Math 3081 (Probability and Statistics)

Lecture #18 of 27 $\sim$ August 4th, 2021

Hypothesis Testing and z Tests

- Overview of Hypothesis Testing
- Hypothesis Testing Terminology
- z Tests

This material represents §4.1.1-4.1.2 from the course notes, and problems 1-7 from WeBWorK 6.

Overview of §4, I

We now move into the fourth chapter of the course, which introduces the fundamentals of hypothesis testing.

- In the previous chapter, we discussed methods for estimating parameters, and for constructing confidence intervals that quantify the precision of the estimate.
- The overarching goal of this chapter is to use similar ideas to quantify the plausibility of a particular hypothesis.

Overview of §4, II

In many cases, parameter estimations and confidence intervals can help us assess the plausibility of a hypothesis directly.

- To decide how plausible it is that a given coin is fair, we can flip the coin several times, examine the likelihood of obtaining those 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.
- 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.

Overview of §4, III

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.

- Our goal in this chapter is first to give a formal description of how to set up such hypothesis tests and introduce the relevant terminology.
- We will then illustrate how to set up and interpret hypothesis tests using a variety of z-tests, which are hypotheses about the mean of an (approximately) normally-distributed variable with a known standard deviation.
- Once we have run through some concrete examples to get a sense of how hypothesis testing works, we will discuss some conceptual topics, such as errors in hypothesis tests, the power of a test, statistical significance versus practical significance, and uses and misuses of p-values.

Framework for Hypothesis Tests, I

If we are making a binary decision, our first step is to explicitly identify the two possible results.

Examples:

1. "The coin is fair" versus "The coin is not fair".
2. "The coin has probability 2/3 of landing heads" versus "The coin does not have probability 2/3 of landing heads".
3. "Class 1 has the same average exam score as Class 2" versus "Class 1 does not have the same average score as Class 2".
4. "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 random-variable model. In order to do this, we must formulate the hypothesis in a way that allows us to analyze the underlying variable's distribution.

Framework for Hypothesis Tests, II

In the four examples from the previous slide, only one of the two possible hypotheses provides grounds for a random-variable model:

Framework for Hypothesis Tests, II

In the four examples from the previous slide, only one of the two possible hypotheses provides grounds for a random-variable model:

Examples:

1. "The coin is fair" versus "The coin is not fair".
2. "The coin has probability 2/3 of landing heads" versus "The coin does not have probability 2/3 of landing heads".
3. "Class 1 has the same average exam score as Class 2" versus "Class 1 does not have the same average score as Class 2".
4. "Treatment A is more effective than a placebo" versus "Treatment A is not more effective than a placebo".

In each case, the hypothesis in red provides a specific assumption that allows us to set up a statistical model.

Framework for Hypothesis Tests, II

In the four examples from the previous slide, only one of the two possible hypotheses provides grounds for a random-variable model:

Examples:

1. "The coin is fair" versus "The coin is not fair".
2. "The coin has probability 2/3 of landing heads" versus "The coin does not have probability 2/3 of landing heads".
3. "Class 1 has the same average exam score as Class 2" versus "Class 1 does not have the same average score as Class 2".
4. "Treatment A is more effective than a placebo" versus "Treatment A is not more effective than a placebo".

In each case, the hypothesis in red provides a specific assumption that allows us to set up a statistical model.

Framework for Hypothesis Tests, III

Example: "The coin is fair" versus "The coin is not fair".

- "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 (which is a binomial distribution).
- 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.

Framework for Hypothesis Tests, IV

Example: "The coin has probability 2/3 of landing heads" versus "The coin does not have probability 2/3 of landing heads".

- The coin has probability 2/3 of landing heads" likewise provides us a model we can analyze explicitly (it is binomial, like the previous example).
- However, the hypothesis "The coin does not have probability 2/3 of landing heads" does not give a specific model: there are lots of possible models.

Framework for Hypothesis Tests, V

Example: "Class 1 has the same average exam score as Class 2" versus "Class 1 does not have the same average score as Class 2".

- "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.

Framework for Hypothesis Tests, VI

Example: "Treatment A is more effective than a placebo" versus "Treatment A is not more effective than a placebo".

- "Treatment A is not more effective than a placebo" provides us a model we can analyze, at least if we make the same sorts of presumptions as in the previous example (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 as a placebo" in order to test it using the model.

