MONT 105Q – Mathematical Journeys
Lecture Notes on Statistical Inference and Hypothesis Testing
March-April, 2016

Background

For the purposes of this introductory discussion, imagine first that you are a scientist. You have collected some data from an experiment and you believe that it shows that a certain pattern exists in the real world situation you are studying. For instance, you might think you have shown that a particular non-till farming method reduces soil erosion as compared with conventional farming methods. *How do you demonstrate that the measurements you have made provide evidence for claiming that that pattern actually exists? Your arguments should be in a form that can be checked and evaluated by other scientists, and your goal is essentially to convince those others that your claims are true, using some accepted mathematical tools.*

Similarly, now suppose you are a public opinion pollster, and you think you have found some interesting pattern in the answers to a question on a survey you have designed. For instance, this pattern could be that a particular political candidate seems to be leading in the run-up to an election–a situation we are all probably sick of(!) Again, *how do you demonstrate that this pattern is there, in a way that is verifiable and that will convince others of the correctness of your claims?*

In both cases, you would need to be able to show convincingly that whatever you saw in the data you collected was not simply the result of some naturally occurring variation or some random glitch in the measurement process. In both cases, the issue is whether the same pattern observed in a *sample* (that is, the measurements made) is also true of the whole population from which the samples were drawn (that is, all possible instances of using the farming method, or all the people in the population from which the respondents of the questionnaire were chosen).

Key Idea: The typical methods used by most scientists (including social scientists) would include performing a statistical analysis on the data aimed at showing that *the observed results would be extremely rare if the claimed pattern did not actually hold true.*

If that can be done successfully, the fact that the pattern in the measurements made was observed can be due to one of two possible reasons: Either it is

1) due to the fact that the pattern is *truly there* in the whole population, or else,
2) due essentially to some "bad luck" in how your experiment turned out.

In the second case, you "hit the jackpot," but in a negative way. You observed a very rare event that does not agree with the true state of things.

This probably sounds rather convoluted at first, but it is the logic behind the *hypothesis tests* we will discuss, and it is the most important idea behind this part of our course.

Null and Alternative Hypotheses

The way the testing process is described typically involves two *competing explanations* for the results of the experiment:

- A *null hypothesis* (often denoted H_0 if we need a symbolic abbreviation) that says essentially "there's nothing there – the results were just the result of naturally occurring variation or some random glitch in the measurement process." Of course, the exact statement is usally more precise than this – it usually involves a more specific statement about the thing(s) you are measuring.
- An *alternative hypothesis* (denoted H_a) that is is some assertion that the claimed pattern is really there. This is also usually stated in more precise form(!)

(These can be different from the scientific hypotheses that were made in designing the experiment that produced the data – these statistical hypotheses are proposed explanations for the data that was observed.) The goal is to decide whether the weight of the evidence contained in the measurements supports H_a, or whether that evidence could reasonably be explained by H_0. From the point of view of a researcher, of course, it is typically the alternative hypothesis H_a that is the "preferred" alternative. Not being able to rule out the null hypothesis is usually taken as a negative outcome, since in that case we are saying the results we saw could just be due to chance.

Here are two examples that should make the distinction between the null and alternative hypotheses clearer.

Example 1. *A random sample of $n = 50$ compact fluorescent light bulbs was chosen from the production line of the manufacturer and the mean weight of mercury per bulb was measured to be 5.3 mg, with an SD of .5 mg. Assume the mercury amounts in all such bulbs are normally distributed. Looking at the 5.3, it might be tempting to say: "The average amount of mercury in these bulbs is > 5 mg." But does the data support that? Might the particular sample chosen have just contained especially mercury-heavy bulbs, not typical of the whole population? In this case the null hypothesis could be H_0 : the actual population average mercury level μ satisfies $\mu \leq 5$ mg per bulb. And the alternative hypothesis could be: H_a : the actual population average mercury level is $\mu > 5$ mg per bulb.*

