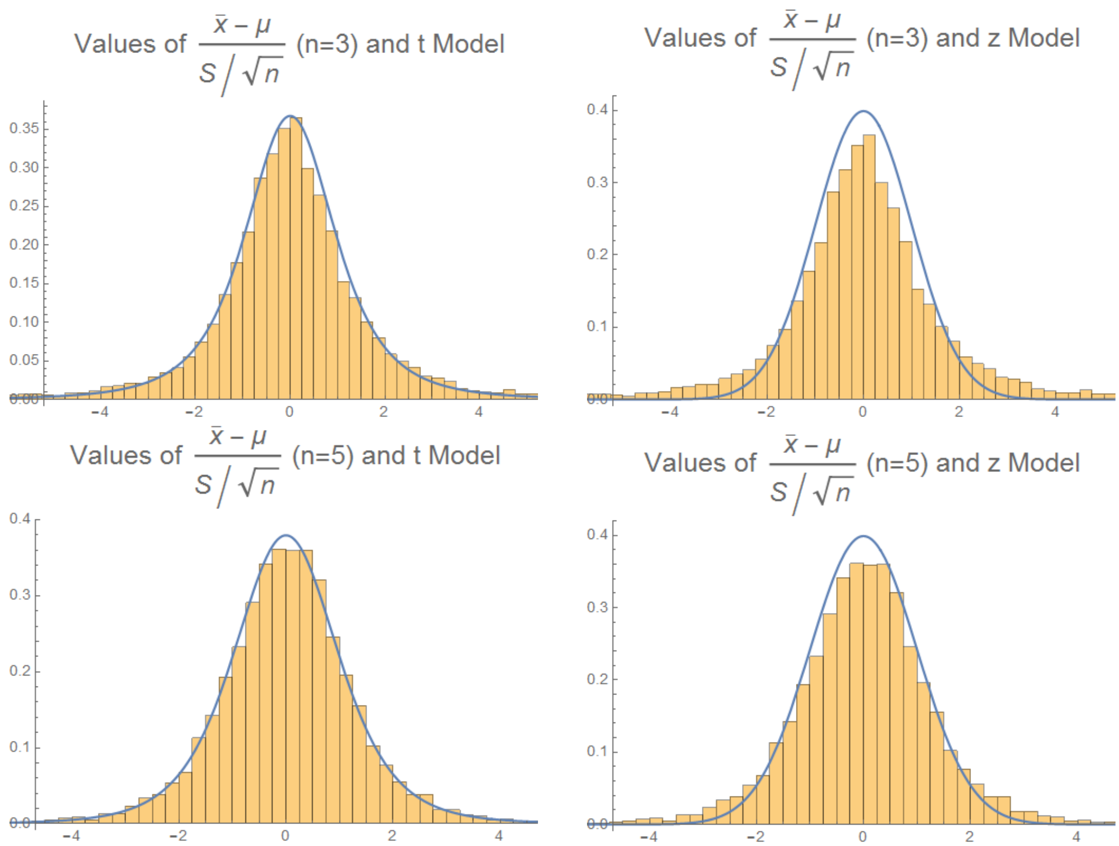
[7] The standard version of the story holds that Guinness wanted all its staff to publish using pseudonyms to protect its brewing methods and related data, since a paper had been previously published by one of its statisticians that inadvertently revealed some of its trade secrets.

- It is more difficult to compute the variance, but by manipulating the integrals appropriately, one can eventually show that the variance is undefined for $k = 1$ (since the expected value is not defined), ∞ for $k = 2$, and $\dfrac{k}{k-2}$ for $k > 2$.
- As $k \to \infty$, the probability density function $p_{T_k}(x)$ approaches the standard normal distribution: using the fact that $\lim_{k\to\infty}(1+y/k)^k = e^y$, we can see that $\lim_{k\to\infty}(1+x^2/k)^{-(k+1)/2} = e^{-x^2/2}$, and thus[8] $\lim_{k\to\infty} \dfrac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(\frac{k}{2})} \cdot (1+x^2/k)^{-(k+1)/2} = \lim_{k\to\infty} \dfrac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(\frac{k}{2})} \cdot \lim_{k\to\infty}(1+x^2/k)^{-(k+1)/2} = \dfrac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

- Theorem (t Distribution): Suppose $n \geq 2$ and that $X_1, X_2, \dots, X_n$ are independent, identically normally distributed random variables with mean μ and standard deviation σ. If $\overline{X} = \dfrac{1}{n}(X_1 + \cdots + X_n)$ denotes the sample mean and $S = \sqrt{\dfrac{1}{n-1}\left[(X_1-\overline{X})^2 + (X_2-\overline{X})^2 + \cdots + (X_n-\overline{X})^2\right]}$ denotes the sample standard deviation, then the distribution of the normalized test statistic $\dfrac{\overline{X}-\mu}{S/\sqrt{n}}$ is the t distribution T_{n-1} with $n-1$ degrees of freedom.
  - We will only outline the full proof, since most of the actual calculations are lengthy and unenlightening.
  - Proof (outline): First, we show that the sample mean $\overline{X}$ and the sample standard deviation S are independent. This is relatively intuitive, but the proof requires the observation that orthogonal changes of variable preserve independence.
  - Next, we compute the probability density functions for the numerator $\overline{X}-\mu$ (which is normal with mean 0 and standard deviation $\sigma/\sqrt{n}$) and the denominator.
  - For the latter, we compute the probability density of $\dfrac{n-1}{\sigma^2}S^2 = \dfrac{1}{\sigma^2}\left[(X_1-\overline{X})^2 + (X_2-\overline{X})^2 + \cdots + (X_n-\overline{X})^2\right] = (\frac{X_1-\mu}{\sigma})^2 + (\frac{X_2-\mu}{\sigma})^2 + \cdots + (\frac{X_n-\mu}{\sigma})^2$: this last expression can be shown to be equal to the sum of the squares of $n-1$ independent standard normal distributions, which is known as a χ^2 distribution (and which we will discuss in more detail later).
  - The pdf of the denominator $\dfrac{1}{S/\sqrt{n}}$ can then be computed using the pdf above, using standard techniques for computing the pdf of a function of a random variable.
  - Then, because since $\overline{X}-\mu$ and $S/\sqrt{n}$ were shown to be independent, the joint pdf for $\overline{X}-\mu$ and $S/\sqrt{n}$ is simply the product of their individual pdfs.
  - Then, at last, we can the probability density function for $\dfrac{\overline{X}-\mu}{S/\sqrt{n}} = (\overline{X}-\mu)\cdot\dfrac{1}{S/\sqrt{n}}$ can be calculated by evaluating an appropriate integral of the joint pdf of $\overline{X}-\mu$ and $S/\sqrt{n}$.

- We can illustrate this result using the sampled data from earlier.
  - Here are the same histograms with $n = 3$ and $n = 5$ as before, comparing the t distribution model to the normal distribution model:

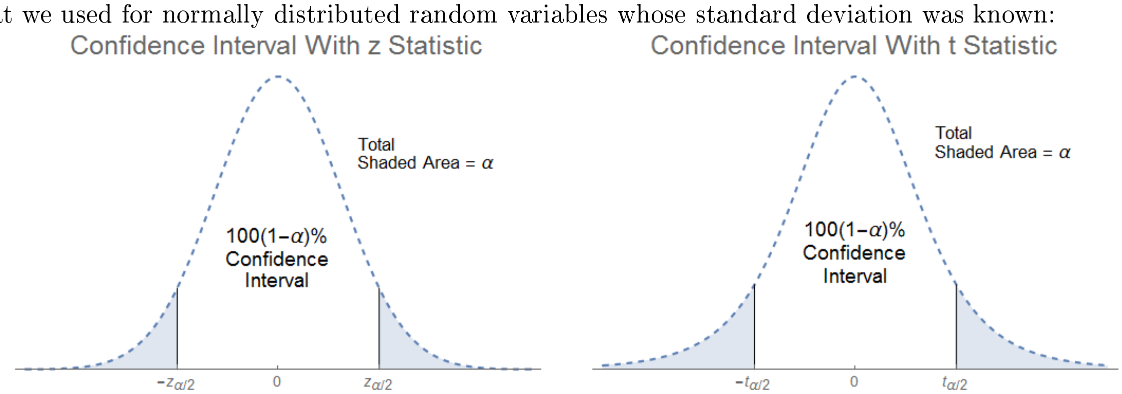
[8]We can compute the limit $\lim_{k\to\infty} \dfrac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(\frac{k}{2})}$ of the constant either using Stirling's approximation $k! \approx k^k e^{-k}\sqrt{2\pi k}$, which also extends to the gamma function, or simply by observing that it must be the constant that makes the resulting function a probability density function.



- ◦ It is quite obvious from the plots that the t distribution is a far superior model for these data samples.

4.2.2 Confidence Intervals Using t Statistics

- Before we discuss how to use the t distribution for hypothesis testing, we will mention how to use t statistics for finding confidence intervals.
  - ◦ The idea is quite simple: if we want to find a confidence interval for the unknown mean of a normal distribution whose standard deviation is also unknown, we can estimate the mean using the t distribution.
  - ◦ Specifically, since the normalized statistic $\dfrac{\overline{x}-\mu}{S/\sqrt{n}}$ is modeled by the t-distribution T_n with $n-1$ degrees of freedom, we can compute a $100(1-\alpha)\%$ confidence interval using a t-statistic in place of the z-statistic that we used for normally distributed random variables whose standard deviation was known:



  - ◦ Like with the normal distribution, we usually want to select the narrowest possible confidence interval, which will also be the one that is symmetric about our sample mean.
  - ◦ If we compute the constant $t_{\alpha/2,n}$ such that $P(-t_{\alpha/2,n} < T_{n-1} < t_{\alpha/2,n}) = 1-\alpha$, then this yields the $100(1-\alpha)\%$ confidence interval $\hat{\mu} \pm t_{\alpha/2,n} \cdot \frac{S}{\sqrt{n}} = (\hat{\mu} - t_{\alpha/2,n}\frac{S}{\sqrt{n}}, \hat{\mu} + t_{\alpha/2,n}\frac{S}{\sqrt{n}})$.

- Using the symmetry of the t distribution, $P(-t_{\alpha/2,n} < T_{n-1} < t_{\alpha/2,n}) = 1-\alpha$ is equivalent to $P(T_{n-1} < -t_{\alpha/2,n}) = \alpha/2$, or also to $P(t_{\alpha/2,n} < T_{n-1}) = 1-(\alpha/2)$, which allows us to compute the value of $t_{\alpha/2,n}$ by evaluating the inverse cumulative distribution function for T_{n-1}.
- We can summarize this discussion as follows:

- Proposition (t Confidence Intervals): A $100(1-\alpha)\%$ confidence interval for the unknown mean μ of a normal distribution with unknown standard deviation is given by $\hat{\mu} \pm t_{\alpha/2,n}\dfrac{S}{\sqrt{n}} = (\hat{\mu} - t_{\alpha/2,n}\dfrac{S}{\sqrt{n}}, \hat{\mu} + t_{\alpha/2,n}\dfrac{S}{\sqrt{n}})$ where n sample points $x_1, \dots, x_n$ are taken from the distribution, $\hat{\mu} = \frac{1}{n}(x_1 + \cdots + x_n)$ is the sample mean, $S = \sqrt{\frac{1}{n-1}[(x_1-\hat{\mu})^2 + \cdots + (x_n - \hat{\mu})^2]}$ is the sample standard deviation, and $t_{\alpha/2,n}$ is the constant satisfying $P(-t_{\alpha/2,n} < T_{n-1} < t_{\alpha/2,n}) = 1-\alpha$.
  - Some specific values of $t_{\alpha/2,n}$ for various common values of n and α are given in the table below (note that the last row for $n = \infty$ represents the entry for the normal distribution):

Entries give $t_{\alpha/2,n-1}$ such that $P(-t_{\alpha/2,n} < T_{n-1} < t_{\alpha/2,n}) = 1-\alpha$								
$n \ \backslash\ 1-\alpha$	50%	80%	90%	95%	98%	99%	99.5%	99.9%
2 (1 df)	1	3.0777	6.3138	12.706	31.820	63.657	127.32	636.62
3 (2 df)	0.8165	1.8856	2.9200	4.3027	6.9646	9.9248	14.089	31.599
4 (3 df)	0.7649	1.6377	2.3534	3.1824	4.5407	5.8409	7.4533	12.924
5 (4 df)	0.7407	1.5332	2.1318	2.7764	3.7469	4.6041	5.5976	8.6103
10 (9 df)	0.7027	1.3830	1.8331	2.2622	2.8214	3.2498	3.6897	4.7809
15 (14 df)	0.6924	1.3450	1.7613	2.1448	2.6245	2.9768	3.3257	4.1405
20 (19 df)	0.6876	1.3277	1.7291	2.0930	2.5395	2.8609	3.1737	3.8834
50 (49 df)	0.6795	1.2991	1.6766	2.0096	2.4049	2.6800	2.9397	3.5004
100 (99 df)	0.6770	1.2902	1.6604	1.9842	2.3646	2.6264	2.8713	3.3915
∞	0.6745	1.2816	1.6449	1.9600	2.3263	2.5758	2.8070	3.2905

- Example: The exam scores in a statistics class are expected to be normally distributed. 15 students' scores are sampled, and the average score is 78.2 points with a sample standard deviation of 9.1 points. Find 90%, 95%, and 99.5% confidence intervals for the true average score on the exam.
  - We have $\hat{\mu} = 78.2$ and $S = 9.1$, so the desired confidence interval is given by $\hat{\mu} \pm t_{\alpha/2,n}(S/\sqrt{n})$, where $n = 15$ here.
  - From the proposition and the table of values below it, we obtain the 90% confidence interval $\hat{\mu} \pm 1.7613 \cdot S/\sqrt{n} = \boxed{(74.06, 82.34)}$, the 95% confidence interval $\hat{\mu} \pm 2.1448 \cdot S/\sqrt{n} = \boxed{(73.16, 83.24)}$, and the 99% confidence interval $\hat{\mu} \pm 3.3257 \cdot S/\sqrt{n} = \boxed{(70.39, 86.01)}$.

