A Brief Review

- Confidence intervals look like

$$\text{mean} - z \cdot SD/\sqrt{n} \ \ldots \ \text{mean} + z \cdot SD/\sqrt{n}$$

  Usually $z = 2$ (for 95% confidence). For other levels of confidence, or to account for estimating the SD, take $z =$ something else.

- Confidence intervals can be used to test predictions.
  - Make a prediction about a sample or experiment, like "the mean should be $\cdots$"
  - Make a confidence interval from the data in the sample or the experiment
  - If the confidence interval contains the prediction, you are happy.
  - If the confidence interval doesn't contain the prediction, you suspect [you are "95% confident"] that something was wrong with the assumptions underlying the prediction.

This is a round-about way of doing things. *Hypothesis testing* is a more direct way of testing hypotheses or predictions about data.

Hypothesis Testing

In *hypothesis testing* you directly calculate the probability that a re-run of the sample (or experiment) could be farther from the prediction than what you observe, *assuming the prediction is correct*.

> The probability that a re-run of the sample or experiment could yield a result farther from the prediction than what we have observed, *assuming the prediction is correct*, is called a p-value.

- p-value small $\Rightarrow$ Prediction does not match data
- p-value large $\Rightarrow$ Prediction and data match

When the prediction does not match the data, you suspect the assumptions underlying the prediction.

- *Sometimes you are happy that the prediction doesn't work out*
- *Sometimes you are happy that the prediction does work out*

…it depends on the problem!

Example

Let's return to the SAT example:

The SAT is constructed so that scores have a national average of of 500 and a national standard deviation of 100. A random sample of 64 students from a recent entering class at Carnegie Mellon have an SAT verbal average of 555.

Central Question:
How likely is it that we would see an SAT average even farther from 500 if we re-ran the sample *from the national population*?

$$\begin{aligned} P[\overline{X} > 555] &= P\left[\frac{\overline{X} - 500}{100/\sqrt{64}} > \frac{555 - 500}{100/\sqrt{64}}\right] \\ &= P[Z > 4.4] \\ &\approx 0 \end{aligned}$$

without the help of any tables.
Formal pieces:

- The *Null Hypothesis* (H_0) is

  "The average CMU entering freshman SATV score is equal to the national average"

Note that this says there is "no change" from the national population to the CMU population.
This allows us to *predict* that the mean of a sample of CMU freshmen should have average SATV score around 500.

- The *Alternative Hypothesis (H_a)* is

  > "The average CMU entering reshman SATV score <u>is greater than</u> the national average"

  This is what we will be led to believe if H_0 doesn't agree with the data. It is a *one-sided alternative* because we only care if CMU's SAT average is on one side of the national average.
- The *Significance Test* is the calculation of the p-value, or probability than another CMU sample would yield an average SATV higher than 555, assuming that CMU students are like the national population of students. The summary of the data we focus on when we calculate the p-value is the *test statistic*.
- The *p-value* tells us how the prediction matches the data. A *small* p-value like ours suggests that the prediction does not match the data ("*statistically significant difference*").

Another example

A paint manufacturer tries to ensure that the amount of color in one line of paint is exactly 8 g/kg. A sample of 36 cans of paints from a particular paint mixer has an average of 8.3 g/kg with an SD of 1.02. Should the manufacturer be worried?

- The *Null Hypothesis (H_0)* is

  > "The average color content of paints from this mixer is *the same as* the manufacturer's goal."

  In other words "nothing unusual is happening" in that mixer.

- The *Alternative Hypothesis (H_a)* is

  > "The average color content of paints from this mixer is *different from* the manufacturer's goal."

  This is what we will be led to believe if H_0 doesn't agree with the data. It is a *two-sided alternative* because we don't care which side of the manufacturer's goal we end up on, we want to know about all differences.

- The *Significance Test* is the calculation of the p-value, or probability than another sample from this mixer would yield an average color content farther from 8 g/kg than our sample did. The sample average is the *test statistic*. We calculate:

$$\begin{aligned} P[|\overline{X} - 8| > |8.3 - 8|] &= \\ P\left[\left|\frac{\overline{X} - 8}{1.02/\sqrt{36}}\right| > \left|\frac{8.3 - 8}{1.02/\sqrt{36}}\right|\right] &= P[|Z| > 1.76] \end{aligned}$$

We know to expect this to be between 32% and 5%. Using the Minitab command

```
MTB > cdf 1.76;
SUBC> normal.
Normal with mean = 0 and
standard deviation = 1.00000

         x        P( X <= x)
    1.7600            0.9608
```

or looking this up in a table, we discover that $P[Z < 1.76] \approx 0.96$, so it must be that $P[|Z| > 1.76] \approx 0.08$. Thus, the p-value is approx. 0.08.