Example 2. *StarLink (a registered trademark) corn is a genetically engineered variety that was approved as a source of feed for animals, but never approved for human consumption. A study by the USDA shows that 99 out of 1100 samples of corn taken from US (human) corn-based food products were contaminated by traces of StarLink corn. At the same time the corresponding agency of the Mexican government does a similar study and finds that 100 out of 1200 samples of corn-based food products taken from Mexican manufacturers contain traces of StarLink. Note that the contaminated proportions are $\widehat{p_{US}} = 99/1100 = .09$, while $\widehat{p_M} = 100/1200 \doteq .083$ is smaller. Does the data support the conclusion that StarLink contamination of human food products is different in the US than in Mexico? Here we could have $H_0 : p_{US} = p_M$ (the actual proportions in the two*

countries are just the same and the difference observed in the samples is due to chance). The corresponding alternative hypothesis might be $H_a : p_{US} \neq p_M$. It would also be possible to use $H_a : p_{US} > p_M$ if the goal is to decide if the evidence shows that the proportion in the US is actually greater.

The conclusions from statistical hypothesis tests are usually phrased in very cautious language. One usually says something like "there is enough evidence to reject the null hypothesis" rather than saying "the alternative hypothesis is definitely true." Similarly a negative result might be described by saying "there is not enough evidence to reject the null hypothesis" rather than saying "the alternative hypothesis is definitely false" or "the null hypothesis is definitely true." The point here, of course, is that the results from one experiment, no matter how strong or weak, can only provide limited evidence one way or the other. In addition, the scientific hypotheses underlying the experiment or the statistical forms H_0 and H_a are always subject to revision if further evidence indicates that previous thinking was incorrect.

Type I and Type II Errors, α and β

A statistical test aimed at choosing between the two hypotheses H_0, H_a can "go wrong" in two different ways:

- **Type I Error:** We could reject H_0 (and say the evidence favors accepting H_a) when H_0 is actually true.
- **Type II Error:** We could fail to reject H_0 when it is actually false.

One often says a Type I error involves a *false positive result*. The situation is analogous to a medical test that indicates the presence of a condition *when you actually do not have it*. On the other hand, a Type II error is a *false negative result* – a case where the results of the test indicate that you are "in the clear" but you actually do have the condition.

Both types of errors are of concern, but the Type I error is usually considered to be, if anything, more serious (at least in non-medical situations(!)). That is because making an incorrect conclusion when an apparent pattern is due only to chance variation can throw off subsequent research, can lead to inappropriate recommendations for real-world action, and can have other undesirable consequences. Making a Type II error, on the other hand, means essentially that we *missed* a pattern that *is there* (perhaps by being too cautious in assessing the data). A common view would be that since there is always the chance of *catching* the pattern with another experiment later, Type II errors are, in a sense, easier to correct. If there were some urgency or time pressure involved in the real-world situation under study (for example, the medical testing situation!), a Type II error could also lead to serious real-world consequences, though.

For another example, consider a study of the effectiveness of a vaccine for a disease (informally: H_0 : no benefit from the vaccine and H_a : there is a benefit). If the testing was being done while an epidemic was in progress, then a Type II error leading to a decision not to use the vaccine might lead to loss of lives that could have been saved by using the vaccine. In real life, in fact, people are often willing to try unproven medical treatments in situations where they have nothing to lose, or in which any chance of a positive outcome, even a small one, outweighs other risks.

In any case, it is important to realize that most statistical tests are *set up to make the chance of a Type I error small*. This is exactly the *Key Idea* stated at the beginning of these notes, restated in a more precise form. The chance of making a Type I error is usually denoted by the Greek letter α ("alpha"), and typical values for α used in the design of statistical hypothesis tests are .05, or .01, or perhaps even smaller values. Saying $\alpha = .05$, for instance, says that we want to set up our test so that if H_0 is actually true, then roughly 95 times out of 100 the results of the test will correctly indicate that H_0 *should not* be rejected. Or equivalently, Type I errors would happen roughly only 5 out of 100 times when H_0 is true. In other words, as we said before, *results indicating that we should reject H_0 would be extremely rare if the claimed pattern did not actually hold true (that is if H_0 is actually true)* .

