Statistical inference is inferring information about the distribution of a population from information about a sample. We're generally talking about one of two things:

1. ESTIMATING PARAMETERS (CONFIDENCE INTERVALS)
2. ANSWERING A YES/NO QUESTION ABOUT A PARAMETER (HYPOTHESIS TESTING)

1 Confidence intervals

A confidence interval is the interval within which a population parameter is believed to lie with a measurable level of confidence.

We want to be able to fill in the blanks in a statement like the following:

We estimate the parameter to be between _____ and _____ and this interval will contain the true value of the parameter approximately _____% of the times we use this method.

What does confidence mean? A CONFIDENCE LEVEL OF C INDICATES THAT IF REPEATED SAMPLES (OF THE SAME SIZE) WERE TAKEN AND USED TO PRODUCE LEVEL C CONFIDENCE INTERVALS, THE POPULATION PARAMETER WOULD LIE INSIDE THESE CONFIDENCE INTERVALS C PERCENT OF THE TIME IN THE LONG RUN.

Where do confidence intervals come from? Let's think about a generic example. Suppose that we take an SRS of size n from a population with mean μ and standard deviation σ. We know that the sampling distribution for sample means ($\bar{x}$) is (approximately) $\underline{\mathrm{N}(\mu, \frac{\sigma}{\sqrt{n}})}$.

By the 68-95-99.7 rule, about 95% of samples will produce a value for $\bar{x}$ that is within about $\underline{2}$ standard deviations of the true population mean μ. That is, about 95% of samples will produce a value for $\bar{x}$ that is within $\underline{2\frac{\sigma}{\sqrt{n}}}$ of μ. (Actually, it is more accurate to replace 2 by $\underline{1.96}$, as we can see from a Normal Distribution Table.)

But $\boxed{\textbf{if } \bar{x} \textbf{ is within } (1.96)\frac{\sigma}{\sqrt{n}} \textbf{ of } \mu\textbf{, then } \mu \textbf{ is within } (1.96)\frac{\sigma}{\sqrt{n}} \textbf{ of } \bar{x}.}$ This won't be true of every sample, but it will be true of 95% of samples.

So we are fairly confident that μ is between $\underline{\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}}$ and $\underline{\bar{x} + 1.96\frac{\sigma}{\sqrt{n}}}$. We will express this by saying that the 95% CI for μ is $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$

Example. A sample of 25 Valencia oranges weighed an average (mean) of 10 oz per orange. The standard deviation of the population of weights of Valencia oranges is 2 oz. Find a 95% confidence interval (CI) for the population mean.

$10 \pm 1.96\ \left(\frac{2}{5}\right)$, OR $(9.216, 10.784)$

Of course, we are free to choose any level of confidence we like. The only thing that will change is the **critical value** 1.96. We denote the critical value by z^*. If we want a confidence level of C, we choose the critical value z^* so that

THE PROBABILITY $P(Z > z^*) = C/2$

and then

$$\boxed{\text{the level C confidence interval is } \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}}$$

$z^* \dfrac{\sigma}{\sqrt{n}}$ is called THE MARGIN OF ERROR

Example. Now compute a 99% CI for the mean weight of the Valencia oranges based on our previous sample of size 25. ($n = 25$, $\bar{x} = 10$, $\sigma = 2$)

$z^* = 2.576$ CI: $(8.970, 11.030)$

Example. Now compute a 90% CI for the mean weight of the Valencia oranges based on our previous sample of size 25. ($n = 25$, $\bar{x} = 10$, $\sigma = 2$)

$z^* = 1.645$ CI: $(9.342, 10.658)$

Example. A random sample of 100 batteries has a mean lifetime of 15 hours. Suppose the standard deviation for battery life in the population of all batteries is 2 hours. Give a 95% CI for the mean battery life.

$z^* = 1.96$ $n = 100$ $\sigma = 2$ $\bar{x} = 15$ CI: $(14.608, 15.392)$

1.1 What can we learn from the formula?

1. As n increases, the width of the CI DECREASES
2. As σ increases, the width of the CI INCREASES
3. As the confidence level increases, z^* INCREASES , so the width INCREASES

1.2 Determining the required sample size for estimating

Part of designing an experiment is determining what margin of error you can live with. For example, a manufacturer of toothpicks may not need to know as accurately the mean width of their product as a manufacturer of eyeglass screws.