Null and Alternative Hypotheses, I

Let's get a bit more precise about our terminology:

- 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[1] captures the notion that "something is occurring", while the null hypothesis generally captures the notion that "nothing is occurring".

[1] Of course, there are occasional exceptions. Our hypotheses are actually set up in such a way that the null hypothesis makes a definitive statement about a statistical model we can apply.

Null and Alternative Hypotheses, II

Because of the structure of our statistical approach, we are only able to test the null hypothesis directly. We have two options:

1. Reject the null hypothesis in favor of the alternative hypothesis: we do this in the event that our analysis indicates that the observed data set was too unlikely to arise by random chance.
2. Fail to reject the null hypothesis: we do this in the event that the data set could plausibly have arisen by chance.

You can think of these two options as similar to "guilty" (rejecting the null hypothesis) or "not guilty" (failing to reject the null hypothesis) in a courtroom.

Null and Alternative Hypotheses, III

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.

Null and Alternative Hypotheses, IV

- 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.
- The idea is similar in flavor to our interpretation of confidence intervals: just because a particular value doesn't lie in our confidence interval doesn't mean it cannot be the true value of the parameter, it is just very unlikely.
- Likewise, when we reject the null hypothesis, it does not mean the null hypothesis is impossible, it just means it is very unlikely.

Null and Alternative Hypotheses, V

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.

Examples:

- "The probability of obtaining heads when flipping a coin is $1/2$".
- "The probability of obtaining heads when flipping a coin is $2/3$".
- "The difference in the average scores of Class 1 and Class 2 is zero".
- "The difference between the average outcome using Treatment A and the average outcome using a placebo is zero".

I left the formatting this way just because I thought it looked ridiculous to have this many bad linebreaks.

Null and Alternative Hypotheses, VI

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.

Null and Alternative Hypotheses, VII

Examples:

- "The probability of obtaining heads when flipping a coin is not $1/2$" is a two-sided alternative hypothesis.
- "The probability of obtaining heads when flipping a coin is not $2/3$" is also two-sided.
- "The difference in the average scores of Class 1 and Class 2 is not zero" is two-sided.
- In contrast, "The difference in the average scores of Class 1 and Class 2 is positive" is one-sided.
- "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.

Null and Alternative Hypotheses, VIII

We usually write the null and alternative hypotheses in algebraic shorthand, because algebra is great and we should all aspire to use it as often as possible.

- We usually label the null hypothesis H_0 and the alternative hypothesis H_a.
- Some authors (e.g., the course textbook) instead label the alternative hypothesis H_1. This has the virtue of making perfect sense to computer scientists, but I prefer H_a.

Null and Alternative Hypotheses, IX

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.

Null and Alternative Hypotheses, IX

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.

Null and Alternative Hypotheses, X

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.

Null and Alternative Hypotheses, X

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.

Null and Alternative Hypotheses, XI

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.

Null and Alternative Hypotheses, XI

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.

Null and Alternative Hypotheses, XII

Example: We wish to test whether a particular hockey player performs better in the playoffs than during the regular season, which we do by comparing the player's points-per-game average P_r during regular-season games to their points-per-game average P_p during playoff games. Give the null and alternative hypotheses for this test.

Null and Alternative Hypotheses, XII

Example: We wish to test whether a particular hockey player performs better in the playoffs than during the regular season, which we do by comparing the player's points-per-game average P_r during regular-season games to their points-per-game average P_p during playoff games. Give the null and alternative hypotheses for this test.

- The null hypothesis is H_0: $P_r = P_p$, since this represents the result that the points-per-game averages do not differ.
- The alternative hypothesis is H_a: $P_r < P_p$, since this represents the result that the playoff points-per-game average is better than the regular-season points-per-game average.
- Here, the alternative hypothesis is one-sided.

Test Statistics and Decisions, I

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.

Test Statistics and Decisions, II

In situations where the projected test statistic has a discrete distribution, we could, in principle, compute the exact probability of obtaining the test statistic if the null hypothesis were true.

- However, for continuous distributions, the likelihood of observing any particular data sample will always be zero.
- Furthermore, even for a discrete distribution, this exact probability may be extremely small and not especially useful for deciding how plausible the null hypothesis is.
- For example, if the distribution is binomial with $n = 10000$ and $p = 1/2$, since there are so many possible outcomes, it is hard to decide from a raw probability what "plausible" really means.