- Example: A normal distribution with unknown mean and standard deviation is sampled five times, yielding the values 1.21, 4.60, 4.99, −2.21, and 3.21. Find 80%, 90%, 95%, and 99.9% confidence intervals for the true mean of the distribution. Compare the results to the corresponding confidence intervals for a normal distribution whose standard deviation is the same as this sample estimate.
  - First, we compute the sample mean $\hat{\mu} = \frac{1}{5}(1.21+4.60+4.99-2.21+3.21) = 2.36$ and the sample standard deviation $S = \sqrt{\frac{1}{4}[(1.21-2.36)^2 + (4.60-2.36)^2 + (4.99-2.36)^2 + (-2.21-2.36)^2 + (3.21-2.36)^2} = 2.9523$.
  - The desired confidence interval is given by $\hat{\mu} \pm t_{\alpha/2,n}(S/\sqrt{n})$, where $n = 5$ here.
  - From the proposition and the table of values below it, we obtain the 80% confidence interval $\hat{\mu} \pm 1.5332 \cdot S/\sqrt{n} = \boxed{(0.3357, 4.3843)}$, the 90% confidence interval $\hat{\mu} \pm 2.1318 \cdot S/\sqrt{n} = \boxed{(-0.4546, 5.1746)}$, the 95% confidence interval $\hat{\mu} \pm 2.7764 \cdot S/\sqrt{n} = \boxed{(-1.3057, 6.0257)}$, and the 99.9% confidence interval $\hat{\mu} \pm 8.6103 \cdot S/\sqrt{n} = \boxed{(-9.0083, 13.7283)}$.
  - The confidence interval estimates for a normal distribution are given by using the z-statistic (from the row with $n = \infty$) in place of the t-statistic.

- We obtain the 80% confidence interval $\hat{\mu} \pm 1.2816 \cdot \sigma/\sqrt{n} = \boxed{(0.6679, 4.0521)}$, the 90% confidence interval $\hat{\mu} \pm 1.6449 \cdot \sigma/\sqrt{n} = \boxed{(0.1882, 4.5118)}$, the 95% confidence interval $\hat{\mu} \pm 1.9600 \cdot \sigma/\sqrt{n} = \boxed{(-0.2278, 4.9478)}$, and the 99.9% confidence interval $\hat{\mu} \pm 3.2905 \cdot \sigma/\sqrt{n} = \boxed{(-1.9845, 6.7045)}$.
- Note how much narrower the normal confidence intervals are than the correct t confidence intervals, especially for the larger confidence percentages.
- For example, if we erroneously quoted the 80% normal confidence interval, by using the cdf for the t distribution we can see that it is actually only a 64% confidence interval for the t statistic: quite a bit lower!
- Similarly, if we erroneously quoted the 99.9% normal confidence interval, it would actually only be a 97% confidence interval for the t statistic.

- <u>Example</u>: To estimate the reaction yield, a new chemical synthesis is run three times, giving yields of 41.3%, 52.6%, and 56.1%. Find 50%, 80%, 90%, and 95% confidence intervals for the true reaction yield, under the assumption that the reaction yield is approximately normally distributed.
  - Since the reaction yield is approximately normally distributed, but we do not know the standard deviation, it is appropriate to use the t distribution here.
  - First, we compute the sample average $\hat{\mu} = \frac{1}{3}(41.3\% + 52.6\% + 56.1\%) = 50\%$, and the sample standard deviation $S = \sqrt{\frac{1}{2}\left[(41.3\% - 50\%)^2 + (52.6\% - 50\%)^2 + (56.1\% - 50\%)^2\right]} = 7.7350\%$.
  - Then the desired confidence interval is given by $\hat{\mu} \pm t_{\alpha/2,n}(S/\sqrt{n})$, where here $n = 3$.
  - From the proposition and the table of values below it, we obtain the 50% confidence interval $\hat{\mu} \pm 0.8165 \cdot S/\sqrt{n} = \boxed{(46.35\%, 53.65\%)}$, the 80% confidence interval $\hat{\mu} \pm 1.8856 \cdot S/\sqrt{n} = \boxed{(41.58\%, 58.42\%)}$, the 90% confidence interval $\hat{\mu} \pm 2.9200 \cdot S/\sqrt{n} = \boxed{(36.96\%, 63.04\%)}$, and the 95% confidence interval $\hat{\mu} \pm 4.3027 \cdot S/\sqrt{n} = \boxed{(30.79\%, 69.21\%)}$.

4.2.3 One-Sample t Tests

- In a similar way to how we adapted the procedure for constructing confidence intervals with the normal distribution to construct confidence intervals using t statistics, we can also adapt our procedures for z tests to do hypothesis testing with the t distribution: we call these <u>t tests</u>.

- We first describe <u>one-sample t tests</u>, in which we want to perform a hypothesis test on the unknown mean of a normal distribution with unknown standard deviation, based on an independent sampling of the distribution yielding n values $x_1, x_2, \dots, x_n$.
  - The key difference here is that the standard deviation of the normal distribution is unknown, rather than given to us as is always the case with z tests.
  - As usual with hypothesis tests, we first select appropriate null and alternative hypotheses and a significance level α.
  - Our null hypothesis will be of the form H_0: $\mu = c$ for some constant c that is our hypothesized value for the mean of the normal distribution.
  - We take the test statistic $t = \dfrac{\overline{x} - \mu}{S/\sqrt{n}}$, where $\overline{x}$ is the sample mean and S is the sample standard deviation.
  - From our results about the t distribution, the distribution of the test statistic will be the t distribution T_{n-1} with $n - 1$ degrees of freedom.
  - We can then calculate the p-value based on the alternative hypothesis.
  - If the hypotheses are $H_0 : \mu = c$ and $H_a : \mu > c$, then the p-value is $P(T_{n-1} \geq t)$.
  - If the hypotheses are $H_0 : \mu = c$ and $H_a : \mu < c$, then the p-value is $P(T_{n-1} \leq t)$.
  - If the hypotheses are $H_0 : \mu = c$ and $H_a : \mu \neq c$, then the p-value is $P(|T_{n-1}| \geq |t|) = \begin{cases} 2P(T_{n-1} \geq t) & \text{if } t \geq \mu \\ 2P(T_{n-1} \leq t) & \text{if } t < \mu \end{cases}$.

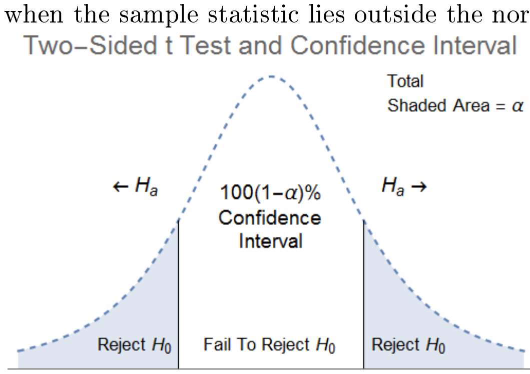
- We then compare the p-value to the significance level and then either reject or fail to reject the null hypothesis, as usual.

- Example: Suppose four values 9, 18, 7, 10 are sampled from a normal distribution with unknown mean and standard deviation. Test at the 20%, 11%, 2%, and 0.6% significance levels that the mean is (i) greater than 10, (ii) greater than 0, (iii) less than 25, (iv) less than 5, (v) equal to 10, and (vi) equal to 16.
  - First, we compute the sample mean $\hat{\mu} = \frac{1}{4}(9 + 18 + 7 + 10) = 11$ and sample standard deviation $S = \sqrt{\frac{1}{3}\left[(9-11)^2 + (18-11)^2 + (7-11)^2 + (10-11)^2\right]} = 4.8305$.
  - For (i), our hypotheses are $H_0 : \mu = 10$, $H_a : \mu > 10$; we want this one-sided alternative hypothesis since the actual sample mean is greater than 10.
  - The value of our test statistic is $t = \dfrac{11-10}{4.8305/\sqrt{4}} = 0.4140$, giving p-value $P(T_{n-1} \geq 0.4140) = 0.3533$.
  - Since this is greater than all four significance levels, we fail to reject the null hypothesis in all cases.
  - For (ii), our hypotheses are $H_0 : \mu = 0$, $H_a : \mu > 0$; we want this one-sided alternative hypothesis since the actual sample mean is greater than 0.
  - The value of our test statistic is $t = \dfrac{11-0}{4.8305/\sqrt{4}} = 4.5544$, giving p-value $P(T_{n-1} \geq 4.5544) = 0.00992$.
  - Since this is less than the first three significance levels, we reject the null hypothesis in those cases. However, it is greater than 0.6%, so we fail to reject the null hypothesis at that significance level.
  - For (iii), our hypotheses are $H_0 : \mu = 25$, $H_a : \mu < 25$; we want this one-sided alternative hypothesis since the actual sample mean is less than 25.
  - The value of our test statistic is $t = \dfrac{11-25}{4.8305/\sqrt{4}} = -5.7966$, giving p-value $P(T_{n-1} \leq -5.7966) = 0.00511$.
  - Since this is less than all four significance levels, we reject the null hypothesis in all cases.
  - For (iv), our hypotheses are $H_0 : \mu = 5$, $H_a : \mu > 5$; we want this one-sided alternative hypothesis since the actual sample mean is greater than 5.
  - The value of our test statistic is $t = \dfrac{11-5}{4.8305/\sqrt{4}} = 2.4842$, giving p-value $P(T_{n-1} \geq 2.4842) = 0.0445$.
  - Since this is less than 20% and 11%, we reject the null hypothesis in those cases. However, it is greater than 2% and 0.6%, so we fail to reject the null hypothesis at those significance levels.
  - For (v), our hypotheses are $H_0 : \mu = 10$, $H_a : \mu \neq 10$; we want this two-sided alternative hypothesis since we are only testing whether the mean equals 10 or not.
  - The value of our test statistic is $t = \dfrac{11-10}{4.8305/\sqrt{4}} = 0.4140$, giving p-value $P(|T_{n-1}| \geq 0.4140) = 2P(T_{n-1} \geq 0.4140) = 0.7067$.
  - Since this is (much!) greater than all of the listed significance levels, we fail to reject the null hypothesis in each case.
  - For (vi), our hypotheses are $H_0 : \mu = 16$, $H_a : \mu \neq 16$; we want this two-sided alternative hypothesis since we are only testing whether the mean equals 16 or not.
  - The value of our test statistic is $t = \dfrac{11-16}{4.8305/\sqrt{4}} = -2.0702$, giving p-value $P(|T_{n-1}| \geq |-2.0702|) = 2P(T_{n-1} \geq 2.0702) = 0.1302$.
  - Since this is less than 20%, we reject the null hypothesis in that case. However, it is greater than 11%, 2% and 0.6%, so we fail to reject the null hypothesis at those significance levels.

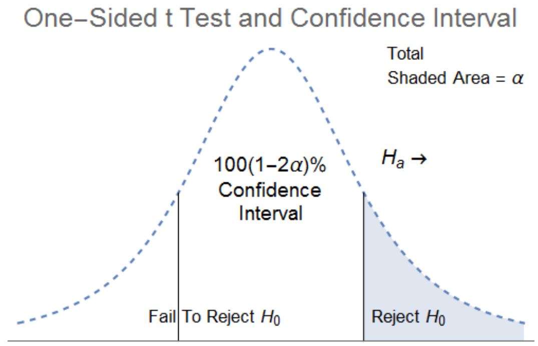
- Example: To estimate the reaction yield, a new chemical synthesis is run three times, giving yields of 41.3%, 52.6%, and 56.1%. It is expected that the yield should be approximately normally distributed. Test at the 20%, 8%, and 1% significance levels the hypotheses that (i) the average yield is above 45%, (ii) the average yield is below 57%, (iii) the average yield is above 64%.

- First, we compute the sample average $\hat{\mu} = \frac{1}{3}(41.3\% + 52.6\% + 56.1\%) = 50\%$, and the sample standard deviation $S = \sqrt{\frac{1}{2}\left[(41.3\% - 50\%)^2 + (52.6\% - 50\%)^2 + (56.1\% - 50\%)^2\right]} = 7.7350\%$.
- For (i), our hypotheses are $H_0 : \mu = 45\%$, $H_a : \mu > 45\%$; we want this one-sided alternative hypothesis since the actual sample mean is greater than 45%.
- The value of our test statistic is $t = \dfrac{50\% - 45\%}{7.7350\%/\sqrt{3}} = 1.1196$, giving p-value $P(T_{n-1} \geq 1.1196) = 0.1896$.
- Since this is less than the first significance level 20%, we reject the null hypothesis in that case. However, it is greater than 8% and 1%, so we fail to reject the null hypothesis at those significance levels.
- For (ii), our hypotheses are $H_0 : \mu = 57\%$, $H_a : \mu < 57\%$; we want this one-sided alternative hypothesis since the actual sample mean is less than 57%.
- The value of our test statistic is $t = \dfrac{50\% - 57\%}{7.7350\%/\sqrt{3}} = -1.5675$, giving p-value $P(T_{n-1} \leq -1.5675) = 0.1288$.
- Since this is less than the first significance level 20%, we reject the null hypothesis in that case. However, it is greater than 8% and 1%, so we fail to reject the null hypothesis at those significance levels.
- For (iii), our hypotheses are $H_0 : \mu = 64\%$, $H_a : \mu < 64\%$; we want this one-sided alternative hypothesis since the actual sample mean is less than 60%.
- The value of our test statistic is $t = \dfrac{50\% - 64\%}{7.7350\%/\sqrt{3}} = -3.1349$, giving p-value $P(T_{n-1} \leq -3.1349) = 0.0442$.
- Since this is less than the first two significance levels 20% and 8%, we reject the null hypothesis in those cases. However, it is greater than 1%, so we fail to reject the null hypothesis at that significance level.

