MATH 1070 Introductory Statistics
Lecture notes
Confidence Intervals and Hypothesis Testing

Objectives:

1. Learn the concepts of estimating a parameter and how to measure the confidence of that estimate
2. Point estimates and interval estimates
3. Characteristics of good estimators
4. Learn how to compute confidence intervals
5. Computing necessary sample size for a given margin of error
6. Hypothesis testing and significance (Large samples - z test)
7. Type I and Type II error and what they are
8. Using p-values in hypothesis testing
9. Hypothesis testing and significance (Small samples - t test)
10. Testing the difference between two means (Matched pairs)

Estimating a Parameter

With small populations, less than 30 to 50 observations, it is relatively easy to compute the actual value for the mean (μ). It can be done, it might be tedious and boring, but it can be done. But if the population is 300 or 3000 or 300,000 observations the possibility of computing the mean becomes less likely. Still possible, but with decreasing chance of it actually happening.

When the population becomes as large as the population of the United States (recently 300,000,000) the possibility is so minimal as to be nonexistent. Gathering all the data is an onerous chore, and crunching that much data would take a signficant amount of time. It's would be easier to compute with a smaller group of the data, something like a sample.

When we use a sample to compute a statistic ($\bar{x}$) we can then use that statistic to estimate the population parameter (μ). That's the hope. This is why we want to eliminate as much bias as possible from our sample. A biased sample cannot be generalized back to the population at large. Quite simply, the answer would be wrong.

This type of statistic is called a **point estimate** since it is a single number. Formally we would say that

> A point estimate of a parameter is the value of a statistic that is used to estimate the parameter. (Weiss, p. 445)

Alternatively we could say

> A point estimator draws inferences about a population by estimating the value of an unknown parameter using a single value or point. (Keller and Warrack, p. 304)

The problem with this is that the sample statistic ($\bar{x}$) is just one estimate of the parameter and while it might get the answer correct it could just as easily have been wrong. (The sample mean $\bar{x}$ is, however, the best estimator of the population mean μ.) This means there is error in the point estimate.

So what makes a good estimator? There are three characteristics for a good estimator which are:

1. Unbiased - the expected value should be equal to the parameter being estimated.
2. Consistent - as sample size increases, the value of the statistic should approach the value of the parameter.
3. Relatively efficient - smallest variance of all possible estimators.

Unbiased means that, if you were to take an infinite number of samples, calculate the value of the estimator in each sample, then average these values, the average value would equal the parameter. This amounts to saying that, on average, the sample statistic is equal to the parameter. (Keller and Warrack, p. 305)

We would like to know how good the estimator is and how close it comes to the real parameter. We figure this out as a **confidence interval** (**CI**). We try to bound the real value within a range of error to either side (+/-) of the estimate. A **confidence interval estimate** of a parameter consists of an interval of numbers obtained from a point estimate of a parameter together with a percentage that specifies how confident we are that the parameter lies in the interval. The confidence percentage is called the **confidence level**.

Law of Large Numbers

> If an experiment is repeated again and again, the probability of an event obtained from the relative frequency approaches the actual or theoretical probability. *Mann, p. 162*

> If a situation, trial, or experiment is repeated again and again, the proportion of successes will tend to approach the probability that any one outcome will be a success. *Freund and Simon, p. 124*

Critical Values of z

With a confidence interval we get to choose the margin of error we will accept and how confident we want to be. This is great flexibility, but it is not free.

Why would we ever want to be less than 99% confident? Why would anyone? The more confident you want to be the more data you will need. The law of large numbers says that the larger the sample the closer to the real value of the parameter. But the key word is "large". We have to have more data to be more confident about the interval. And data isn't always without cost. Time or money will most likely be needed.

We use different values for z in the computation of the confidence interval depending upon the level of confidence we require. There are three very common levels of confidence - 90%, 95%, and 99%. Each has a different value of z assigned to it. Where does this value come from, though?