Test Statistics and Decisions, III

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".
- However, if H_a is one-sided, we only care about deviation from the null hypothesis in the corresponding direction of H_a.

Once we have computed the *p*-value, we must decide whether we believe this deviation in the test statistic plausibly occurred by chance.

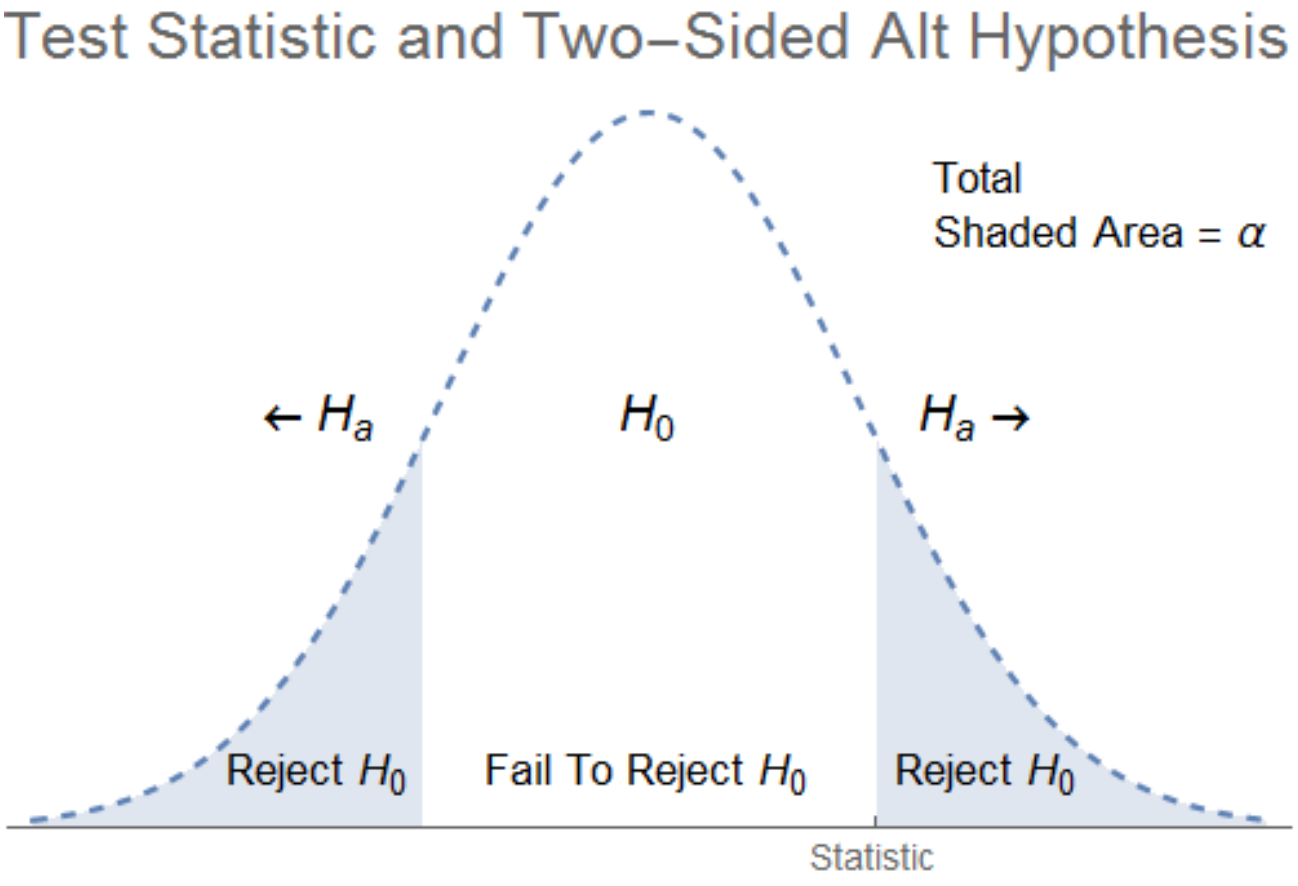
Test Statistics and Decisions, IV

To decide whether to reject the null hypothesis, we adopt a decision rule of the following nature:

- We select a significance level α (many people pick $\alpha = 0.1$, 0.05, 0.01, or 0.001, but we can – and will! – choose lots of other values).
- Then we 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.