- Just as with z tests, we can also interpret one-sample t tests using confidence intervals. The idea is exactly the same as before, except the underlying distribution is now a t distribution rather than a normal distribution.
  - Since we work with the normalized test statistic, we have to compare to the corresponding normalized confidence interval, which is $(-t_{\alpha/2,n}, t_{\alpha/2,n})$.
  - For a two-sided alternative hypothesis, if we give a $100(1-\alpha)\%$ confidence interval around the mean of a distribution under the conditions of the null hypothesis, then we will reject the null hypothesis with significance level α precisely when the sample statistic lies outside the normalized confidence interval:



  - We can do the same thing with a one-sided alternative hypothesis, but because of the lack of symmetry in the rejection region, we instead need to use a $100(1-2\alpha)\%$ confidence interval to get the correct area:



- Example: The online list prices for four randomly-chosen statistics textbooks are \$193.95, \$171.89, \$221.80, and \$215.32. Assuming that the prices of statistics textbooks are approximately normally distributed, find 80%, 90%, 95%, 98%, 99%, and 99.5% confidence intervals for the average list price of a statistics textbook online. Then test at the 10% and 1% significance levels the hypotheses that (i) the average price is \$200, (ii) the average price is \$230, (iii) the average price is \$275, (iv) the average price is above \$170, (v) the average price is above \$270.
  - The sample mean is $\hat{\mu} = \frac{1}{4}(\$193.95+\$171.89+\$221.80+\$215.32) = \$200.74$ with sample standard deviation $S = \sqrt{\frac{1}{3}\left[(\$193.95-\$200.74)^2+(\$171.89-\$200.74)^2+(\$221.80-\$200.74)^2+(\$215.32-\$200.74)^2\right]} =$ \$22.617.
  - For the confidence levels, we look up or calculate the appropriate t-statistics for the given confidence levels and $n = 4$ (3 degrees of freedom).
  - The confidence intervals are as follows:

α	80%	90%	95%	98%	99%	99.5%
Conf Int (\$)	(182.22, 219.26)	(174.13, 227.35)	(164.75, 236.73)	(149.39, 252.09)	(134.69, 266.79)	(116.46, 285.02)

  - For the hypothesis tests, we just need to identify whether or not the hypothesized average is in the appropriate confidence interval (depending on the alternative hypothesis).
  - For (i), we take $H_0 : \mu = 200$, $H_a : \mu \neq 200$. This is a two-sided confidence interval, and so we want to look at the $100(1-\alpha)\%$ confidence intervals for $\alpha = 0.10$ and $\alpha = 0.01$. Since 200 lies in both the 90% and 99% confidence intervals, we fail to reject the null hypothesis in both cases.
  - Explicitly, the value of the normalized test statistic is $t = \dfrac{\$200.74 - \$200}{\$22.617/\sqrt{3}} = 0.0654$, and so our p-value is $2P(T_{n-1} \geq 0.0654) = 0.9519$: well above the 10% significance level.
  - For (ii), we take $H_0 : \mu = 230$, $H_a : \mu \neq 230$. As before, we want to look at the 90% and 99% confidence intervals.
  - Since 230 lies outside the 90% confidence interval, we reject the null hypothesis at the 10% significance level. But since 230 lies inside the 99% confidence interval, we fail to reject the null hypothesis at the 1% significance level.
  - Explicitly, the value of the normalized test statistic is $t = \dfrac{\$200.74 - \$230}{\$22.617/\sqrt{3}} = -2.5875$, and so our p-value is $2P(T_{n-1} \leq -2.5875) = 0.0813$: below the 10% significance level but above the 1% significance level.
  - For (iii), we take $H_0 : \mu = 275$, $H_a : \mu \neq 275$. As before, we want to look at the 90% and 99% confidence intervals. Since 275 lies outside both the 90% and 99% confidence intervals, we reject the null hypothesis in both cases.
  - Explicitly, the value of the normalized test statistic is $t = \dfrac{\$200.74 - \$275}{\$22.617/\sqrt{3}} = -6.5669$, and so our p-value is $2P(T_{n-1} \leq -6.5669) = 0.00718$: below the 1% significance level.
  - For (iv), we take $H_0 : \mu = 170$, $H_a : \mu > 170$. Now we have a one-sided alternative hypothesis, so we want to look at the $100(1-2\alpha)\%$ confidence interval. Since 170 lies below the 80% confidence interval, we reject the null hypothesis at the 10% significance level. However, 170 does lie inside the 98% confidence interval, so we fail to reject the null hypothesis at the 1% significance level.

- Explicitly, the value of the normalized test statistic is $t = \dfrac{\$200.74 - \$170}{\$22.617/\sqrt{3}} = 2.7184$, and so our p-value is $P(T_{n-1} \geq 2.7184) = 0.02653$: below the 10% significance level but above the 1% significance level.
- For (v), based on the statement we could try taking $H_0 : \mu = 270$, $H_a : \mu > 270$. As above, we want to look at the $100(1 - 2\alpha)\%$ confidence interval.
- Notice that 270 lies outside both the 80% and 98% confidence intervals. However, the confidence intervals themselves are below 270, meaning that our deviation away from the hypothesized value falls into the null hypothesis tail of the distribution, rather than the alternative hypothesis tail.
- Thus, we fail to reject the null hypothesis at either the 10% significance level or the 1% significance level.
- Explicitly, the value of the normalized test statistic is $t = \dfrac{\$200.74 - \$270}{\$22.617/\sqrt{3}} = -6.1247$, and so our p-value is $P(T_{n-1} \geq -6.1247) = 0.9956$.
- Here, we actually ought to have tested the alternative hypothesis $H_a : \mu < 270$, since the sample mean was less than 270. In this case, 270 would still lie outside both the 80% and 98% confidence intervals, but now 270 would land in the alternative hypothesis tail rather than the null hypothesis tail, so we would reject the null hypothesis at both significance levels.
- In that situation, the p-value is $P(T_{n-1} \leq -6.1247) = 0.00438$, which is indeed quite small.

4.2.4 Two-Sample t Tests

- Now that we have treated the situation of one-sample t tests, we discuss the thornier issue of two-sample t tests, in which we want to compare the unknown means of two normally-distributed populations with unknown standard deviations.
- Let us first review the setup for a two-sample z test:
  - Suppose the two populations are labeled A and B, with respective means μ_A and μ_B and population standard deviations σ_A and σ_B. We sample population A a total of n_A times, and population B a total of n_B times.
  - Like with two-sample z tests, we would like to take our test statistic as the difference $\mu_A - \mu_B$ in the two population means.
  - First suppose that we are testing whether $\mu_A = \mu_B$, which we can equivalently phrase as asking whether $\mu_A - \mu_B = 0$.
  - Then, per our assumptions, the sample mean $\hat{\mu}_A$ will be normally distributed with mean μ_A and standard deviation $\sigma_A/\sqrt{n_A}$, while $\hat{\mu}_B$ will be normally distributed with mean μ_B and standard deviation $\sigma_B/\sqrt{n_B}$.
  - The key piece of information here is that since $\hat{\mu}_A$ and $\hat{\mu}_B$ are independent and normally distributed, their difference is also normally distributed with mean $\mu_A - \mu_B$ and standard deviation $\sqrt{\dfrac{\sigma_A^2}{n_A} + \dfrac{\sigma_B^2}{n_B}}$.
  - Therefore, if we are testing the null hypothesis H_0: $\mu_A - \mu_B = c$, then under the assumption of the null hypothesis, our test statistic $\hat{\mu}_A - \hat{\mu}_B$ will be normally distributed with mean c and standard deviation $\sqrt{\dfrac{\sigma_A^2}{n_A} + \dfrac{\sigma_B^2}{n_B}}$.
- Now we analyze the situation of a two-sample t test, in which we do not know the standard deviations σ_A and σ_B.
  - If we do not know σ_A and σ_B, then we must use the sample standard deviation estimates S_A and S_B to estimate the standard deviation of the quantity $\hat{\mu}_A - \hat{\mu}_B$.
  - However, just as we discussed before, using the sample standard deviation in place of the population standard deviation changes the underlying distributions: although $\dfrac{\hat{\mu}_A - \mu_A}{\sigma_A/\sqrt{n_A}}$ has the standard normal distribution, $\dfrac{\hat{\mu}_A - \mu_A}{S_A/\sqrt{n_A}}$ has the distribution of the random variable T_{n_A-1}.

- If we solve for the distribution of $\hat{\mu}_A$, we see it is no longer given by the normal random variable $\mu_A + \frac{\sigma}{\sqrt{n_A}} N_{0,1}$ (normal with with mean μ_A and standard deviation $\sigma_A/\sqrt{n_A}$), but rather a "rescaled" t distribution $\mu_A + \frac{S_A}{\sqrt{n_A}} T_{n_A-1}$.
- Likewise, the random variable $\hat{\mu}_B$ has a rescaled t distribution $\mu_B + \frac{S_B}{\sqrt{n_B}} T_{n_B-1}$.
- Then the quantity $\hat{\mu}_A - \hat{\mu}_B$ is modeled by the random variable $\left[\mu_A + \frac{S_A}{\sqrt{n_A}} T_{n_A-1}\right] - \left[\mu_B + \frac{S_B}{\sqrt{n_B}} T_{n_B-1}\right] = (\mu_A - \mu_B) + \frac{S_A}{\sqrt{n_A}} T_{n_A-1} - \frac{S_B}{\sqrt{n_B}} T_{n_B-1}$.
- Equivalently, $\mu_A - \mu_B$ is modeled by the random variable $(\hat{\mu}_A - \hat{\mu}_B) + \frac{S_A}{\sqrt{n_A}} T_{n_A-1} - \frac{S_B}{\sqrt{n_B}} T_{n_B-1}$, where we absorbed the minus sign into the two t-distributions (since they are symmetric about 0).

- The problem here is that we do not have a nice description of what the difference between two (scaled) t distributions looks like.
  - For normal distributions, we can use the very convenient fact that the sum or difference of normal random variables is also normal; that is not the case for t distributions!
  - In principle, because we know the probability density functions of T_{n_A-1} and T_{n_B-1}, and they are independent, we could calculate the probability density function of the random variable listed above for particular values of all of the parameters.
  - But that does not solve our problem, because we need to write down a test statistic whose distribution is independent of the test parameters (i.e., that does not depend on S_A and S_B, which is not the case in the expression above).
  - In general, solving this problem of finding an appropriate statistic for testing the equality of two sample means from normally distributed random samples is known as the Behrens-Fisher problem. (Various generalizations are also often named with this moniker as well.)

- The first approximation is to construct a pooled standard deviation, much like our approach previously when we did two-sample z tests for binomially distributed data. The approach comes from the following theorem:

- Theorem (t Distribution With Pooled Variance): Suppose that $X_1, \dots, X_n$ are independent and identically normally distributed with mean μ_X and standard deviation σ, and that $Y_1, \dots, Y_m$ are independent and identically normally distributed with mean μ_Y and standard deviation σ. If $\hat{\mu}_X$, $\hat{\mu}_Y$, S_X, and S_Y denote the sample means and sample standard deviations of $\{X_1, \dots, X_n\}$ and $\{Y_1, \dots, Y_n\}$, then for the pooled standard deviation $S_{\text{pool}} = \sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{m+n-2}} = \sqrt{\frac{\sum_{i=1}^n (X_i - \hat{\mu}_X)^2 + \sum_{j=1}^m (Y_j - \hat{\mu}_Y)^2}{m+n-2}}$, the distribution of the test statistic $\frac{(\hat{\mu}_X - \hat{\mu}_Y) - (\mu_X - \mu_Y)}{S_{\text{pool}}\sqrt{\frac{1}{n} + \frac{1}{m}}}$ is T_{m+n-2}, the t distribution with $m+n-2$ degrees of freedom.
  - The idea of the proof is similar to the theorem we proved earlier for the probability density function of the t distribution, and we will in fact reduce to applying the arguments of that theorem in a special case.
  - Proof (outline): Observe that the test statistic is the quotient of $\frac{(\hat{\mu}_X - \hat{\mu}_Y) - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}$ by $\frac{S_{\text{pool}}}{\sigma}$.
  - The first term $\frac{(\hat{\mu}_X - \hat{\mu}_Y) - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}$, from our discussion above, is normally distributed with mean 0 and standard deviation 1 (i.e., its distribution is simply the standard normal).
  - The other term has square $\frac{S_{\text{pool}}^2}{\sigma^2} = \frac{1}{m+n-2}\left[\sum_{i=1}^n \left(\frac{X_i - \hat{\mu}_X}{\sigma}\right)^2 + \sum_{j=1}^m \left(\frac{Y_j - \hat{\mu}_Y}{\sigma}\right)^2\right]$.

- As noted in the proof of our earlier theorem, the sum $\sum_{i=1}^{n}\left(\frac{X_i-\hat{\mu}_X}{\sigma}\right)^2$ can be rewritten as the sum of squares of $n-1$ standard normal variables, and by the same argument, $\sum_{i=1}^{n}\left(\frac{X_i-\hat{\mu}_X}{\sigma}\right)^2$ can be rewritten as the sum of squares of $m-1$ standard normal variables.
- Therefore, the sum of these terms is the sum of squares of $m+n-2$ standard normal variables.
- But as we showed in our earlier theorem, the distribution of a ratio $\dfrac{\overline{x}-\mu}{S/\sqrt{n}} = \dfrac{(\overline{x}-\mu)/(\sigma/\sqrt{n})}{S/\sigma} = \dfrac{N_{0,1}}{S/\sigma}$ is the t distribution T_{n-1} with degrees of freedom equal to the number of squares of standard normal variables summed in the denominator.
- Since there are $m+n-2$ standard normal variables, that means that our quotient of $\dfrac{(\hat{\mu}_X-\hat{\mu}_Y)-(\mu_X-\mu_Y)}{\sigma\sqrt{\dfrac{1}{n}+\dfrac{1}{m}}}$ by $\dfrac{S_{\text{pool}}}{\sigma}$ is t-distributed with $m+n-2$ degrees of freedom, as claimed.