- The *p-value* is often compared to standard "thresh-hold values". The three most commonly used values are 0.01, 0.05 and 0.10. We could say:
  - *There is no significant difference* between the paint content for this mixer, and the company goal of 8 g/kg, *at the 5% level.*
  - *There is a significant difference at the 10% level.*

Notes:

- *Two sided tests give the same answers as confidence interval tests.* The 95% confidence interval here would have been $8.3 \pm 2 \cdot 1.02/\sqrt{36}$, or from 7.96 to 8.64; this contains the target value of 8 g/kg, so we *cannot reject the null hypothesis* with 95% confidence.
  Note that "significant at the p% level" is the same as "(1-p)% confidence".
- *A shortcut for computing the p-value of a two-sided test* is to do the one-sided test and then double the p-value.
- *One-sided or two-sided?* Use the one that goes along with the story. Some problems seem to demand one or the other. If you aren't sure, use a two-sided interval.

Refinements: The T Distribution Again

In the paint example, we used the sample SD instead of a population SD.

To adjust for estimating the SD from the sample, we should have used the t-distribution with 35 ($35 = 36-1 = n-1$) "degrees of freedom". (this allows the adjustment to depend on the sample size):

$$\begin{aligned} P[|\overline{X} - 8| > |8.3 - 8|] &= \\ P\left[\left|\frac{\overline{X} - 8}{1.02/\sqrt{36}}\right| > \left|\frac{8.3 - 8}{1.02/\sqrt{36}}\right|\right] &= P[|Z| > 1.76] \end{aligned}$$

Using Minitab, we see

```
MTB > cdf 1.76;
SUBC> t 35.
Student's t distribution with 35 DF
         x      P( X <= x)
    1.7600          0.9564
```

So, since $P[Z < 1.76] \approx 0.9564$, we see that $P[|Z| > 1.76] \approx 0.0872$. (compared with p-value $= 0.08$ that we got before).

You should not be surprized if you see the t distribution in Minitab output. The interpretation is virtually the same as with the Normal distribution.

Two samples

In the sleeping pill example from last time, we

- gave 36 subjects the sleeping pill: mean extra sleep 0.5h, SD=1h
- gave 20 more subjects a sugar pill: mean extra sleep 0.25h,SD=.25h

- The *Null Hypothesis (H_0)* is

  "There is no difference between the mean extra sleep for the two groups."

  In other words "no difference" between the groups.

- The *Alternative Hypothesis (H_a)* is

  "There is a difference."

  This is a *two-sided alternative*.

- The *Significance Test* is the calculation of the p-value, or probability than a re-run of the experiment would yield a test statistic more different from zero than the one we observed. Our *test statistic* is the difference in the two means, divided by an appropriate standard error.

If we ask Minitab to do this. . .

```
Two Sample T-Test and Confidence Interval

Two sample T for drug vs placebo
          N      Mean     StDev   SE Mean
drug     36      0.50      1.00      0.17
placebo  20     0.250     0.999      0.22

95% CI for mu drug - mu placebo: ( -0.31,  0.81)
T-Test mu drug = mu placebo (vs not =): T= 0.90
P=0.38  DF=  39
```

- The *p-value* , 0.38 is not “statistically significant”, i.e. it is not less than any threshhold like 0.10, 0.05 or 0.001.

Hypothesis Testing: Vocabulary and Summary

(1) A Hypothesis in statistics refers to a statement about a feature (parameter) of the population.

(2) The Null Hypothesis, (H_0), is a statement that the parameter is equal to some specified number. Some examples are:

"The average CMU entering freshman SATV score is equal to the national average"

"The paint concentration is 8 g/kg",

"The mean increase in sleep for patients on a new drug is the same as for patients who get a placebo." [two-sided alternative]

(3) The Alternative Hypothesis, (H_a), is also called the *motivating or research hypothesis*. It is the opposite of the null hypothesis. Some examples are:

"The average CMU entering reshman SATV score is greater than the national average" [one-sided alternative]

"The paint concentration is different from 8 g/kg" [two-sided alternative]

"The mean increase in sleep for patients on a new drug is NOT the same as for patients who get a placebo."

(4) A Significance Test is based on a test statistic that shows whether or not the data provide evidence against the null hypothesis.

(5) The p-value is a number between 0 and 1 that measures the weight of the evidence in the data *against* H_0. Small p-values indicate strong evidence against H_0. If the p-value of a test is smaller than a specific value, such as 0.05, then the data are said to provide statistically significant evidence *against* H_0.