Example. A manufacturer of toothpicks wonders what the mean width of a toothpick is under a new manufacturing method. How many toothpicks must the manufacturer measure to be 90% confident that the sample mean is no farther from the true mean than 0.10 mm? Assume normality, and note that toothpicks produced under the old method had a standard deviation of 0.4 mm.

$z^* = z_{0.05} = 1.645$. (Where did 1.645 come from?)

$$0.1 \geq 1.645 \left(\frac{0.4}{\sqrt{n}}\right) \Rightarrow n \geq \left[\frac{(1.645)(0.4)}{0.1}\right]^2 = 43.29.$$

SO, n SHOULD BE AT LEAST 44.

If the manufacturer wants to be 99% confident, we must use $z^* = z_{0.005} = \underline{2.576}$.

$$0.1 \geq 2.576 \frac{0.4}{\sqrt{n}} \Rightarrow n \geq \left[\frac{(2.576)(0.4)}{0.1}\right]^2 = 106.17.$$

SO, n SHOULD BE AT LEAST 107.

General Method: For margin of error no more than b, simply solve the inequality

$$b \geq z^* \frac{\sigma}{\sqrt{n}}$$

for n, so $n \geq \left(\frac{z^*\sigma}{b}\right)^2$

1.3 An unreasonable assumption

Notice that in all of the confidence intervals above we were required to know σ, the standard deviation of the population distribution. **In practice, one almost never knows** σ. So what can we do?

Fortunately, in many situations, we can estimate that σ is probably pretty close to s, the standard deviation of the sample. **Once we use** s **in place of** σ**, however, the sampling distribution is no longer normal.** When n is quite large, the sampling distribution will still be quite close to normal, however. For smaller values of n we will have to work with a new family of distributions, called Student's t-distributions. But we're getting a little ahead of ourselves. Stay tuned for more information on the t-distributions.

1.4 Some Examples

Here are some additional examples. In these examples we will give the sample standard deviation (s) rather than the population standard deviation σ which would probably be unavailable in most cases. For now we will just proceed using s in place of σ and a normal distribution. **In several of the cases below, this simplification is unwarranted because the sample sizes are too small**, but we will learn how to deal appropriately with these small sample sizes when we learn more about t-distributions. In these cases, our confidence intervals will be smaller using a normal distribution than they should be.

For each of these examples compute a 95% confidence interval and a confidence interval at some other confidence level.

1. In a sample of 45 circuits, the mean breakdown voltage (in kV) under controlled circumstances was 54.7 and the sample standard deviation was 5.23.

   95% CI: $(53.172, 56.228)$, 99% CI: $(52.692, 56.708)$
2. Ten new Discraft 175 gram discs were tested to see how much water they could hold. The mean volume of water was 1.936 liters and the standard deviation was 0.0259.

   90% CI: $(1.933, 1.939)$, 95% CI: $(1.932, 1.940)$
3. Ten old Discraft 175 gram discs were also tested to see how much water they could hold. The mean volume of water was 1.778 liters and the standard deviation was 0.0582.

   95% CI: $(1.769, 1.787)$, 99% CI: $(1.767, 1.789)$
4. An experiment was conducted to see whether it is easier to learn a list of words by looking at textual or graphical representations of the words. 10 people were given a list of words, and 10 were given a list of pictures. The subjects were given 30 seconds to study the list and then asked to recall as many of the items as they could. For those with picture lists, the mean was 8.9 with a standard deviation of 1.595. For those with word lists, the mean was 7.8 with a standard deviation of 1.874.

   95% CI FOR PICTURE: $(7.911, 9.889)$, 98% CI FOR WORD: $(6.422, 9.178)$

We will return to these examples and do them right once we know how to deal with such small samples.

1.5 One-Sided Confidence Intervals

Sometimes we are only interested in bounding our estimate for a parameter in one direction. That is, we want to be able to say with some level of confidence that we believe the parameter θ is greater than some value (or less than some value).

