

## Two-sample t-tests.

- Independent samples
- Pooled standard deviation
- The equal variance assumption



Last time, we used the mean of one sample to test against the hypothesis that the true mean was a particular value.

$$H_0: \mu = 20$$

$$H_A: \mu \neq 20$$

One-sided test:

$$H_0: \mu \geq 20$$

$$H_A: \mu < 20$$

Two-sided test:

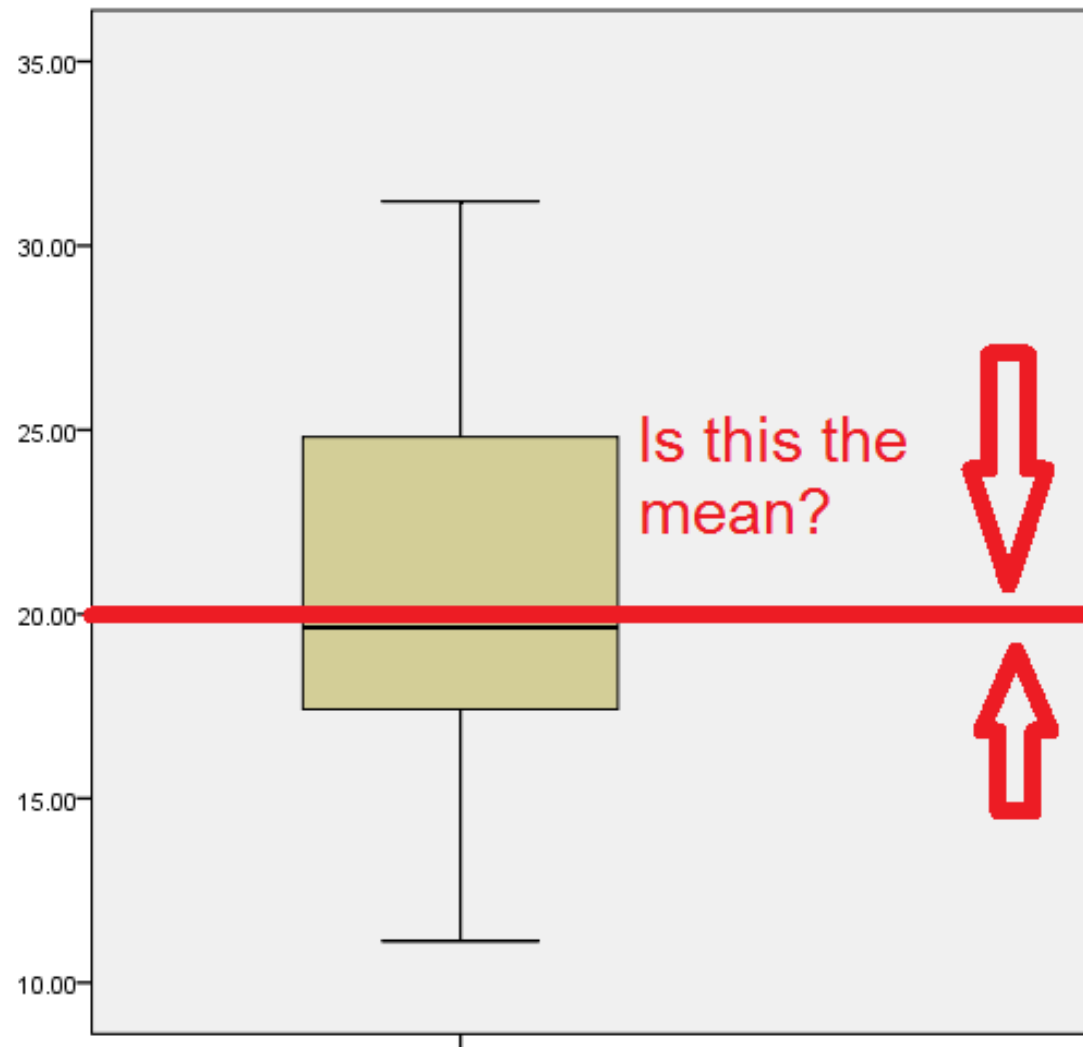
We also applied the idea of testing against a specific value to a proportion.