- This theorem gives us an explicit procedure for performing a two-sample t test with a pooled standard deviation where the population variances are assumed to be equal. This test is known as Student's equal-variances t test.
  - First, we select appropriate null and alternative hypotheses and a significance level α.
  - Our null hypothesis will be of the form H_0: $\mu_A-\mu_B=c$ for some constant c that is our hypothesized value for the difference of the means (usually 0).
  - We take the test statistic $t=\dfrac{(\hat{\mu}_A-\hat{\mu}_B)-c}{S_{\text{pool}}\sqrt{\dfrac{1}{n_A}+\dfrac{1}{n_B}}}$, where $S_{\text{pool}}=\sqrt{\dfrac{(n_A-1)S_A^2+(n_B-1)S_B^2}{n_A+n_B-2}}$ is the pooled standard deviation estimate.
  - From our theorem above, the distribution of the test statistic will be the t-distribution $T_{n_A+n_B-2}$ with n_A+n_B-2 degrees of freedom.
  - We can then calculate the p-value based on the alternative hypothesis.
  - If the hypotheses are $H_0:\mu_A-\mu_B=c$ and $H_a:\mu_A-\mu_B>c$, then the p-value is $P(T_{n_A+n_B-2}\geq t)$.
  - If the hypotheses are $H_0:\mu_A-\mu_B=c$ and $H_a:\mu_A-\mu_B<c$, then the p-value is $P(T_{n_A+n_B-2}\leq t)$.
  - If the hypotheses are $H_0:\mu_A-\mu_B=c$ and $H_a:\mu_A-\mu_B\neq c$, then the p-value is $P(|T_{n_A+n_B-2}|\geq|t|)=\begin{cases}2P(T_{n_A+n_B-2}\geq t) & \text{if } t\geq\mu\\ 2P(T_{n_A+n_B-2}\leq t) & \text{if } t<\mu\end{cases}$.
  - We then compare the p-value to the significance level and then either reject or fail to reject the null hypothesis, as usual.

- Before doing an example, we can also give some brief motivation for why these particular choices of parameters (the pooled standard deviation, the test statistic, and the number of degrees of freedom) are logical.
  - The denominator of the test statistic is analogous to the standard deviation $\sqrt{\dfrac{\sigma^2}{n_A}+\dfrac{\sigma^2}{n_B}}$ for the difference of the two normal distributions, and is what we would divide by if we were doing a two-sample z test to get a normalized statistic.
  - The pooled standard deviation can be thought of as arising from the pooled variance: if the two sample means were actually equal to the same number μ, then the pooled variance of the set $\{X_1,X_2,\dots,X_n,Y_1,\dots,Y_m\}$ would be $\dfrac{1}{m+n}\left[(X_1-\mu)^2+\cdots+(X_n-\mu)^2+(Y_1-\mu)^2+\cdots+(Y_m-\mu)^2\right]$.
  - However, since the two sets don't have the same mean, we instead measure them relative to their own means. Furthermore, the resulting variance estimate is biased (for the same reason that the estimate for the sample variance for a single sample is biased), so we must divide by $\dfrac{1}{m+n-2}$ rather than $\dfrac{1}{m+n}$ to unbias it.
  - We can then rewrite the complicated sum above more simply in terms of the sample standard deviations S_X and S_Y: this is precisely the pooled standard deviation S_{pool}.

- The number of degrees of freedom of the t distribution is $m+n-2$, because we have $m+n$ data points, but we lose one degree of freedom by comparing A to its mean, and we lose another by comparing B to its mean.

- Example: A statistics instructor wants to determine whether students do better on exams in a morning class or in an evening class. They randomly sample 11 exams from the morning class, which have an average score of 84 and a sample standard deviation of 13, and compare to a random sample of 11 exams from the evening class, which have an average score of 77 and a sample standard deviation of 9. Assuming that the population variances are equal, test at the 11%, 3%, and 0.7% significance levels the hypotheses that (i) the average score in the morning class is higher, (ii) the average score in the two classes are different, and (iii) the average score in the morning class is at least 2 points higher than the evening class.

  - Since the population variances are assumed to be equal, we use Student's equal-variances t test.
  - For (i), our hypotheses are $H_0 : \mu_m - \mu_e = 0$, $H_a : \mu_m - \mu_e > 0$; we want this one-sided alternative hypothesis since the morning class average is higher than the evening class average.
  - The pooled standard deviation is $S_{\text{pool}} = \sqrt{\dfrac{(11-1)\cdot 13^2 + (11-1)\cdot 9^2}{11+11-2}} = 11.1803$.
  - The test statistic is $t = \dfrac{(84-77)-0}{11.1803 \cdot \sqrt{\dfrac{1}{11}+\dfrac{1}{11}}} = 1.4683$, giving p-value $P(T_{20} \geq 1.4683) = 0.07878$.
  - Since the p-value is less than the first significance level, we reject the null hypothesis in that case. However, it is greater than the other two significance levels, so we fail to reject the null hypothesis in those cases.
  - For (ii), our hypotheses are $H_0 : \mu_m - \mu_e = 0$, $H_a : \mu_m - \mu_e \neq 0$; we want this two-sided alternative hypothesis since now we want only to test whether the scores are equal.
  - The parameters are the same as in (i) above; the only difference is that the p-value is now $2P(T_{20} \geq 1.4683) = 0.1576$.
  - Since the p-value is above all three significance levels, we fail to reject the null hypothesis in each case.
  - For (iii), our hypotheses are $H_0 : \mu_m - \mu_e = 2$, $H_a : \mu_m - \mu_e > 2$; we want this one-sided alternative hypothesis since the morning class actually did score more than two points above the evening class.
  - The pooled standard deviation is the same as before.
  - The test statistic is $t = \dfrac{(84-77)-2}{11.1803 \cdot \sqrt{\dfrac{1}{11}+\dfrac{1}{11}}} = 1.0488$, giving p-value $P(T_{20} \geq 1.0488) = 0.1534$.
  - Since the p-value is above all three significance levels, we fail to reject the null hypothesis in each case.

- In most situations when we are comparing two populations, it is not reasonable to assume that the population variances are the same. For this reason, various unpooled two-sample t tests have been developed.

  - The most popular such test is known as Welch's unequal-variances t test. It is generally more accurate than Student's equal-variances t test (described above) in the situation where the two sample variances are far apart, or when the sample sizes differ drastically.
  - With null hypothesis H_0: $\mu_A - \mu_B = c$, the test statistic is $t = \dfrac{(\hat{\mu}_A - \hat{\mu}_B) - c}{S_{\text{unpool}}}$, where $S_{\text{unpool}} = \sqrt{\dfrac{S_A^2}{n_A} + \dfrac{S_B^2}{n_B}}$ is the natural standard deviation estimate for the difference in the sample means.
  - As we discussed earlier, the resulting test statistic does not actually have an exact distribution we can describe in any convenient way.
  - However, as proven by Welch, it is approximately t-distributed by the t distribution with the number of degrees of freedom equal to the rather complicated formula $df = \dfrac{(S_A^2/n_A + S_B^2/n_B)^2}{\frac{1}{n_A-1}(S_A^2/n_A)^2 + \frac{1}{n_B-1}(S_B^2/n_B)^2}$.

- Most computer systems allow the number of degrees of freedom to be an arbitrary positive real number (in which case one may use the exact value given above); otherwise, such as when using tables, one usually rounds to the nearest integer.

- Welch's result is quite technical, but we can describe roughly where the formula for the degrees of freedom comes from.
  - The idea is to rewrite the quotient in the test statistic (in the same way we did in the theorem earlier) and try to write the denominator ratio as the sum of squares of independent standard normals.
  - This cannot be done exactly, but if it could, we would then be able to find the number of terms by using the method of moments to compare the means and variances of the two expressions.
  - Carefully going through the calculations eventually yields the degree-of-freedom formula given above.

- We can use Welch's unequal-variances t test to compare the sample means for populations whose variances are not assumed to be equal.

- Example: Use Welch's unequal-variances t test with the previous example (morning class with 11 exams of average score 84 and sample standard deviation 13, evening class with 11 of average score 77 and sample standard deviation 9) to test at the 11%, 3%, and 0.7% significance levels the hypotheses that (i) the average score in the morning class is higher, (ii) the average score in the two classes are different, and (iii) the average score in the morning class is at least 2 points higher than the evening class.
  - For (i), our hypotheses are $H_0 : \mu_m - \mu_e = 0$, $H_a : \mu_m - \mu_e > 0$; we want this one-sided alternative hypothesis since the morning class average is higher than the evening class average.
  - The unpooled standard deviation is $S_{\text{unpool}} = \sqrt{\dfrac{13^2}{11} + \dfrac{9^2}{11}} = 4.7673$.
  - The test statistic is $t = \dfrac{(84-77)-0}{4.7673} = 1.4683$, and the number of degrees of freedom is $df = \dfrac{(13^2/11 + 9^2/11)^2}{\frac{1}{11-1}(13^2/11)^2 + \frac{1}{11-1}(9^2/11)^2} = 17.7951$ giving p-value $P(T_{17.7951} \geq 1.4683) = 0.07973$.
  - Since the p-value is less than the first significance level, we reject the null hypothesis in that case. However, it is greater than the other two significance levels, so we fail to reject the null hypothesis in those cases.
  - Remark: Note from before that the pooled p-value estimate was 0.07878, which is quite close.
  - For (ii), our hypotheses are $H_0 : \mu_m - \mu_e = 0$, $H_a : \mu_m - \mu_e \neq 0$; we want this two-sided alternative hypothesis since now we want only to test whether the scores are equal.
  - The parameters are the same as in (i) above; the only difference is that the p-value is now $2P(T_{17.7951} \geq 1.4683) = 0.1594$.
  - Since the p-value is above all three significance levels, we fail to reject the null hypothesis in each case.
  - For (iii), our hypotheses are $H_0 : \mu_m - \mu_e = 2$, $H_a : \mu_m - \mu_e > 2$; we want this one-sided alternative hypothesis since the morning class actually did score more than two points above the evening class.
  - The unpooled standard deviation and degrees of freedom are the same as before.
  - The test statistic is $t = \dfrac{(84-77)-2}{4.7673} = 1.0488$, giving p-value $P(T_{17.7951} \geq 1.0488) = 0.1542$.
  - Since the p-value is above all three significance levels, we fail to reject the null hypothesis in each case.
  - Remark: Note from before that the pooled p-value estimate was 0.1534, which is again quite close.

- We can also adapt the two testing methods to give confidence intervals for the difference of two sample means.
  - Using either Student's or Welch's procedure, we simply compute the appropriate t-statistic for the t distribution with the number of degrees of freedom indicated by the method, and take as the standard deviation either $S = S_{\text{pool}}\sqrt{\dfrac{1}{n_A} + \dfrac{1}{n_B}}$ or $S = S_{\text{unpool}}$ respectively.
  - Then the desired $100(1-\alpha)\%$ confidence interval will be given by $(\hat{\mu}_A - \hat{\mu}_B) \pm t_{\alpha/2,df} S$.

- Example: The salaries of six randomly-chosen male faculty in a university's math department are \$51000, \$90500, \$46000, \$97000, \$108000, \$85000 and the salaries of five randomly-chosen female faculty in the same department are \$56600, \$55000, \$104000, \$70500, \$87000. Test at the 20%, 11%, and 2% significance levels whether the average salary of male faculty is equal to the average salary of female faculty using (i) both Student's equal-variances t test and (ii) Welch's unequal-variances t test. Also, find 80% and 95% confidence intervals for the difference between the average salaries of male and female faculty.

  - For the male faculty ($n_m = 6$) the sample mean is $\hat{\mu}_m = \$79583$ and the sample standard deviation is $S_m = \$25315$, while for the female faculty ($n_f = 5$) the sample mean is $\hat{\mu}_f = \$74620$ and the sample standard deviation is $S_f = \$20875$.
  - Our hypotheses are $H_0 : \mu_m - \mu_f = 0$ and $H_a : \mu_m - \mu_f \neq 0$.
  - For (i), the pooled standard deviation is $S_{\text{pool}} = \sqrt{\dfrac{(6-1)\cdot \$25315^2 + (5-1)\cdot \$20875^2}{6+5-2}} = \$23446$.
  - The test statistic is $t = \dfrac{(\$79583 - \$74620) - 0}{\$23446 \cdot \sqrt{\frac{1}{6}+\frac{1}{5}}} = 0.3496$, giving p-value $2P(T_9 \geq 0.3496) = 0.7347$.
  - The p-value is quite large so we fail to reject the null hypothesis in all cases.
  - For (ii), the unpooled standard deviation is $S_{\text{unpool}} = \sqrt{\dfrac{\$25315^2}{6} + \dfrac{\$20875^2}{5}} = \$13927$.
  - The test statistic is $t = \dfrac{(\$79583 - \$74620) - 0}{\$13927} = 0.3564$, and the number of degrees of freedom is $df = \dfrac{(\$25315^2/6 + \$20875^2/5)^2}{\frac{1}{6-1}(\$25315^2/6)^2 + \frac{1}{5-1}(\$20875^2/5)^2} = 8.9991$ giving p-value $P(T_{8.9991} \geq 0.3564) = 0.7298$.
  - The p-value is again quite large so we fail to reject the null hypothesis in all cases.
  - To compute the 80% and 95% confidence intervals, we need to compute the appropriate t-statistics.
  - We see that the difference in the average salaries is $\hat{\mu}_m - \hat{\mu}_f = \$4963$.
  - For the pooled estimate, there are 9 degrees of freedom, so using a t table or computer yields $t_{\alpha/2,n} = 1.3830$ for the 80% confidence interval and $t_{\alpha/2,n} = 2.2622$.
  - Then $S = S_{\text{pool}}\sqrt{\frac{1}{6}+\frac{1}{5}} = \$14197$, so our 80% confidence interval is $\$4963 \pm 1.3830 \cdot \$14197 = (-\$14672, \$24598)$ and our 95% confidence interval is $\$4963 \pm 2.2622 \cdot \$14197 = (-\$27153, \$37079)$.
  - For the unpooled estimate, there are $df = 8.9991$ degrees of freedom, so using a t table or computer yields $t_{\alpha/2,n} = 1.3830$ for the 80% confidence interval and $t_{\alpha/2,n} = 2.2622$. (The degrees of freedom are so close to 9 in this case it actually doesn't matter if we just round to 9.)
  - Then $S = S_{\text{unpool}} = \$13927$, so our 80% confidence interval is $\$4963 \pm 1.3830 \cdot \$14197 = (-\$14298, \$24225)$ and our 95% confidence interval is $\$4963 \pm 2.2622 \cdot \$14197 = (-\$26542, \$36469)$.
  - We can see here that the two estimates are quite close, since the sample variances are not far away from each other.