Example. Let's return to our battery example. A producer of batteries would like to make a claim that on average the batteries last *at least* some length of time. (No one will complain, after all, if they happen to last longer than expected.) A random sample of 100 batteries has a mean lifetime of 15 hours. Suppose the standard deviation for battery life in the population of all batteries is known to be 2 hours. Give a **95% confidence lower bound** for the mean battery life.

$z^* = 1.645$ $n = 100$ $\sigma = 2$ $\bar{x} = 15$ lower bound: 14.671 HOURS

For more practice, go back to the previous confidence interval examples and compute one-sided intervals instead.

1.6 Sampling distribution of sample means

For **random samples** from a **normal population** with mean μ and standard deviation σ,

- $\bar{x}$ is normally distributed with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.
- $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ has a normal distribution with mean 0 and standard deviation 1. ($N(0,1)$)
- $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ has a t-distribution with $n-1$ degrees of freedom. ($df = n-1$)

While the family of normal distributions is defined by the mean and standard deviation, the family of t-distributions is defined by **degrees of freedom**, usually abbreviated df.

1.7 The t-distributions

In many ways the t-distributions are a lot like the standard normal distribution. For example, all t-distributions are UNIMODAL and SYMMETRIC with a mean and median both equal to ZERO But the density curve for the t distributions is FLATTER and MORE SPREAD OUT than the density curve for the normal distribution. Statisticians like to say that the t distributions have HEAVIER TAILS , which means more of the values are farther from the center (0).

As df increases, the t-distribution becomes more and more like the standard normal distribution:

We can compute probabilities based on t-distributions just as easily as we can for normal distributions, but we need to use a computer or a different table.

Examples. Suppose the random variable t has a t-distribution with $df = 11$. Use the t-distribution table to determine the following:

- $P(t > 1.363)$ =0.1
- $P(t < 1.363)$ =0.9
- $P(t > -1.363)$ =0.9
- $P(t < -1.363)$ =0.1
- $P(t > 2.201)$ =0.025
- $P(t < -1.796)$ =0.05
- $P(-1.796 < t < 1.796)$ =0.9
- $P(t > 2.00)$ =BETWEEN 0.025 AND 0.05

Now use the table to determine critical values for the following situations.

Examples.

- $df = 27$, 95% confidence, two-sided t^* = `2.052`
- $df = 27$, 95% confidence, one-sided t^* = `1.703`
- $df = 19$, 99% confidence, two-sided t^* = `2.771`
- $df = 19$, 90% confidence, one-sided t^* = `1.314`

1.8 Using the t-distributions for Inference

If **the population is not normal**, then sampling distributions are approximately as indicated above provided the sample size is large enough. In practice, the approximation to the t-distributions is only good enough when one of the following holds

1. *Small Sample Sizes:* If $n < 15$, the population must be normal or close to normal (unimodal, symmetric). If we see that our data are clearly not normal, or if there are outliers in our data, and we don't have good reason to believe that the population is normal, we should not use t-distributions.
2. *Medium Sample Sizes:* The t-procedures will be reasonably accurate even if there is some skewing. As a general rule, we can use the t distributions so long as we do not see strong skewing or outliers in our data.
3. *Large Sample Sizes:* If $n \geq 40$, the t-procedures can be used even if the population is strongly skewed.

1.9 Confidence intervals revisited

$$\text{CI for } \mu \text{ if } \sigma \text{ is known:} \quad \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

$$\text{CI for } \mu \text{ if } \sigma \text{ is unknown:} \quad \bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

Example. Sixteen plots in a farmer's field were randomly selected and measured for total yield. The mean yield of the sample was 100 kg with a standard deviation of 18 kg. Assuming yield per plot is normally distributed, find a 98% CI for the mean yield per plot in the farmer's field.

$$100 \pm 2.602 \frac{18}{\sqrt{16}}, \text{ OR } (88.291, 11.709)$$

When the sample size is very large, the t-distributions and the standard normal distribution are nearly identical. In the t-distribution table, critical values for the standard normal distribution can be found in the row labeled ∞. We will use these values also when df is large.