$$H_0 : \pi \geq .90$$

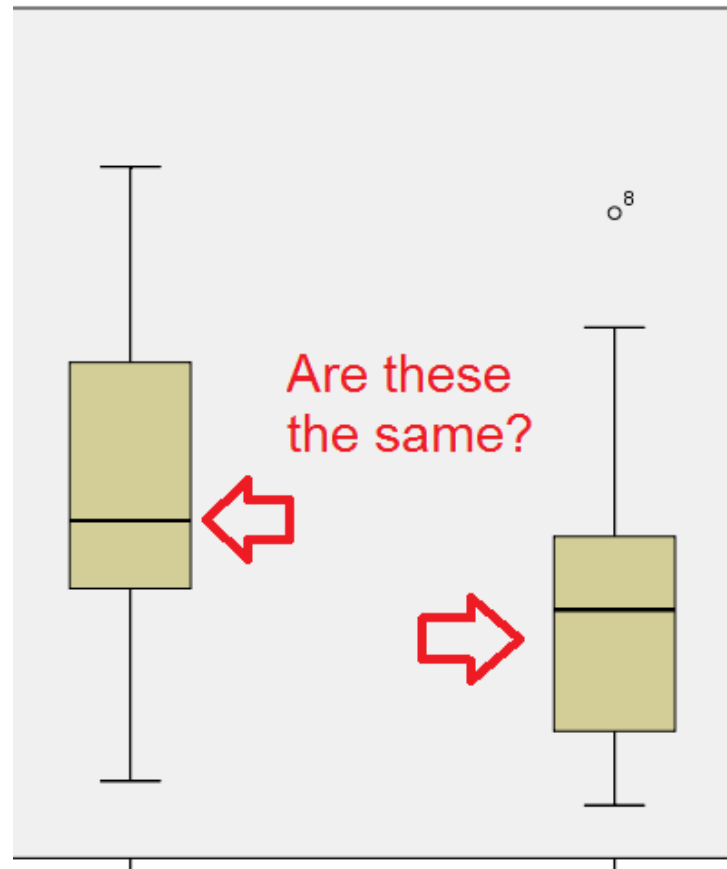
$$H_A : \pi < .90$$

After all, a proportion is just a mean of zeros (nos) and ones (yeses).

In every one sample test, we have a given value we're comparing the sample mean against. The question: Is this given value plausible?



But what if we don't have a specific value to compare against?  
What if, instead, we're comparing the means of two groups  
against *each other*?



That's a job for two-sample testing.

Two independent samples.

The null hypothesis is that the means are the same.

$$H_0 : \mu_1 = \mu_2$$

$$H_A : \mu_1 \neq \mu_2$$

The alternative can be two-sided (not-equal), or one-sided on either side (less than / more than)

$$H_0 : \mu_1 = \mu_2$$

$$H_A : \mu_1 < \mu_2$$

Another way to say “two means at the same” is

“The difference between these means is **zero**.”

$$\mu_1 = \mu_2$$



$$\mu_1 - \mu_2 = 0$$



By using the “difference of zero” mindset, we get a hint on how to handle two sample tests.

Make the null hypothesis about the difference between the means, instead of just a single mean.

Then our t-score formula works just fine.

$$t = \frac{\text{Sample Value of ??} - \text{Null Hypothesis of ??}}{\text{Standard Error of ??}}$$

Let  $\bar{X}_1$  and  $\bar{X}_2$  be the means of each of the samples.

Let  $\mu_1$  and  $\mu_2$  be the true means, (that  $H_0$  says are the same)

The t-score looks like:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{SE}$$

We never deal with the formula in this form. Since the null assumption is that the true means are the same.

$$t = \frac{\begin{array}{c} \text{Sample Mean} \\ \text{Difference} \\ \downarrow \\ (\bar{X}_1 - \bar{X}_2) \end{array} - \begin{array}{c} \text{Population Mean} \\ \text{Difference} \\ \downarrow \\ (\mu_1 - \mu_2) \end{array}}{\text{SE} \leftarrow \text{Standard Error}}$$

This is the formula we really use for the t-score. (But now you know where it comes from)

$$t = \frac{\bar{X}_1 - \bar{X}_2}{SE}$$

The standard error is defined by further assumptions.



