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.

To illustrate hypothesis testing, we develop 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 also apply these ideas to extend our discussion to situations involving the binomial distribution and other distributions that are approximately normal.

We close with a lengthy discussion of topics related to errors in hypothesis testing, and discuss numerous misinterpretations and misuses of various aspects of hypothesis tests.

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 statistical hypothesis testing.
  - 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

- If we are making a binary decision, our first step is to explicitly identify the two possible results.
  - Example: "The coin is fair" versus "The coin is not fair".
  - Example: "The coin has probability 2/3 of landing heads" versus "The coin does not have probability 2/3 of landing heads".
  - Example: "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".
  - Example: "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:
  - Example: "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.
  - Example: "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".
  - Example: "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.

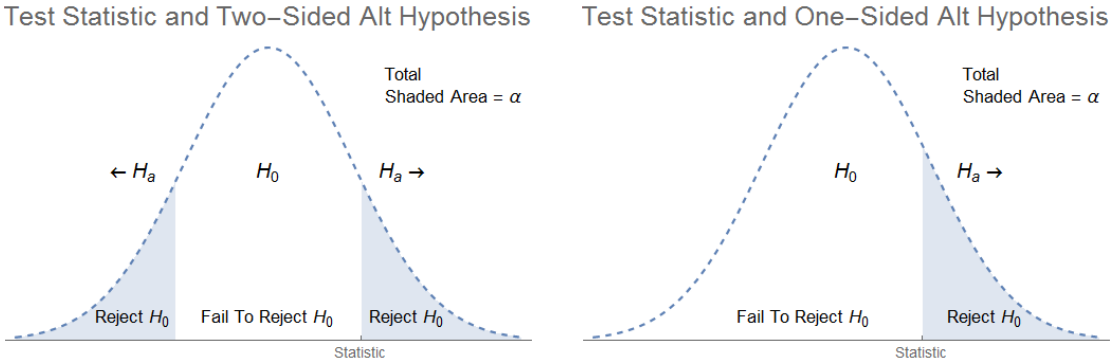
- Example: We wish to test whether a particular baseball player performs better in the playoffs than during the regular season, which we do by comparing the player's hitting percentage h_r during regular-season games to their hitting percentage h_p during playoff games. Give the null and alternative 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.

4.1.2 Test Statistics, p-Values, and Decision Rules

- 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 ("rejection regions"): 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 or rejection regions 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.

4.2 One-Sample and Two-Sample z Tests

- In this section we will illustrate our general framework 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.

4.2.1 One-Sample z Tests

- We begin by discussing the situation of testing whether a normally-distributed random variable has a particular mean: these tests are known as one-sample z tests 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 $\frac{\hat{\mu}-\mu}{\sigma/\sqrt{n}}$, 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: An airline wants to measure how accurate its cross-country travel time predictions are. They believe that their predictions are accurate on average, with a standard deviation of 20 minutes. They collect data from 6 routes, whose errors in travel-time predictions are -39 minutes, $+14$ minutes, -21 minutes, -23 minutes, $+25$ minutes, and -31 minutes (positive values are flights arriving early and negative values are

flights arriving late). Test at the 10% significance level the hypothesis that the true mean error μ is 0 minutes, if (i) the airline is concerned about errors in any direction, and (ii) the airline is only concerned about errors that make flights late.

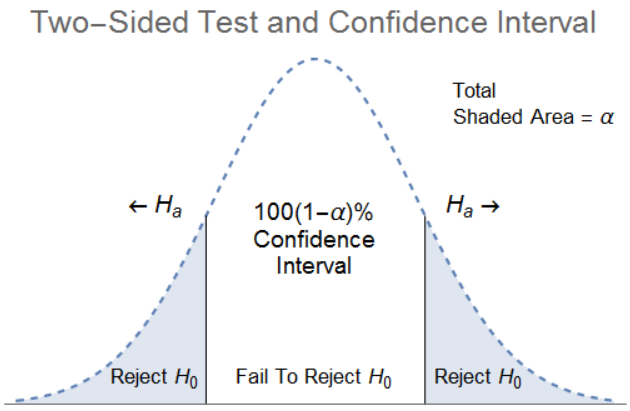
- For (i), our hypotheses are H_0: $\mu = 0$ and $H_a : \mu \neq 0$, since the airline cares about errors in any direction.
- Our test statistic is the average error, which is $-75/6 = -12.5$ minutes.
- The distribution of the test statistic, under the null hypothesis, is normal with mean 0 minutes and standard deviation $30/\sqrt{6} = 8.1650$ minutes.
- 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 $H_a : \mu < 0$, since the airline 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, since the result of the test depends on the specific nature of the alternative hypothesis.
  - 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.