Recall the standard normal distribution. We used it in standardized variables. The area under this curve is 1 and at any point along the curve we know the percentage of area to the left of that point. Also recall the 65-95-99.7 empirical rule where we used the standard deviation to identify where 65%, 95%, and 99.7% of the data would be found.

We use the same concept here to compute the confidence interval.

The formula for the confidence interval is

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

that is, the sample mean plus or minus the margin of error. The α in the computation is the amount of error we are willing to accept in the interval. Why is it divided by 2? Because the distribution is symmetrical (remember?) we have to have the same amount of area (error) in the upper and lower tails. So we have half the error allocated to the lower tail and half the error allocated to the upper tail.

When we subtract this amount of error from the whole (1.0) we get the confidence level we want for the interval estimate. This is also the area under the curve of the normal distribution.

So if we want a 95% confidence interval we are willing to accept 5% error. Divide that error by half and we get 2.5% ($0.05/2 = 0.025$). Now, look inside the standard normal table for a value close to 0.0250. We find this on the row labeled "-1.9" and in the column labeled "0.06", so the critical value for z is -1.96. Since the distribution is symmetrical that means the upper tail is at 1.96. We use the positive value for the computation of the margin of error since we add or subtract the same amount from the sample mean to create the interval.

Selecting Sample Size

How do we know what size sample to select? Randomly guess? Actually, the appropriate sample size can easily be determined from a formula. To use this formula we need two pieces of information

1. The desired confidence level of the interval estimator; and
2. The margin of error, or bound of the error of estimation (B).

The formula for the margin of error is

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = B$$

By use of algebra we can manipulate this equation into the following form:

$$n = \left(\frac{z_{a/2}\sigma}{B}\right)^2$$

where n is the sample size. So given a confidence level and a target value for the margin of error we can then use the reworked equation to compute the desired sample size.

A forester would like to estimate the mean tree diameter of a large tract of trees. He wants to estimate μ to within 0.5 inch, with 99% confidence. A quick survey reveals that the smallest tree has a diameter of two inches, while the largest tree has a diameter of 27 inches. How large a sample should he take?

Taking this apart we inventory what information we have:

- Margin of error = 0.5 in
- Confidence interval = 99%
- $\alpha = 0.01$ (1 - 0.99)
- Range of tree trunk sizes = 27 - 2 = 25

From these values we know that $z_{\alpha/2} = z_{0.01/2} = z_0.005$ so we need a value for z with 0.005 area in the upper and lower tails. Looking in the standard normal table we don't find that exact value, but we do find 0.9949 at 2.57 and 0.9951 at 2.58. Since our desired value is between the two, we split the difference at 2.575.

We do not have the population standard deviation σ, but we can estimate the value if we take the range of the values and divide by 4. So, the range of trunk sizes is 27 - 2 = 25 which when we divide 25 by 4 we get 6.25. Neat trick, eh?

So now it's just "plug and churn" in the formula, like so:

$$
\begin{aligned}
n &= \left(\frac{z_{a/2}\sigma}{B}\right)^2 && (1)\\
&= \left(\frac{2.575 \times 6.25}{0.5}\right)^2 && (2)\\
&= \left(\frac{16.09375}{0.5}\right)^2 && (3)\\
&= (32.18751)^2 && (4)\\
&= 1036.0351 && (5)
\end{aligned}
$$

So the sample size for this experiment should be 1,037 trees.

Note that the answer wasn't 1,037 but 1,036.0351. We had to round up to the next smallest integer. That's because we can't take a partial observation. What is 0.0351 of a tree anyway? If we just threw away the fractional part and used a sample size of 1,036 we wouldn't get the desired margin of error.

Hypothesis Tests in Statistics

In the hard sciences, and even in some of the "soft" sciences, we ask questions that we answer with experiments. We usually have an idea about what should happen, or we hope will happen, and we run experiments to collect data to support or refute these ideas. We call these ideas, or guesses, hypotheses. In statistics we have the same methodology for testing claims or beliefs about populations of data.

The methodology is the same:

1. Formulate a hypothesis about a phenomenon of interest.
2. Design an experiment to test this hypothesis.
3. Collect the data.
4. Analyze the data from the experiment.
5. Draw a conclusion.