- We make a few brief remarks about when to use these various t tests.

  - Most sources still identify Student's t test as the preferred test to use when the sample variances are not far away from each other, and give various approximate rules for deciding what "far away" means (e.g., requiring the variance not to differ by a factor of more than 2).
  - When the sample variances are far apart, Welch's t test tends to give more reliable results (in the sense of having lower type I and type II error probabilities). Even when the sample variances are close, Welch's t test is generally not that much worse than Student's t test (which has a higher power in the situations where it should be used).
  - Neither test is exact (in the sense that it gives exact p-values) except in the case of Student's t test where the population variances are equal. In practice, this means that the type I error rate will deviate somewhat from the desired significance level α.

- Welch's t test tends to maintain a type I error rate closer to the desired significance level α than Student's t test does (although of course there are scenarios in which it is worse).
- It is also worth noting that, as the sample sizes of both groups become large, both tests are very close to the two-sample z test we have previously described. In practice, with samples larger than 100-200 or so, there is a negligible difference between the results of these t tests and the simpler z test.

- We also mention one additional scenario involving t tests and the comparison of two samples, involving matched pairs.
  - In matched-pairs comparisons, we are comparing the means of two sets of paired data.
  - A common situation is to make a before-and-after comparison of measurements taken before applying a treatment to measurements taken after applying the treatment: the goal is to determine whether (and how) the treatment affected the average outcome.
  - Although this scenario involves two data sets, the matched-pairs design means that the initial and later measurements will be correlated, so it is not appropriate to use a two-sample t test.
  - Instead, what we do is compute the difference in the results (for each individual), and use a one-sample t test to compare the average outcome to 0.

- Example: To test whether studying improves students' exam scores, an instructor has 6 students take a pre-assessment, complete a study module, and then a post-assessment. The results are summarized in the table below. Test, at the 9%, 1%, and 0.3% significance levels whether the students' scores improved after studying:

Student	A	B	C	D	E	F
Pre-study	61	71	90	81	55	81
Post-study	74	88	97	80	85	93

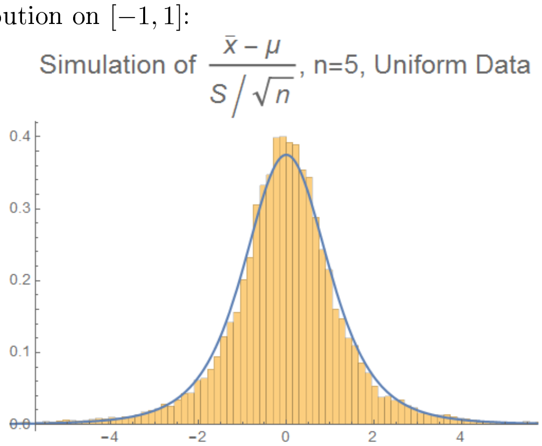
  - Here, we have matched-pair data, because the measurements of the scores are coming from the same students. Since the values in the samples are not independent, but come from matched pairs, we want to use a one-sample t test here.
  - Our hypotheses are $H_0 : \mu_{\text{post}} = \mu_{\text{pre}}$ and $H_a : \mu_{\text{post}} > \mu_{\text{pre}}$, which we can rephrase in terms of the difference in means $\mu_{\text{diff}} = \mu_{\text{post}} - \mu_{\text{pre}}$ as H_0: $\mu_{\text{diff}} = 0$ and $H_a : \mu_{\text{diff}} > 0$.
  - Our test statistic is the difference in means μ_{diff}, which will be t-distributed with $6 - 1 = 5$ degrees of freedom.
  - Our sample data set consists of the six differences of scores $\{13, 17, 7, -1, 30, 12\}$, with mean $\hat{\mu}_{\text{diff}} = 13$ and sample standard deviation $S = 10.3730$, and the value of the sample statistic is $t = \dfrac{\hat{\mu}_{\text{diff}} - 0}{S/\sqrt{n}} = 3.0698$.
  - Thus, the p-values is $P(T_5 \geq 3.0698) = 0.01389$.
  - Since the p-value is below 9% we reject the null hypothesis at that significance level, but since it is above 1% and 0.3% we fail to reject at those significance levels.

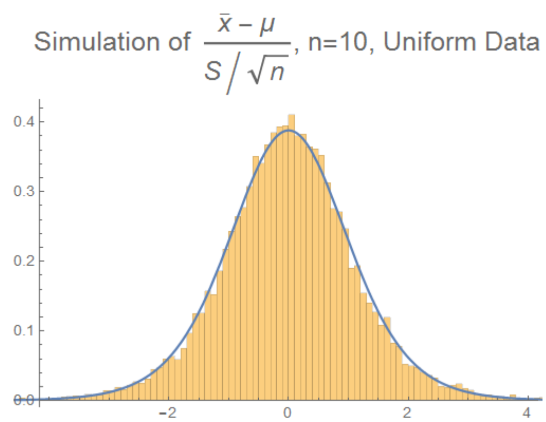
4.2.5 Robustness of t Tests

- We will make a few comments about robustness: the accuracy of the tests when applied to distributions that are not exactly the ones predicted by the model.
  - All of our discussion of z tests and t tests has been predicated on the assumption that the underlying populations we are studying are normally distributed.
  - In reality, except for very rare examples arising in physics with phenomena having exact theoretical models, no population is precisely normally distributed.
  - It is therefore important to understand how well the tests we have developed will perform in situations where the underlying distributions are not exactly normal, but only approximately normal.
  - It is a similar concern to the one that motivated our discussion of the t distribution and t tests: we could simply have tried using z tests but with S in place of σ. The resulting test would then not be exact, but we could hope that it is fairly close.

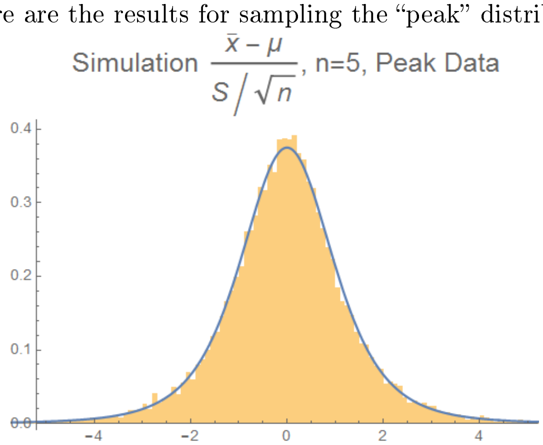
- As we have explained, with small samples using a z place instead of a t test will generally be much less accurate (in the sense that the type I and type II error probabilities will generally be much larger).
- However, with large samples (e.g., n around 100 or more) then the difference between the standard normal distribution and the t distribution is negligible, and so using a z test in place of a t test in such situations does not introduce much error.
- In principle, if we had a different underlying distribution (e.g., a uniform distribution), we could develop analogues of the z test and t test, and in fact there are many other statistical tests that have been developed precisely to allow accurate study of data sets that have very non-normally-shaped distributions.

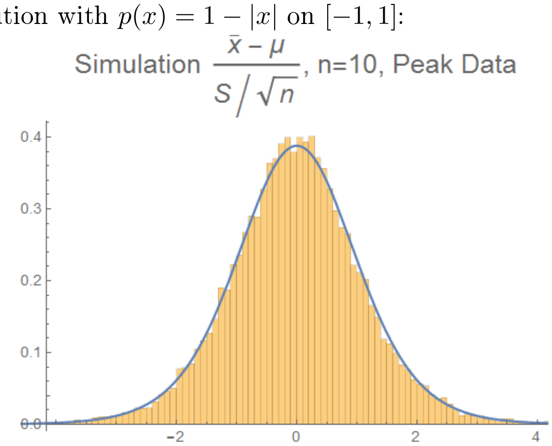
- It turns out that the t test is actually fairly robust, in that it performs fairly well even with distributions that are moderately non-normal.
  - Here are some examples for different simulations of the t-statistic $\dfrac{\overline{x}-\mu}{S/\sqrt{n}}$ for sampling the uniform distribution on $[-1,1]$:



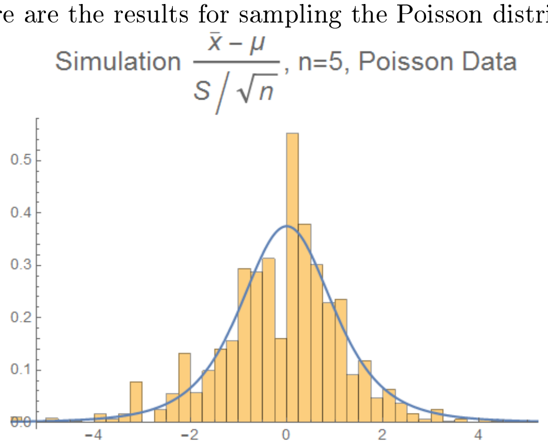


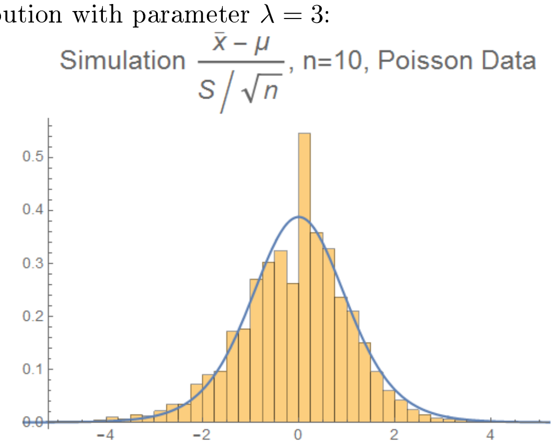
  - Here are the results for sampling the "peak" distribution with $p(x) = 1 - |x|$ on $[-1,1]$:



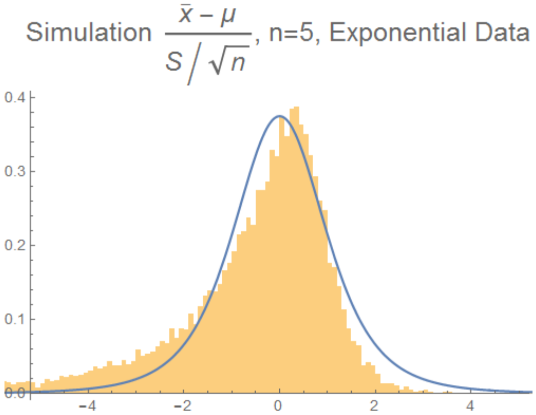


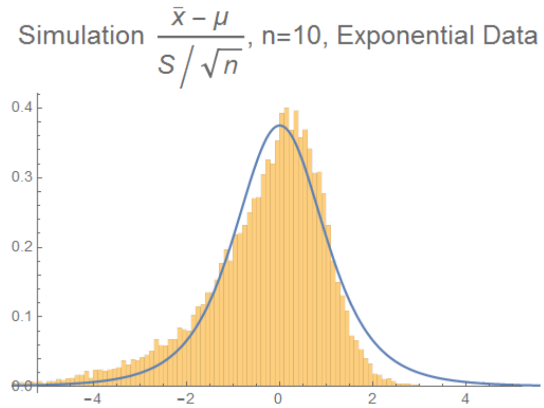
  - Here are the results for sampling the Poisson distribution with parameter $\lambda = 3$:





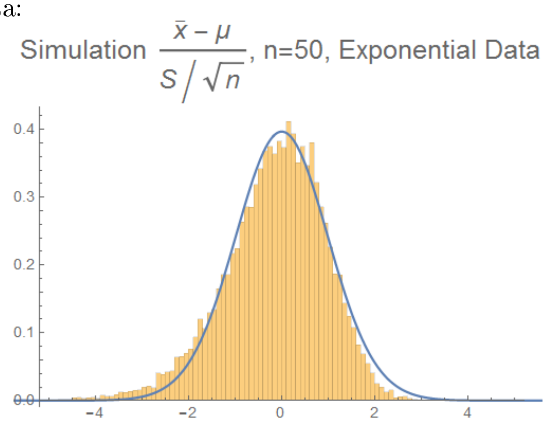
- Here are the results for sampling the exponential distribution with parameter $\lambda = 1/2$:

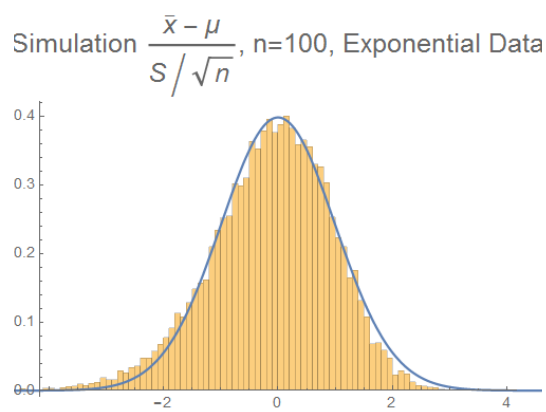




- We can see from the simulations that the t distribution is fairly close for the uniform and peak distributions, it is off a bit for the Poisson, and it is very far off for the exponential.
- The uniform and peak distributions are both symmetric and do not have wide tails.
- The Poisson distribution is more skewed and has a wider tail. It also has the difficulty that it is discrete and that small samples will sometimes yield all identical values (giving a sample standard deviation of 0, yielding an undefined test statistic): this explains the peculiar spike at 0.
- The exponential distribution is very skewed, which causes the resulting test statistic also to be skewed. We can see that the t distribution is not a very good model here even with a sample size $n = 10$.