The chance of making a Type II error is denoted by another Greek letter, β ("beta"). In a sense, the real quantity of interest here is $1 - \beta$, the chance of making a correct conclusion and rejecting H_0 when H_0 is false. This is called the *power* of the test, and ideally we would like the power to be as close to 1 as possible. But it is important to realize that the power, and equivalently, the value of β, are usually somewhat *harder to control* than α because they will typically depend on characteristics of the population from which the measurements are being taken. Since the purpose of making the test may actually be to estimate properties of the population, those properties may not be known exactly.

For instance, if H_0 is the hypothesis that a population mean is equal to some particular number μ_0, then H_0 being false means the population mean is something different from μ_0. The value of β will usually depend on the exact value of that population mean. Thus, the power of the test is actually *a function of μ = true population mean*, not just a constant. Statisticians have developed techniques for studying β and the power of tests as well, but they are somewhat beyond the scope of our treatment of this subject. Without going into the details, we can just say that in order to obtain a test with a given desired power, we would typically have to be able to choose a sample size n that was sufficiently large.

Since making measurements in carrying out experiments will usually incur real-world costs in one fashion or another, this might not always be feasible. For example, a polling organization might not have the time, the manpower, or the access to contact information to carry out a phone survey of a random sample of $n = 5000$ likely voters within a few days before an election in order in a test of voter preferences, even though having n that large might be necessary to to achieve a small β value.

Test Statistics, Rejection Regions

Here is the overall plan for a typical statistical test:

- Using assumptions about the distribution of the possible measurement values from the population, some *test statistic* that has a known probability distribution under the assumption that the null hypothesis H_0 is *true* is identified.
- Some desired Type I error probability α is selected. (Often $\alpha = .05$ is used as a standard choice; smaller values might be used too. Values $\alpha > .05$ would almost never be used in practice since there is something of a consensus that Type I error probabilities that large are unacceptable.)

- Using that probability distribution, a *rejection region* is identified. This is a range R of possible values of the test statistic with the property that, under the assumption H_0 is true, the chance that the test statistic lies in R is α.
- Then starting from the observed measurements, we compute the value of the test statistic, determine whether it lies in the rejection region or not.
- If so, we say "there is evidence indicating H_0 should be rejected" or something similar; if not we say "there is not enough evidence to reject H_0."

This probably seems somewhat abstract without a specific example to think about. Hence we will proceed immediately to a first example test, the "z-test for a mean."

A z-test for a mean ("large sample case")

Suppose we have made a relatively large number $n \geq 30$ of numerical measurements of a particular characteristic from randomly chosen individuals in some fixed population (think something like lengths of individual adult fish from some particular species). Call those measurements $Y_1, \ldots, Y_n$. Then it is reasonable to expect that the sample mean

$$\overline{Y} = \frac{Y_1 + \cdots + Y_n}{n}$$

should give a good estimate of the population average μ. From the Central Limit Theorem, recall that we expect that the values of $\overline{Y}$ from different samples should be normally distributed, and that the conversion to the standard normal should go by this z-score formula:

$$\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}},$$

where σ is the *population* SD. Of course we don't know the exact value of σ in practice, so we would usually estimate that using $S =$ the sample SD. Moreover, we don't know the value of μ either.

Let's consider the following situation. Suppose we have a particular "candidate" value $\mu = \mu_0$ in mind as part of a null hypothesis– $H_0 : \mu = \mu_0$. Moreover, suppose the data seems to indicate, say, that $\mu > \mu_0$. That will be the alternative hypothesis H_a.

Under the assumptions that $n \geq 30$ and that H_0 is true, it can be shown that the test statistic

$$Z = \frac{\overline{Y} - \mu_0}{S/\sqrt{n}}$$

has *approximately* a standard normal distribution. (It actually has one of the so-called t-distributions studied by statisticians, but for n that large, the difference is negligible.)

Hence from the standard normal table, the chance that $Z > 1.65$ would be about .05. This gives the rejection region for our test with $\alpha = .05$ – we reject H_0 whenever $Z > 1.645$ and we do not reject it otherwise.