Example. If we want to compute a 90% CI and $n = 200$ ($df = 199$), then $t^* = \underline{1.660}$

Example. U.S. Public Health Service 10 year hypertension study. The mean change in blood pressure for the 171 people using medication was -9.85, with a standard deviation of 11.55. Give a 95% CI for the true mean change in blood pressure.

$$-9.85 \pm 1.984 \frac{11.55}{\sqrt{171}}, \text{ OR } (-11.602, -8.098)$$

1.10 Some cautions about CIs for the mean of a population

1. Data must be from an SRS.
2. The formula is not correct for more complex designs (e.g., stratified random sampling).
3. Because the mean is not resistant, outliers can have a large effect on the CI.
4. If the sample size is small and the population not normal, we need different procedures.
5. The margin of error in a CI covers only random sampling errors, not errors in design or data collection.

1.11 Prediction Intervals

Prediction intervals are similar to confidence intervals, but they are estimating something different. The confidence intervals we have been looking at give an estimate for **the mean of the population**. But there is another type of estimate that is commonly done. For a prediction interval, we want to give a range in which we expect **one randomly selected individual's** value to be.

Once again, the key to figuring out prediction intervals is understanding the sampling distribution involved. We consider our sample $x_1, x_2, \ldots, x_n$ to be n randomly selected values from a population distribution, and we consider x to be one additional random seletion from this distribution. The idea is to make a prediction for x (in the form of a confidence interval) based on the results of the sample.

The obvious point estimate to use is $\bar{x}$, the sample mean. But what can we say about the quality of this estimate? Let's look at the sampling distribution of $\bar{x} - x$ (the differences between the sample mean and the value of an additional random selection). If the population has mean μ and standard deviation σ, then

$$E(\bar{x} - x) = E(\bar{x}) - E(x) = \mu - \mu = 0$$

and

$$V(\bar{x} - x) = V(\bar{x}) + V(x) = \frac{\sigma^2}{n} + \sigma^2 = \sigma^2(1 + \frac{1}{n})$$

so the standard deviation of the sampling distribution is $\sigma\sqrt{1 + \frac{1}{n}}$.

From this we see that a prediction interval would be

$$\bar{x} \pm z^* \sigma \sqrt{1 + \frac{1}{n}}$$

if we knew σ. As you might guess, when we do not know sigma, we will use instead

$$\bar{x} \pm t^* s \sqrt{1 + \frac{1}{n}}$$

using the sample standard deviation s in place of the population standard deviation σ and using a t-critical value (with $n - 1$ degrees of freedom) instead of a critical value from the normal distribution.

2 Hypothesis Testing

Our goal with tests of significance (hypothesis testing) is to ASSESS THE EVIDENCE PROVIDED BY DATA WITH RESPECT TO SOME CLAIM ABOUT A POPULATION.

A **hypothesis test** is a formal procedure for comparing observed (sample) data with a hypothesis whose truth we want to ascertain.

The **hypothesis** is a STATEMENT about the PARAMETERS in a POPULATION or MODEL. The results of a test are expressed in terms of a probability that measures how well the data and the hypothesis agree. In other words, it helps us decide if a hypothesis is reasonable or unreasonable based on the likelihood of getting sample data similar to our data.

Just like confidence intervals, hypothesis testing is based on our knowledge of SAMPLING DISTRIBUTIONS.

2.1 4 step procedure for testing a hypothesis

Step 1: Identify parameters and state the null and alternate hypotheses.

State the hypothesis to be tested. It is called the **null hypothesis**, designated H_0.

The **alternate hypothesis** describes what you will believe if you reject the null hypothesis. It is designated H_a. This is the statement that we hope or suspect is true instead of H_0.

Example. The average LDL of healthy middle-aged women is 108.4. We suspect that smokers have a higher LDL (LDL is supposed to be low). $H_0 : \mu = 108.4, \quad H_a : \mu > 108.4$

Example. The average Beck Depression Index (BDI) of women at their first post-menopausal visit is 5.1. We wonder if women who have quit smoking between their baseline and post visits have a BDI different from the other healthy women. $H_0 : \mu = 5.1, \quad H_a : \mu \neq 5.1$

Step 2: Compute the test statistic

A test statistic measures how well the sample data AGREES WITH THE NULL HYPOTHESIS.