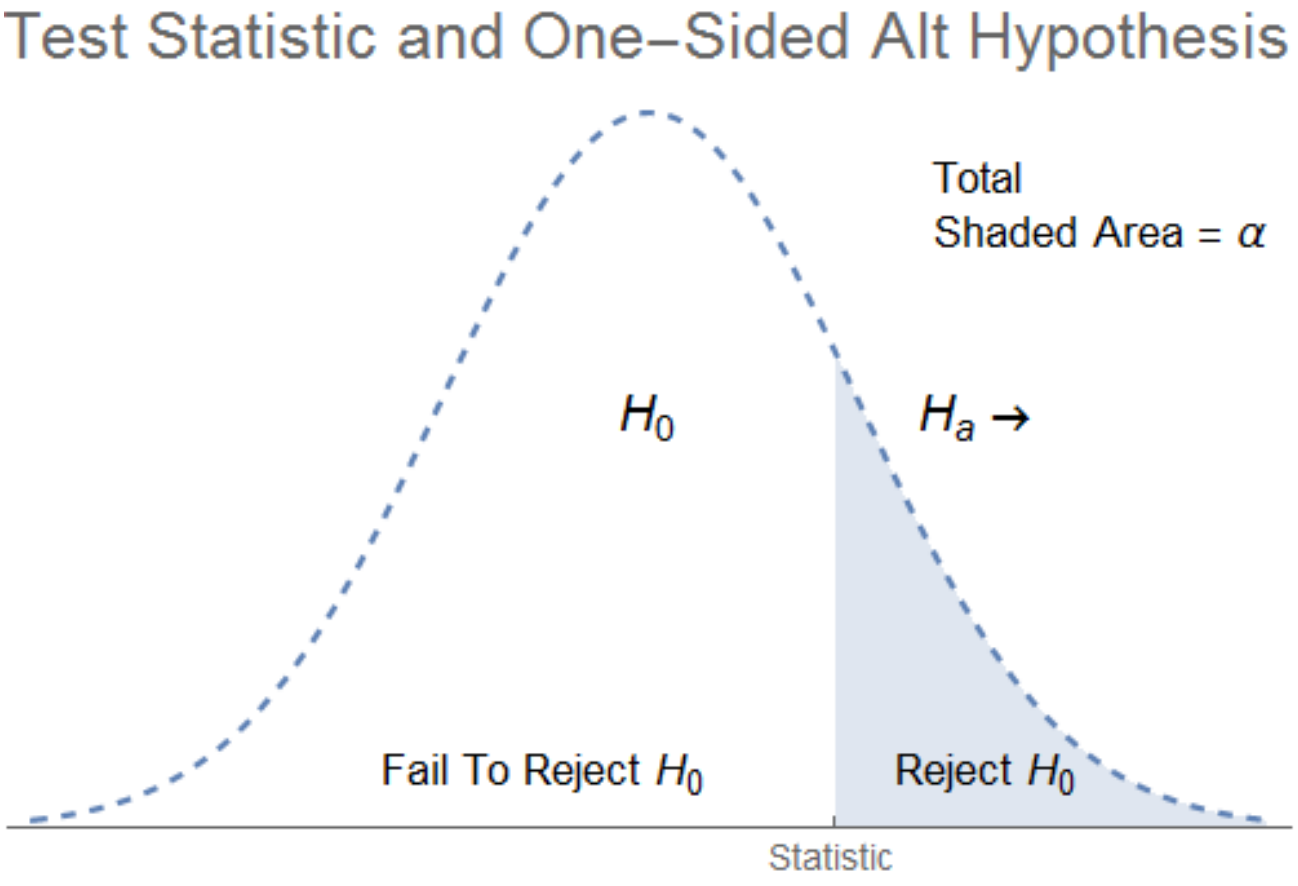
Test Statistics and Decisions, V

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:



Test Statistics and Decisions, VI

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:



Test Statistics and Decisions, VII

Explicitly:

- 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 $1 - \alpha$.
- 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 $1 - \alpha$.

I *highly encourage* you to draw a picture like the ones I just showed to help you remember what the tails of the distribution should look like.

Test Statistics and Decisions, VIII

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.

- Some of the WeBWorK questions may ask you to calculate these things.
- The critical value (or values) will be the values on the borderline, where the total area "more extreme" than the critical value is exactly α.
- The critical region (or regions) will be the range of test statistic values where you would reject the null hypothesis (i.e., past the critical value).