Example 3. *Now let's illustrate how this would be applied. Suppose we had measurement data that looked like this ($n = 30$ measurements):*

$$5.2, 5.3, 5.9, 8.8, 8.9, 7.1, 5.9, 6.3, 7.2, 5.0,$$
$$5.3, 4.9, 3.2, 5.8, 6.2, 6.0, 7.2, 7.2, 4.2, 5.1,$$
$$6.8, 7.8, 6.2, 5.2, 5.5, 4.3, 4.7, 4.7, 6.7, 6.3$$

We have $\overline{Y} = 5.96$ and $S = 1.30$ (rounding to two decimal places). For our example test, let's take:

- $H_0 : \mu = \mu_0 = 5.75$ *and*
- $H_a : \mu > \mu_0 = 5.75$.

Then, we compute:

$$Z = \frac{\overline{Y} - \mu_0}{S/\sqrt{n}} = \frac{5.96 - 5.75}{1.30/\sqrt{30}} = .8677$$

Since this is not > 1.645, we cannot reject H_0 on the basis of the evidence in this data.

(Comments: The above numbers were generated from a normal population with true mean $\mu = 6.4$ and true $\sigma = 1.2$(!) As you should have noticed in Problem Set 3, there is a lot of possible variability in sampling or measuring. In this case, the sample mean $\overline{Y}$ is especially small relative to the actual μ, and the sample SD S is larger than the actual σ. Both of these contributed to a test statistic Z that was "smaller than expected." This is a case where we are actually making a Type II error! Although we won't discuss how this would be derived, the power of this test with $\mu = 6.4$ would be about $1 - \beta = .83$ – that is the probability of a Type II error would be about $\beta = 1 - .83 = .17$. The above data set "beat the odds" in a way, but that sort of thing would happen roughly 17% of the time.)

There are a number of related forms of z-tests for means based on the form of the alternative hypothesis. For a test with $\alpha = .05$, for instance,

- if $H_a : \mu > \mu_0$ (an "upper-tail" test), then we would reject H_0 if $Z > 1.645$ as above
- if $H_a : \mu < \mu_0$ (a "lower-tail" test), then we would reject H_0 if $Z < -1.645$
- if $H_a : \mu \neq \mu_0$ (a "two-tail" test), then we would reject H_0 if $Z < -1.96$ or $Z > 1.96$.

These rejection regions all come from the areas of regions under the standard normal curve. Note that the area between 0 and 1.96 is about .475, so the area in the upper tail from 1.96 to $+\infty$ is about .025. Similarly the area in the lower tail from $-\infty$ to -1.96 is also about .025 by the symmetry of the standard normal curve. This means that the rejection region for the two-tail test has total area about $.025 + .025 = .05$ and the chance that Z lands in that region under the assumption that $H_0 : \mu = \mu_0$ is true is $\alpha = .05$ (approximately).

Hypothesis tests and confidence intervals[1]

[1] You can safely omit this section on a first reading and we will not discuss this topic in our course.

There is very close connection between the rejection region for a two-tail hypothesis test and a confidence interval for the mean. In this large sample case, the endpoints of the 95% confidence interval for μ would be computed from the sample mean $\overline{Y}$ and $S =$ the sample SD by the formula

$$\mu = \overline{Y} \pm 1.96 \times \frac{S}{\sqrt{n}}$$

If we add the assumption that $\mu = \mu_0$ from the null hypothesis, when you look at this formula, it is not too difficult to see that it is saying the following. The values of $\overline{Y}$ for which we would *not* reject $H_0 : \mu = \mu_0$ are exactly the $\overline{Y}$ such that

$$\mu_0 - 1.96 \times \frac{S}{\sqrt{n}} < \overline{Y} < \mu_0 + 1.96 \times \frac{S}{\sqrt{n}}$$

(since it is exactly those values of $\overline{Y}$ for which $Z = \frac{\overline{Y} - \mu_0}{S/\sqrt{n}}$ satisfies $-1.96 < Z < 1.96$).