When we are testing a hypothesis about the mean of a population, the test statistic has the form

$$\frac{\text{(DATA VALUE) - (HYPOTHESIS VALUE)}}{\text{SD OR SE}}$$

For our HWS examples (LDL: 159 smokers, $\bar{x} = 115.7$, sd=29.8; BDI: 27 quitters, $\bar{x} = 5.6$, sd=5.1):

LDL: $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{115.7-108.4}{29.8/\sqrt{159}} = 3.089$

BDI: $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{5.6-5.1}{5.1/\sqrt{27}} = 0.509$

Step 3: Compute the p-value

The **p-value** (for a given hypothesis test and sample) **is the probability, if the null hypothesis is true, of obtaining a test statistic as extreme or more extreme than the one actually observed.**

How do we compute the p-value?

Here is how hypothesis testing works based on the t-distribution:

Suppose that an SRS of size n is drawn from a population having unknown mean μ. To test the hypothesis H_0: $\mu = \mu_0$ based on an SRS of size n, compute the one-sample t statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

The p-value for a test of H_0 against

- H_a: $\mu > \mu_0$ is $P(T > t)$
- H_a: $\mu < \mu_0$ is $P(T < t)$
- H_a: $\mu \neq \mu_0$ is $P(T > |t|) + P(T < -|t|) = 2P(T > |t|)$

where T is a random variable with a t distribution ($df = n - 1$).
These p-values are exact if the population distribution is normal and are approximately correct for large enough n in other cases.

The first two kinds of alternatives lead to **one-sided** or **one-tailed** tests. The third alternative leads to a **two-sided** or **two-tailed** test.

Example. Find the p-value for testing our hypotheses regarding LDL in the HWS. $t = 3.09$, one-tailed alternate hypothesis, $n = 159$ smokers.

THOUGH $df = 158$ IS NOT ON TABLE D, WE CAN TELL FROM THE VALUES THAT ARE THERE THAT $P(t \geq 3.09) < 0.0025$.

Example. Find the p-value for testing our hypothesis regarding BDI in the HWS. t=.51, two-tailed alternate hypothesis, $n = 27$ quitters.

SINCE OFF THE CHART TO THE LEFT OF TABLE D, THE ONE-SIDED P-VALUE WOULD BE GREATER THAN 0.25. SINCE WE DOUBLE THIS FIGURE FOR THE TWO-SIDED ALTERNATIVE HYPOTHESIS, WE KNOW $P(|t| > 0.51) > 0.5$.

Step 4: State a conclusion

A decision rule is simply a statement of the conditions under which the null hypothesis is or is not rejected. This condition generally means choosing a **significance level** α. **If** p **is less than the significance level** α**, we reject the null hypothesis**, and decide that the sample data do not support our null hypothesis and the results of our study are then called **statistically significant at significance level** α.

If $p > \alpha$**, then we do not reject the null hypothesis.** This doesn't necessarily mean that the null hypothesis is true, only that our data do not give strong enough evidence to reject it. It is rather like a criminal trial. If there is strong enough evidence, we convict (guilty), but if there is reasonable doubt, we find the defendant "not guilty". This doesn't mean the defendant is innocent, only that we don't have enough evidence to claim with confidence that he or she is guilty.

Looking back at our previous examples, if we select a significance level of $\alpha = 0.05$, what do we conclude?

- LDL (p-value < 0.0025)

  Since $0.0025 < 0.05$, we REJECT the null hypothesis and conclude that the smokers have a significantly higher LDL. "Smokers have higher LDL ($t = 3.09$, $df = 158$, $p < .0025$)".

- BDI (p-value $> .5$)

  This p-value is higher than .05, so we DO NOT REJECT the null hypothesis. It is quite possible that the null hypothesis is true and our data differed from the hypothesized value just based on random variation. In this case we say that the difference between the mean of 5.6 (those who quit smoking) and 5.1 (average healthy women) is not statistically significant. "Those who quit smoking do not have a significantly different average BDI than middle-aged healthy women ($t = .51$, $df = 26$, $p > .50$)."