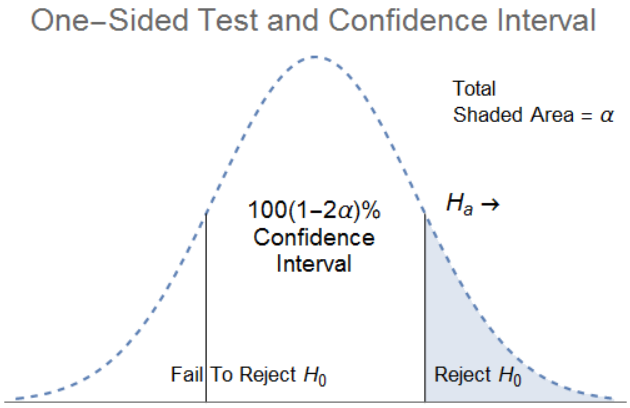
4.2.2 Two-Sample z Tests

- 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 for the one-sample z test 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 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.

4.2.3 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 $P(a \leq B_{n,p} \leq b)$ 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 $P(B_{n,p} \geq k) \approx P(N_{np,\sqrt{np(1-p)}} > k - 0.5)$.
  - If the hypotheses are $H_0 : p = c$ and $H_a : p < c$, the associated p-value is $P(B_{n,p} \leq k) \approx P(N_{np,\sqrt{np(1-p)}} < k + 0.5)$.
  - Finally, if the hypotheses are $H_0 : p = c$ and $H_a : p \neq c$, the associated p-value is $P(|B_{n,p} - nc| \geq |k - nc|) \approx \begin{cases} 2P(N_{np,\sqrt{np(1-p)}} > k - 0.5) & \text{if } k > nc \\ 2P(N_{np,\sqrt{np(1-p)}} < k + 0.5) & \text{if } k < nc \end{cases}$. (For completeness[2], in the trivial case $k = c$ the p-value is 1.)

[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

- 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.

- 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.)

---

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.

- 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 $\sqrt{\dfrac{\hat{p}_A(1-\hat{p}_A)}{n_A} + \dfrac{\hat{p}_B(1-\hat{p}_B)}{n_B}}$ 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 $\sigma_{t-r,\text{pool}} = \sqrt{p_{\text{pool}}(1-p_{\text{pool}})\left[\dfrac{1}{n_A} + \dfrac{1}{n_B}\right]} = 0.03927$.
  - 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 ♣.)

4.3 Errors and Misinterpretations of Hypothesis Testing

- We now discuss various forms of errors in hypothesis testing, and mention some of the common misinterpretations and misuses of hypothesis testing.

4.3.1 Type I and Type II Errors

- 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).

[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.

- 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 $20/\sqrt{400} = 1$.
  - 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 $P(N_{201,1} \leq 202) = P(N_{0,1} \leq 1) = \boxed{0.8413}$: 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 $P(N_{202,1} \leq 202) = P(N_{0,1} \leq 0) = \boxed{0.5}$: 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 $P(N_{203,1} \leq 202) = P(N_{0,1} \leq -1) = \boxed{0.1587}$: 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 $P(N_{204,1} \leq 202) = P(N_{0,1} \leq -2) = \boxed{0.02275}$: 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 $P(N_{205,1} \leq 202) = P(N_{0,1} \leq -3) = \boxed{0.00135}$: 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:

- Example: 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) $n = 1600$.