In other words, the confidence interval gives the range of "believable" values for μ based on the the mean and SD of the sample. The rejection region for the test is something like the *complement of (that is, the part of the number line outside)* the confidence interval computed using μ_0 from the null hypothesis.

If we only have $n < 30$ measurements, then the actual properties of the t-distributions must be taken into account. We will see what to do in those "small sample" cases later.

z-tests for a proportion

Say we have asked a random sample of n people from a particular population a "yes-or-no" question and some number Y of them answer "yes." Suppose P is the proportion of the whole population who would answer "yes" if asked and we want to estimate this proportion. From the sample, we can estimate P using $\widehat{P} = Y/n$.

The theoretical basis for the test in this case is the following mathematical result:

Under $H_0 : P = P_0$,

a) when n is relatively large (the cutoff value $n \geq 30$ is often used), *or more generally*
b) when nP_0 and $n(1 - P_0)$ are both relatively large (a typical "rule of thumb" is both are ≥ 5),

then (by using the Central Limit Theorem in an appropriate way) it can be shown that the test statistic

$$Z = \frac{\widehat{P} - P_0}{\sqrt{\frac{\widehat{P}(1-\widehat{P})}{n}}}$$

has an approximately standard normal distribution.

So we can set up hypothesis tests with $\alpha = .05$ using the same rejection regions as above in the case of a large-sample z-test for a mean:

- if $H_a : P > P_0$ (an "upper-tail" test), then we would reject H_0 if $Z > 1.645$

- if $H_a : P < P_0$ (a "lower-tail" test), then we would reject H_0 if $Z < -1.645$
- if $H_a : P \neq P_0$ (a "two-tail" test), then we would reject H_0 if $Z < -1.96$ or $Z > 1.96$.

Example 4. *Supppose that $n = 1000$ likely voters are surveyed and $Y = 550$ of them say they will vote for candidate Jones in the next city council election. With $\alpha = .05$, is there sufficient evidence to say that Jones will win the coming election? We take $H_0 : P = P_0 = .5$ and $H_a : P > .5$. (Note that this is one case where the statistical null hypothesis is somewhat different than we might expect – Jones will only lose the election if $P < .5$, and the boundary case $P = .5$ would be a "dead heat" in the election.) In any case, with the values of n, Y as given, we have $n \geq 30$ and $nP_0 = 500 = n(1 - P_0)$, so both conditions a) and b) above are satisfied and we can use this approach. We compute*

$$Z = \frac{.55 - .50}{\sqrt{\frac{(.55)(.45)}{1000}}} \doteq 3.17 > 1.645.$$

This is strong evidence to indicate that we should reject H_0. Note that any value $P_0 < .5$ would yield an even larger Z. So taking $P_0 = .5$ in this case is justified since it is the boundary case between a loss for Jones and a win for Jones.

z-tests for differences of means and differences of proportions

There are similar tests of hypotheses about the difference of two population means or two proportions. The ones we will discuss are valid *only in the large sample case* (both groups of samples of size at least 30). The analysis is based on the assumption that the samples are random and independent.

Say we have measured some quantity in a random sample of size $n_1 \geq 30$ from a first group, and measured the same quantity in a random sample of size $n_2 \geq 30$ from a second group. (The two numbers n_1 and n_2 can be different, but both should be in the large sample range.) Call the two groups of measurements $Y_1, \ldots, Y_{n_1}$ and $X_1, \ldots, X_{n_2}$. So we have two sample means $\overline{Y}$ and $\overline{X}$, as well as sample SD's S_1 (from the Y_i) and S_2 (from the X_j).

Say μ_1 and μ_2 are the population means of the two groups. To test $H_0 : \mu_1 = \mu_2$ versus $H_a : \mu_1 \neq \mu_2$, we would use the test statistic:

$$Z = \frac{\overline{Y} - \overline{X}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

since under the assumption $\mu_1 = \mu_2$, Z has a standard normal distribution. The rejection region would be set up according to the value of α exactly as before.