A hairy topic, to be sure. How about an example?

Example: New textbook vs. old textbook.

Say we wanted to test if a new brand of textbook is a **better** resource as measured by provincial exam scores. (alpha = .05)

We'll get two classrooms of the same grade and similar skill level. We'll then flip a coin, if heads:

Class A gets the current textbook, Class B get the new one.

Otherwise, Class A gets the new one.

At the end of the semester, these are the summary stats:

	Old text class	New text class
Mean	64.3	68.8
Std. Dev. s	7.1	7.4
Sample Size n	21	23

First, identify:

We are comparing two mean, a two-sample test is appropriate.

We want to know if the new textbook is ***better*** so this is a one-sided test.

But what do we do about the standard error?



There are values for  $n$ , and two values for standard deviation.

	Old text class	New text class
Mean	64.3	68.8
Std. Dev. s	7.1	7.4
Sample Size n	21	23

We can get a measure that collects ( ***pools*** ) the standard deviation of the whole system.

It's called the ***pooled standard deviation***.

The pooled standard deviation,  $S_p$ , works by way of a weighted average of the variances (standard deviation squared) that we already have.

$$S_p^2 = \frac{S_1^2(n_1 - 1) + S_2^2(n_2 - 1)}{(n_1 - 1) + (n_2 - 1)}$$

It's a weighted average because if one sample is much larger (has more *degrees of freedom*), it should count for more.

In this case, the samples are roughly the same size, so the pooled standard deviation  $S_P$  should be near the *middle* of the two standard deviations (7.1 and 7.4).

$$S_P^2 = \frac{7.1^2(20) + 7.4^2(22)}{20 + 22} = 7.26$$

The degrees of freedom of the pooled standard deviation  $S_P$  is the total of the df from each sample.  $(n_1 - 1) + (n_2 - 1)$

$S_p = 7.26$  is used to find the standard error.

$$SE = S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

This standard error formula is used when there are two samples, hence it has two sample sizes  $n_1$  and  $n_2$ .

$$SE = 7.26 \sqrt{\frac{1}{21} + \frac{1}{23}} = 2.19$$

Both  $n_1$  and  $n_2$  have to be large to make the standard error small. This reflects that added uncertainty of having two unknown means instead of one.

$$SE = 7.26 \sqrt{\frac{1}{21} + \frac{1}{23}} = 2.19$$

Now we can get the t-score.

$$SE = 2.19 \quad X_{\text{new}} = 68.8 \quad X_{\text{old}} = 64.3$$

$$t = \frac{68.8 - 64.3}{2.19} = 2.054$$

Sample 1 has 20 df, Sample 2 has 22 df, so the total df is 42.

We compare this t-score, 2.054 , to the critical value in the t-table, one-sided, 0.05 significance at 40 df (always round down if it's between values).

$$t^* = \mathbf{1.684} , t\text{-score} = 2.054 > t^*$$

Reject  $H_0$

The new book classroom did do better.





Puzzled? Let's try another.

## Example 2: Bio-equivalence.

When one drug is being tested to replace another, it's important to check that the new drug has the same effects on the body as the old drug.

This is typically done with a two-sample t-test.

Expenza\*, a name-brand drug is being used to lower blood pressure. We've been hired to test if Thriftubin\*, a cheaper generic drug, has the same effect on blood pressure.

\*Made up drug names

The name-brand drug has been used for a while, so we have lots of data on it, so its sample size is large.