In statistical tests, however, we don't always have to run a physical experiment. We can analyze data from an observational study which we have questions about.

Hypothesis testing is an area of statistics where the art comes into play. It is true that using the same data set and slightly different, but equally valid, criteria one can prove the exact opposite of an argument. In this area it is much more about the interpretation of the numbers than the numbers themselves.

Test statistics used in hypothesis testing - the z statistic

The **z test** is a statistical test for the mean of a population and is used when the population is normally distributed and σ is known or n is greater than or equal to 31.

The formula for the z test is

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

where

$$\begin{aligned} \bar{x} &= \text{sample mean} \\ \mu &= \text{population mean} \\ \sigma &= \text{population standard deviation} \\ n &= \text{sample size} \end{aligned}$$

The Null and Alternative Hypothesis

Just as any argument has two sides, our tests have two hypotheses. These hypotheses have specific names and notation. We also have to pay attention to how we word and formulate the hypotheses.

The first hypothesis is the **null hypothesis**. This is a claim or statment about a population parameter that is assumed to be true until it is declared false (Mann p. 423). The notation we use for the null is H_0, read "H naught" or "H zero". In most cases we want the null hypothesis to be proven false. There are certain cases where we don't want to disprove the null, and it is important to understand that we don't prove the null. We said it is assumed to be true, not that we know it to be true for certain.

The other side of the argument is the **alternative hypothesis**. We write the alternative as H_a or H_1. This is read the way it appears. This is a claim about a population parameter that will be true if the null hypothesis is false. This is the side of the argument that we are usually trying to prove to be true. We want the alternative to be true, if the data support it.

When discussing tests of hypothesis it is important to remember that we do not prove the null hypothesis to be true. We can prove the null to be false but we cannot prove it to be true. Recall that we said the null was assumed to be true. So what's the difference?

We will use the US legal system to illustrate the difference. In the legal system a defendant, the person being prosecuted, is presumed innocent. It is the job of the prosecutor to prove the guilt of the defendant, not prove his innocence. We assume the defendant is innocent until proven guilty. In situations where the guilt of the defendant seems obvious but the defendant goes free we often hear the phrase, "the prosecution failed to make the case." Insufficient evidence was presented to prove the guilt so the presumption of innocence was not refuted.

We have to word our conclusions accordingly. Some authors will say that we accept the null hypothesis, but this is not the wording that is commonly used. The phrase more often used is "fail to reject" which is not the same as "accept" for the reasons stated above. "Fail to reject" the null hypothesis conveys the idea that although we may not be happy about it we have to live with the null hypothesis.

Error in Hypothesis Testing

Sometimes we still make an error and draw the wrong conclusion despite the data. We could reject a true null hyopthesis or accept a false alternative hypothesis. Of course, this is bad and we want to minimize the possibility of this happening. We have to admit, though, that the two possible errors are not the same in severity. Rejecting a true null hypothesis is worse than accepting a false alternative.

Again, think about the legal system. Convicting an innocent person is much worse than allowing a criminal go free. True, letting a criminal escape punishment is bad for society but convicting an innocent person for a crime and punishing them is worse. This is a major argument used against capital punishment (the death penalty). The consequences are so great it is considered too risky by its opponents.

The two types of errors are represented by different Greek letters and are given different names. **Type I** error occurs when a true null hypothesis is rejected. We use the letter alpha (α) to represent this error. We want this error to be as small as possible. We call this the **level of significance** of the test. **Type II** error occurs when a false null hypothesis is not rejected. The Greek letter beta (β) is used to represent this error.

Stating the Hypotheses

How to state the hypotheses depends on what the goal of the test is. Depending upon the desired outcome we will formulate either a one-tailed or two-tailed null and alternative hypothesis. Remember, the null hypothesis is what we have to live with if the data don't suport the alternative. The null is what we assume to be true in the absence of conclusive proof to the contrary.