Similarly, for a test on a difference of proportions, say we have asked the same "yes-or-no" question to random samples from two different groups. Say Y_1 out of $n_1 \geq 30$ in the first group say "yes" while Y_2 out of $n_2 \geq 30$ in the second group say "yes." (There is also a more general "rule of thumb" parallel to b) in the discussion of the one-proportion

test above that is sometimes used; we will not discuss that, however.) We take $\widehat{P_1} = Y_1/n_1$ estimating P_1 and $\widehat{P_2} = Y_2/n_2$ estimating P_2 (the two population proportions). Then to test $H_0 : P_1 = P_2$ versus $H_a : P_1 \neq P_2$, we need some estimate of the common value under the null hypothesis in order to construct our test statistic. The one that is commonly used is a "pooled" estimator of the common proportion (assuming the null hypothesis):

$$P_p = \frac{Y_1 + Y_2}{n_1 + n_2}.$$

Then the test statistic is

$$Z = \frac{\widehat{P_1} - \widehat{P_2}}{\sqrt{\widehat{P_p}(1 - \widehat{P_p})} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

and the rejection regions are as before.

Note that the "yes-or-no" question can be anything with only those two possible answers. This applies much more generally than to only the polling situation. For instance,

Example 5. *Refer back to Example 2 above. Is there evidence to show that different proportions of corn products contain traces of StarLink in the US and in Mexico? Note that the contaminated proportions are* $\widehat{p_{US}} = 99/1100 = .09$, *while* $\widehat{p_M} = 100/1200 \doteq .083$. *So* $P_p = \frac{199}{2300} = .087$. *We have*

$$Z = \frac{.09 - .83}{\sqrt{(.087)(.913)}\sqrt{\frac{1}{1100} + \frac{1}{1200}}} \doteq .595$$

For the two-tail test with $\alpha = .05$, *we would be looking for* $Z > 1.96$. *So this is not strong enough evidence to reject* H_0. *The proportions in the US and in Mexico could be the same and the difference could be due to random variation.*

t-tests for a mean ("small sample case")

When the number of measurements $n < 30$, then the tests for the mean presented above must use different rejection regions coming from the "Student's t-distributions" developed by W.S. Gosset in the early 1900's. There is a rather interesting story behind this contribution. Gosset was employed by the Guinness brewery in Ireland at the time. The Guinness company sponsored a fair amount of research on scientific and mathematical topics related to agriculture and fermentation. But they had a policy of forbidding their employees from publishing results that might be useful to competitors. Gosset was able to convince his bosses that his statistical work had no practical implications for brewing (although it certainly could and has been used to analyze data from various manufacturing quality-control situations). So he eventually got their approval to publish his work on the t-distributions. Nevertheless, as a condition, he was not allowed to place his own name on the article, which appeared under the pseudonym "Student."

When $n < 30$, we use the same test-statistic as before, although it is now commonly denoted T rather than Z:

$$T = \frac{\overline{Y} - \mu_0}{S/\sqrt{n}}.$$

But now the rejection regions are found from a t-table and will depend on n:

- if $H_a : \mu > \mu_0$ (an "upper-tail" test), then we would reject H_0 if $T > t_{.05}$ for $n-1$ degrees of freedom
- if $H_a : \mu < \mu_0$ (a "lower-tail" test), then we would reject H_0 if $T < -t_{.05}$ for $n-1$ degrees of freedom
- if $H_a : \mu \neq \mu_0$ (a "two-tail" test), then we would reject H_0 if $T < -t_{.025}$ or $T > t_{.025}$ for $n-1$ degrees of freedom.

The needed information to set up the two-tail rejection region for $\alpha = .05$ is given in the t-table on the course homepage. For example, if we had $n = 10$ sample measurements, then the rejection region for an upper-tail test with $\alpha = .05$ would be $T > 1.833$ from the $n = 10$ row of the table. (The t-distribution involved is the one that has 9 degrees of freedom, though – note that the number of degrees of freedom is always $n - 1$.) The corresponding two-tail test would have rejection region $T < -2.262$ or $T > 2.262$. Note that these are more restrictive than the corresponding rejection regions for the large-sample z-test. This is because the distribution of $\overline{Y}$ and the test statistic T *has even more variability when* n *is small*. So the rejection region must be smaller than the corresponding rejection region for a large-sample z-test with the same $\alpha = .05$.