Blood pressure (Systolic)	Name Brand	Generic
Mean	130.0	123.5
Std. Dev. s	12.6	13.5
Sample Size n	144	16

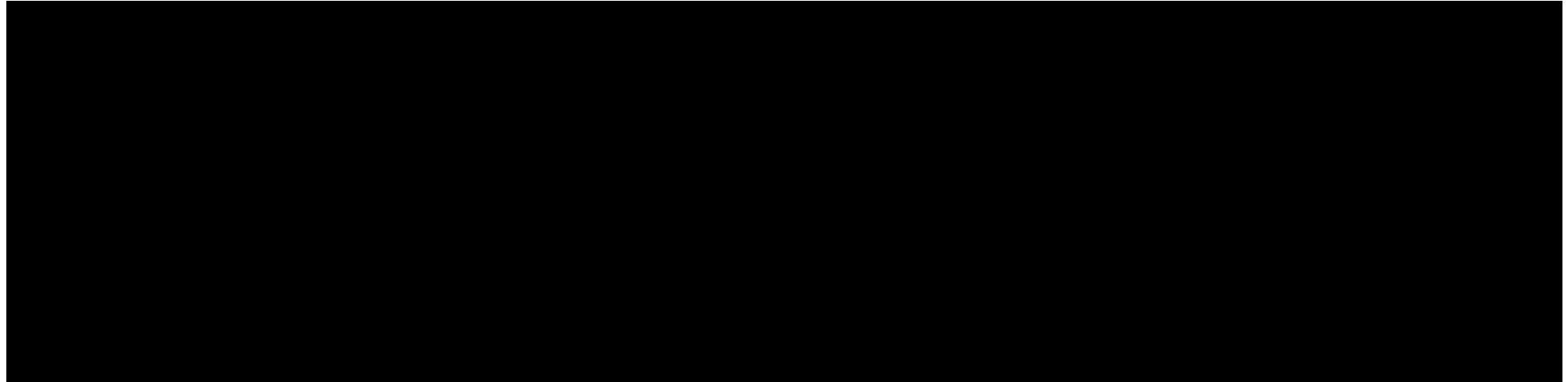
Comparing two means  $\rightarrow$  Two-sample t-test.

We want to show equal vs. not equal.  $\rightarrow$  Two-sided test.