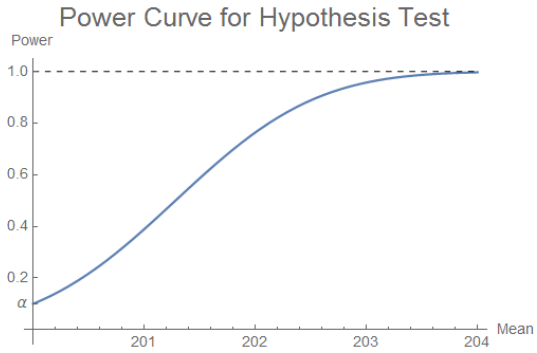
- For (i), to find the probability of a type I error we assume $\mu = 200$. Then the sample mean $\hat{\mu}$ is normally distributed with mean 200 and standard deviation $\sigma = 20/\sqrt{100} = 2$, so the probability of a type I error is $P(N_{200,2} > 202) = P(N_{0,1} > 1) = \boxed{0.1587}$.
- For 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 $P(N_{203,2} \leq 202) = P(N_{0,1} \leq -0.5) = \boxed{0.3085}$.
- 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 $P(N_{203,1} \leq 202) = P(N_{0,1} \leq -1) = \boxed{0.1587}$.
- 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 $P(N_{203,0.5} \leq 202) = P(N_{0,1} \leq -2) = \boxed{0.02275}$.

- 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 $P(N_{203,20/\sqrt{n}} \leq c) = P(N_{0,1} \leq 2.3263 - \frac{3\sqrt{n}}{20})$.
  - For (i), evaluating this probability for $n = 100$ yields the type-II error probability as $P(N_{0,1} \leq 0.8263) = \boxed{0.7957}$.
  - For (ii), evaluating this probability for $n = 400$ yields the type-II error probability as $P(N_{0,1} \leq -0.6737) = \boxed{0.2503}$.
  - For (iii), evaluating this probability for $n = 900$ yields the type-II error probability as $P(N_{0,1} \leq -2.1737) = \boxed{0.01486}$.
  - For (iv), evaluating this probability for $n = 1600$ yields the type-II error probability as $P(N_{0,1} \leq -3.6737) = \boxed{0.0001195}$.

- 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 $P(N_{x,1} \geq 201.282 = P(N_{0,1} \geq 201.282 - x)$, 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.

4.3.2 Misinterpretations and Misuses of Hypothesis Testing

- 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% probability that the obsered result could have arisen by chance if there is no real result to find) 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 p-hacking 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 data dredging: 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. The situation is similar to testing for a rare disease: positive tests are much more likely to be a false positive than a correct positive.

- Example: Suppose we are trying to identify genes that are linked to breast cancer, and hypothesize that about 100 of the 20,000 protein-coding genes in the genome will, if suitably mutated, increase the risk of developing breast cancer. We collect data for all 20,000 of these genes and perform a hypothesis test with $\alpha = 0.05$ for each of them individually. Assuming that the test correctly identifies all 100 of the truly related genes, what is the probability that a gene identified by the test is actually linked to breast cancer?
  - There are 19,900 genes that are not related to breast cancer, so if we use a significance level $\alpha = 0.05$ for each of them individually, we would expect a total of roughly $19900 \cdot 0.05 = 999.5$ type I errors overall (i.e., false positives).
  - Thus, of the $100 + 999.5 = 1099.5$ genes identified by the test, only $100/1099.5 \approx 9.1\%$ of them are actually linked to breast cancer.

- As illustrated by the example, 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 Bonferroni correction, 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 be true creates a substantial waste of resources.
  - The mildest possible consequences are that unnecessary replication studies must be performed to identify and weed out incorrect results. More broadly, until such results are identitied, there is also the likelihood of building additional research on a faulty foundation.

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

- But there are more serious consequences that can occur, such as in medical testing: if, for example, it is purportedly shown using faulty statistics that a new and expensive drug is better at treating a serious disease when it is in fact no better than a placebo (or perhaps actually worse), then tremendous amounts of money and resources could be wasted on manufacturing and delivering the drug.
- This is one of the reasons for the strictness in testing requirements in the development of medical treatments: they exist to ensure that only treatments that are shown to be effective, and that do not have serious side effects, will actually pass the screenings.

- 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 an appropriate significance level α (which balances the possibility of making a type I error against the possibility of making a type II error) and 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.

- There is a great deal more to be said about the proper interpretation and presentation of experimental data, and we will not extend our discussion much further.
  - As a final remark, we note the importance of examining the power of the proposed hypothesis tests[6]. With low power, repeated experiments are likely to yield a wide spread of different p-values, even when the effect size is very small.
  - This provides another reason it is very important to repeat experiments, even ones with a very small p-value or that seem to suggest a large effect size: seemingly-compelling results may merely be an artifact of a test with low power and the presence of an unusual data set.

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

[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 .

[6] https://www.nature.com/articles/nmeth.3288