Example 6. *A random sample of ten 1-square kilometer plots in a forest are chosen and the number of robin nests in each area is counted, yielding the following data:*

$$310, 311, 412, 368, 447, 376, 303, 410, 365, 350$$

The sample mean is $\overline{Y} = 365.2$ *and the sample SD is* $S = 48.417$. *Is there sufficient evidence to claim that the average number of robin nests per square kilometer in this forest is less than 380? We compute*

$$T = \frac{365.2 - 380}{48.417/\sqrt{10}} \doteq -.967$$

Since $T > -1.833 = -t_{.05}$ *for 9 degrees of freedom, we do not have sufficient evidence to reject* $H_0 : \mu = 380$ *with* $\alpha = .05$.

The p-value of a hypothesis test

When reporting the results of a statistical hypothesis test, it is now common to provide an additional (or sometimes alternate) piece of information called the *p-value* of the test. Another name is the *attained significance level*. The reason for that name is the fact that the Type I error probability α is often called the *significance level* of the test. So, for instance if we have decided on $\alpha = .05$ and the result of the test is to reject H_0, we might say that the result is "significant at the .05-level." (This is the precise meaning of statements like "the results of the experiment were statistically significant.")

Look back at Example 4. There we had a z-test where the test statistic had the value $Z = 3.17$. We could say "there is evidence to reject H_0 at the $\alpha = .05$ significance level. But in fact since $Z = 3.17$ is quite a bit larger than $z_{.05} = 1.645$, in a way we are *understating the strength of our evidence* if we stop there. The attained significance level p is, by definition, *the smallest value of α for which H_0 would be rejected with this value of Z*. Since the area under the standard normal curve to the right of 3.17 is approximately .00076, we would reject H_0 with this data for any $\alpha > .00076$. We say the p-value is $p = .00076$.

An alternative way to think about what the p-value of a test is telling us is this: the p-value is the *chance of observing the given value of the test statistic, or something "more extreme," if H_0 is true*. So *very small* values of p indicate *very strong evidence for rejecting* H_0 according to the Key Idea at the start of these notes. Conversely, larger values of p (often any $p > .05$) is taken as indicating the evidence is too weak to reject H_0. Since there is so much arbitrariness in the $\alpha = .05$ value, nowadays, in fact, it is actually common just to report the p-value of a test and leave the interpretation of whether to accept or reject H_0 up to the reader(!)

Example 7. *Refer back to Example 1. Since $n = 50$, we use the large-sample formulas based on the standard normal curve. Our test statistic is*

$$Z = \frac{5.3 - 5}{.5/\sqrt{50}} \doteq 4.24$$

For the upper-tail z-test of $H_0 : \mu = 5$ versus $H_a : \mu > 5$, we have a p-value $p = 1.1 \times 10^{-5} =$ .000011. This would be interpreted as very strong evidence to reject H_0. Note that it is the sample size $n = 50$ that is making it come out this way. If we knew the population SD was .5, under the null hypothesis a single measurement of 5.3 would have a z-score of

$$\frac{5.3 - 5}{.5} = .6$$

This is not very large (less than one SD above 5), so nothing remarkable. But the fact that we have the average of $n = 50$ measurements being 5.3 means that $Z = .6 \times \sqrt{50} \doteq 4.24$.