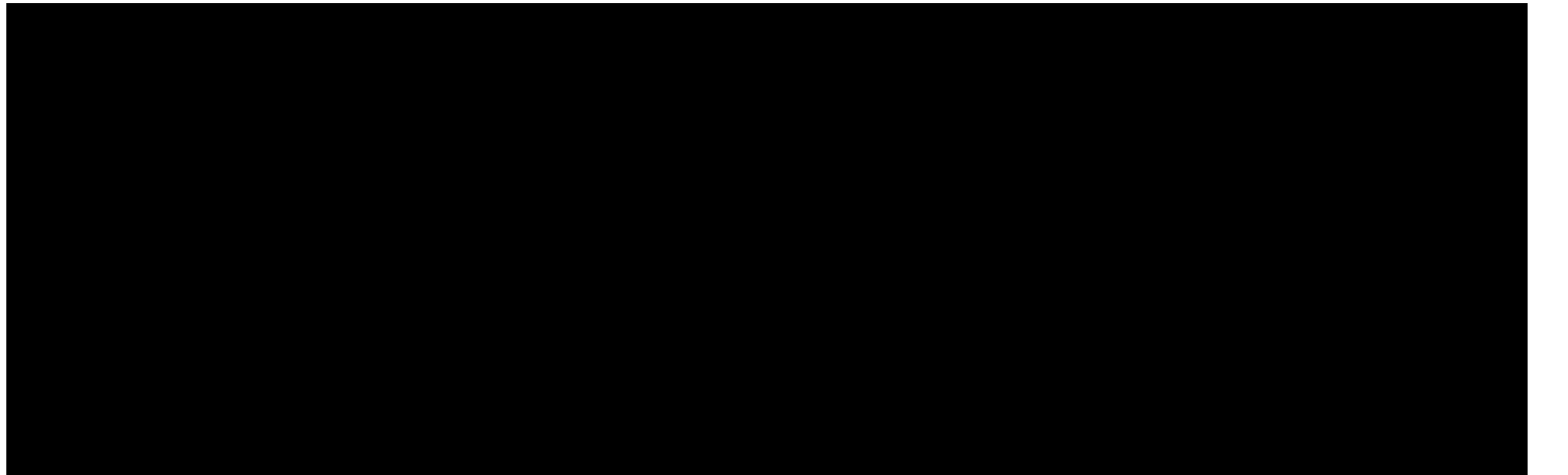
$$H_0 : \mu_1 = \mu_2$$

$$H_A : \mu_1 \neq \mu_2$$

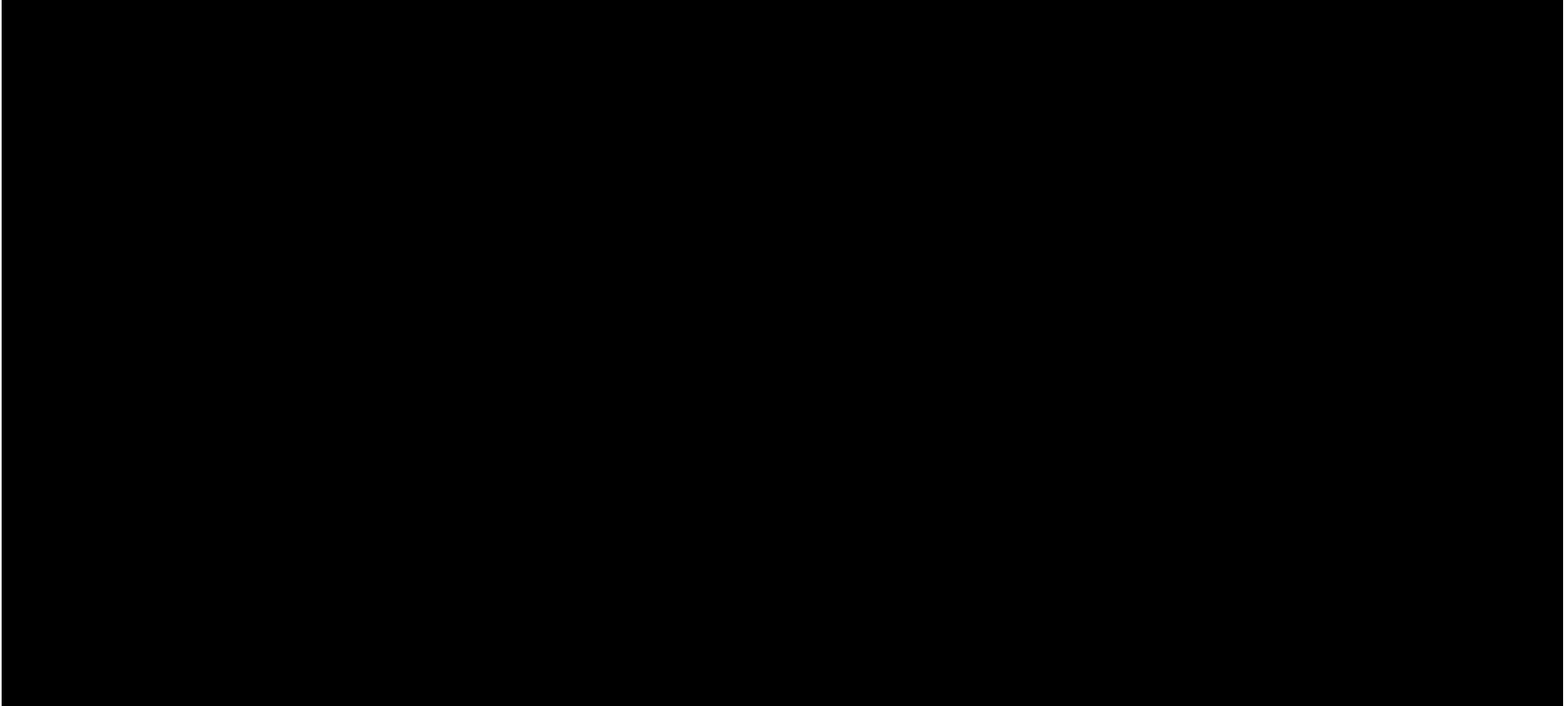
First, get the pooled standard deviation.