Let's look at an example where the hypotheses are one-tailed: Suppose building specifications in a certain city require that the average breaking strength of residential sewer pipe be more than 2,400 pounds per foot of length. Each manufacturer who wants to sell pipe in this city must demontrate that its product meets the specification. From the point of view of the city conducting the tests, the null hypothesis is that the manufacturer's pipe does not meet specifications unless the tests provide evidence otherwise. The hypotheses are then

$$H_0 : \mu \leq 2,400 \qquad H_a : \mu > 2,400$$

The following example shows how to state a two-tailed hypothesis: A food processor wants to check whether the average amount of coffee that goes into his 4-ounce jars is indeed 4 ounces. Since the food processor cannot afford to put much less than 4 ounces into each jar for fear of losing customer acceptance, nor can he afford to put much more than 4 ounces into each jar for fear of losing part of his profit, the appropriate alternative hypothesis is $\mu \neq 4$. (Miller and Freund, p. 237)

In this example the null hypothesis is that the jars are filled exactly right with 4 ounces each. The alternative is that they aren't. Maybe less, maybe more, but not 4 ounces. So they would be written as

$$H_0 : \mu = 4 \qquad H_a : \mu \neq 4$$

Decision Making with p-values

We need some way of determining if we reject or fail to reject the null hypothesis. We do that by comparing the calculated test value against the critical value in the appropriate table. But that means checking different tables for different statistics. There's a simpler way.

A **p-value** is "the probability of getting a difference between $\bar{x}$ and μ_0 greater than or equal to that actually observed." (Miller and Freund, p. 243). The p-value is the area under the normal curve. By comparing the p-value to the alpha level we can easily decide to reject or fail to reject.

1. If $p > \alpha$, then fail to reject H_0.
2. If $p \leq \alpha$, the reject H_0.

How do we calculate the p-value? One method from McClave and Dietrich (p. 328) is as follows:

1. Determine the value of the test statistic z corresponding to the result of the sampling experiment.
2. (a) If the test is one-tailed, the p-value is equal to the tail area beyond z in the same direction as the alternative hypothesis. Thus, if the alternative hypothesis is of the form $>$, the p-value is the area to the right of, or above the observed z value. The same is true in the case of $<$.
   (b) If the test is two-tailed, the p-value is equal to twice the tail area beyond the observed z value in the direction of the sign of z. That is, if z is positive,the p-value is twice the area to the right of, or above the observed z value. The same holds true in the case where z is negative.

To illustrate, let us look at the example from McClave and Dietrich (p. 327-328):

In their example, they calculate the value for the test statistic for $n = 50$ sections of sewer pipe. This was calculated as $z = 2.12$. The question was whether the mean breaking strength of the sewer pipe exceeds 2,400 pounds per foot. Since we are asking if the strength *exceeds* a certain value the test is *one-tailed*. Our alternative hypothesis would be $H_a > 2,400$. We want to test for values greater than $z = 2.12$ as those would be even more contradictory to H_0. So the observed p-value for this test is

$$P(z \geq 2.12)$$

This is equivalent to the area under the standard normal curve to the right of $z = 2.12$. Looking in a standard normal table we find that $z = 2.12$ corresponds to 0.9830. The area to the right of that value, then, is $1 - 0.9830 = 0.0170$. This is quite small and we therefore say these results are "very significant", that is, they disagree with the null hypothesis and favor the alternative.

How do we know whether this means "reject" or "fail to reject"? That depends upon the alpha level we chose at the beginning of the experiment. If we chose $\alpha = 0.05$ (a standard value) then the p-value is less than α and we reject the null. If we chose $\alpha = 0.01$ then the p-value is greater than α and we fail to reject the null.

It is not "fair" to wait until the p-value is calculated and then choose the α level that gives you the answer you want. It is acceptable, however, to choose a more stringent value if the original value allows you to reject the null. In other words, you can always reject the null more strongly.

Looking at another example (McClave and Dietrich, p. 329) we can illustrate the process for a two-tailed test. With a two-tailed test we originally said that the parameter (μ) was exactly equal to a given value. In a two-tailed test it does not matter if the statistic is more or less than the established parameter value, only that is isn't the same (= versus $\neq$).