Test Statistics and Decisions, IX

A few minor notes:

- 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.
- Also, for various reasons, 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.
- Since in principle we could work with any test statistic, we will use whichever one is most convenient. (However, the WeBWorK will often require you to use normalized test statistics, since they are unique.)

Hypothesis Testing Procedure

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).

One-Sample z Tests, I

Okay, that was all theoretical. Now let's do some actual hypothesis testing.

- We will start with one of the simplest possible situations: testing whether a normally-distributed quantity with a known standard deviation has a particular mean.
- This is known as a one-sample z test after the letter z traditionally used for normally-distributed quantities.
- Notice that this is the exact same scenario we used two days ago when we started discussing confidence intervals.

One-Sample z Tests, II

Here is the more detailed procedure for a one-sample z-test:

- 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 true population mean μ.
- Then the null hypothesis will be of the form $H_0 : \mu = c$, for some specific value of c.
- Under the assumption that H_0 is true, the test statistic is normally distributed with mean c (the true mean postulated by the null hypothesis) and standard deviation $\sigma/\sqrt{n}$ (which we must be given).

One-Sample z Tests, III

Once we have written down the test statistic, we can compute the p-value:

- 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 $H_a : \mu \neq c$, it 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.

One-Sample z Tests, IV

Finally, once we compute the p-value, we compare it to the significance level α.

- If $p < \alpha$, we reject the null hypothesis. Our interpretation is that the test statistic is so far away from the prediction that it could not reasonably have happened by chance (for "reasonable" as defined by the significance level α).
- If $p \geq \alpha$, we fail to reject the null hypothesis. Our interpretation is that the test statistic is close enough from the prediction that it could reasonably have happened by chance (again, for "reasonable" as defined by the significance level α).

One-Sample z Tests, V

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. Assume that the standard deviation for the new method is also 10 widgets for a 9-hour shift.

1. State the null and alternative hypotheses.
2. Identify the test statistic and its distribution.
3. Calculate the p-value.
4. Test the claim at the 10%, 5%, and 1% levels of significance.

One-Sample z Tests, VI

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. Assume that the standard deviation for the new method is also 10 widgets for a 9-hour shift.

1. State the null and alternative hypotheses.

One-Sample z Tests, VI

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. Assume that the standard deviation for the new method is also 10 widgets for a 9-hour shift.

1. State the null and alternative hypotheses.

- 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 $H_a : \mu > 180$.

One-Sample z Tests, VII

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. Assume that the standard deviation for the new method is also 10 widgets for a 9-hour shift.

2. Identify the test statistic and its distribution.

One-Sample z Tests, VII

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. Assume that the standard deviation for the new method is also 10 widgets for a 9-hour shift.

2. Identify the test statistic and its distribution.

- Our test statistic is $z = 197$ widgets.
- By assumption, the number of widgets on a shift is normally distributed with standard deviation 10 widgets.
- Under the assumption of the null hypothesis, the mean will be 180 widgets.

One-Sample z Tests, VIII

Example: The production of an assembly line is normally distributed, with mean 180 widgets and standard deviation 10 widgets for a 9-hour shift. A new method is used for a 9-hour shift and produces a total of 197 widgets.

3. Calculate the p-value.

One-Sample z Tests, VIII

Example: The production of an assembly line is normally distributed, with mean 180 widgets and standard deviation 10 widgets for a 9-hour shift. A new method is used for a 9-hour shift and produces a total of 197 widgets.

3. Calculate the p-value.

- Our test statistic is normally distributed with $\mu = 180$ and $\sigma = 10$, and our observed value is $z = 197$.
- 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 we can calculate the p-value $P(N_{180,10} \geq 197) = P(N_{0,1} \geq 1.7) = 0.04457$.

One-Sample z Tests, IX

Example: The production of an assembly line is normally distributed, with mean 180 widgets and standard deviation 10 widgets for a 9-hour shift. A new method is used for a 9-hour shift and produces a total of 197 widgets.

4. Test the claim at the 10%, 5%, and 1% levels of significance.

- We have $p = 0.04457$.

One-Sample z Tests, IX

Example: The production of an assembly line is normally distributed, with mean 180 widgets and standard deviation 10 widgets for a 9-hour shift. A new method is used for a 9-hour shift and produces a total of 197 widgets.