- In general, the t distribution models the sample statistic $\frac{\overline{x}-\mu}{S/\sqrt{n}}$ well when the underlying distribution is symmetric, but not as well when the underlying distribution is asymmetric or skewed to one side.
  - When the underlying distribution is asymmetric or skewed, using a t test will not generally give reliable results with small sample sizes, and it is necessary to use different tests that are more robust for skewed data.
  - With large sample sizes (the exact definition of large, of course, depends on the scenario, but as we have seen in our discussion of the central limit theorem, usually $n = 100 - 200$ or so is quite sufficient), the central limit theorem will eventually take over and cause the sample average to be approximately normally distributed, even if the original distribution was asymmetric or skewed.
  - In such cases, since the t distribution is so close to the normal distribution, either the t test or the z test will be fairly reliable.
  - For example, here are the results of simulating the test statistic for larger n with exponentially distributed data:





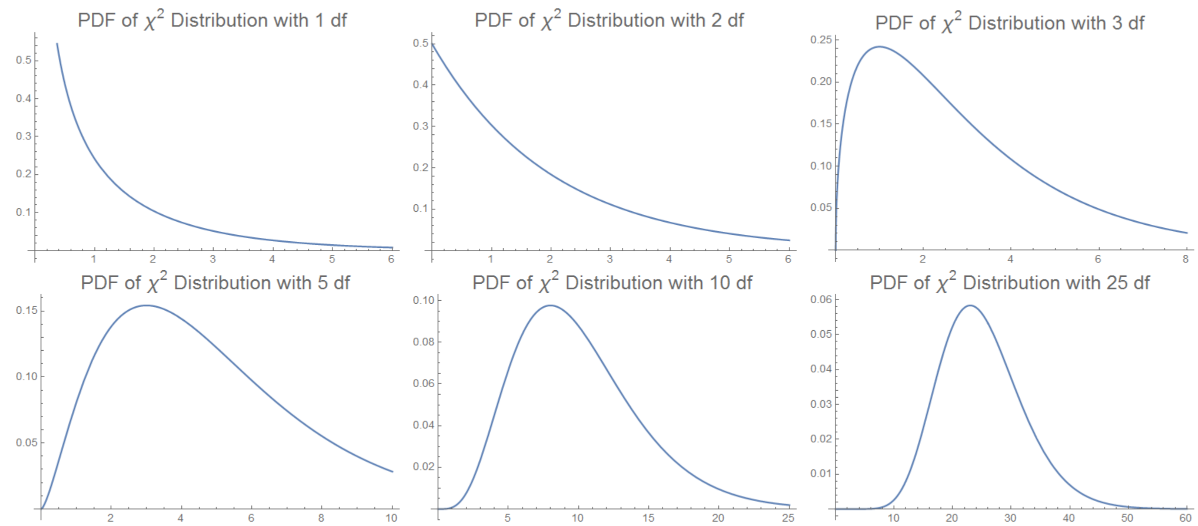
  - We can see here that although there is still some skewness in the histogram, it is now better approximated by the t distribution.

4.3 The χ^2 Distribution and χ^2 Tests

- Our goal in this section is to discuss the χ^2 distribution and two different χ^2 tests, which allow us to expand our hypothesis tests to testing statements about the variance (and standard deviation) of a distribution.
  - All of our hypothesis tests so far have essentially focused on testing statements about the mean of a distribution.
  - However, in certain scenarios, some of which we will discuss now, we might also want to test hypotheses about the variance of a distribution.
  - If the underlying distribution is normal, or obtained as a sum of normal distributions, we can use the χ^2 distribution to construct such tests.

4.3.1 The χ^2 Distribution

- We have previously discussed (at length) methods for constructing confidence intervals for the mean μ of a normally-distrbuted random variable with (known or unknown) standard deviation σ, given a random sample $x_1, \ldots, x_n$ from this normal distribution.
- Our present goal is to apply the same ideas to construct confidence intervals for the variance σ^2 (or equivalently the standard deviation σ) of the normal distribution.
  - Of course, the problem is only interesting when we do not already know σ, which is to say, when we are estimating it from the sample.
  - As we have also discussed at length, the sample variance $S^2 = \dfrac{1}{n-1}\left[(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2\right]$ gives an unbiased estimator for σ^2.
  - In order to construct confidence intervals for σ^2, it is enough to write down the underlying distribution of the statistic $\dfrac{(n-1)S^2}{\sigma^2} = \left(\dfrac{x_1 - \overline{x}}{\sigma}\right)^2 + \cdots + \left(\dfrac{x_n - \overline{x}}{\sigma}\right)^2$.
- Definition: The χ^2 distribution with k degrees of freedom is the continuous random variable Q_k whose probability density function $p_{Q_k}(x) = \dfrac{1}{2^{k/2}\Gamma(k/2)} \cdot x^{(k/2)-1}e^{-x/2}$ for all real numbers $x > 0$.
  - As we will outline in a moment, the χ^2 distribution with $n-1$ degrees of freedom is the proper model for the test statistic $\dfrac{(n-1)S^2}{\sigma^2}$.
  - Example: The χ^2 distribution with 1 degree of freedom has probability density function $p_{Q_1}(x) = \dfrac{1}{\sqrt{2\pi x}}e^{-x/2}$ for $x > 0$.
  - Example: The χ^2 distribution with 2 degrees of freedom has probability density function $p_{Q_1}(x) = \dfrac{1}{2}e^{-x/2}$ for $x > 0$, which is the exponential distribution with parameter $\lambda = 1/2$.
  - Example: The χ^2 distribution with 3 degrees of freedom has probability density function $p_{Q_1}(x) = \dfrac{\sqrt{x}}{\sqrt{2\pi}}e^{-x/2}$ for $x > 0$.
  - It is not hard to show using the probability density function that the χ^2 distribution with k degrees of freedom has mean k and variance $2k$.
  - We also emphasize that the χ^2 distribution, unlike the normal and t distributions, is quite skewed to the right, but the skewness decreases with more degrees of freedom. Here are plots of some of these pdfs:



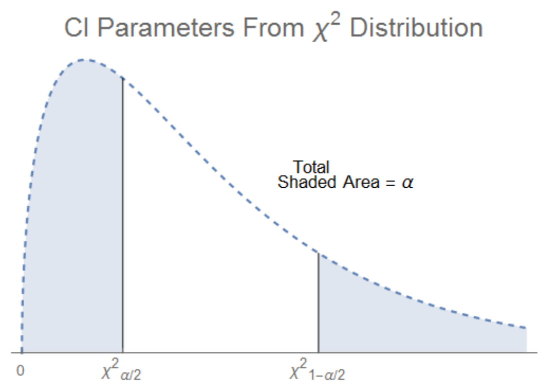
- Proposition (χ^2 Distribution From Normals): If $X_1, \ldots, X_n$ are independent standard normal random variables (i.e., with mean 0 and standard deviation 1), then the random variable $Q_n = X_1^2 + \cdots + X_n^2$ has a χ^2 distribution with n degrees of freedom.
  - The proof is a relatively straightforward calculation using the joint pdf of $X_1, \ldots, X_n$ (which is simply the product of the one-variable pdfs, since these variables are independent).
  - We then just have to set up and evaluate the appropriate n-dimensional integral to compute the probability density function of $Q_n = X_1^2 + \cdots + X_n^2$.
  - We will omit the explicit details of the calculations, although we will mention that the main idea in the computation of the integral is to convert to n-dimensional spherical coordinates.
  - As a corollary, since the χ^2 distribution is obtained by summing independent, identically-distributed random variables, by the central limit theorem it approaches the appropriate normal distribution (with the same mean and variance) as $k \to \infty$.

- Theorem (χ^2 Distribution As Sampling Distribution): Suppose $n \geq 2$ and that $X_1, X_2, \ldots, X_n$ are independent, identically normally distributed random variables with mean μ and standard deviation σ. If $\overline{X} = \dfrac{1}{n}(X_1 + \cdots + X_n)$ denotes the sample mean and $S^2 = \dfrac{1}{n-1}\left[(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2\right]$ denotes the sample variance, then the distribution of the test statistic $\dfrac{(n-1)S^2}{\sigma^2}$ is the χ^2 distribution Q_{n-1} with $n-1$ degrees of freedom.
  - Proof: Let $W = \sum_{i=1}^n \left[\dfrac{X_i - \mu}{\sigma}\right]^2$. Then

$$W = \sum_{i=1}^{n}\left[\frac{(X_i - \overline{X}) + (\overline{X} - \mu)}{\sigma}\right]^2 = \sum_{i=1}^{n}\left[\frac{X_i - \overline{X}}{\sigma}\right]^2 + 2\sum_{i=1}^{n}\left[\frac{X_i - \overline{X}}{\sigma}\right]\left[\frac{\overline{X} - \mu}{\sigma}\right] + \sum_{i=1}^{n}\left[\frac{\overline{X} - \mu}{\sigma}\right]^2.$$

  - In this last expression, the first term is $\dfrac{(n-1)S^2}{\sigma^2}$, the middle term is zero by evaluating the sum (since $\sum_{i=1}^n X_i = \sum_{i=1}^n \overline{X}$), and the last term is $n\left[\dfrac{\overline{X} - \mu}{\sigma}\right]^2 = \left[\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right]^2$. Thus, we see that $W = \dfrac{(n-1)S^2}{\sigma^2} + \left[\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right]^2$.
  - Note that W is the sum of squares of n independent standard normal variables, so it has a χ^2 distribution with n degrees of freedom.
  - Also, S and $\overline{X}$ are independent (as we previously noted in our derivation of the properties of the t distribution).

- Thus, since $\left[\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right]^2$ is the square of a standard normal variable, and S is independent from it, this means the distribution of $\frac{(n-1)S^2}{\sigma^2} = W - \left[\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right]^2$ is given by the sum of squares of $n-1$ independent standard normal variables[9].
- This means $\frac{(n-1)S^2}{\sigma^2}$ has a χ^2 distribution with $n-1$ degrees of freedom, as claimed.

- The theorem above tells us that we can use the χ^2 distribution as a model for the ratio between the sample variance and the population variance, after rescaling appropriately.
  - Thus, we can construct confidence intervals for the population variance using χ^2-statistics and the sample variance.
  - Specifically, since the statistic $\frac{(n-1)S^2}{\sigma^2}$ is modeled by the χ^2 distribution Q_{n-1} with $n-1$ degrees of freedom, we can compute a $100(1-\alpha)\%$ confidence interval using χ^2-statistics in place of the z- and t-statistics that we used for the confidence intervals for the mean of a normally distributed random variable.
  - 

  - Here, we want the parameters $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ to satisfy $P(Q_{n-1} \le \chi^2_{\alpha/2}) = \alpha/2 = P(Q_{n-1} \ge \chi^2_{1-\alpha/2})$, so that the total area in each tail of the distribution is α, leaving an area $1-\alpha$ in the middle.
  - In other words, we have $P(\chi^2_{\alpha/2} \le Q_{n-1} \le \chi^2_{1-\alpha/2}) = 1-\alpha$. Since $\frac{(n-1)S^2}{\sigma^2}$ is χ^2-distributed, this is equivalent to saying that $P(\chi^2_{\alpha/2} \le \frac{(n-1)S^2}{\sigma^2} \le \chi^2_{1-\alpha/2}) = 1-\alpha$.
  - We can then rewrite the above equation to get the desired $100(1-\alpha)\%$ confidence interval for σ:

- Proposition (χ^2 Confidence Intervals): A $100(1-\alpha)\%$ confidence interval for the unknown variance σ^2 of a normal distribution with unknown mean and standard deviation is given by $\left(\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n}}, \frac{(n-1)S^2}{\chi^2_{\alpha/2,n}}\right)$ where n sample points $x_1, \dots, x_n$ are taken from the distribution, $\hat{\mu} = \frac{1}{n}(x_1 + \cdots + x_n)$ is the sample mean, $S = \sqrt{\frac{1}{n-1}[(x_1-\hat{\mu})^2 + \cdots + (x_n-\hat{\mu})^2]}$ is the sample standard deviation, and $\chi^2_{\alpha/2,n}$ and $\chi^2_{1-\alpha/2,n}$ are the constants satisfying $P(Q_{n-1} \le \chi^2_{\alpha/2,n-1}) = \alpha/2 = P(Q_{n-1} \ge \chi^2_{1-\alpha/2,n-1})$ where Q_{n-1} is χ^2-distributed with $n-1$ degrees of freedom.
  - For a $100(1-\alpha)\%$ confidence interval for σ we just take the square root: $\left(\sqrt{\frac{n-1}{\chi^2_{1-\alpha/2,n-1}}}S, \sqrt{\frac{n-1}{\chi^2_{\alpha/2,n-1}}}S\right)$.
  - In order to compute the necessary χ^2 statistics, we must (as with the normal distribution or t distribution) either use a table of values or a computer to evaluate the inverse cumulative distribution function.

[9] Technically, this step requires additional justification: one may make this argument completely precise using moment-generating functions.

Here is a small table of such values:

Inverse-CDF entries give $\chi^2_{\beta,n}$ such that $P(Q_n < \chi^2_{\beta,n}) = \beta$.										
df	0.005	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99	0.995
1	0.0000	0.0002	0.0010	0.0039	0.0158	2.7055	3.8415	5.0239	6.6349	7.8794
2	0.0100	0.0201	0.0506	0.1026	0.2107	4.6052	5.9915	7.3778	9.2103	10.5966
3	0.0717	0.1148	0.2158	0.3518	0.5844	6.2514	7.8147	9.3484	11.3449	12.8382
4	0.2070	0.2971	0.4844	0.7107	1.0636	7.7794	9.4877	11.1433	13.2767	14.8603
5	0.4117	0.5543	0.8312	1.1455	1.6103	9.2364	11.0705	12.8325	15.0863	16.7496
6	0.6757	0.8721	1.2373	1.6354	2.2041	10.6446	12.5916	14.4494	16.8119	18.5476
7	0.9893	1.2390	1.6899	2.1673	2.8331	12.0170	14.0671	16.0128	18.4753	20.2777
8	1.3444	1.6465	2.1797	2.7326	3.4895	13.3616	15.5073	17.5345	20.0902	21.9550
9	1.7349	2.0879	2.7004	3.3251	4.1682	14.6837	16.9190	19.0228	21.6660	23.5894
10	2.1559	2.5582	3.2470	3.9403	4.8652	15.9872	18.3070	20.4832	23.2093	25.1882
15	4.6009	5.2293	6.2621	7.2609	8.5468	22.3071	24.9958	27.4884	30.5779	32.8013
20	7.4338	8.2604	9.5908	10.8508	12.4426	28.4120	31.4104	34.1696	37.5662	39.9968

- We need to compute both $\chi^2_{\alpha/2,n}$ and $\chi^2_{1-\alpha/2,n}$, since the χ^2 distribution is not symmetric.