In this example, McClave and Dietrich computed the test statistic to measure the mean response time for drug-injected rats. They originally stated that the response time was 1.2 seconds. The hypotheses would then be

$$H_0 : \mu = 1.2 \text{ seconds} \qquad H_a : \mu \neq 1.2 \text{ seconds}$$

The observed value of the test statistic was $z = -3.0$. Any value greater or less than this would lead to a rejection of the hypothesis. (Again, we don't care whether the resonse time is greater or less, just that it isn't 1.2 seconds.) The observed significance level for the test is

$$P(z < -3.0 \text{ or } z > 3.0)$$

Looking in the standard normal table for the area under the curve at $z = -3.0$ we find 0.0013. No subtraction is necessary. But since it is a two-tailed test the total p-value is twice this area, so

$$2 \times P(z < -3.0) = 2 \times 0.0013 = 0.0026$$

which again is quite strong. So if α was 0.05 or 0.01 we would reject the null and conclude the mean response time was not 1.2 seconds.

Test statistics used in hypothesis testing - the t statistic

The **t test** is a statistical test for the mean of a population and is used when the population is normally distributed, σ is unknown, and n is less than 31.

The formula for the t test is

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

where

$$\begin{aligned} \bar{x} &= \text{sample mean} \\ \mu &= \text{population mean} \\ s &= \text{sample standard deviation} \\ n &= \text{sample size} \end{aligned}$$

The degrees of freedom (d.f.) are $n - 1$.

t Distribution versus Standard Normal Distribution

Similarities:

1. Bell shaped
2. Symmetric around the mean
3. Mean, median, and mode are equal to 0 and located at the center of the distribution
4. Never touches the x axis (asymptotic)

Differences:

1. Variance greater than 1
2. Actually a family of curves based on the concept of *degrees of freedom*, which is related to sample size
3. As the sample size increases, the t distribution approaches the normal distribution

Testing the Difference Between Two Sample Means

So far all we have discussed are tests of one sample taken from a population. But what if we want to test the effectiveness of a new drug or teaching method? We need to run an experiment where we have two groups, one the experimental subjects and one the control subjects, and then compare the results.

To control as much as possible for outside influences, we try to make sure the samples are as alike as possible. To do this we use the **matched pairs** design of experiments. In this design we takes samples where the subjects in one group match as closely as possible the subjects in the other group. This can sometimes be accomplished with the same set of subjects. In this case we refer to this as **pre-test, post-test** since we test the subjects before we perform the experiment and then test again after the experiment. We then perform a hypothesis test on the difference of the two sets of measurements.

When we are testing two samples in a matched pairs design we are testing for the presence of some change in the measured phenomenon. We take the difference of the two measurements, compute the mean, and test against zero (0). Why zero? If the mean of the differences is statistically close to zero then there is no effect. The test we use is the t test and the formula for the t test is similar to the one we already have seen and used, with one exception:

$$t = \frac{\bar{x} - 0}{\frac{s}{\sqrt{n}}}$$

In this formula we have replaced the population mean (μ) with zero. Our other variables in the equation remain the same and the hypothesis testing methods are the same.

The computation of the difference can be affected by which sample is of interest. That is, if we are testing if men make more than women in the workplace we want to subtract the salary of the woman from the salary of the man in the observation. We then want to test the hypothesis that $H_a : \mu > 0$ since the question was, "Do men make more than women?"

Hypothesis Formulation Table (Mann p. 430)

	Two-tailed Test	Left-tailed Test	Right-tailed Test
For the null (H_0)	$=$	$\geq$	$\leq$
For the alternative (H_a)	$\neq$	$<$	$>$
Rejection region	In both tails	Left tail	Right tail

Tips to Remember for Hypothesis Testing

1. Proper sample size selection is required for tests to be effective.
2. H_a can be $<$, $>$, or $\neq$.
3. If $p > \alpha$, then fail to reject H_0.
4. If $p \leq \alpha$, the reject H_0.
5. An α of 0.05 is typical.