4. Test the claim at the 10%, 5%, and 1% levels of significance.

- We have $p = 0.04457$.
- At the 10% level of significance ($\alpha = 0.10$), we have $p < \alpha$: the result is statistically significant, and we reject the null hypothesis.
- At the 5% level of significance ($\alpha = 0.05$), we have $p < \alpha$: the result is statistically significant, and we reject the null hypothesis.
- At the 1% level of significance ($\alpha = 0.01$), we have $p > \alpha$: the result is not statistically significant, and we fail to reject the null hypothesis.

One-Sample z Tests, X

Example: The production of an assembly line is normally distributed, with mean 180 widgets and standard deviation 10 widgets for a 9-hour shift. A new method is used for a 9-hour shift and produces a total of 197 widgets.

4. Test the claim at the 10%, 5%, and 1% levels of significance.

- Overall, since we reject the null hypothesis at the 10% and 5% levels of significance, we have fairly strong evidence that the new method is actually better than the old one.
- However, it is not significant at the 1% level, so it is not incredibly strong.
- As you can see, the p-value is a more fine-grained measure of the strength of the evidence against the null hypothesis. (But of course, you have to get a sense of what a p-value of 0.04457 really tells you!)

One-Sample z Tests, XV

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 80.05 points, while in Class B's sample, the average is 81.76 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

1. Whether the average in Class A is equal to 80 points.
2. Whether the average in Class B is equal to 80 points.
3. Whether the average in Class B is greater than 81 points.

One-Sample z Tests, XVI

Example: A sample from Class A has 64 students and average score 80.05 points. A sample from Class B has 100 students and average 81.76 points. Assume the standard deviation is known to be 5 points. Test at the 10% and 3% significance levels

1. Whether the average in Class A is equal to 80 points.

One-Sample z Tests, XVI

Example: A sample from Class A has 64 students and average score 80.05 points. A sample from Class B has 100 students and average 81.76 points. Assume the standard deviation is known to be 5 points. Test at the 10% and 3% significance levels

1. Whether the average in Class A is equal to 80 points.

- Let μ_A and μ_B be the respective class averages.
- 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 $z = 80.05$ 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.

One-Sample z Tests, XVII

Example: A sample from Class A has 64 students and average score 80.05 points. A sample from Class B has 100 students and average 81.76 points. Assume the standard deviation is known to be 5 points. Test at the 10% and 3% significance levels

1. Whether the average in Class A is equal to 80 points.

- 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.

One-Sample z Tests, XVII

Example: A sample from Class A has 64 students and average score 80.05 points. A sample from Class B has 100 students and average 81.76 points. Assume the standard deviation is known to be 5 points. Test at the 10% and 3% significance levels

1. Whether the average in Class A is equal to 80 points.

- 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 $P(|N_{80,0.625} - 80| \geq |80.05 - 80|) = 2 \cdot P(N_{80,0.625} \geq 80.05) = 0.9362$.
- 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.

One-Sample z Tests, XVIII

Example: A sample from Class A has 64 students and average score 80.05 points. A sample from Class B has 100 students and average 81.76 points. Assume the standard deviation is known to be 5 points. Test at the 10% and 3% significance levels

2. Whether the average in Class B is equal to 80 points.

One-Sample z Tests, XVIII

Example: A sample from Class A has 64 students and average score 80.05 points. A sample from Class B has 100 students and average 81.76 points. Assume the standard deviation is known to be 5 points. Test at the 10% and 3% significance levels

2. Whether the average in Class B is equal to 80 points.

- Our hypotheses are H_0: $\mu_B = 80$ and $H_a : \mu_B \neq 80$, as (like before) we do not care about a particular direction of error.
- Our test statistic is $z = 81.76$ 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.

One-Sample z Tests, XIX

Example: A sample from Class A has 64 students and average score 80.05 points. A sample from Class B has 100 students and average 81.76 points. Assume the standard deviation is known to be 5 points. Test at the 10% and 3% significance levels

2. Whether the average in Class B is equal to 80 points.

- 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.

One-Sample z Tests, XIX

Example: A sample from Class A has 64 students and average score 80.05 points. A sample from Class B has 100 students and average 81.76 points. Assume the standard deviation is known to be 5 points. Test at the 10% and 3% significance levels

2. Whether the average in Class B is equal to 80 points.

- 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 $P(|N_{80,0.5} - 80| \geq 1.16) = 2 \cdot P(N_{80,0.5} \geq 81.76) = 0.00043$.
- Since the p-value is quite small, it is significant at both the 10% and 3% significance levels, and we accordingly reject the null hypothesis in both cases.