- Example: A normal distribution is sampled six times yielding values −3, 1, 5, −2, 7, and 8. Find 80%, 90%, and 99% confidence intervals for the standard deviation of the distribution.
  - We first compute the sample mean $\mu = 2.6667$ and sample standard deviation $S = 4.6762$.
  - Since there are 6 values, the number of degrees of freedom for the underlying χ^2 statistics is 5.
  - For the 80% confidence interval, the required values are $\chi^2_{0.9,5} = 9.2364$ and $\chi^2_{0.1,5} = 1.6103$, and so the confidence interval for σ is $\left(\sqrt{\frac{5}{9.2364}} \cdot 4.6762, \sqrt{\frac{5}{1.6103}} \cdot 4.6762\right) = \boxed{(3.4405, 8.2400)}$.
  - For the 90% confidence interval, the required values are $\chi^2_{0.95,5} = 11.0705$ and $\chi^2_{0.05,5} = 1.1455$, and so the confidence interval for σ is $\left(\sqrt{\frac{5}{11.0705}} \cdot 4.6762, \sqrt{\frac{5}{1.1455}} \cdot 4.6762\right) = \boxed{(3.1426, 9.7697)}$.
  - For the 99% confidence interval, the required values are $\chi^2_{0.995,5} = 16.7496$ and $\chi^2_{0.005,5} = 0.4117$, and so the confidence interval for σ is $\left(\sqrt{\frac{5}{16.7496}} \cdot 4.6762, \sqrt{\frac{5}{0.4117}} \cdot 4.6762\right) = \boxed{(2.5549, 16.2962)}$.

- We can also adapt our characterization to give a procedure for doing a hypothesis test about the unknown variance of a normal distribution based on an independent sampling of the distribution yielding n values $x_1, x_2, \ldots, x_n$.
  - As usual with hypothesis tests, we first select appropriate null and alternative hypotheses and a significance level α.
  - Our null hypothesis will be of the form H_0: $\sigma^2 = c$ for some constant c, with an appropriate one-sided or two-sided alternative hypothesis.
  - We take the test statistic $\chi^2 = \frac{(n-1)S^2}{c}$, where S is the sample standard deviation.
  - From our results about the χ^2 distribution, the test statistic is χ^2-distributed with $n-1$ degrees of freedom.
  - If the test is one-sided, we can calculate the p-value based on the alternative hypothesis.
  - If the hypotheses are $H_0 : \sigma^2 = c$ and $H_a : \sigma^2 > c$, then the p-value is $P(Q_{n-1} \geq \chi^2)$.
  - If the hypotheses are $H_0 : \sigma^2 = c$ and $H_a : \sigma^2 < c$, then the p-value is $P(Q_{n-1} \leq \chi^2)$.
  - If the hypotheses are $H_0 : \sigma^2 = c$ and $H_a : \sigma^2 \neq c$, then it is not as obvious how to compute a p-value because of the asymmetry of the χ^2 distribution. We will take the convention of doubling the appropriate one-sided tail probability (as we did with z tests and t tests).

- We then compare the p-value to the significance level and then either reject or fail to reject the null hypothesis, as usual.

- Example: A normal distribution is sampled six times yielding values -3, 1, 5, -2, 7, and 8. Test at the 10% and 1% significance levels the hypothesis that the variance of this distribution is (i) greater than 16, and (ii) less than 225.
  - We calculated the sample standard deviation $S = 4.6762$ earlier, and the number of degrees of freedom is still 5.
  - For (i), our hypotheses are $H_0 : \sigma^2 = 16$ and $H_a : \sigma^2 > 16$, since in fact the sample standard deviation is greater than 16.
  - Our test statistic is $\chi^2 = \dfrac{(n-1)S^2}{\sigma^2} = \dfrac{5 \cdot 4.6762^2}{16} = 6.8333$, and so the p-value is $P(Q_5 > 6.8333) = 0.2333$.
  - Since the p-value is greater than both significance levels, we fail to reject the null hypothesis in both cases.
  - This result is reasonable, since the sample variance is not that much greater than 16. We can also see that $\sigma = 4$ lies well inside the 80% confidence interval we computed earlier.
  - For (ii), our hypotheses are $H_0 : \sigma^2 = 225$ and $H_a : \sigma^2 < 225$, since in fact the sample standard deviation is less than 225.
  - Our test statistic is $\chi^2 = \dfrac{(n-1)S^2}{\sigma^2} = \dfrac{5 \cdot 4.6762^2}{225} = 0.4859$, and so the p-value is $P(Q_5 < 0.4859) = 0.00737$.
  - Since the p-value is less than both significance levels, we fail reject the null hypothesis in both cases.
  - This result is also reasonable, since the sample variance is quite a bit less than 225. We can also see that $\sigma = 15$ lies well outside the 80% confidence interval we computed earlier, but it is inside the 99% confidence interval (corresponding to the fact that the p-value is greater than 0.005).

4.3.2 The χ^2 Test For Goodness of Fit

- We often have reasons to believe that sample data should adhere to a particular shape or distribution. However, in many cases, we need to verify whether a particular model actually fits the data set we have collected.
  - In situations where we have a single variable of interest, we can often use the hypothesis tests we have already developed to test the reasonableness of a model.
  - For example, our z-test for unknown proportion is, very directly, testing whether a particular Bernoulli random variable is a good model for the observed data set (i.e., the collection of successes and failures observed in a sequence of Bernoulli trials).
  - However, most situations have a wider array of data values that we will want to compare to a prediction, and the hypothesis tests we have previously developed are not suitable for that more complicated task.
  - For example, we might want to test whether a die is fair by rolling it many times and tabulating the number of times each of the outcomes 1-6 is observed.
  - Of course, when we roll the die, we do not expect to get a proportion of precisely $1/6$ for each possible outcome (indeed, the distribution of the number of each roll will be binomially distributed).
  - What we want is a way to combine these results into a single test statistic to determine whether all of the results are collectively reasonable or unreasonable.

- The following theorem of Pearson gives a χ^2 test statistic for precisely this type of scenario where values are drawn from a discrete random variable:

- Theorem (χ^2 Goodness of Fit): Suppose that a discrete random variable E has outcomes $e_1, e_2, \dots, e_k$ with respective probabilities $p_1, p_2, \dots, p_k$. If we sample this random variable n times, obtaining the respective outcomes $e_1, e_2, \dots, e_k$ a total of $x_1, x_2, \dots, x_k$ times, then as $n \to \infty$ the random variable $D = \dfrac{(x_1 - np_1)^2}{np_1} + \dfrac{(x_2 - np_2)^2}{np_2} + \cdots + \dfrac{(x_k - np_k)^2}{np_k}$ is χ^2-distributed with $k-1$ degrees of freedom.

- Note that each individual total $x_1, x_2, \ldots, x_k$ is binomially distributed (n trials, success probability p_i). The precise joint distribution of all of these totals is called a multinomial distribution.
- Thus, the quantity np_i represents the expected number of times we would expect to see the outcome e_i if we sample the random variable n times.
- As a practical matter, the approximation will be good whenever the expected frequencies np_i are all at least 5 or so.
- We will not prove this theorem, as the actual details are quite technical (the idea relies on using moment-generating functions).
- However, we can give some brief motivation: since x_i is binomially distributed, in the scenario where the normal approximation to the binomial is good, then x_i is approximately normally distributed with mean np_i and standard deviation $\sqrt{np_i(1-p_i)}$.
- Equivalently, that means $\dfrac{x_i - np_i}{\sqrt{np_i}}$ is approximately normally distributed with mean 0 and standard deviation $\sqrt{1-p_i}$, and so the quantity $(1-p_i)\dfrac{(x_1-np_1)^2}{np_1}$ is approximately χ^2-distributed with 1 degree of freedom.
- Summing over all of the random variables and noting that $(1-p_1)+(1-p_2)+\cdots+(1-p_n) = n-1$ shows that D is essentially the sum of $n-1$ χ^2-distributed variables each with 1 degree of freedom, which is equivalent to saying that it is a χ^2-distributed variable with $n-1$ degrees of freedom.
- This argument is not rigorous because it does not account for the non-independence of the totals; it is only intended as an approximate outline of the main ideas.

- Using this theorem, we can give a hypothesis testing procedure for analyzing the goodness of fit of a model:
  - We take our test statistic as $d = \dfrac{(x_1-np_1)^2}{np_1} + \dfrac{(x_2-np_2)^2}{np_2} + \cdots + \dfrac{(x_k-np_k)^2}{np_k} = \displaystyle\sum_{\text{data}} \dfrac{[\text{Observed} - \text{Expected}]^2}{\text{Expected}}$.
  - Our hypotheses are usually $H_0 : d = 0$ and $H_a : d > 0$, since the value $d = 0$ means the model is perfect and a positive value of d indicates deviation from the model.
  - In order to apply Pearson's result above, we must verify that most of the predicted observation sizes np_i are at least 5. Again, this is a heuristic estimate, so many different versions of a criterion are possible here. We will adopt the convention that at least 80% of the entries should be at least 5 or larger. Another option is to combine some of these small entries into groups that have a predicted size greater than 5.
  - If that is the case, then the test statistic is χ^2-distributed with $k-1$ degrees of freedom, and we can calculate the p-value as $P(Q_{k-1} \geq d)$.
  - We then compare the p-value to the significance level and then either reject or fail to reject the null hypothesis, as usual.
  - Remark: In some situations, we may instead want to test whether a model is "too good to believe" (e.g., if we are investigating whether it is reasonable to think that the data have been falsified or altered to adhere too closely to a model). In those situations we would instead want the hypotheses to be $H_0 : d = c$ and $H_a : d < c$ for (an arbitrary) positive c, and we would compute the p-value instead as $P(Q_{k-1} \leq d)$.

- Example: To test for fairness, a six-sided die is rolled 2000 times, yielding the results below. Test at the 10%, 3%, and 0.4% significance levels whether the die is fair.

Outcome	1	2	3	4	5	6
Observed	354	347	318	312	333	336

  - If the die is fair, we would expect each outcome to occur with probability 1/6, meaning that the expected totals are $2000/6 = 333.\overline{3}$ for each of the six possibilities.
  - Our test statistic is $d = \dfrac{(354-333.\overline{3})^2}{333.\overline{3}} + \dfrac{(347-333.\overline{3})^2}{333.\overline{3}} + \dfrac{(318-333.\overline{3})^2}{333.\overline{3}} + \dfrac{(312-333.\overline{3})^2}{333.\overline{3}} + \dfrac{(333-333.\overline{3})^2}{333.\overline{3}} + \dfrac{(336-333.\overline{3})^2}{333.\overline{3}} = 3.934$.

- We can tabulate the test statistic a bit more conveniently by adding two extra rows to the table:

Outcome	1	2	3	4	5	6
Observed	354	347	318	312	333	336
Expected	$333.\overline{3}$	$333.\overline{3}$	$333.\overline{3}$	$333.\overline{3}$	$333.\overline{3}$	$333.\overline{3}$
$(O-E)^2/E$	$1.281\overline{3}$	$0.560\overline{3}$	$0.705\overline{3}$	$1.365\overline{3}$	$0.000\overline{3}$	$0.021\overline{3}$

- Since there are 6 possible outcomes, there are $6-1=5$ degrees of freedom.
- Thus, the p-value is $P(Q_5 \geq 3.934) = 0.5590$. Since this is well above each of our significance levels, we fail to reject the null hypothesis in each case.
- Remark: The values were obtained by actually simulating a fair die roll, so it is not surprising that the p-value is large!

• Example: To determine whether a pollster is actually conducting their polls, the tenths-place digits from a random sample of 200 of their reported results are tabulated. The results are given below. It is expected that the tenths-place digit from poll percentages of thousands of people should be essentially uniformly distributed. Test at the 10%, 1%, and 0.02% significance levels whether the data appear to adhere to a uniform model.

Tenths Digit	0	1	2	3	4	5	6	7	8	9
Observed	7	26	13	44	25	10	9	41	12	13

- Here are the expected and χ^2-statistic values added to the table:

Tenths Digit	0	1	2	3	4	5	6	7	8	9
Observed	7	26	13	44	25	10	9	41	12	13
Expected	20	20	20	20	20	20	20	20	20	20
$(O-E)^2/E$	8.35	1.8	2.45	28.8	1.25	5	6.05	22.05	3.2	2.45

- Our test statistic is $d = 8.45 + 1.8 + 2.45 + 28.8 + 1.25 + 5 + 6.05 + 22.05 + 3.2 + 2.45 = 81.5$.
- There are 10 possible outcomes hence $10-1=9$ degrees of freedom.
- Thus, the p-value is $P(Q_9 \geq 81.5) = 8.13 \cdot 10^{-14}$. This is extremely small, so we reject the null hypothesis at all of the indicated significance levels.
- Remark: We can see here that the digits 3 and 7 were substantially overused, while 0 was underused. This sort of tendency to overuse certain digits and underuse others is common when humans try to generate lists of random digits.

• Example: It is believed that a Poisson model is appropriate to model the number of collisions at a particular busy intersection in a given week. The collisions are tabulated over a 5-year period (a total of 261 weeks), and the results are given below. Test at the 9% and 1% significance levels the accuracy of the model with parameter (i) $\lambda = 2.2$, and (ii) $\lambda = 2.9$.