The p-values of hypothesis tests are often reported with other information such as computation of equations of regression lines. When we use the regression functions from the Data Analysis package in Excel, for example, we get information like the following:

Example 8. *Suppose we enter this data in Excel: in cells A1 - A8,* $1, 2, 3, 4, 5, 6, 7, 8$ *and in cells B1 - B8,* $2.3, 3.6, 4.1, 4.3, 5.2, 5, 6.3, 7$. *Highlight those cells and from the Data/Data Analysis menu, select regression. The output generated (look on the tabs below for a new output sheet) includes something like this:*

SUMMARY OUTPUT

Regression Statistics
Multiple R – 0.97434237
R Square – 0.949343055
Adjusted R Square – 0.94090023
Standard Error – 0.363787396
Observations – 8

$\vdots$

	Coeff′s	Std Error	t Stat	P − value	Lower 95%	Upper95%
Int.	2.046428	0.283460	7.219439	0.00035	1.352824	2.740032
XVar.	0.595238	0.056133	10.60395	$4.142E-05$	0.457884	0.732592

What is going on here? The R^2 *value is related to a correlation coefficient: an indication of how close the data was to lying on a single straight line (pretty close here!) The block at the bottom shows that the equation of the regression line is*

$$y = b + mx = 2.046428 + 0.595238x$$

The rest of that block reports the results of two statistical tests. Namely, the line for the $b =$ *Intercept ("Int.") shows that in a test of* $H_0 : b = 0$ *versus* $H_a : b \neq 0$, *a standard test statistic (a t-test as in the small sample test for a mean), the p-value* $p = .00035$ *indicates pretty strong evidence to reject* H_0. *Similarly, the last line gives the result of a similar test of* $H_0 : m = 0$ *versus* $H_a : m \neq 0$. *Here the results are even more striking, and we would conclude that there is very strong evidence for saying* $m \neq 0$ *from this data (in other words, that* y *depends on* x*). The last two columns in the block give 95% confidence bounds for intervals containing* m, b *in an actual linear model describing the relation between* y *and* x *based on this data.*

Statistical significance and practical significance

Some caution is certainly a good thing in applying the methodology of hypothesis testing. As we indicated before, the significance level of a test α (or the p-value) is based entirely on the probability of rejecting the null hypothesis when it is actually true. So we are really only taking the Type I errors into account. For that reason, the statistical significance level of a test is only an imperfect measure of how much information we get from the test.

Moreover, there are certainly times when a statistically very significant test (say one that gives a very small p-value) has little *practical* significance. For example, a regression analysis on a large data set might yield the result that the slope of the regression line was nonzero with an attained significance level of $p = .0001$. But the corresponding confidence interval of values for the slope could contain only very small numbers (e.g. .01 to .03). There might not be much of a practical difference between

- saying a unit change in x produces a change of between .01 and .03 in y, or
- saying a unit change in x produces no change in y

if the values of y are in the 1000's, for instance. Statistical significance and real-world practical significance are not the same things!

Some Practice Questions

Directions: For each question, first identify the relevant null and alternative hypotheses. Then carry out the appropriate hypothesis test, estimate the p-value if possible with the information you have, and answer the other questions.

1) The average CO_2 emissions in lb/Mwh from the 100 largest coal-fired power plants in the US in 1998 was $\overline{Y} = 2223.1$ with $SD = 211.3$ lb/Mwh.
   a) If this were a random sample, would there be evidence at the $\alpha = .05$ level to say that the average coal-fired power plant emits an amount different from 2200 lb/Mwh?
   b) Is this a random sample? Does the calculation in part a) really make sense in this setting?

2) A simple random sample of 2000 Germans in 2001 showed that 840 thought that all electricity should be generated by "green" energy sources such as wind and solar power. Is there evidence at the $\alpha = .05$ level to say that at least 40% of all Germans thought the same about sources of electric power?

3) In a study of $n = 17$ starfish arm lengths, the average was $\overline{Y} = 6.8$ cm with an SD $= .5$ cm. Is there evidence at the $\alpha = .05$ level to say that the population arm length is different from 7 cm?

4) A medical study compared the resting pulse rates of random, independent samples of 100 smokers and 100 nonsmokers. The smokers had an average pulse rate of $\overline{Y} = 86$ beats per minute with $SD_1 = 5.4$, while the nonsmokers had an average pulse rate of $\overline{X} = 80$ and $SD_2 = 4.9$. Is there evidence at the $\alpha = .05$ level to say that nonsmokers have a lower average pulse rate?