One-Sample z Tests, XX

Example: A sample from Class A has 64 students and average score 80.05 points. A sample from Class B has 100 students and average 81.76 points. Assume the standard deviation is known to be 5 points. Test at the 10% and 3% significance levels

3. Whether the average in Class B is greater than 81 points.

One-Sample z Tests, XX

Example: A sample from Class A has 64 students and average score 80.05 points. A sample from Class B has 100 students and average 81.76 points. Assume the standard deviation is known to be 5 points. Test at the 10% and 3% significance levels

3. Whether the average in Class B is greater than 81 points.

- Like before, we want $H_0 : \mu_B = 81$.
- Because the actual average is 81.76 points (greater than 81), and we also want to test whether the average is greater than 81, we take the alternative hypothesis $H_a : \mu_B > 81$.
- The test statistic will be normally distributed with mean 81 (per H_0) and standard deviation $5/\sqrt{100} = 0.5$.
- Thus, the p-value is $P(N_{81,0.5} \geq 81.76) = 0.0643$.
- This is statistically significant at the 10% significance level (so we reject the null there) but not at the 3% significance level (so we fail to reject the null there).

One-Sample z Tests, XI

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

1. the Institute is concerned about errors of any kind.
2. the Institute is only concerned about errors that make people late.

One-Sample z Tests, XII

Example: The Bad Timing Institute believe that watch errors have 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 $=$ runs fast). Test at the 10% significance level the Bad Timing Institute's hypothesis that the true mean error μ is 0 seconds, if

1. the Institute is concerned about errors of any kind.

One-Sample z Tests, XII

Example: The Bad Timing Institute believe that watch errors have 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 = runs fast). Test at the 10% significance level the Bad Timing Institute's hypothesis that the true mean error μ is 0 seconds, if

1. the Institute is concerned about errors of any kind.

- 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 $20/\sqrt{6} = 8.1650$ seconds.

One-Sample z Tests, XIII

Example: The Bad Timing Institute believe that watch errors have 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 = runs fast). Test at the 10% significance level the Bad Timing Institute's hypothesis that the true mean error μ is 0 seconds, if

1. the Institute is concerned about errors of any kind.

- 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 significant at the 10% significance level, and we accordingly fail to reject the null hypothesis.

One-Sample z Tests, XIV

Example: The Bad Timing Institute believe that watch errors have a standard deviation of 20 seconds. They poll 6 people, whose watches have mean error -12.5 seconds (positive $=$ runs fast). Test at the 10% significance level the Bad Timing Institute's hypothesis that the true mean error μ is 0 seconds, if the Institute

2. is only concerned about errors that make people late.

One-Sample z Tests, XIV

Example: The Bad Timing Institute believe that watch errors have a standard deviation of 20 seconds. They poll 6 people, whose watches have mean error -12.5 seconds (positive $=$ runs fast). Test at the 10% significance level the Bad Timing Institute's hypothesis that the true mean error μ is 0 seconds, if the Institute

2. is only concerned about errors that make people late.

- 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 before.
- 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 $p < 0.10$, the result is significant at the 10% significance level, and we accordingly reject the null hypothesis.

Interlude: One-Sided Versus Two-Sided, I

This example illustrates a peculiar situation: the decision here about whether or not to reject the null hypothesis depends solely on the choice of alternative hypothesis.

- Here, we did not reject the null hypothesis with the two-sided alternative hypothesis, but we did reject it with the one-sided alternative hypothesis.
- On its face, this seems very bizarre: it says, simultaneously, that we *do* have strong evidence that the error is nonzero (and in the negative direction) and also that we *do not* have strong evidence that the error is nonzero (without regard to direction).

Interlude: One-Sided Versus Two-Sided, II

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.

Interlude: One-Sided Versus Two-Sided, III

This particular ambiguity also demonstrates one reason it is poor form simply to state the result of a test without clearly identifying the hypotheses and the p-value.

- Specifically, the result of the binary decision ("significant" / "reject the null hypothesis" versus "not significant" / "fail to reject the null hypothesis") provides very little information by itself.
- In this last example, 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 (e.g., $p = 0.0001$), the factor of 2 would not have affected the statistical significance.

Summary

We introduced the framework for hypothesis testing along with the relevant terminology.

We discussed one-sample z tests and gave several examples.

Next lecture: More z-tests, z tests for unknown proportion