# Collisions	0	1	2	3	4	5	6	7+
Observed	17	45	66	55	38	21	12	7

- If the Poisson model is accurate, we would expect the proportion of outcomes yielding d collisions to be $\frac{\lambda^d e^{-\lambda}}{d!}$, so the expected number of occurrences would be 261 times this quantity.
- For (i), here are the results for $\lambda = 2.2$ added to the table:

# Collisions	0	1	2	3	4	5	6	7+
Observed	17	45	66	55	38	21	12	7
Expected	28.92	63.63	69.99	51.32	28.23	12.42	4.55	1.95
$(O-E)^2/E$	4.9128	13.7927	0.2270	0.2635	3.3833	5.9271	12.1744	13.1090

- Here, we have 2 entries out of 8 that are less than 5. This is a sufficiently large percentage that we can use our χ^2 test.
- Our test statistic is $d = 4.9128 + 13.7927 + 0.2270 + 0.2635 + 3.3833 + 5.9271 + 12.1744 + 13.1090 = 53.7898$.
- Since there are 8 possible outcomes, there are $8-1=7$ degrees of freedom.
- Thus, the p-value is $P(Q_7 \geq 53.7898) = 2.588 \cdot 10^{-9}$. Since this is far below our significance levels, we reject the null hypothesis in both cases.

- For (ii), here are the results for $\lambda = 2.9$ added to the table:

# Collisions	0	1	2	3	4	5	6	7+
Observed	17	45	66	55	38	21	12	7
Expected	14.36	41.65	60.39	58.38	42.32	24.55	11.86	7.50
$(O-E)^2/E$	0.4849	0.2699	0.5215	0.1952	0.4414	0.5125	0.0016	0.0327

- Our test statistic is $d = 0.4849 + 0.2699 + 0.5215 + 0.1952 + 0.4414 + 0.5125 + 0.0016 + 0.0327 = 2.4597$.
- As above there are 7 degrees of freedom, so the p-value is $P(Q_7 \geq 2.4597) = 0.9301$. This is quite large, so we fail to reject the null hypothesis.
- Remark: The data set was generated by sampling a Poisson distribution whose actual parameter was $\lambda = 2.9$, so it is not so surprising that the null hypothesis is rejected here!

- In this last example, we could have performed a maximum likelihood estimation for the Poisson parameter to find the ideal λ fitting the observed data.
  - The maximum likelihood estimator for that example ends up being $\hat{\lambda} = 2.7586$, which is not far from the actual value.
  - However, if we do this sort of "tuning" of the model to fit the data, we would expect to get somewhat better agreement than without being able to adjust a parameter to get a better fit.
  - In order to correct for this, if we use a model with r unknown parameters that have been calculated to obtain optimal fit to the observed data, we should use a χ^2 test with $k - 1 - r$ degrees of freedom.
  - Roughly speaking, each unknown parameter removes one degree of freedom from the hypothesis test, since each parameter value we are allowed to choose will allow us to model one additional outcome from the list of k correctly.

4.3.3 The χ^2 Test for Independence

- As a final application of the χ^2 test, we will apply it to study the independence of discrete random variables.
  - Recall that we can test whether two discrete random variables X and Y are independent by checking whether $p_{X,Y}(x, y) = p_X(x) \cdot p_Y(y)$.
  - If we construct a joint probability distribution table, we can check whether X and Y are independent by computing the row and column sums, and then testing whether each entry $p_{X,Y}(x, y)$ in the table is the product of its associated row sum $p_X(x)$ and its associated column sum $p_Y(y)$.
  - Now suppose we are computing the joint distribution table for two random variables X and Y by sampling a population. We would expect the entries in the resulting table (which are now counts of individual observations) to show some random variation in their values away from the true proportion $p_{X,Y}(x, y)$.
  - Thus, if we try to determine whether X and Y are independent using the criterion $p_{X,Y}(x, y) = p_X(x) \cdot p_Y(y)$, it is very unlikely that we would see exact independence.
  - We can, however, adapt Pearson's χ^2 test for goodness-of-fit to give a hypothesis test for independence: the scenario we are describing is essentially identical to the one we just analyzed.

- Theorem (χ^2 Independence): Suppose that the discrete random variables X and Y have outcomes $x_1, \ldots, x_a$ and $y_1, \ldots, y_b$. Suppose that (X, Y) is sampled n times, such that the outcome x_i occurs a proportion p_i times, the outcome y_j occurs a proportion q_j times, and the outcome pair (x_i, y_j) occurs $a_{i,j}$ times for each $1 \leq i \leq a$ and $1 \leq j \leq b$. Then, as $n \to \infty$, the random variable $D = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{(a_{i,j} - np_i q_j)^2}{np_i q_j}$ is χ^2-distributed with $(a-1)(b-1)$ degrees of freedom.
  - The idea is that if X and Y are independent, then $np_i q_j$ is the expected number of times we should obtain the outcomes x_i (probability p_i) and y_j (probability q_j) together.
  - Thus, we are computing the same sum $D = \sum_{\text{data}} \frac{[\text{Observed} - \text{Expected}]^2}{\text{Expected}}$ as before.

- The proof of this result is similar to the one we gave earlier for goodness-of-fit: for large n, each of the ratios $\dfrac{(a_{i,j} - np_i q_j)^2}{np_i q_j}$ will behave like a scaled χ^2 distribution with 1 degree of freedom.

- We will briefly explain the non-obvious fact about why the number of degrees of freedom is $(a-1)(b-1)$.
  - Essentially, the idea is that if we are filling entries into the joint pdf table of X and Y, then all of the entries in the $a \times b$ table are completely determined once we fill in the upper left $(a-1) \times (b-1)$ table, under the presumption that we also know the row and column sums p_i and q_j (because we extract p_i and q_j from the data, we view them as parameters that we have selected).
  - We can fill in all the entries because once we have all but one entry in a given row, we can fill in the last entry since we know the row sum. The same holds true for the columns, so applying this for each row and column (including the bottom row that we just filled) allows us to fill the entire grid.
  - On the other hand, if we have fewer than $(a-1)(b-1)$ entries, we cannot fill the entire grid. Thus, the total number of independent values is $(a-1)(b-1)$, so this is the number of degrees of freedom.
  - An equivalent (and more highbrow) way to make this observation is that the entries in the upper $(a-1) \times (b-1)$ subgrid form a basis for the vector space consisting of the entries of the grid with fixed row and column sums.

- Using this theorem, we can give a hypothesis testing procedure for analyzing the independence of two random variables X and Y:
  - First, we write down the $a \times b$ joint probability distribution table for the observed values of X and Y, and compute the row proportions p_i and column proportions q_j.
  - Then we compute the expected value of each entry $np_i q_j$, and calculate the test statistic as $d = \displaystyle\sum_{i=1}^{a}\sum_{j=1}^{b} \frac{(a_{i,j} - np_i q_j)^2}{np_i q_j} = \sum_{\text{data}} \frac{[\text{Observed} - \text{Expected}]^2}{\text{Expected}}$.
  - We take as our hypotheses $H_0 : d = 0$ and $H_a : d > 0$, since the value $d = 0$ means that the model is perfect (indicating that all of the entries are exactly equal to the predicted value, which means X and Y are independent) and a positive value of d indicates deviation from independence.
  - In order to apply Pearson's result above, we must verify that most of the predicted observation sizes np_i are at least 5. We will adopt the same convention as above, that at least 80% of the entries should be at least 5 or larger.
  - If that is the case, then the test statistic is χ^2-distributed with $(a-1)(b-1)$ degrees of freedom, and we can calculate the p-value as $P(Q_{(a-1)(b-1)} \geq d)$.
  - We then compare the p-value to the significance level and then either reject or fail to reject the null hypothesis, as usual.

- <u>Example</u>: The faculty members in a university mathematics department are either tenure-track or non-tenure-track. These categories are broken down further by gender as indicated below. Test at the 9% and 0.8% significance levels whether the two variables of tenure track status and gender are independent.

Observed	Tenure-Track	Non-Tenure-Track
Male	20	8
Female	4	8

  - There are 30 faculty in total, so we can compute the row and column proportions and then fill in the table of expected values as follows:

Expected	Tenure-Track	Non-Tenure-Track	Proportion
Male	$40 \cdot 0.42 = 16.8$	$40 \cdot 0.28 = 11.2$	0.7
Female	$40 \cdot 0.18 = 7.2$	$40 \cdot 0.12 = 4.8$	0.3
Proportion	0.6	0.4	

  - Then the test statistic is given by $\dfrac{(20-16.8)^2}{16.8} + \dfrac{(8-11.2)^2}{11.2} + \dfrac{(4-7.2)^2}{7.2} + \dfrac{(8-4.8)^2}{4.8} = 5.0794$.

- The total number of degrees of freedom is $(2-1)(2-1) = 1$, so the p-value is given by $P(Q_1 \geq 5.0794) = 0.02421$.
- Since the p-value is below the 9% significance level but above the 0.8% significance level, we reject the null hypothesis in the first case but not in the second case.
- Our interpretation of the test is that we have moderately strong evidence that the variables are not independent.

- Example: A survey is taken of 400 households asking about the number of children and the number of TVs in the household. Test at the 11% and 2% significance levels whether the number of TVs is independent of the number of children.

Observed	0 Children	1 Child	2 Children	3+ Children
0 TVs	10	25	29	16
1 TV	19	88	104	29
2+ TVs	9	24	29	18

  - We compute the row and column proportions and then fill in the table of expected values as follows:

Expected	0 Children	1 Child	2 Children	3+ Children	Proportion
0 TVs	7.6	27.4	32.4	12.6	0.2
1 TV	22.8	82.2	97.2	37.8	0.6
2+ TVs	7.6	27.4	32.4	12.6	0.2
Proportion	0.095	0.3425	0.405	0.1575	

  - Then the test statistic is given by $\frac{(10-7.6)^2}{7.6} + \frac{(25-27.4)^2}{27.4} + \cdots + \frac{(18-12.6)^2}{12.6} = 9.1602$.
  - The total number of degrees of freedom is $(4-1)(3-1) = 6$, so the p-value is given by $P(Q_6 \geq 9.1602) = 0.1648$.
  - Since the p-value is above the 11% and 2% significance levels, we fail reject the null hypothesis in both cases
  - Our interpretation is that we have fairly weak evidence that the variables are not independent: the number of TVs and the number of children do not appear to be far off independence.

- Example: A poll is taken on a trenchant political issue and the support is broken down by age group, as shown below. Test at the 8%, 2%, and 0.3% significance levels whether the level of support is independent of the age group.

Observed	Age 18-29	Age 30-49	Age 50-64	Age 65+
Support	20	13	12	8
Oppose	7	9	14	17

  - There are 100 responses in total, so we can compute the row and column proportions and then fill in the table of expected values as follows:

Expected	Age 18-29	Age 30-49	Age 50-64	Age 65+	Proportion
Support	14.31	11.66	13.78	13.25	0.53
Oppose	12.69	10.34	12.22	11.75	0.47
Proportion	0.27	0.22	0.26	0.25	

  - Then the test statistic is given by $\frac{(20-14.31)^2}{14.31} + \frac{(13-11.66)^2}{11.66} + \cdots + \frac{(17-11.75)^2}{11.75} = 10.057$.
  - The total number of degrees of freedom is $(4-1)(2-1) = 3$, so the p-value is given by $P(Q_3 \geq 10.057) = 0.01809$.
  - Since the p-value is below the 8% and 2% significance levels, we reject the null hypothesis in those cases. However, it is above the 0.3% significance level, so we fail to reject the null hypothesis in that case.
  - Our interpretation of the test is that we have fairly strong evidence that the variables are not independent: the support does appear to depend on the age group.

- We will remark that for 2×2 tables (i.e., the situation of 1 degree of freedom), there does exist an exact test due to Fisher, known as Fisher's exact test, that allows for performing a hypothesis test associated to a given table without the need for using a χ^2 approximation.

- The idea is that if the row and column totals are known, then (as we have noted above) only the single upper-left entry is required to determine the full table.
- Fisher's original example was of the "lady tasting tea", who claimed to be able to decide, solely by the flavor, whether a cup of tea with milk had the milk poured into the tea or the tea poured into the milk.
- Eight cups were poured, four with milk first and four with tea first; the lady tasted each and decided whether the tea or the milk had been poured first. Suppose that the results were as follows:

Observed	Lady: Milk first	Lady: Tea first
Milk poured first	a	b
Tea poured first	c	d

- Under the null hypothesis of random guessing, we assume that the lady would guess exactly 4 cups of each type, since she was aware that there were 4 of each type.
- Thus, to obtain the table above the lady will always guess $a+c$ of the cups to have milk first and $b+d$ to have tea first, so there are a total $\binom{a+b+c+d}{a+c}$ possible tables satisfying this condition.
- To obtain the specific table above, exactly a of the $a+c$ cups the lady says have milk must actually have milk, and exactly d of the cups the lady says have tea must actually have tea. There are $\binom{a+c}{a} \cdot \binom{b+d}{d}$ ways of making these selections, so the total probability of obtaining the given table is $\binom{a+c}{a}\binom{b+d}{d} / \binom{a+b+c+d}{a+c}$.
- We can then compute the probability of obtaining a result at least as extreme (in the direction of accuracy) by summing over the possible tables with upper-left entry at least as large as the observed value.
- For example, if the results had been

Observed	Lady: Milk first	Lady: Tea first
Milk poured first	3	1
Tea poured first	1	3

then the probability of obtaining this precise table is $\binom{4}{3}\binom{4}{3}/\binom{8}{4} = \frac{16}{70} \approx 0.2286$. The only result yielding more correct responses would be the table with entries $(4,0),(0,4)$ which occurs with probability $\binom{4}{4}\binom{4}{4}/\binom{8}{4} = \frac{1}{70} \approx 0.0143$. Thus, the tail probability is the sum $\frac{16}{70} + \frac{1}{70} \approx 0.2429$. We would likely not view this as conclusive evidence.
- In fact, the results of the actual test were that the lady correctly identified all 8 cups. In that case, the probability of obtaining the result by random guessing is $\binom{4}{4}\binom{4}{4}/\binom{8}{4} = \frac{1}{70} \approx 0.0143$: much more compelling!

Well, you're at the end of my handout. Hope it was helpful.