Use  $S_p = 12.69$  to get the standard error.



Finally, use  $SE = 3.34$ ,  $X_{NB} = 130.0$ ,  $X_G = 123.5$  to find the t-score and compare that to the critical value.



t-score = 1.944,  $t^* = 1.984$  (using  $df=100$ , .05 significance, two-sided), so we fail to reject, but just barely.

We can't detect at the 5% level that the generic drug has a different effect on blood pressure than the name-brand drug.

But it's close, (t-score was almost as large as  $t^*$ ) so we'll want to mention that to whomever hired us.

If we had a larger sample, we'd be better able to detect a difference. Which sample size would we increase to do this?

We can't detect at the 5% level that the generic drug has a different effect on blood pressure than the name-brand drug.

But it's close, (t-score was almost as large as  $t^*$ ) so we'll want to mention that to whomever hired us.

If we had a larger sample, we'd be better able to detect a difference. Which sample size would we increase to do this?

***The generic drug sample (it's smaller, so that's where a lot of our uncertainty is coming from)***





I hope you're not feeling in the dark.

Is Syria more dangerous now than two years ago?

The WHO (World Health Organization) is interested in knowing if the fighting in Syria is affecting the general health of the population.

They give you the death records to two days in the capital, Damascus; one in 2010, and one in the same day in 2012 (to avoid seasonal effects).

\*2010 Mean from world factbook. All else made up.

Age at Death	Syria Then	Syria Now
Mean	74.1	71.2
Std. Dev.	8.3	22.6
Sample Size	52	45

We take the average age of the deaths on each day to estimate life expectancy.

Is life expectancy in 2012 less than in 2010? ( $\alpha = .01$ )

(Two sample, one side test.)

The pooled  $S_p$  assumes that the true standard deviation of both groups is the same, and that we can use both sample standard deviations together to get a better estimate of that true standard deviation.

In this case, the standard deviation (and hence the variance) is much larger in the 2012 death ages.

The ***assumption of equal variance*** isn't reasonable, so we can't get a pooled estimate.

Instead, we use a standard error formula that includes both standard deviations, but separately.

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Just make sure  $S_1$  and  $n_1$  are referring to the same sample.

$$SE = \sqrt{\frac{8.3^2}{52} + \frac{22.6^2}{45}} = 3.56$$

The rest is getting the t-score and comparing to a critical.

$$t = \frac{71.2 - 74.1}{3.56} = -0.81$$

$t^* = 2.414$  for a one sided test,  $\alpha = 0.01$ , and  $df=44$ .

t-score = -0.81, which is between -2.414 and 2.414, so we fail to reject.

We cannot find evidence that life expectancy as a whole has been reduced in Damascus, Syria.

By using two different standard deviations, we lose some degrees of freedom. SPSS computes the degrees of freedom exactly. (formula available on request)

However, whenever we're doing this by hand, we'll use the worst case scenario of the lower of  $(n_1 - 1)$  and  $(n_2 - 1)$

So that's why  $df=45 - 1 = 44$ , instead of

$$(45 - 1) + (52 - 1) = 95$$



How do we know if the standard deviations are too different to use the pooled estimate  $S_p$  ?

SPSS has a test to see if the pooled standard deviation is reasonable, and will give us both answers. (Levene's test for equal variances)

To do it ourselves, we have the F-statistic and F-test, which we'll cover in the ANOVA section near the end of the semester.

Also, if you don't assume equal variance and use the standard error formula with separate standard deviations, you'll get an answer very close to the standard error from  $S_p$ .

Textbook SE using $S_p$	2.191
Textbook SE using $S_1, S_2$	2.187
Drugs using $S_p$	3.34
Drugs using $S_1, S_2$	3.53

Next day: Paired t-tests

Two sample t-tests for proportions

SPSS two samples