It's good to report the actual p-value, rather than just the level of significance at which it was or was not significant. This is because:

IT IS USEFUL TO KNOW BY HOW MUCH THE NULL HYPOTHESIS AS BEEN REJECTED OR NOT REJECTED. IT ALSO LETS SOMEONE ELSE WHO HAS SET THE SIGNIFICANCE LEVEL α DIFFERENTLY TO DRAW HER OWN CONCLUSION.

Example. Suppose that it costs \$60 for an insurance company to investigate accident claims. This cost was deemed exorbitant compared to other insurance companies, and cost-cutting measures were instituted. In order to evaluate the impact of these new measures, a sample of 26 recent claims was selected at random. The sample mean and sd were \$57 and \$10, respectively. At the 0.01 level, is there a reduction in the average cost? Or, can the difference of three dollars (\$60-\$57) be attributed to the sample we happened to pick?

Hypotheses: $H_0 : \mu = 60, \quad H_a : \mu < 60$

Test statistic: $t = \dfrac{57 - 60}{10/\sqrt{26}} = -1.530$ p-value: $0.1 < P(|t| \geq 1.530) < 0.2$

Conclusion: THE DATA IS NOT SIGNIFICANT AT THE 1% LEVEL TO REJECT THE NULL HYPOTHESIS. RATHER, IT IS CONSISTENT WITH THE BELIEF THAT, DESPITE THE COST-CUTTING MEASURES ALREADY TAKEN, THE AVERAGE COST PER ACCIDENT CLAIM IS STILL \$60. $(t = 1.53, df = 25, P > 0.1)$

Example. The amount of lead in a certain type of soil, when released by a standard extraction method, averages 86 parts per million (ppm). Developers of a new extraction method wondered if their method would extract a significantly different amount of lead. 41 specimens were obtained, with a mean of 83 ppm lead and a sd of 10 ppm.

Hypotheses: $H_0 : \mu = 86, \quad H_a : \mu \neq 86$

Test statistic: $t = \dfrac{83 - 86}{10/\sqrt{41}} = -1.921$ p-value: $0.05 < P(|t| \geq 1.921) < 0.1$

Conclusion: SAMPLES AS EXTREME AS THE ONE THESE DEVELOPERS TOOK OCCUR LESS THAN 10%, BUT MORE THAN 5% OF THE TIME WHEN THE MEAN FOR THE METHOD IS STILL 86 PARTS PER MILLION. THUS, WE COULD REJECT H_0 AT THE 5% LEVEL, BUT NOT THE 10% LEVEL. $(t = 1.921, df = 40, 0.05 < P < 0.1)$

Example. The final stage of a chemical process is sampled and the level of impurities determined. The final stage is recycled if there are too many impurities, and the controls are readjusted if there are too few impurities (which is an indication that too much catalyst is being added). If it is concluded that the mean impurity level=.01 gram/liter, the process is continued without interruption. A sample of $n = 100$ specimens is measured, with a mean of 0.0112 and a sd of 0.005 g/l. Should the process be interrupted? Use a level of significance of .05.

Hypotheses: $H_0 : \mu = 0.01, \quad H_a : \mu \neq 0.01$

Test statistic: $t = \dfrac{0.0112 - 0.01}{0.005/\sqrt{100}} = 2.4$ p-value: $0.01 < P(|t| \geq 2.4) < 0.02$

Conclusion: THE RESULTS ARE SIGNIFICANT AT THE 5% LEVEL TO CONCLUDE THE PROCESS MEAN IS NOT AT THE TARGET LEVEL (0.01 G/L) AND SHOULD BE INTERRUPTED FOR ADJUSTMENTS. $(t = 2.4, df = 99, P < 0.02)$

2.2 Comparing significance tests and confidence intervals

Example. The drained weights for a sample of 30 cans of fruit have mean 12.087 oz, and standard deviation 0.2 oz. Test the hypothesis that on average, a 12-oz. drained weight standard is being maintained.

Now let's compute a CI.

What is the relationship between confidence intervals and two-sided hypothesis tests?